Anomaly detection in multimodal MRI identifies rare individual phenotypes among 20,000 brains

Zhiwei Ma[1*], Daniel S. Reich[2], Sarah Dembling[1], Jeff H. Duyn[1], Alan P. Koretsky[1*]

[1]Laboratory of Functional and Molecular Imaging, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892-1065, USA

[2]Translational Neuroradiology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892-1400, USA

* Corresponding author.

Email: koretskya@ninds.nih.gov; maz4@nih.gov

Abstract

The UK Biobank (UKB) is a large-scale epidemiological study and its imaging component focuses on the pre-symptomatic participants. Given its large sample size, rare imaging phenotypes within this unique cohort are of interest, as they are often clinically relevant and could be informative for discovering new processes and mechanisms. Identifying these rare phenotypes is often referred to as “anomaly detection”, or “outlier detection”. However, anomaly detection in neuroimaging has usually been applied in a supervised or semi-supervised manner for clinically defined cohorts of relatively small size. There has been much less work using anomaly detection on large unlabeled cohorts like the UKB. Here we developed a two-level anomaly screening methodology to systematically identify anomalies from ~19,000 UKB subjects. The same method was also applied to ~1,000 young healthy subjects from the Human Connectome Project (HCP). In primary screening, using ventricular, white matter, and gray matter-based imaging phenotypes derived from multimodal MRI, every subject was parameterized with an anomaly score per phenotype to quantitate the degree of abnormality. These anomaly scores were highly robust. Anomaly score distributions of the UKB cohort were all more outlier-prone than the HCP cohort of young adults. The approach enabled the assessments of test-retest reliability via the anomaly scores, which ranged from excellent reliability for ventricular volume, white matter lesion volume, and fractional anisotropy, to good reliability for mean diffusivity and cortical thickness. In secondary screening, the anomalies due to data collection/processing errors were eliminated. A subgroup of the remaining anomalies were radiologically reviewed, and a substantial percentage of them (UKB: 90.1%; HCP: 42.9%) had various brain pathologies such as masses, cysts, white

matter lesions, infarcts, encephalomalacia, or prominent sulci. The remaining anomalies of the subgroup had unexplained causes and would be interesting for follow-up. Finally, we show that anomaly detection applied to resting-state functional connectivity did not identify any reliable anomalies, which was attributed to the confounding effects of brain-wide signal variation. Together, this study establishes an unsupervised framework for investigating rare individual imaging phenotypes within large heterogeneous cohorts.

Keywords: Machine learning; Big data; Multimodal MRI; Individual-level analysis; Radiological findings.

Abbreviations

VV: ventricular volume

WMLV: white matter lesion volume

FA: fractional anisotropy

MD: mean diffusivity

CTh: cortical thickness

RSFC: resting-state functional connectivity

UKB: UK Biobank

HCP: Human Connectome Project

SD: standard deviation

Introduction

Identifying image-based biomarkers for neurological and psychiatric disorders has been an important goal of neuroimaging. A common approach is to recruit diagnosed patients for assembling clinically defined cohorts. This strategy helps identify biomarkers for disease progression, but searching for pre-symptomatic biomarkers needs to image individuals before disease onset (Miller et al., 2016). Therefore, there is growing interest in collecting large pre-symptomatic cohorts to help search for relevant imaging biomarkers over a broad range of diseases. For example, UK Biobank (UKB) is enrolling 500,000 subjects 40-69 years of age for extensive phenotyping and subsequent long-term monitoring of health outcomes (Allen et al., 2012). One hundred thousand subjects in this cohort will also be imaged by MRI, making it the largest multimodal MRI cohort in the world (Littlejohns et al., 2020).

Given the large sample size, the UKB cohort enables a unique opportunity to discover rare brain imaging phenotypes. These rare imaging phenotypes are of interest because they are often clinically relevant and could also be informative for discovering new processes and mechanisms. These rare observations are only expected to constitute a very small portion of the dataset. Rare observations are quantitative imaging phenotypes that differ by a large amount from most other observations around the group average. They are defined as phenotypes that are many standard deviations (SD) away from the average. Identifying them is often commonly referred to as “anomaly detection”, or “outlier detection” (Tan et al., 2006). Anomaly detection in neuroimaging has been applied as a data cleaning method to remove artefactual observations. For example, the built-in outlier detection feature in the FSL package is used to identify motion-corrupted

functional MRI timepoints via one of the unidimensional motion metrics (Jenkinson et al., 2012). More sophisticated machine learning algorithms such as one-class support vector machine, Gaussian process regression, and autoencoders have been used to identify deviations of healthy individuals or groups of diagnosed patients' from the normal subjects (Marquand et al., 2016; Mourao-Miranda et al., 2011; Pinaya et al., 2019; van Hespen et al., 2021). These studies usually relied on a relatively small cohort using only one imaging phenotype, making them difficult to generalize. Notably, the methods used in these studies were either supervised or semi-supervised, which requires diagnostic labels for all subjects (supervised) or at least the labels of normal subjects (semi-supervised) in advance (Goldstein and Uchida, 2016). However, diagnostic labels are not always available for imaging cohorts designed to capture pre-symptomatic participants such as the UKB, making these approaches challenging to implement.

To address these issues and to identify rare imaging phenotypes in individual subjects, in the present study, we developed a two-level anomaly screening methodology that was applied to the UKB cohort of about 19,000 individuals. The same method was also applied to the Human Connectome Project (HCP) cohort of about 1,000 individuals. The HCP cohort, composed of healthy young adults aged 22-37, was used as a supplement to the UKB cohort, in which the latter includes much older subjects with different and often undetermined pathologies. Both cohorts have had the datasets curated with established procedures to review and decide on the inclusion of individual brain imaging data (Alfaro-Almagro et al., 2018; Glasser et al., 2013). Thus, these datasets should be of high quality, contain relatively few acquisition/processing errors, and give a wide range of ages to detect anomalies and to infer their causes. Here we made use of

the multimodal MRI data to derive ventricular, white matter, and gray matter-based imaging phenotypes of the brain (Fig. 1a). The first step was to parameterize each subject with an "anomaly score" per imaging phenotype in an unsupervised manner without any prior labels (Fig. 1b). This anomaly score quantifies how far an individual deviates from most other subjects. Anomaly subjects were defined as having an anomaly score greater than 3, which is equivalent to 4.7 times the SD above the average for a standard normal distribution. The robustness of anomaly scores for each imaging phenotype was examined. Anomaly score reliability was characterized in the subjects that had repeat MRI scans. Correlations of anomaly scores between different imaging phenotypes were also evaluated. The next step was to validate these anomalies (Fig. 1c). The anomalies were categorized according to whether there were data collection/processing errors, or whether the individual had positive radiological findings determined by a board-certified neuroradiologist. Finally, some individuals were considered to be novel anomalies because there was no specific reason to explain their large deviations from the average.

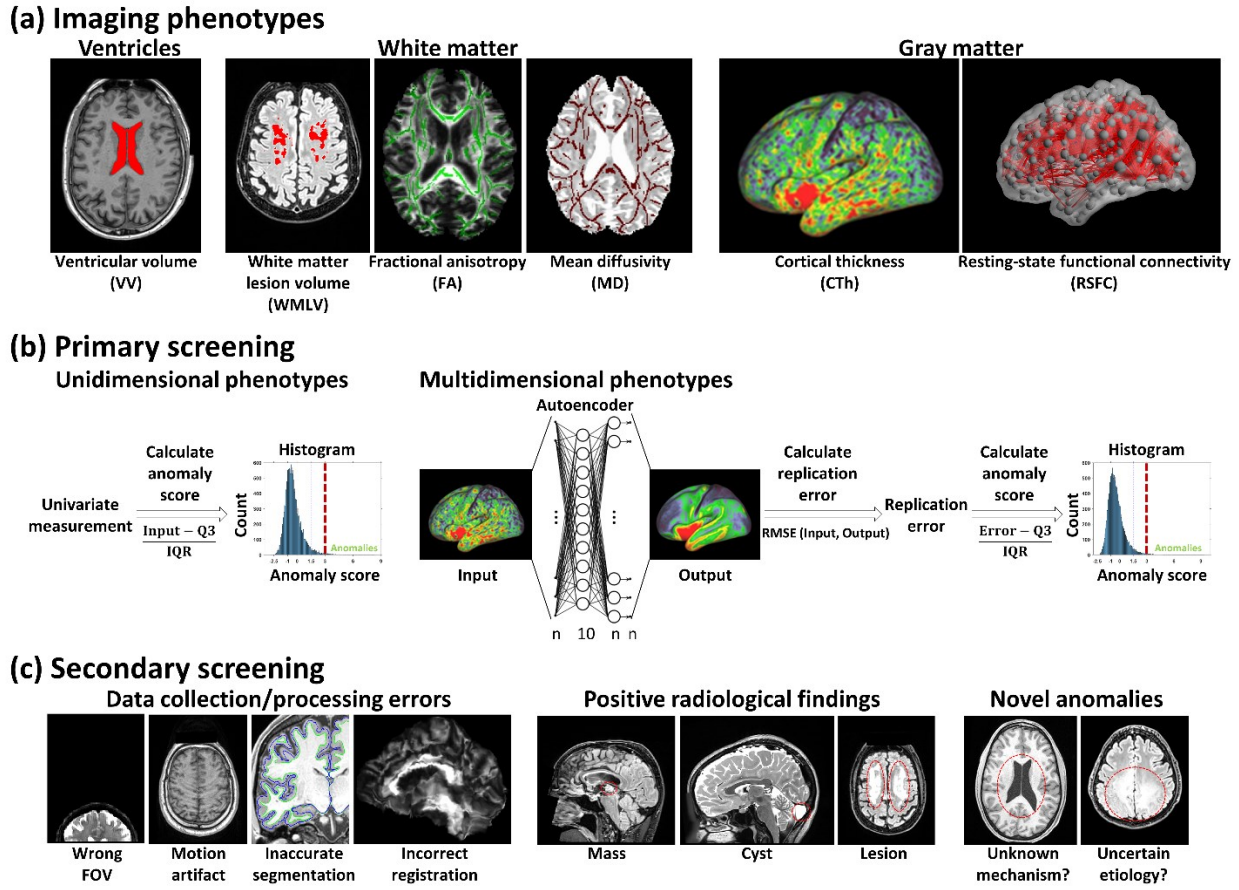


Fig. 1. Schematic illustration of the anomaly detection pipeline. (**a**) Brain imaging phenotypes used for anomaly detection. (**b**) Primary screening. For a unidimensional phenotype (VV, WMLV), the anomaly score of an individual was obtained from the volume measurement by subtracting the third quartile (Q3) and dividing the difference by the interquartile range (IQR) of the volume measurements. For a multidimensional phenotype (FA, MD, CTh, RSFC), an autoencoder was trained to replicate the input at its output. The input was a matrix (matrix size = dimensionality of the given phenotype × number of subjects). As an example, for display purposes, a grayordinate-wise CTh map, which was a column from the input matrix for CTh anomaly detection, is shown at the input. The subject-specific replication errors were measured by the root mean square errors (RMSE) between each input and the replicated output. The anomaly score of an individual was obtained from this subject-specific replication error by subtracting Q3 and dividing the difference by IQR of the replication errors. The subjects with an anomaly score > 3 were considered as anomaly subjects. (**c**) Secondary screening. The anomalies identified in primary screening were first checked for association with data collection/processing errors (left column). For the remaining anomalous subjects without the errors, a subgroup was reviewed by a board-certified neuroradiologist (middle column). The remaining anomalous subjects without positive radiological findings were considered as novel anomalies (right column).

Materials and Methods

Datasets

Brain imaging data were obtained from two cohorts: the UKB (Miller et al., 2016) (https://www.ukbiobank.ac.uk) and HCP (Van Essen et al., 2013) (Young Adult, https://www.humanconnectome.org). For the UKB cohort, the initial imaging visit data of 19411 subjects (9172 males and 10239 females; age 44-80) were used in the present study (UKB "discovery" group). If available (not marked as "unusable" or "incompatible" by the UKB), each UKB subject's T1w (3D magnetization-prepared rapid acquisition gradient echo [MPRAGE]) and T2w FLAIR structural MRI, spin echo (SE) echo-planar imaging (EPI) diffusion MRI (dMRI), and gradient echo (GE) EPI rsfMRI data were used. For the HCP cohort, the 3T data from the 1200 Subjects Release (1113 subjects: 550 males and 656 females; age 22-37) were used in the present study. If available, each HCP subject's T1w (3D-MPRAGE) and T2w (3D sampling perfection with application-optimized contrast using different flip-angle evolutions [SPACE]) structural MRI, SE-EPI dMRI, and GE-EPI rsfMRI data were used. For both cohorts, some subjects only had usable structural MRI data, resulting in a reduced sample size of dMRI and rsfMRI data. For the HCP rsfMRI data, a smaller group of 795 subjects using an improved image reconstruction algorithm "r227" was used. The detailed demographic information is summarized in Table S1.

The UKB data were acquired on identical 3T Siemens Skyra MRI scanners, and the HCP data were acquired on a 3T Siemens Connectome Skyra MRI scanner. The detailed UKB and HCP data acquisition protocols can be found elsewhere (Alfaro-Almagro et al., 2018; Glasser et al., 2013). The UKB was approved by the North West

Multi-centre Research Ethics Committee. The HCP project was approved by the Institutional Review Board of Washington University. For each cohort, informed consent was obtained from all participants. The present study was approved by the Office of Human Subjects Research Protections at the National Institutes of Health (ID#: 18-NINDS-00353).

Image preprocessing and extraction of imaging phenotypes

The following imaging phenotypes were extracted from imaging preprocessing outputs: ventricular volume (VV), white matter lesion volume (WMLV), fractional anisotropy (FA), mean diffusivity (MD), cortical thickness (CTh), and resting-state functional connectivity (RSFC). The detailed procedures are described as follows.

T1w MPRAGE and T2-FLAIR images (T2w SPACE images if from the HCP cohort) were preprocessed by the HCP structural pipeline (v4; https://github.com/Washington-University/HCPpipelines) (Glasser et al., 2013) based on FreeSurfer (v6.0.0; https://surfer.nmr.mgh.harvard.edu) (Fischl, 2012). To obtain the ventricular segmentation from the subjects without usable T2-FLAIR images or the subjects who failed the HCP pipeline, these subjects' data were preprocessed with FreeSurfer v6.0.0 directly using their T1w images (one subject failed FreeSurfer v6.0.0 was reprocessed with FreeSurfer v7.1.0). Also, the subjects with large segmentation defects in their enlarged ventricles were reprocessed with "-bigventricles" flag in FreeSurfer v6.0.0 directly using their T1w images. The quality of ventricular segmentation was manually inspected. Each subject's VV was calculated by summing up the volumes of lateral ventricles, temporal horns of the lateral ventricles, choroid plexuses, third ventricle, and

fourth ventricle. WMLV In the UKB cohort was calculated by the Brain Intensity Abnormality Classification Algorithm (BIANCA) (Griffanti et al., 2016), a k-nearest-neighbor-based automated supervised method, using T2-FLAIR images but also T1w images as its inputs. Because of the lack of HCP T2-FLAIR data, WMLV In the HCP cohort was obtained from the volumes of T1w white matter hypointensities segmented by FreeSurfer, which uses probabilistic information estimated from a built-in set of manually segmented images (Fischl et al., 2002). For both cohorts, CTh values in the standard CIFTI grayordinate space (with folding-related effects corrected) from only the subjects preprocessed successfully by the HCP structural pipeline, were used for primary screening.

In both cohorts, dMRI data underwent FSL eddy-current and head-movement correction (Andersson and Sotiropoulos, 2016), gradient distortion correction, diffusion tensor model fitting using the b = 1000 shell (Basser et al., 1994), and Tract-Based Spatial Statistics (TBSS) analyses (Smith et al., 2006). The TBSS skeletonized images were averaged within the ROIs of the John Hopkins University white matter atlas (Mori et al., 2008). Here the original MD values were multiplied by 10000 to convert to the unit of 10^{-4} mm^2/s. The FA or MD maps of 27 major white matter ROIs (Table S2) were used for primary screening.

UKB rsfMRI data were preprocessed by the UKB rsfMRI pipeline (v1; https://git.fmrib.ox.ac.uk/falmagro/UK_biobank_pipeline_v_1) (Alfaro-Almagro et al., 2018), and the ICA + FIX denoised data (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014) were brought to the HCP standard surface space using Ciftify (v2.3.2; https://github.com/edickie/ciftify) (Dickie et al., 2019). For each subject, the SD of percent

change time series of each grayordinate was calculated, and the grayordinates with this SD greater than 0.1 were considered as noisy grayordinates. These noisy grayordinates were masked from further analyses. HCP rsfMRI data were preprocessed by HCP functional pipeline (v3) (Glasser et al., 2013) and were also denoised by ICA + FIX. The two runs (left-to-right and right-to-left phase encoding directions) of the same session were demeaned, variance normalized, and then concatenated temporally, so each HCP subject had two sessions of preprocessed rsfMRI data. Using a well-established RSFC-based parcellation scheme (333 parcels) (Gordon et al., 2016), RSFC was quantified by the Pearson cross-correlation coefficient between the ROI-averaged time series of each pair of parcels, with or without global signal regression, respectively. RSFC was also quantified using partial correlations with Tikhonov regularization (UKB: ρ = 0.5; HCP: ρ = 0.01) (Pervaiz et al., 2020). Due to the symmetry of the RSFC matrices, the upper triangular parts of these matrices (333*332/2 = 55278 elements) from each of these three RSFC evaluation methods were used for primary screening respectively.

Primary screening

Two-level anomaly screening was performed for each dataset separately. In primary screening, each imaging phenotype was screened separately. In a given imaging phenotype, every subject was parameterized with an anomaly score. This anomaly score quantified the degree of abnormality in that imaging phenotype.

For a unidimensional imaging phenotype (VV, WMLV), using VV as an example, the anomaly score of an individual was transformed from the VV value of this individual:

$$Anomaly\ score = \frac{VV - Q3}{IQR} \quad (1)$$

where *Q3* was the third quartile of the VV distribution of all subjects, and *IQR* was the interquartile range of this distribution. The anomaly score of the other unidimensional imaging phenotype, WMLV, was calculated similarly.

For each multidimensional imaging phenotype (FA, MD, CTh, RSFC), an autoencoder was used to calculate the anomaly scores (Hawkins et al., 2002). Setting the dimensionality of the imaging phenotype as M and the number of subjects in a cohort as N, the inputs to the autoencoder were the values of that imaging phenotype across the whole cohort (M * N), and the autoencoder was trained to replicate this input at its output. Here, this autoencoder was comprised of an input layer (M dimensions), a hidden layer of 10 neurons, and an output layer (M dimensions). A sparsity proportion of 0.05 was used, and the sparsity regularization coefficient was set to 1. The L2 weight regularization coefficient was set to 0.001. The sigmoid function was used as the activation function, and the mean squared error function adjusted for sparse autoencoder was used as the loss function. A scaled conjugate gradient descent algorithm (Moller, 1993) was used for training this autoencoder. Regional deviation values (M * N) were calculated by subtracting the autoencoder-predicted output from input, and the value in the i^{th} row and j^{th} column of this matrix characterizes deviations from the value predicted by the autoencoder. For each subject, these values can be plotted in the white matter ROIs (FA or MD) or in grayordinates on the brain surface (CTh), respectively, to visualize regional deviations. The subject-specific replication errors (also known as “reconstruction error” in the context of autoencoder) were measured by the root mean square errors between each input and the replicated output. The anomaly score of an individual was obtained by transforming this subject-specific replication error:

$$Anomaly\ score = \frac{error - Q3}{IQR} \quad (2)$$

where *Q3* was the third quartile of the replication error distribution of all subjects, and *IQR* was the interquartile range of this distribution. The autoencoders were implemented using the 'trainAutoencoder' function in the MATLAB and were trained using a GPU cluster (https://hpc.nih.gov). Multiple autoencoders were used for an imaging phenotype when the input (UKB CTh, UKB RSFC) was too large to fit into the GPU memory: In these scenarios, the input data were split into 9 to 10 smaller groups in a stratified manner, which preserved the ratio of age and sex in each group. For each group, an autoencoder was trained using the data of that group as the input. The trained autoencoders were then applied to the full dataset and the output of the cohort was obtained by averaging the outputs from each of these autoencoders. For HCP RSFC, because each subject had two sessions, the RSFC data of the first sessions were used to train the autoencoder.

In the above analyses, to control the effects of two covariates (age, brain volume) on anomaly detection, their correlations with VV, WMLV, and the autoencoder replication errors of multidimensional imaging phenotypes were evaluated. The covariates with correlation >0.3 were regressed out from VV, WMLV, or the replication errors before applying Eq. (1) or (2). Therefore, brain volume and age were regressed out from UKB VV, but only brain volume was regressed out from HCP VV (Fig. S1). Age was regressed out from UKB WMLV, and brain volume was regressed out from HCP WMLV (Fig. S1).

The outlying subjects only comprised a small portion of the cohort used for training the autoencoder, therefore the trained autoencoder cannot replicate these rare anomalies as well as the commonly seen normal subjects. This contributed to the larger replication errors and subsequently larger anomaly scores for the outlying subjects. In statistics, Q3

+ 3 * IQR is commonly used to define extreme outliers in that distribution (Tukey, 1977). This is equivalent to an anomaly score of 3. Here, ((Q3 + 3 * IQR) – Q3)/IQR = 3. Therefore, subjects with an anomaly score of greater than 3 were considered anomalies in primary screening.

Secondary screening

In secondary screening, the anomaly subjects identified in primary screening were first checked to see if the anomalies were associated with data collection/processing errors.

For each VV anomaly subject, ventricle segmentation quality was visually inspected by overlaying the border of the segmented ventricle mask on the T1w image. For each WMLV anomaly subject, white matter lesion segmentation quality was visually inspected by overlaying the border of the segmented lesion mask on the T2-FLAIR image (T1w image if from the HCP cohort). The anomaly subjects with incorrect segmentation were deemed to be associated with data collection/processing errors.

For the subjects with usable dMRI data, the dMRI motion parameters (*.eddy_restricted_movement_rms) were calculated by FSL's eddy tool (Andersson and Sotiropoulos, 2016), and the head motion of each subject was summarized by the mean and largest values of the volumetric movements between adjacent frames. The subjects with at least one of these two summary parameters above the upper inner fence (Q3 + 1.5 * IQR, commonly used to define mild outliers in statistics (Tukey, 1977)) of the cohort distribution were flagged with severe head motion. The registration quality was assessed by each subject's mean deformation of the TBSS nonlinear registration, and the subjects

with this parameter above the upper inner fence (Q3 + 1.5 * IQR) of the cohort distribution were flagged with bad registration. The FA or MD anomaly subjects were also visually checked for registration quality and FOV coverage. The FA or MD anomaly subjects with incorrect FOV coverage, bad head motion, or bad registration were deemed to be associated with data collection/processing errors.

For CTh, volume registration quality was quantified by the number of suprathreshold voxels in the Jacobian map of nonlinear registration. Surface registration quality was quantified by areal and shape distortion maps of folding alignment (MSMSulc) surface registration (Robinson et al., 2018). Using the aforementioned multidimensional anomaly detection method and these distortion maps as inputs, an anomaly score for the areal distortion map and an anomaly score for the shape distortion map were calculated for each subject. T1w/T2w ratio myelin maps (Glasser and Van Essen, 2011) were further used to detect potential surface segmentation issues that could be caused by the subject's anatomy, and an anomaly score for the T1w/T2w ratio myelin map was calculated for each subject via multidimensional anomaly detection. White/pial surface segmentation quality of the CTh anomaly subjects was checked via HCP pipeline structural quality control scenes (https://github.com/Washington-University/StructuralQC; v1.4.0), and the CTh anomaly subjects with poor surface segmentation were flagged. T1w structural images of the CTh anomaly subjects were also inspected visually, and the subjects with visible motion artifacts such as ringing artifacts were flagged. Taken together, the CTh anomaly subjects with bad volume or surface registration quality, anomalous T1w/T2w ratio map, bad white/pial surface segmentation, or visible motion

artifacts in T1w images were deemed to be associated with data collection/processing errors.

For RSFC, because the global signal has been recognized as a controversial confounding factor in rsfMRI data processing (Liu et al., 2017), the relationship between RSFC anomaly score and the global signal was assessed for each RSFC evaluation method. In previous studies of volume-based analyses (Wong et al., 2016; Wong et al., 2013), percent change time series at each voxel was calculated by dividing the demeaned preprocessed rsfMRI time series by the mean, and a global mean percent change time series was obtained by averaging percent change time series across all brain voxels. The SD of this global mean percent change time series was defined as the global signal amplitude in these volume-based analyses. In the present study of surface-based analyses, a percent change time series was calculated at each grayordinate instead by dividing the demeaned ICA-FIX denoised CIFTI time series by the mean, and a global mean percent change time series was obtained by averaging percent change time series across all cortical grayordinates. The SD of this grayordinate-based global mean percent change time series was defined as the global signal amplitude in the present study.

The anomaly subjects without data collection/processing errors were radiologically reviewed. T1w MPRAGE and T2-FLAIR images (T2w SPACE images if from the HCP cohort), as well as subjects' age, were provided for a board-certified neuroradiologist (D.S.R.) for radiological review. The instructions to the neuroradiologist were to identify major findings that might plausibly account for the outlying anomaly score — not to identify subtle abnormalities that would have required dedicated review on clinical-grade display systems. When the neuroradiologist had any uncertain diagnoses, UKB health outcomes

data (UKB Category 1712), which recorded the first occurrence of various diseases, including neuropsychiatric and neurological disorders, were used in an attempt to determine the diagnoses. The anomaly subjects with findings of uncertain etiology, or without any positive radiological findings were considered as "novel" anomalies. VV anomalies of big ventricles but without any other noticeable pathology were also considered as novel anomalies. WMLV anomalies were considered as "novel" only when the distribution of white matter lesions was atypical.

Evaluation of robustness of anomaly scores

The initial imaging visit data of another 19350 subjects (9005 males and 10345 females; age 47-82) were used in the present study (UKB "replication" group; see Table S3 for detailed demographic information) to evaluate the robustness of anomaly scores. This group had no overlapping subjects with the UKB discovery group.

For VV anomaly score, because it's unidimensional, its robustness was assessed by directly comparing the VV distribution of the discovery group against the distribution of the replication group via a two-sample Kolmogorov-Smirnov test. The robustness of WMLV anomaly score was assessed similarly.

For FA anomaly score, first, the robustness of the discovery group subjects' anomaly scores was evaluated by the intraclass correlation (Shrout and Fleiss, 1979) (ICC) between two sets of their anomaly scores calculated separately using two different autoencoders: one autoencoder trained using the discovery group itself, and another autoencoder trained using the replication group. Second, the robustness of the replication group subjects' anomaly scores was evaluated by the ICC between two sets of their

anomaly scores calculated separately using two different autoencoders: one autoencoder trained using the replication group itself, and another autoencoder trained using the discovery group. For each of these ICCs, a one-way random effects model was used:

$$ICC(1,1) = \frac{MSb - MSw}{MSb + (k-1)MSw} \quad (3)$$

where *MSb* is the between-subject mean square, *MSw* is the within-subject mean square, and *k* is the number of observations per subject (McGraw and Wong, 1996). The robustness of the anomaly scores of other multidimensional imaging phenotypes was assessed similarly.

Evaluation of reliability of anomaly scores

A subgroup (1427 subjects) of the UKB discovery group subjects had a repeat MRI session (aka “retest”) two to three years after the initial imaging visit (aka “test”). The test and retest data of these subjects were used to evaluate long-term reliability of anomaly scores. Short-term (~1 day) reliability of RSFC anomaly score was also assessed using the two sessions of the HCP rsfMRI data. For each unidimensional imaging phenotype, unlike the primary screening, their measurements in the reliability analysis were no longer adjusted for covariates. The Q3 and IQR were calculated from the full test data and applied to calculate anomaly scores for both test data and, for subjects who were scanned twice, retest data. For each multidimensional imaging phenotype, anomaly scores of the test visit calculated in the primary screening were used directly. To calculate the anomaly scores of the retest visit, the autoencoders that were trained on the full test data were applied to the retest data. The reliability was quantified by ICC between the anomaly scores of the test and retest data using Eq. (3). Reliability was defined as excellent (ICC >

0.8), good (0.6 < ICC < 0.8), moderate (0.4 < ICC < 0.6), fair (0.2 < ICC < 0.39), or poor (ICC < 0.2) (Guo et al., 2012) in the present study.

Evaluation of the relationships between anomaly scores of different imaging phenotypes

In the UKB discovery group, the relationships between anomaly scores of different imaging phenotypes were quantified using Pearson cross-correlation coefficients. The anomalies due to data collection/processing errors were excluded from this analysis. Two representative relationships of anomaly scores, WMLV versus VV, and WMLV versus FA, were also visualized using scatterplots. In each scatterplot, three zones were defined to categorize anomaly subjects. For WMLV versus VV, zone I covered the subjects who were VV anomalies but with normal WMLV (WMLV anomaly score < 1.5), zone II covered the subjects who were both VV and WMLV anomalies, and zone III covered the subjects who were WMLV anomalies but with normal VV (VV anomaly score < 1.5). The density of subjects in each zone was calculated by dividing the number of subjects by the area of the zone as follows:

$$Density_{zone\ I} = \frac{Number\ of\ subjects\ in\ Zone\ I}{(1.5-\min(WMLV\ anomaly\ score))*(\max(VV\ anomaly\ score)-3)} \quad (4)$$

$$Density_{zone\ II} = \frac{Number\ of\ subjects\ in\ Zone\ II}{(\max(WMLV\ anomaly\ score)-3)*(\max(VV\ anomaly\ score)-3)} \quad (5)$$

$$Density_{zone\ III} = \frac{Number\ of\ subjects\ in\ Zone\ III}{(\max(WMLV\ anomaly\ score)-3)*(1.5-\min(VV\ anomaly\ score))} \quad (6)$$

To evaluate the differences in densities across the three zones, a bootstrap procedure with replacement on subjects was used to generate 100,000 bootstrap samples of the original sample size. For each bootstrap sample, the density of each zone

was re-computed. A one-way ANOVA was then performed to evaluate the differences across the zones using the bootstrap samples. Similar analyses were also carried to evaluate the relationship between WMLV and FA anomaly scores.

Results

Anomaly detection was successfully performed for the following brain imaging phenotypes: VV, WMLV, FA, MD, and CTh, respectively. The robustness, distribution properties, and reliability of anomaly scores were evaluated. Individual anomalous patterns were examined for each imaging phenotype. Anomaly score relationships across these imaging phenotypes were assessed. In addition to these imaging phenotypes, anomaly detection was also performed for RSFC, but no anomalies that withstood the retest session were identified. One possible cause was the confounding effects of the global signal change. For this reason, RSFC results are reported separately as a failure example for detecting reliable anomalies.

Robustness of anomaly scores

The robustness of anomaly scores was tested using group comparisons between two large groups, a discovery group and a replication group. These two groups were of comparable size from the UKB cohort and had no overlapping subjects. For each unidimensional imaging phenotype, the anomaly score distribution of the discovery group was highly similar to that of the replication group (Fig. S2ab), and there was no significant difference between these two distributions (two-sample Kolmogorov-Smirnov tests: for VV, $p = 0.36$; for WMLV, $p = 0.66$). For each multidimensional imaging phenotype, the

discovery group subjects had highly reproducible anomaly scores regardless of whether the discovery group or the replication group was used for training the autoencoder. The ICC between the anomaly scores calculated using the autoencoder trained with either one of the two groups ranged from a lowest of 0.86 in MD, to a highest of 0.99 in CTh is 0.98 (Fig. S2c-e). This also held for the anomaly scores of replication group subjects (ICC ranged from 0.90 to 0.98. Fig. S2c-e). Taken together, these results indicate that for each of these imaging phenotypes, anomaly scores were highly robust.

Properties of anomaly score distributions

The results presented throughout the rest of the manuscript were obtained using the UKB discovery group unless otherwise specified. The anomaly score histogram of each imaging phenotype is shown in the panels of Figs. 2 (VV, WMLV, FA, CTh) and S3a (MD), respectively. These distributions were all right-skewed and more leptokurtic than a standard normal distribution (see Table 1 for skewness and kurtosis values). In statistics, the third quartile plus three times the interquartile range (Q3 + 3 * IQR) of the distribution is commonly used to define extreme outliers (Tukey, 1977). This criterion was adopted in the present study to find anomalies, which corresponds to an anomaly score threshold of 3. Based on this threshold, the percentage of anomalies ranged from a lowest of 0.6% in CTh, to a highest of 3.9% in WMLV (see Table 1 for details). These anomaly percentages are all much higher than a standard normal distribution predicts, because this threshold is equivalent to about 4.7 times the SD plus the mean in a standard normal distribution, which would only have 0.0001% of data above it. In addition, as a negative control, the anomaly score distributions for the HCP cohort (Fig. S4), composed of healthy young

adults, were also evaluated, and they were less right-skewed and less leptokurtic than the UKB cohort, as indicated by the lower skewness, kurtosis, and anomaly percentage values in the HCP cohort (Table S4). Taken together, the results suggest that the anomaly score distributions of the UKB cohort were all more outlier-prone than a standard normal distribution, and they were also more outlier-prone than the healthy young cohort of the HCP.

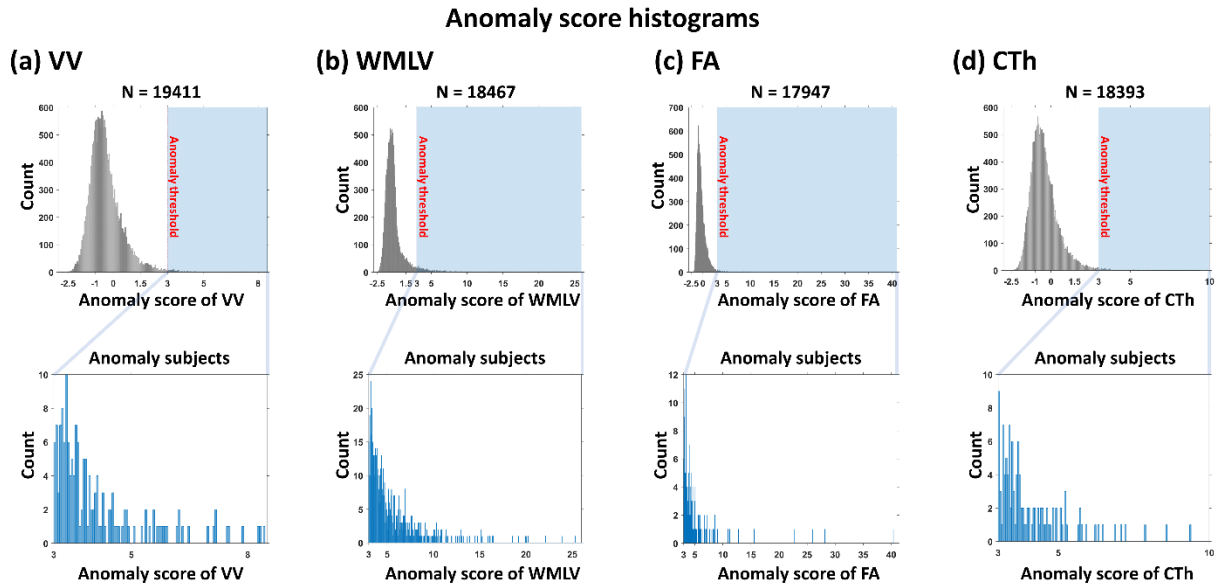


Fig. 2. Anomaly score histograms. (**a**) VV. (**b**) WMLV. (**c**) FA. (**d**) CTh. The zoom panels on the second row show the histograms of anomaly subjects (anomaly score > 3).

Table 1. Summary of anomaly subjects in the UKB discovery group.

Phenotype		VV	WMLV	FA	MD	CTh
Number of Subjects		19411	18467	17947	17947	18393
Skewness		1.77	4.36	7.54	9.04	1.69
Kurtosis		10.06	37.27	191.12	178.31	9.54
Number of Anomalies		158 (0.8%)	716 (3.9%)	189 (1.1%)	258 (1.4%)	119 (0.6%)
Anomalies w/o data issue		158	645	128	177	5
Anomalies read by neuroradiologist		39	63	37	38	5
Summary of radiological review results	Large ventricles	36	18	9	9	1
	White matter lesions	26	63	29	33	2
	Mass	2	1			
	Cyst	4	1	1	2	
	Infarct	6	16	9	12	1
	Encephalo-malacia			3	3	
	Prominent sulci	2	1	3	1	4
	Other findings	4	9	11	8	
	Normal			2		

Note: Empty entries are zeros.

Long-term test-retest reliability of anomaly scores

A subgroup of the discovery group subjects had a repeat MRI session two to three years after the initial visit. The anomaly scores of test versus retest of each imaging phenotype are visualized in the scatterplots of Figs. 3 (VV, WMLV, FA, CTh) and S3b (MD), respectively. VV anomaly scores had excellent test-retest reliability, as indicated by the close-to-one value of anomaly score ICC (ICC = 0.98) between test and retest. The reliabilities of WMLV and FA anomaly scores were lower than VV but still excellent (WMLV ICC = 0.82; FA ICC = 0.87). The reliabilities of MD and CTh anomaly scores were

lower than the former three but still in the range of good reliability (MD ICC = 0.75; CTh ICC = 0.62). This can also be seen in the distributions of test-retest anomaly score change. Fig. S5a shows the distribution fits of anomaly score change of each imaging phenotype. These distributions were near-symmetrical and centered around zero, indicating the means of anomaly score changes were near zero. Larger dispersion of an anomaly score change distribution indicates the anomaly scores changed more in the test-retest and thus had lower reliability. Indeed, the dispersions (Fig. S5b) were consistent with the ICC analysis results. Taken together, the results suggest anomaly scores had good-to-excellent long-term, test-retest reliability.

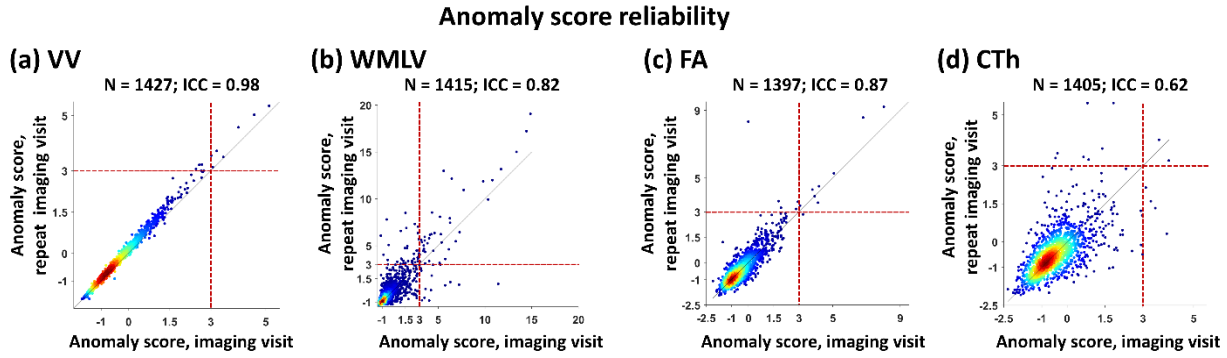


Fig. 3. Long-term test-retest reliability of anomaly scores. (**a**) VV. (**b**) WMLV. (**c**) FA. (**d**) CTh. In each scatterplot, each subject's anomaly score of the initial imaging visit (aka "test"; year 2014+) is plotted against the anomaly score of the first repeat imaging visit (aka "retest"; year 2019+). The UKB subjects that had both test and retest data available are shown in these scatterplots. ICC: intraclass correlation between anomaly scores of the two visits. Red dashed line: anomaly threshold (anomaly score = 3).

Summary percentages of anomalies

The total number of anomalies in the UKB discovery group across all imaging phenotypes (excluding RSFC) was 1440. Because there were anomaly subjects who were anomalies in more than one imaging phenotype (Fig. S11c), of these 1440

anomalies, there were 1110 distinct subjects. Eight hundred ninety-nine (899/19411, 4.6%) of them were not associated with data collection/processing errors. One hundred eleven (111) of these 899 subjects were reviewed by a neuroradiologist (Fig. S6). The subjects with the most extreme anomaly scores not caused by data issues were all included (Fig. S6). Ninety and one-tenths percent (90.1%, 100/111) of these 111 subjects had positive radiological findings.

In comparison, there were 21 distinct anomaly subjects in the HCP cohort, and each of them was anomalous in only one imaging phenotype. Fourteen (14/1113, 1.3%) of them were not associated with data collection/processing errors. All these 14 subjects were read by a neuroradiologist, and 42.9% (6/14) of them had positive radiological findings.

Representative individual anomaly subjects are reported in the next few subsections per their imaging phenotype.

Individuals with anomalous ventricular volume

As an example, Fig. 4a shows a VV anomaly subject versus a normal subject. This subject had significantly enlarged lateral ventricles compared to the latter one (~8.2 difference in anomaly score). None of the VV anomalies were associated with data collection/processing errors. Thirty-nine of the UKB VV anomalies were reviewed by a board-certified neuroradiologist. A substantial percentage (31/39, 79.5%) of the UKB VV anomalies being read were identified with positive radiological findings of different brain pathologies. The major pathologies identified were mass, cyst, infract, and white matter lesions (Table 1), and some were directly linked to ventriculomegaly. For example, a third

ventricle mass (possibly a choroid plexus papilloma), a fourth ventricle mass (possibly an ependymoma), and a colloid cyst, all causing obstructive hydrocephalus, were found in three VV anomaly subjects (Fig. 4b). Other examples include a frontoparietal arachnoid cyst (Fig. 4b), a mega cisterna magna, an infarct, intraventricular nodules, and partial agenesis of the corpus callosum (Fig. S7a). The remaining VV anomalies that were read (UKB: 8/39, 20.5%) were referred to as the “novel” anomalies. In these cases, they had either large ventricles with the pathology of uncertain etiology (Fig. S7b), or large ventricles without any noticeable pathology (Figs. 4c, d, and S7c). In addition to the UKB subjects, a few interesting HCP anomaly subjects are also reported herein. HCP VV anomalies had monozygotic twins in this “novel” group (Figs. 4d and S7c). In one family (Fig. 4d), the female monozygotic twins were both VV anomalies, but their non-twin brother had normal VV. In another family (Fig. S7c), one twin of a male monozygotic twin pair was a VV anomaly, but the other twin and his non-twin brother both had normal VV. These twin data open the possibility of probing genetic and environmental causes underlying the anomalously large VV. Taken together, the results indicate VV anomalies were either associated with brain pathologies or were novel.

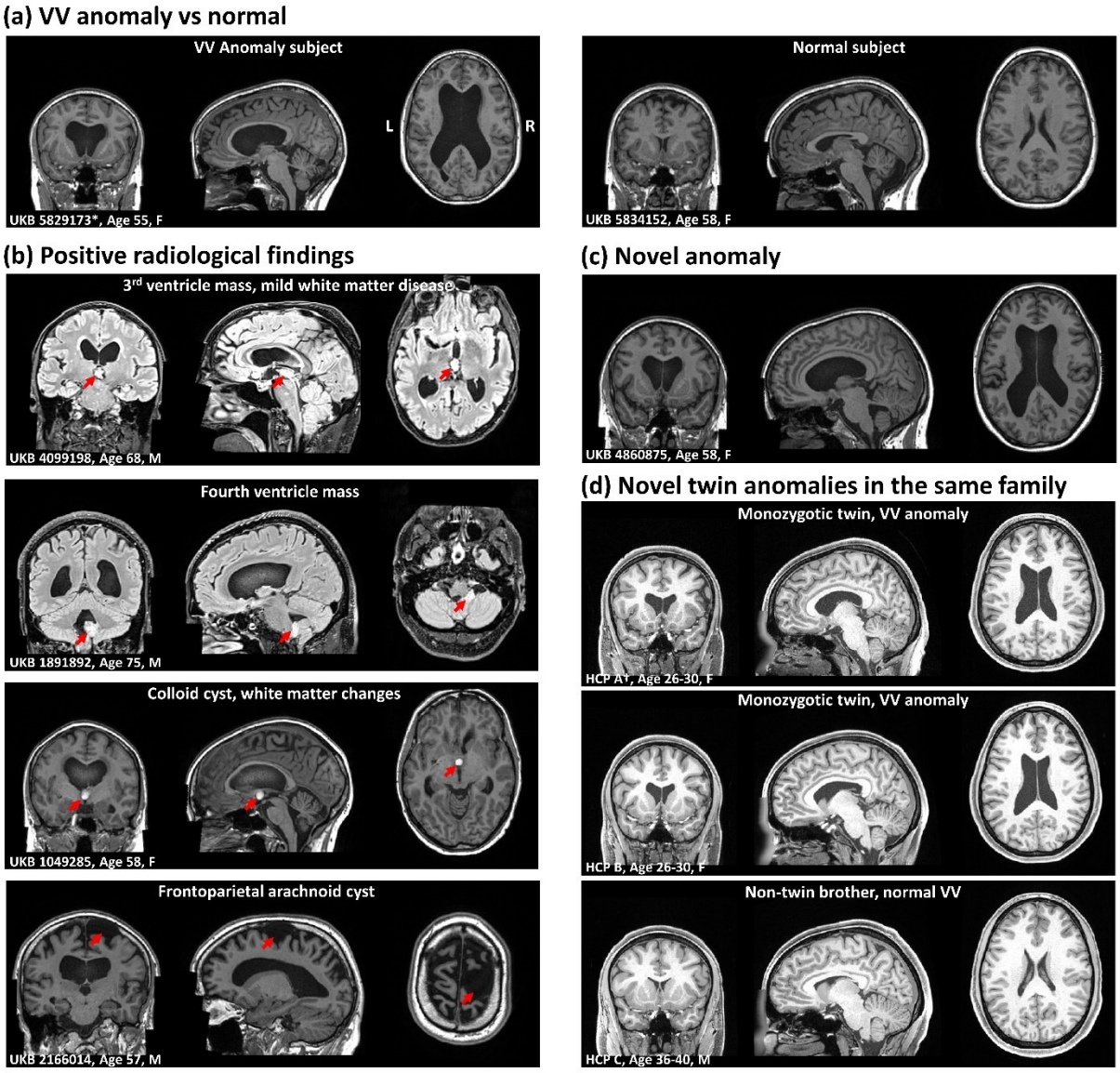


Fig. 4. Individuals with anomalous VV. (**a**) Structural images of an example of a VV anomaly subject (left column, anomaly score = 7.2) and an example of a normal VV subject (right column, anomaly score = -1.0). (**b**) Structural images showing positive radiological findings in four representative anomaly subjects (anomaly scores: UKB subject 4099198, 4.0; UKB subject 1891892, 8.3; UKB subject 1049285, 7.5; UKB subject 2166014, 8.3). (**c**) Structural images of a novel VV anomaly (anomaly score = 7.0). (**d**) Structural images of a family (monozygotic twins and their non-twin brother). The twins (first and second rows, anomaly scores: 3.8, 3.4) were novel VV anomalies, but their non-twin brother (third row, anomaly score: -0.5) had normal VV. *Note: UKB Subject IDs in this study were pseudonymized and unique to the UKB application 22875. A "bridging" tool could be used to relate these pseudonymized IDs to the UKB datasets supplied to other researchers (https://biobank.ndph.ox.ac.uk/showcase/help.cgi?cd=bridging). †Note: For HCP subjects, a key that maps their alphabet IDs to HCP-assigned numeric subject IDs will be available in the ConnectomeDB (https://db.humanconnectome.org) upon publication as per the HCP Restricted Data Use Terms.

Individuals with anomalous patterns of white matter-based imaging phenotypes

Anomaly detection on the white matter was performed with WMLV, FA, and MD, respectively. As an example, Fig. 5a shows a WMLV anomaly subject versus a normal subject (~26.5 difference in anomaly score). The anomaly subject had irregular periventricular white matter lesions extending into the deep white matter with large confluent areas, whereas an example normal subject had only tiny lesions on the periventricular caps. Fig. 5b shows regional FA deviation maps of an FA anomaly subject versus a normal subject (~9.5 difference in anomaly score). For this representative anomaly subject, regional FA negatively deviated in all 27 white matter ROIs used in this study, whereas the FA of the representative normal subject had almost no deviations. Fig. S3c shows regional MD deviation maps of an MD anomaly subject versus a normal subject (~5.5 difference in anomaly score), in which a large positive MD deviation was observed in the left superior longitudinal fasciculus of the anomaly subject.

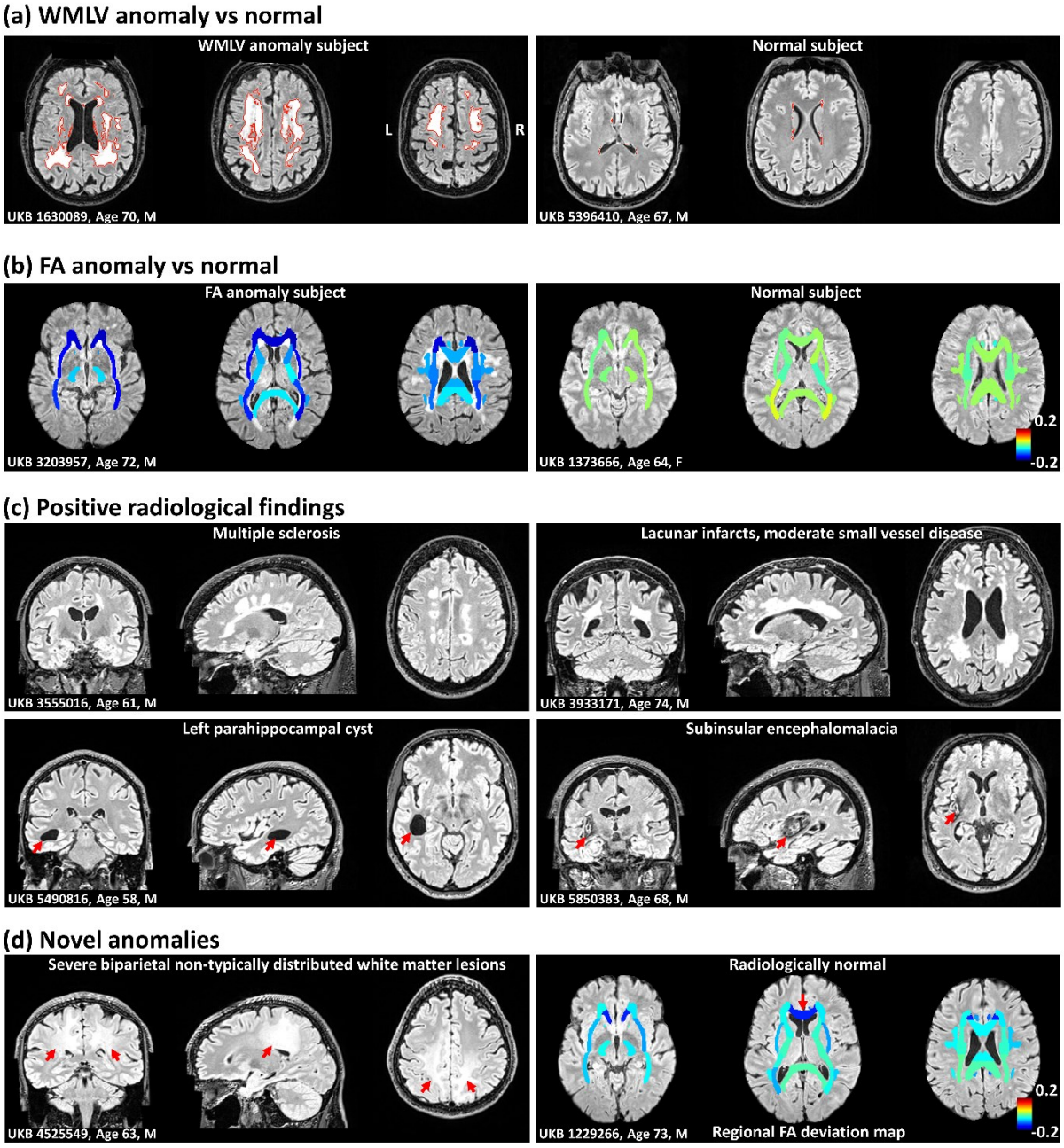


Fig. 5. Individuals with anomalous patterns of white matter-based imaging phenotypes. (**a**) T2 FLAIR images of an example of a WMLV anomaly subject (left column, anomaly score = 25.3) and an example of a normal WMLV subject (right column, anomaly score = -1.2). The red line represents the boundary of white matter lesion regions segmented using BIANCA. (**b**) Regional FA deviation maps (overlaid on T2 FLAIR images) of an example of an FA anomaly subject (left column, anomaly score = 8.5) and an example of a normal FA subject (right column, anomaly score = -1.0). (**c**) Structural images showing positive radiological findings in representative anomaly subjects of multiple sclerosis (anomaly scores: WMLV 7.7, FA 6.5, MD 10.1), lacunar infarcts with moderate small vessel disease (anomaly scores: WMLV 10.3, FA 5.6, MD 5.0), cyst (anomaly scores: MD 9.8), and encephalomalacia (anomaly scores: FA 3.5, MD 10.0). (**d**) Novel anomalies. Left column: T2 FLAIR images of an anomaly subject with severe biparietal non-typical distributed white matter lesions of uncertain etiology (anomaly scores: WMLV 16.4, FA 5.6, MD 10.1). Right column:

regional FA deviation map (overlaid on T2 FLAIR images) of an anomaly subject that was radiologically normal. The cause of anomalous FA was unknown (FA anomaly score: 3.2). An FA deviation map visualizes how the FA values in a subject deviate from the autoencoder-predicted FA values. For display purposes, in FA deviation maps, each white matter ROI is displayed in its full size instead of only the TBSS skeleton.

More frequent data collection/processing errors were found in these white matter anomalies as compared to the VV anomalies (Table 1). Some of these errors occurred at the data acquisition stage, due to head motion artifacts (Figs. S8a and S9b) or the selection of a wrong FOV (Fig. S9a). Others occurred at the data processing stage, such as incorrect segmentation (Fig. S8bc) or incorrect registration (Figs. S9c).

A proportion of the white matter anomalies without the data errors were reviewed by the neuroradiologist and there were many positive radiological reads. 98.4% (62/63) of the reviewed WMLV anomalies, 91.9% (34/37) of the reviewed FA anomalies, and 97.4% (37/38) of the reviewed MD anomalies were identified with positive radiological findings (Table 1). For instance, likely multiple sclerosis was identified in a subject who was anomalous in WMLV, FA, and MD (Fig. 5c). The diagnosis of multiple sclerosis was confirmed by the UKB health outcomes data. Lacunar infarcts and moderate small vessel disease were identified in another subject who was also anomalous in all three of the white matter-based imaging phenotypes (Fig. 5c). A parahippocampal cyst was identified in an MD anomaly subject (Fig. 5a). Encephalomalacia (Fig. 5a) was identified in a subject who was anomalous in both FA and MD. There were also white matter anomalies that were not explained by data collection/processing error nor positive radiological reads. Fig. 5d shows two examples of these “novel” anomalies. In one subject with anomalous patterns in WMLV, FA, and MD (Fig. 5d, left panel), severe biparietal atypically distributed white matter lesions of uncertain etiology were identified. In another FA anomaly subject

with no noticeable pathology (Fig. 5d, right panel), anomalously low FA value was found specifically in the genu of corpus callosum (Fig. S3d). Taken together, these results indicate that the anomalies of white matter-based imaging phenotypes had more frequent data errors and were associated with a large variety of different positive radiological findings. Novel anomalies, each with unique patterns, only constituted a small fraction of these anomalies.

Individuals with anomalous patterns of cortical thickness

We next examined the individuals with anomalous CTh. As an example, Fig. 6a shows regional CTh deviation maps of an anomaly subject versus a normal subject (~4.9 difference in anomaly score). Widespread negative CTh deviations, representing thinner cortices in these regions, were observed in this anomaly subject. Data collection/processing errors were found to be most abundant in CTh anomalies, constituting 95.8% (114/119) of the anomaly subjects, indicating that CTh is very sensitive to data collection/processing errors. Similar to the errors found in the white matter anomalies, these errors were due to head motion during data collection (Fig. S10a), incorrect segmentation/registration in data processing (Figs. S10b and S10c), or the combination of these issues (Fig. S10d). For the anomaly subjects with good data collection/processing quality, all were identified with positive radiological findings, such as prominent sulci or atrophy (Fig. 6b). Taken together, these results suggest that most CTh anomalies were associated with data collection/processing errors.

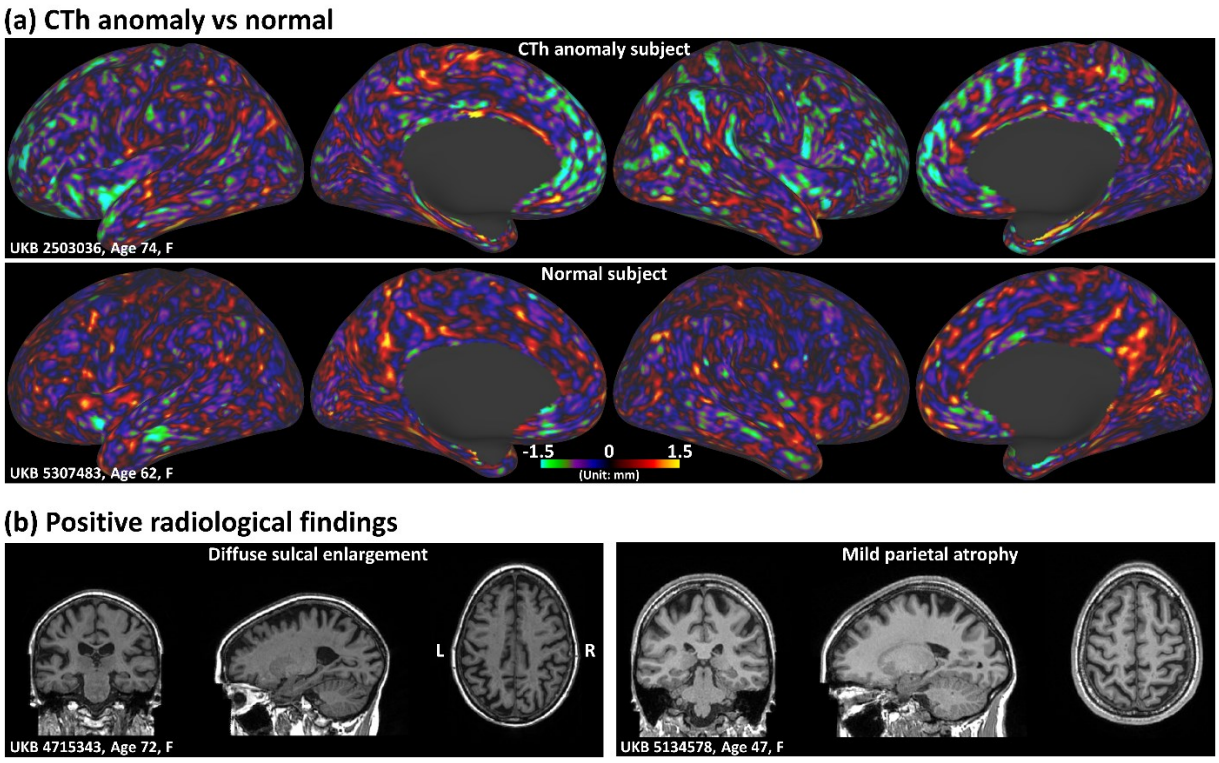


Fig. 6. Individuals with anomalous patterns of CTh. (**a**) Regional CTh deviation maps (displayed on inflated cortical surfaces) of an example of a CTh anomaly subject (first row, anomaly score = 4.3) and an example of a normal CTh subject (second row, anomaly score = -0.6). A CTh deviation map visualizes how the CTh values in a subject deviate from the autoencoder-predicted CTh values. (**b**) Structural images showing positive radiological findings in two representative CTh anomaly subjects (anomaly scores: UKB subject 4715343, 3.7; UKB subject 5134578, 3.2).

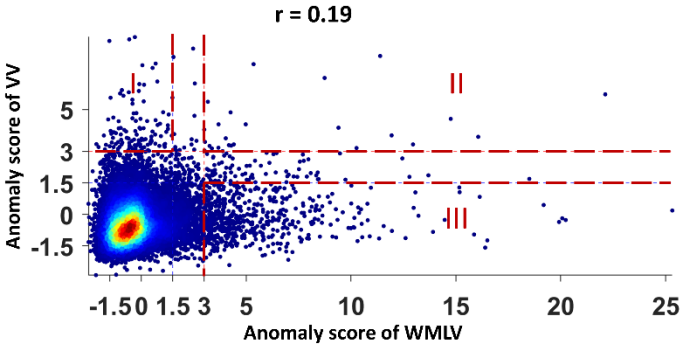
Anomaly score relationships across imaging phenotypes

The relationship of anomaly scores across different imaging phenotypes was assessed via pairwise Pearson correlation coefficients (Fig. 7a). Correlations between some white matter-based imaging phenotypes (FA versus MD; WMLV versus MD) were moderate ($0.4 < r < 0.6$), indicating they can capture similar anomalous patterns in the white matter. All the other correlations were weak ($0.2 < r < 0.4$) or very weak ($r < 0.2$), indicating they were complementary and provided independent information.

(a) Anomaly score correlations across imaging phenotypes

	WMLV	FA	MD	CTh
VV	0.19	0.16	0.24	0.12
WMLV		0.34	0.47	0.12
FA			0.41	0.14
MD				0.22

(b) WMLV anomaly score vs. VV anomaly score



(c) WMLV anomaly score vs. FA anomaly score

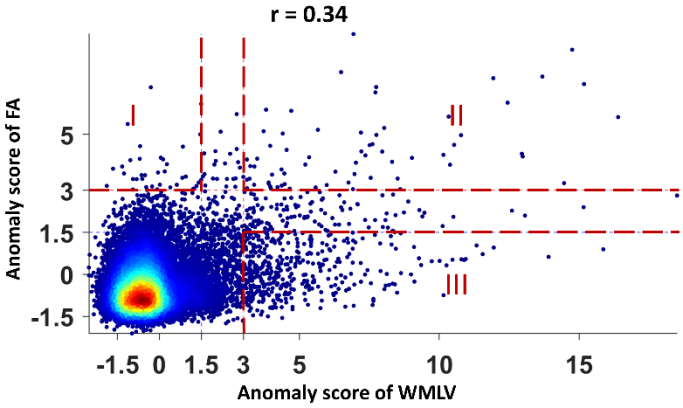


Fig. 7. Relationship between anomaly scores of different imaging phenotypes. (**a**) Correlations between the anomaly scores of different imaging phenotypes in the UKB cohort. The subjects with data collection/processing errors were not included in this analysis. The two representative relationships shown in (b) and (c) are encircled with red boxes. (**b**) WMLV anomaly score plotted against VV anomaly score. Zone I covered the VV anomalies with normal WMLV (WMLV anomaly score < 1.5). Zone II covered the subjects who were both VV and WMLV anomalies. Zone III covered the WMLV anomalies with normal VV (VV anomaly score < 1.5). (**c**) WMLV anomaly score plotted against FA anomaly score. Zone I covered the FA anomalies with normal WMLV (WMLV anomaly score < 1.5). Zone II covered the subjects who were both FA and WMLV anomalies. Zone III covered the WMLV anomalies with normal FA (FA anomaly score < 1.5).

To further illustrate these relationships, Fig. 7b shows a scatterplot of WMLV versus VV anomaly scores, which were poorly correlated (r = 0.19). Very few subjects

were both VV and WMLV anomalies, as evidenced by the sparser data in Zone II than Zone I or III (Fig. 7b). Indeed, the density of Zone II was significantly lower than the other two zones ($p \approx 0$, one-way analysis of variance [ANOVA] of 100000 bootstrap samples. Fig. S11a). It is therefore likely that the biological processes that led to large increases in WMLV are commonly independent of those that led to very enlarged VV. To illustrate another weak correlation, Fig. 7c shows a scatterplot of two white matter-based imaging phenotypes, WMLV versus FA anomaly scores ($r = 0.34$). The density of Zone II was significantly lower than Zone III ($p \approx 0$, one-way ANOVA of 100000 bootstrap samples. Fig. S11b) but was close to Zone I. Fig. S11c shows two examples of these anomalies of anomalies. The upper panel of Fig. S11c shows a subject that was an anomaly in both VV and WMLV. This subject, diagnosed with ventriculomegaly and moderate white matter disease, had both periventricular and deep white matter lesions. The lower panel of Fig. S11c shows a subject that was an anomaly in VV, WMLV, FA, and MD. The radiological read determined there was small vessel disease, evidenced by the white matter lesions, and probable Alzheimer's disease, evidenced by the parieto-temporal atrophy.

Anomaly detection for resting-state functional connectivity did not identify any reliable anomalies

Finally, we show a case where anomaly detection failed. Sixty UKB subjects' RSFC anomaly scores were above the anomaly threshold (Fig. S12a), however, they were later found to be confounded by global signal amplitude and no individual remained an RSFC anomaly in both test and retest sessions. RSFC anomaly scores were only moderately reliable overall (ICC = 0.42. Fig. S12b). At an individual level, the larger the

anomaly score in the initial imaging visit, the more the score decreased in the repeat imaging visit, as shown by the negative correlation between the anomaly score in the initial imaging visit and test-retest score change (r = -0.53. Fig. S12c). Because of this low reliability, among the subjects with available test-retest data, none had both visits identified as anomalies. This change in test-retest anomaly scores was found to be correlated with the change of global signal amplitude (r = 0.42. Fig. S12d); indeed, the RSFC anomaly score itself was found to be moderately correlated with the global signal amplitude (r = 0.48. Fig. S12e). The association was not due to head motion, because the moderate correlation persisted after excluding subjects with large head motion (r = 0.51. Fig. S12f). The association between RSFC anomaly score and global signal amplitude also persisted when using partial correlations to evaluate RSFC, although they became negatively correlated in this case (r = -0.70. Fig. S12g). Global signal regression reduced the association, but RSFC anomaly score (using full correlations) was still weakly correlated with global signal amplitude (r = 0.39). Remarkably, when we carried out similar analyses on the HCP cohort, the results were very similar (Fig. S13). Thus, determining if there are people with anomalous RSFC requires data processing improvements, especially those strategies that can better remove global signal fluctuations.

Discussion

In this study, a semi-automated, two-level screening methodology was used to detect anomalies in MRI imaging phenotypes of VV, WMLV, FA, MD, and CTh (Fig. 1). We demonstrated that anomaly scores of these imaging phenotypes were highly robust (Fig. S2). Anomaly score distributions of the UKB cohort were all more outlier-prone than

a standard normal distribution (Figs. 2 and S3a) and were also more outlier-prone than the HCP cohort of young adults (Fig. S4). We showed that anomaly scores had good-to-excellent long-term test-retest reliability (Figs. 3 and S3b). VV anomalies were associated with positive radiological findings or were novel (Figs. 4bcd and S7). The white matter-based anomalies were associated with more data collection/processing errors (Figs. S8 and S9) or positive radiological findings (Fig. 5c), and a small fraction of them were novel (Figs. 5d and S3e). CTh anomalies were mostly due to data collection/processing errors (Fig. S10a-d). The anomaly scores of different imaging phenotypes were mostly independent (Fig. 7). Finally, we also showed that no reliable anomalies could be detected for RSFC, which was associated with the confounding effects of global signal fluctuations (Figs. S12 and S13).

The approach to screen anomalies at an individual level in large neuroimaging cohorts

Large-scale neuroimaging datasets have emerged in recent years, with anywhere from 1,000 (Di Martino et al., 2014; Holmes et al., 2015; Van Essen et al., 2013) to more than 10,000 subjects (Hagler et al., 2019; Miller et al., 2016). Most studies using these datasets generally focus on the average imaging characteristics at a group level. Meanwhile, there has been much less work on anomaly detection in neuroimaging (Marquand et al., 2016; Mourao-Miranda et al., 2011; Pinaya et al., 2019; van Hespen et al., 2021). To fill this gap, we set out to investigate individual anomalous patterns from the two large-scale cohorts: the UKB and HCP.

To show generalizability, anomaly detection was performed for six commonly used, well-established brain imaging phenotypes. White matter-based imaging phenotypes were derived by the UKB and were readily available, so our approach can be conveniently applied for any new UKB subjects' WMLV, FA, and MD data. For the other imaging phenotypes (VV, CTh, RSFC), because they were initially obtained via different preprocessing pipelines between the UKB and HCP, the UKB structural MRI data were reprocessed by the HCP pipeline (Glasser et al., 2013), and the UKB resting-state fMRI (rsfMRI) data were further processed by Ciftify (Dickie et al., 2019), to make the preprocessing more uniform. This brought them into the same HCP standard surface space for the convenience of future comparisons with many other HCP-style studies (Harms et al., 2018; Lewandowski et al., 2020), but with the nontrivial computational expense of reprocessing the UKB data. The imaging phenotypes derived in the present study will be made available to researchers (e.g., via the UKB), so they will be accessible without the need to repeat the reprocessing done here.

The data of most imaging phenotypes were curated very well, as evidenced by weak or very weak correlations between their anomaly scores and confounding factors (Fig. S14). The head motion and brain registration-related confounding factors evaluated here were either suggested or equivalent to the ones described in recent work on confound modeling of the UKB brain imaging data (Alfaro-Almagro et al., 2021). However, there was inevitably a small fraction of data with acquisition or processing errors. It is therefore critical to identify the anomalies associated with such issues. This was achieved by screening the anomalies via visual inspection and multiple different data quality control metrics (head motion level, brain registration quality, etc.; see *Materials and Methods* for

details) to ensure the capture of different types of errors including wrong FOV (Fig. S9a), head motion artifact (Figs. S8a, S9b, and S10a), incorrect segmentation (Figs. S8bc and S10b), and incorrect registration (Figs. S9c and S10c). Thus, as one of the major uses of anomaly detection (Goldstein and Uchida, 2016), our method is valuable for curating large neuroimaging datasets.

The next screen was for a neuroradiologist to read the anomaly individuals that did not have data collection/processing errors. One hundred eleven (111) UKB anomaly subjects were reviewed by a neuroradiologist. Although these individuals were only a subgroup of all the UKB anomalies, they still covered a wide range of anomaly scores above the threshold (Fig. S6) and were sufficient to capture a diverse category of anomalous phenotypes. Indeed, a large percentage (90.1%, 100/111) had positive radiological findings, such as masses, cysts, white matter lesions, infarcts, encephalomalacia, and prominent sulci. Most of these brain pathologies likely would have led to a recommendation to see a physician for follow-up. For example, a VV anomaly subject (Fig. 4b) was diagnosed with a colloid cyst causing hydrocephalus and the neuroradiologist's read recommended this individual see a neurosurgeon for follow-up. Thus, our method is useful for detecting participants with clinical issues.

The anomalies that did not have data issues or positive radiological findings were considered "novel," which is potentially valuable for investigating underlying mechanisms. It should be noted that some WMLV anomaly subjects, categorized as the anomalies associated with positive radiological findings, were identified with white matter disease. However, this is mostly due to the large amount of WMLV. The underlying etiology may still be ambiguous, requiring further evaluations to see if there was a known clinical cause

for the large WMLV. Therefore, in the present study, the WMLV anomalies were considered as “novel” only when the distribution of white matter lesions was atypical of a specific etiology.

Potential underlying mechanisms of novel anomalies

Of the anomalies that were read by the neuroradiologist, eleven UKB anomalies and eight HCP anomalies had no radiological determination of a known clinical phenotype. These anomaly subjects had no brain-related disorders reported previously: none of the eleven UKB subjects had any prior diagnosis of mental and behavioral disorders, nervous system disorders, or circulatory brain disorders according to their health outcomes data, and all HCP subjects recruited were healthy young adults. Eight UKB and five HCP subjects were novel VV anomalies. The VV of these UKB subjects, ranging between 87.4 mL and 142.4 mL, was comparable to the upper range of VV in Alzheimer’s disease patients (Schott et al., 2005). The VV of these HCP subjects were between 45.4 mL and 56.2 mL, which were still much larger than the volumes of normal young healthy subjects. Our data also showed unexplained variations of VV between two monozygotic twin pairs in novel anomalies. In terms of VV abnormality, the two female individuals within a monozygotic twin pair both had anomalously large VV (Fig. 4d), suggesting a shared congenital, developmental, or environmental causes. In the other monozygotic twin pair, only one twin had anomalous large VV (Fig. S7c). This is probably due to environmental influences or a de novo mutation early in development. Further genetic or clinical investigations will be required to elucidate the underlying mechanisms.

One HCP VV anomaly and one UKB triple-anomaly of WMLV, FA, and MD were novel anomalies due to uncertain etiology. In that HCP VV anomaly subject, unexplained signal abnormalities were identified in the thalamus and brainstem (Fig. S7b). A follow-up T2-weighted fluid-attenuated inversion recovery (FLAIR) scan could be helpful to diagnose this case. In that UKB triple-anomaly, severe bilateral, confluent, and symmetrical white matter lesions were identified in the parietal white matter (Fig. 5d, left panel). The pattern of lesions was different from small vessel disease or multiple sclerosis, but was similar to reported cases of X-linked adrenoleukodystrophy (Geraldes et al., 2018). In the health outcomes data, this male subject was also reported to have hearing loss, a possible symptom of X-linked adrenoleukodystrophy, again indicating the possibility of this rare genetic disorder in this novel anomaly subject.

Two UKB FA anomalies, two HCP FA anomalies, and one HCP MD anomaly were novel because they had no data collection/processing errors and they were radiologically normal. Anomalously low FA values were found in the corpus callosum, superior longitudinal fasciculus, cingulum, posterior thalamic radiation, and limbs of the internal capsule (Figs. 5d, right panel, and S3e). A previous study showed that low FA in normal-appearing white matter preceded the conversion of these low FA regions into white matter lesions (de Groot et al., 2013). These novel FA anomaly subjects may be at risk to develop lesions later in these regions of anomalously low FA. For the HCP MD anomaly, many small perivascular spaces were presented on the structural image. These perivascular spaces were not abnormal, but could raise the MD. All of these novel anomaly subjects would benefit from follow-up assessments to study underlying mechanisms and to see if they progress to any specific clinical diagnosis.

Anomaly detection for resting-state functional connectivity was confounded by global signal fluctuations

Anomaly detection was also performed for RSFC because the rsfMRI data in part may report on subject-specific patterns in brain activity. Unfortunately, interpretation of the signal is confounded by the fact that it also contains contributions from instrumental and autonomic sources, and these, as well as any neuronal sources, may be strongly dependent on the experimental conditions and the subject's mental state. One of these confounds is the so-called "global signal," i.e. the brain-averaged rsfMRI signal, which may vary with changes in autonomic activity, the subject's vigilance state, the amount of head motion during the experiment, and even time of day (Orban et al., 2020; Ozbay et al., 2019; Power et al., 2015; Wong et al., 2013). While rsfMRI confound removal has been an active area of research for some time, no consensus yet exists on the optimal way to remove the global signal (Liu et al., 2017; Murphy et al., 2009; Murphy and Fox, 2017). Many studies evaluated RSFC with or without global signal regression in parallel, and some other work used partial correlations for RSFC instead (Pervaiz et al., 2020). In the present study, we opted to evaluate RSFC in these three ways and perform anomaly detection for each of them respectively. Our results showed that RSFC anomaly score was correlated with global signal amplitude in all these three ways and also regardless of the cohort used (Figs. 12 and 13), indicating global signal amplitude is a large confound in RSFC anomaly detection. RSFC anomaly score, which quantifies each individual in the cohort, may be a useful index for assessing the effectiveness of different processing strategies of the global signal. No matter the exact cause, it is remarkable that no

individual remained an RSFC anomaly in both test and retest. Considering the strong dependence of the rsfMRI signal on various factors that may change between repeat scans, including contributions from both neuronal and non-neuronal sources, it remains uncertain whether RSFC anomalies are reproducible in the human population. This failure case is helpful for researchers who wish to use RSFC for individual-level analysis.

Potential limitations

Although diverse neurological pathologies were presented in the subjects with large anomaly scores, it should be noted that the goal of the present study was not to outperform or replace human neuroradiologists, nor was it designed to compete with many machine learning approaches for diagnosing specific diseases. Instead, the present study aimed at discovering “interesting” anomalous imaging phenotypes, albeit rare, from the largely unlabeled cohort of the UKB. Therefore, we did not quantitively evaluate the effectiveness of our automatic anomaly detection in terms of identifying pathologies. For the same reason, we did not compare our method with other unsupervised anomaly detection algorithms. Despite this, the quality of our method can be appreciated by comparing it with manual radiological screening in the first 1000 UKB subjects (Gibson et al., 2017). In Gibson et al.’s study, 1.8% of the subjects screened via systematic radiologist review had incidental findings in their brain MRIs, whereas 90.1% of the anomaly subjects reviewed radiologically in our study had incidental findings. This indicates that our method can effectively identify a subgroup that is greatly enriched with incidental findings from a large cohort.

Another limitation is that the imaging phenotypes used here were derived from the outputs of conventional neuroimaging data processing pipelines, instead of directly using raw structural images, which could potentially compromise the richness of the input data. It would be interesting to use brain images directly in future studies of anomaly detection. Although some of the imaging phenotypes, for example, total VV, would seem to be global and not sensitive to focal problems, many of the verified abnormalities such as mass, cyst, infract, nodules, and partial agenesis of the corpus callosum were found in the VV anomaly subjects (Figs. 4 and S7). Indeed, it has been found enlargement of lateral ventricles is associated with various neurological and psychiatric disorders (Kempton et al., 2011; Kuller et al., 2016; Kuller et al., 2005; Mak et al., 2017; Nestor et al., 2008; Wright et al., 2000). We also found that the VV anomaly score was significantly higher in the repeat imaging visit than in the initial imaging visit ($p \approx 9*10^{-205}$, one-sample Wilcoxon test. Fig. S15a), and VV outlying subjects usually presented larger long-term test-retest changes than normal subjects ($r = 0.57$. Fig. S15b). The larger change in the VV outlying subjects is most likely due to faster progression of the processes underlying the anomaly, either because of the development of pathology or the rate of change due to aging. For example, the UKB subject with the largest test-retest change in VV anomaly score had bifrontal subdural hematomas on the initial imaging visit that resolved on the repeat imaging visit. The hematomas may cause ventricle compression at the initial time point. Alternatively, or in addition, the tissue injury related to subdural hematomas could cause atrophy and ventricular expansion afterward. As the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort has clearly shown, MRI as a function of age can be useful to distinguish pathology from more "normal" aging (Weiner et al., 2010; Weiner et al., 2012).

T2-FLAIR data was not available for the HCP Young Adult cohort, therefore, HCP subjects' white matter lesions were segmented via their T1w images using FreeSurfer. However, false positive lesions caused by FreeSurfer segmentation errors for T1w hypointensities were noticed in the periventricular regions or near the gray matter-white matter boundaries (Fig. S8c), potentially reducing sensitivity to white matter anomalies in this young cohort. Future work on more robust algorithms for detecting white matter lesions in T1w images will certainly improve the anomaly detection in the cohorts without T2-FLAIR data.

Conclusions

The present study characterized individual anomalous patterns across multiple imaging phenotypes in two large imaging cohorts. Every subject was parameterized with an anomaly score per phenotype to quantitate the abnormality. These anomaly scores were highly robust. Anomaly score distributions of the UKB cohort were all more outlier-prone than the HCP cohort of young adults. The approach enabled the assessments of test-retest reliability via the anomaly scores, which ranged from excellent reliability for VV, WMLV, and FA, to good reliability for MD and CTh. The individual-level analyses of the anomalies revealed their association with data collection/processing errors or different brain pathologies. A number of the novel anomalies can be used as candidates for future screening of potential underlying biological mechanisms. Finally, no consistent anomalies were detected for RSFC. The variability of RSFC anomaly scores was associated with the variations in global signal amplitude that are difficult to remove. Taken together, anomaly detection of large neuroimaging datasets was valuable for data curation,

reliability assessment, and the identification of individuals for medical follow-up or further study of novel mechanisms. Anomaly detection methods should contribute to the effort of developing automatic processes to analyze and interpret brain imaging data in large population cohorts.

Author contributions

Z.M. and A.P.K. designed the study; Z.M., D.S.R., and S.D. performed the analyses; Z.M., D.S.R., J.H.D., and A.P.K. wrote the paper.

Competing interests

None.

Acknowledgments

This study was supported by NIH/NINDS Intramural Research Program (Grant number: NS002989). We would like to thank Dr. Adam Thomas for helping with access to the UK Biobank datasets and for providing data storage resources. This research was conducted using the UK Biobank data under application number 22875. HCP data were provided [in part] by the HCP, WU-Minn Consortium (PIs: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. In addition, the authors would like to thank UKB-Neuroimaging and HCP-Users mailing lists for helpful information. This work used NIH Biowulf high-performance computing resources (https://hpc.nih.gov).

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400-424.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J.L.R., Bastiani, M., Miller, K.L., Nichols, T.E., Smith, S.M., 2021. Confound modelling in UK Biobank brain imaging. Neuroimage 224, 117002.

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., Sprosen, T., Collins, R., Biobank, U., 2012. UK Biobank: Current status and what it means for epidemiology. Health Policy and Technology 1, 123-126.

Andersson, J.L.R., Sotiropoulos, S.N., 2016. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage 125, 1063-1078.

Basser, P.J., Mattiello, J., LeBihan, D., 1994. Estimation of the effective self-diffusion tensor from the NMR spin echo. J Magn Reson B 103, 247-254.

de Groot, M., Verhaaren, B.F., de Boer, R., Klein, S., Hofman, A., van der Lugt, A., Ikram, M.A., Niessen, W.J., Vernooij, M.W., 2013. Changes in normal-appearing white matter precede development of white matter lesions. Stroke 44, 1037-1042.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Muller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19, 659-667.

Dickie, E.W., Anticevic, A., Smith, D.E., Coalson, T.S., Manogaran, M., Calarco, N., Viviano, J.D., Glasser, M.F., Van Essen, D.C., Voineskos, A.N., 2019. Ciftify: A framework for surface-based analysis of legacy MR acquisitions. Neuroimage 197, 818-826.

Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774-781.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341-355.

Geraldes, R., Ciccarelli, O., Barkhof, F., De Stefano, N., Enzinger, C., Filippi, M., Hofer, M., Paul, F., Preziosa, P., Rovira, A., DeLuca, G.C., Kappos, L., Yousry, T., Fazekas, F., Frederiksen, J., Gasperini, C., Sastre-Garriga, J., Evangelou, N., Jacqueline Palace on behalf of the, M.s.g., 2018. The current role of MRI in differentiating multiple sclerosis from its imaging mimics. Nat Rev Neurol 14, 213.

Gibson, L.M., Littlejohns, T.J., Adamska, L., Garratt, S., Doherty, N., Group, U.K.B.I.W., Wardlaw, J.M., Maskell, G., Parker, M., Brownsword, R., Matthews, P.M., Collins, R., Allen, N.E., Sellors, J., Sudlow, C.L., 2017. Impact of detecting potentially serious incidental findings during multi-modal imaging. Wellcome Open Res 2, 114.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M., Consortium, W.U.-M.H., 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105-124.

Glasser, M.F., Van Essen, D.C., 2011. Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. J Neurosci 31, 11597-11616.

Goldstein, M., Uchida, S., 2016. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLoS One 11, e0152173.

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E., 2016. Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. Cereb Cortex 26, 288-303.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsoldos, E., Ebmeier, K.P., Filippini, N., Mackay, C.E., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K.L., Smith, S.M., 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. Neuroimage 95, 232-247.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., Jenkinson, M., 2016. BIANCA (Brain Intensity

AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. Neuroimage 141, 191-205.

Guo, C.C., Kurth, F., Zhou, J., Mayer, E.A., Eickhoff, S.B., Kramer, J.H., Seeley, W.W., 2012. One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. Neuroimage 61, 1471-1483.

Hagler, D.J., Jr., Hatton, S., Cornejo, M.D., Makowski, C., Fair, D.A., Dick, A.S., Sutherland, M.T., Casey, B.J., Barch, D.M., Harms, M.P., Watts, R., Bjork, J.M., Garavan, H.P., Hilmer, L., Pung, C.J., Sicat, C.S., Kuperman, J., Bartsch, H., ..., Dale, A.M., 2019. Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. Neuroimage 202, 116091.

Harms, M.P., Somerville, L.H., Ances, B.M., Andersson, J., Barch, D.M., Bastiani, M., Bookheimer, S.Y., Brown, T.B., Buckner, R.L., Burgess, G.C., Coalson, T.S., Chappell, M.A., Dapretto, M., Douaud, G., Fischl, B., Glasser, M.F., Greve, D.N., Hodge, C., Jamison, K.W., Jbabdi, S., Kandala, S., Li, X., Mair, R.W., Mangia, S., Marcus, D., Mascali, D., Moeller, S., Nichols, T.E., Robinson, E.C., Salat, D.H., Smith, S.M., Sotiropoulos, S.N., Terpstra, M., Thomas, K.M., Tisdall, M.D., Ugurbil, K., van der Kouwe, A., Woods, R.P., Zollei, L., Van Essen, D.C., Yacoub, E., 2018. Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. Neuroimage 183, 972-984.

Hawkins, S., He, H., Williams, G., Baxter, R., 2002. Outlier Detection Using Replicator Neural Networks. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 170-180.

Holmes, A.J., Hollinshead, M.O., O'Keefe, T.M., Petrov, V.I., Fariello, G.R., Wald, L.L., Fischl, B., Rosen, B.R., Mair, R.W., Roffman, J.L., Smoller, J.W., Buckner, R.L., 2015.

Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. Sci Data 2, 150031.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. Neuroimage 62, 782-790.

Kempton, M.J., Salvador, Z., Munafo, M.R., Geddes, J.R., Simmons, A., Frangou, S., Williams, S.C., 2011. Structural neuroimaging studies in major depressive disorder. Meta-analysis and comparison with bipolar disorder. Arch Gen Psychiatry 68, 675-690.

Kuller, L.H., Lopez, O.L., Becker, J.T., Chang, Y., Newman, A.B., 2016. Risk of dementia and death in the long-term follow-up of the Pittsburgh Cardiovascular Health Study-Cognition Study. Alzheimers Dement 12, 170-183.

Kuller, L.H., Lopez, O.L., Jagust, W.J., Becker, J.T., DeKosky, S.T., Lyketsos, C., Kawas, C., Breitner, J.C., Fitzpatrick, A., Dulberg, C., 2005. Determinants of vascular dementia in the Cardiovascular Health Cognition Study. Neurology 64, 1548-1552.

Lewandowski, K.E., Bouix, S., Ongur, D., Shenton, M.E., 2020. Neuroprogression across the Early Course of Psychosis. J Psychiatr Brain Sci 5.

Littlejohns, T.J., Holliday, J., Gibson, L.M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J.D., Boultwood, C., Collins, R., Conroy, M.C., Crabtree, N., Doherty, N., Frangi, A.F., Harvey, N.C., Leeson, P., Miller, K.L., Neubauer, S., Petersen, S.E., Sellors, J., Sheard, S., Smith, S.M., Sudlow, C.L.M., Matthews, P.M., Allen, N.E., 2020. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 11, 2624.

Liu, T.T., Nalci, A., Falahpour, M., 2017. The global signal in fMRI: Nuisance or Information? Neuroimage 150, 213-229.

Mak, E., Su, L., Williams, G.B., Firbank, M.J., Lawson, R.A., Yarnall, A.J., Duncan, G.W., Mollenhauer, B., Owen, A.M., Khoo, T.K., Brooks, D.J., Rowe, J.B., Barker, R.A., Burn, D.J., O'Brien, J.T., 2017. Longitudinal whole-brain atrophy and ventricular enlargement in nondemented Parkinson's disease. Neurobiol Aging 55, 78-90.

Marquand, A.F., Rezek, I., Buitelaar, J., Beckmann, C.F., 2016. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. Biol Psychiatry 80, 552-561.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1, 30-46.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19, 1523-1536.

Moller, M.F., 1993. A Scaled Conjugate-Gradient Algorithm for Fast Supervised Learning. Neural Networks 6, 525-533.

Mori, S., Oishi, K., Jiang, H., Jiang, L., Li, X., Akhter, K., Hua, K., Faria, A.V., Mahmood, A., Woods, R., Toga, A.W., Pike, G.B., Neto, P.R., Evans, A., Zhang, J., Huang, H., Miller, M.I., van Zijl, P., Mazziotta, J., 2008. Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. Neuroimage 40, 570-582.

Mourao-Miranda, J., Hardoon, D.R., Hahn, T., Marquand, A.F., Williams, S.C.R., Shawe-Taylor, J., Brammer, M., 2011. Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine. Neuroimage 58, 793-804.

Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? Neuroimage 44, 893-905.

Murphy, K., Fox, M.D., 2017. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. Neuroimage 154, 169-173.

Nestor, S.M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J.L., Fogarty, J., Bartha, R., Alzheimer's Disease Neuroimaging, I., 2008. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. Brain 131, 2443-2454.

Orban, C., Kong, R., Li, J.W., Chee, M.W.L., Yeo, B.T.T., 2020. Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. PLoS Biol 18.

Ozbay, P.S., Chang, C., Picchioni, D., Mandelkow, H., Chappel-Farley, M.G., van Gelderen, P., de Zwart, J.A., Duyn, J., 2019. Sympathetic activity contributes to the fMRI signal. Commun Biol 2, 421.

Pervaiz, U., Vidaurre, D., Woolrich, M.W., Smith, S.M., 2020. Optimising network modelling methods for fMRI. Neuroimage 211, 116604.

Pinaya, W.H.L., Mechelli, A., Sato, J.R., 2019. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. Hum Brain Mapp 40, 944-954.

Power, J.D., Schlaggar, B.L., Petersen, S.E., 2015. Recent progress and outstanding issues in motion correction in resting state fMRI. Neuroimage 105, 536-551.

Robinson, E.C., Garcia, K., Glasser, M.F., Chen, Z., Coalson, T.S., Makropoulos, A., Bozek, J., Wright, R., Schuh, A., Webster, M., Hutter, J., Price, A., Cordero Grande, L., Hughes, E., Tusor, N., Bayly, P.V., Van Essen, D.C., Smith, S.M., Edwards, A.D., Hajnal, J., Jenkinson, M., Glocker, B., Rueckert, D., 2018. Multimodal surface matching with higher-order smoothness constraints. Neuroimage 167, 453-465.

Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449-468.

Schott, J.M., Price, S.L., Frost, C., Whitwell, J.L., Rossor, M.N., Fox, N.C., 2005. Measuring atrophy in Alzheimer disease: a serial MRI study over 6 and 12 months. Neurology 65, 119-124.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86, 420-428.

Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Matthews, P.M., Behrens, T.E., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31, 1487-1505.

Tan, P.-N., Steinbach, M., Kumar, V., 2006. Introduction to data mining, 1st ed. Pearson Addison Wesley, Boston.

Tukey, J.W., 1977. Exploratory data analysis. Addison-Wesley Pub. Co., Reading, Mass.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.U.-M.H., 2013. The WU-Minn Human Connectome Project: an overview. Neuroimage 80, 62-79.

van Hespen, K.M., Zwanenburg, J.J.M., Dankbaar, J.W., Geerlings, M.I., Hendrikse, J., Kuijf, H.J., 2021. An anomaly detection approach to identify chronic brain infarcts on MRI. Sci Rep 11, 7714.

Weiner, M.W., Aisen, P.S., Jack, C.R., Jr., Jagust, W.J., Trojanowski, J.Q., Shaw, L., Saykin, A.J., Morris, J.C., Cairns, N., Beckett, L.A., Toga, A., Green, R., Walter, S., Soares, H., Snyder, P., Siemers, E., Potter, W., Cole, P.E., Schmidt, M., Alzheimer's Disease Neuroimaging, I., 2010. The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimers Dement 6, 202-211 e207.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q., Alzheimer's Disease Neuroimaging, I., 2012. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimer's & dementia : the journal of the Alzheimer's Association 8, S1-S68.

Wong, C.W., DeYoung, P.N., Liu, T.T., 2016. Differences in the resting-state fMRI global signal amplitude between the eyes open and eyes closed states are related to changes in EEG vigilance. Neuroimage 124, 24-31.

Wong, C.W., Olafsson, V., Tal, O., Liu, T.T., 2013. The amplitude of the resting-state fMRI global signal is related to EEG vigilance measures. Neuroimage 83, 983-990.

Wright, I.C., Rabe-Hesketh, S., Woodruff, P.W., David, A.S., Murray, R.M., Bullmore, E.T., 2000. Meta-analysis of regional brain volumes in schizophrenia. Am J Psychiatry 157, 16-25.