Received: 18 July 2021 | Revised: 25 November 2021 | Accepted: 4 December 2021

DOI: 10.1002/hbm.25756

RESEARCH ARTICLE

WILEY

Outlier detection in multimodal MRI identifies rare individual phenotypes among more than 15,000 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, Maryland, USA

[2]Translational Neuroradiology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA

Correspondence
Zhiwei Ma and Alan P. Koretsky, Laboratory of Functional and Molecular Imaging, NIH/NINDS, 10 Center Dr, Bethesda, MD 20892-1065, USA.
Email: maz4@nih.gov (Z. M.) and koretskya@ninds.nih.gov (A. P. K.)

Funding information
This study was supported by NIH/NINDS Intramural Research Program (Project numbers: NS002989 and NS003119).

Abstract

Outliers in neuroimaging represent spurious data or the data of unusual phenotypes that deserve special attention such as clinical follow-up. Outliers have usually been detected in a supervised or semi-supervised manner for labeled neuroimaging cohorts. There has been much less work using unsupervised outlier detection on large unlabeled cohorts like the UK Biobank brain imaging dataset. 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. Here, we developed a two-level outlier detection and screening methodology to characterize individual outliers from the multimodal MRI dataset of more than 15,000 UK Biobank subjects. In primary screening, using brain ventricles, white matter, cortical thickness, and functional connectivity-based imaging phenotypes, every subject was parameterized with an outlier score per imaging phenotype. Outlier scores of these imaging phenotypes had good-to-excellent test–retest reliability, with the exception of resting-state functional connectivity (RSFC). Due to the low reliability of RSFC outlier scores, RSFC outliers were excluded from further individual-level outlier screening. In secondary screening, the extreme outliers (1,026 subjects) were examined individually, and those arising from data collection/processing errors were eliminated. A representative subgroup of 120 subjects from the remaining non-artifactual outliers were radiologically reviewed, and radiological findings were identified in 97.5% of them. This study establishes an unsupervised framework for investigating rare individual imaging phenotypes within a large neuroimaging cohort.

KEYWORDS
big data, individual-level analysis, machine learning, multimodal MRI, radiological findings

1 | INTRODUCTION

Outliers are defined as observations differing by a large amount from most other observations (Tan, Steinbach, & Kumar, 2006). By this definition, outliers constitute a small portion of a dataset and are

Abbreviations: ANOVA, analysis of variance; CAT, Computational Anatomy Toolbox; CTh, cortical thickness; DC, Dice coefficient; dMRI, diffusion MRI; EPI, echo-planar imaging; FA, fractional anisotropy; FOV, field of view; HCP, Human Connectome Project; ICC, intraclass correlation coefficient; IQR, interquartile range; MD, mean diffusivity; Q3, third quartile; ROI, region of interest; RSFC, resting-state functional connectivity; rsfMRI, resting-state functional MRI; SD, standard deviation; TBSS, Tract-Based Spatial Statistics; UKB, UK Biobank; VV, ventricular volume; WMLV, white matter lesion volume.

exceptional patterns in some sense. Detecting outliers is of interest in brain imaging for two major reasons. First, outliers can occur due to imaging artifacts or noise. For example, head motion adversely affects brain morphometry, diffusion, and connectivity measurements (Power, Schlaggar, & Petersen, 2015; Reuter et al., 2015; Yendiki, Koldewyn, Kakunoori, Kanwisher, & Fischl, 2014) and causes outliers in these data. Second, and more importantly, some outliers represent unusual phenotypes that deserve special attention. For example, an anomalous MRI may indicate the presence of neurological disease that requires clinical attention. Certain unusual phenotypes may also be interesting for follow-up to determine the underlying mechanism for the large deviations of their brain MRI from the population.

Outlier detection methods applied in brain imaging can be categorized in many ways. One common way is based on whether the method makes use of labeled datasets to train the outlier detection model: supervised methods use labeled datasets that contain both labeled outliers and labeled non-outliers for training; semi-supervised methods use labeled datasets that only contain labeled non-outliers for training; and unsupervised methods use unlabeled datasets for training (Goldstein & Uchida, 2016). Using the available diagnostic labels for all subjects or at least the non-outlier subjects, outlier detection studies have employed a variety of algorithms, such as one-class support vector machine, Gaussian process regression, or autoencoders, and these have been applied in a supervised or semi-supervised manner to quantify the outlierness of healthy individuals or patients (Marquand, Rezek, Buitelaar, & Beckmann, 2016; Mourao-Miranda et al., 2011; Pinaya, Mechelli, & Sato, 2019; van Hespen et al., 2021). However, diagnostic labels are not always available, making the supervised or semi-supervised approaches challenging to implement across the board. Unsupervised outlier detection methods are needed for unlabeled brain imaging datasets, for example, the UK Biobank (UKB), an ongoing large epidemiological cohort (Miller et al., 2016).

The 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). In this cohort, 100,000 subjects are currently in the process of being invited back for MRI imaging, making it the largest multimodal MRI cohort in the world (Littlejohns et al., 2020). By enrolling a large population of this age range, this unlabeled brain imaging dataset includes healthy and presymptomatic subjects, as well as a small fraction of subjects with different clinically relevant diseases. Over time, many more subjects in this cohort will become identified with a clinically relevant disease (Miller et al., 2016). Given its large sample size, the UKB cohort enables a unique opportunity for developing unsupervised outlier detection methods to identify rare imaging phenotypes. These rare imaging phenotypes could be clinically relevant or informative for discovering new processes and mechanisms.

In the present study, a two-level outlier detection and screening methodology was developed to characterize individual outlying MRI results across multiple brain imaging phenotypes among more than 15,000 UKB subjects. We made use of the multimodal MRI data to derive ventricular, white matter, and gray matter-based imaging phenotypes of the brain (Figure 1a). Every subject was parameterized with an "outlier score" per imaging phenotype in an unsupervised manner without any prior labels (Figure 1b). This outlier score quantifies how far an individual deviates from most other subjects. Test–retest reliability of outlier scores of each imaging phenotype was characterized in the subjects that had repeat MRI scans, and any less reliable imaging phenotype was not used for further individual-level outlier screening. Individual extreme outlier subjects were categorized according to whether there were data collection/processing errors, or whether the individual had radiological findings or appeared normal as determined by a board-certified neuroradiologist (Figure 1c). Similar outlier detection and screening procedures were also carried out separately in the Human Connectome Project (HCP) dataset (Van Essen et al., 2013), and the extreme outlier subjects from this young adult cohort that might be interesting for follow-up are also described.

2 | MATERIALS AND METHODS

2.1 | Main dataset

The multimodal brain MRI data of 19,406 subjects (9,170 males and 10,236 females; age 44–80) at the initial imaging visit were downloaded from the UKB. This included T1-weighted (T1w) MPRAGE and T2-weighted (T2w) FLAIR structural MRI, spin-echo echo-planar imaging (EPI) diffusion MRI (dMRI), and gradient-echo EPI resting-state functional MRI (rsfMRI) data. Some subjects only had usable T1w data available in this sample, resulting in a reduced initial sample size of other MRI modalities. Following exclusions based on automatic quality control described below in Section 2.3, the final sample size for each imaging modality varied from 15,166 to 19,076 (hereafter referred to as *UKB discovery group* for this final sample). The detailed number of exclusions and the demographic information of the final sample are summarized in Table S1. The data were acquired on identical 3 T Siemens Skyra MRI scanners, and detailed acquisition protocols can be found elsewhere (Alfaro-Almagro et al., 2018). The UKB study was approved by the North West Multi-centre Research Ethics Committee, and 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).

2.2 | Image preprocessing and extraction of imaging phenotypes

The following six commonly used brain 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.

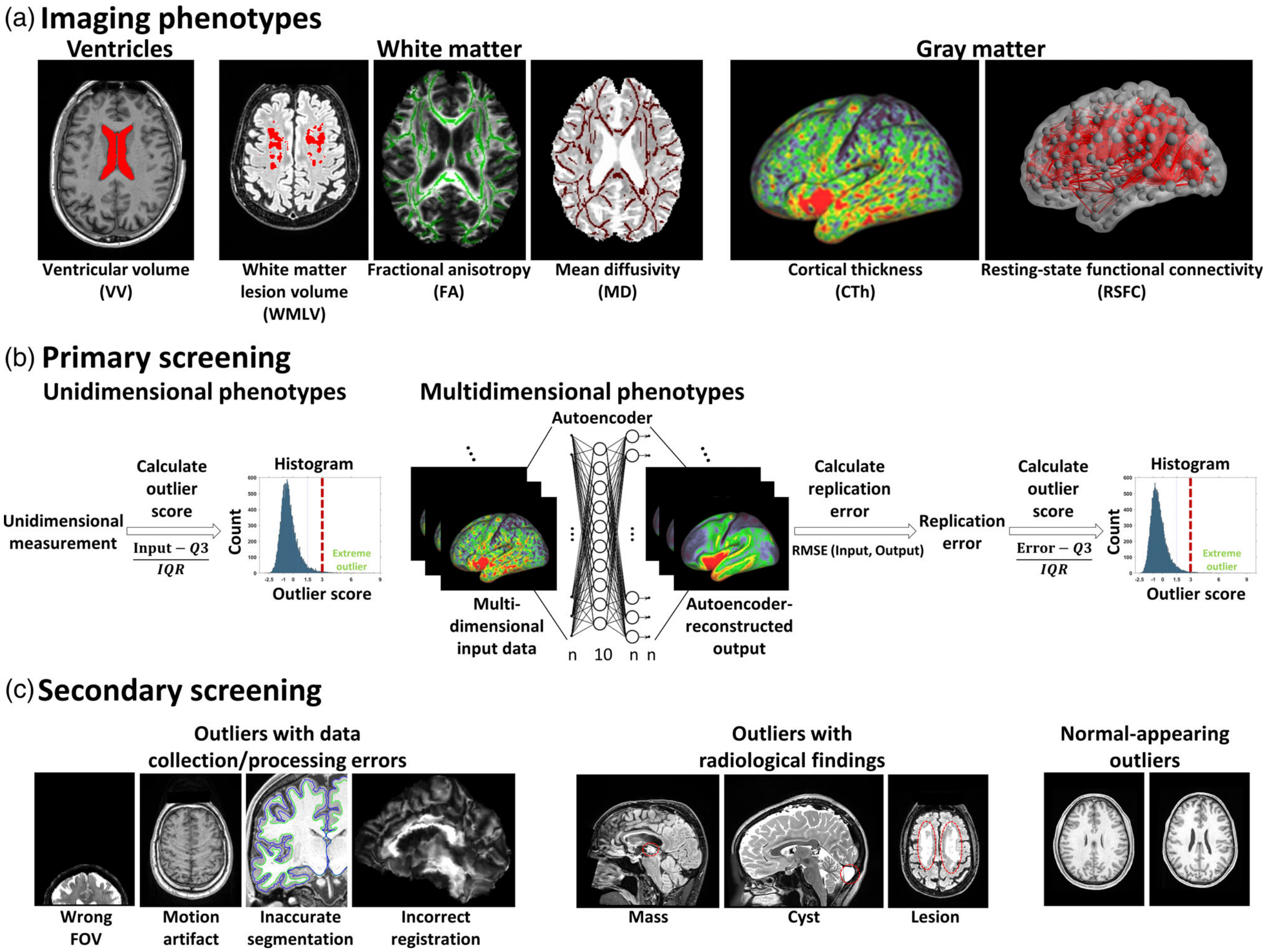


FIGURE 1 Schematic illustration of outlier detection and screening pipeline. (a) Brain imaging phenotypes used for outlier detection. (b) Primary screening: calculation of outlier scores. (c) Secondary screening: investigation of individual extreme outliers

The raw T1w MPRAGE and T2w FLAIR images were preprocessed by the HCP structural pipeline (v4) (Glasser et al., 2013) based on FreeSurfer (v6) (Fischl, 2012). For the subjects without usable T2w FLAIR images, the ventricles were segmented from their T1w images using FreeSurfer. The ventricular segmentations were manually inspected for each subject, and 213 subjects with large segmentation defects in their enlarged ventricles were reprocessed with "-bigventricles" flag in FreeSurfer to correct the defects. Each subject's VV was calculated by summing the volumes of lateral ventricles, temporal horns of the lateral ventricles, choroid plexuses, third ventricle, and fourth ventricle. WMLV was calculated by the Brain Intensity Abnormality Classification Algorithm (BIANCA, Griffanti et al., 2016), a k-nearest-neighbor-based automated supervised method, using T2w FLAIR images but also T1w images as its inputs. Unsmoothed CTh values in the standard CIFTI grayordinate space (with folding-related effects corrected) were averaged within the region of interests (ROIs) of the HCP multimodal parcellation atlas (360 regions) (Glasser et al., 2016), and these ROI-wise CTh values were used for primary screening.

The dMRI data underwent FSL eddy-current and head-movement correction (Andersson & Sotiropoulos, 2016), gradient distortion correction, diffusion tensor model fitting using the $b = 1{,}000$ shell (Basser, Mattiello, & LeBihan, 1994), and Tract-Based Spatial Statistics (TBSS) analyses (Smith et al., 2006). The TBSS skeletonized images were averaged within the ROIs of the Johns Hopkins University white matter atlas (Mori et al., 2008). Here, the original MD values were multiplied by 10,000 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.

The rsfMRI data were preprocessed by the UKB rsfMRI pipeline (v1) (Alfaro-Almagro et al., 2018), and the volumetric FIX-denoised data (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014) were brought to the standard CIFTI grayordinate space using Ciftify (v2.3.2) (Dickie et al., 2019). For each subject, the standard deviation (*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. 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. In addition, RSFC was quantified using partial correlations with Tikhonov regularization ($\rho = 0.5$; FSLNets) (Pervaiz, Vidaurre, Woolrich, & Smith, 2020). Due to the symmetry of the RSFC matrices, the upper triangular parts of these matrices ($333 \times 332/2 = 55{,}278$ elements) from each of these three RSFC evaluation methods were used for primary screening, respectively.

2.3 | Automatic quality control

Recent research has shown the importance of quality control in big neuroimaging datasets (Maximov et al., 2021; Monereo-Sanchez et al., 2021). Exclusion of poor quality data was performed based on eight quality control metrics. First, for all imaging phenotypes, because their preprocessing all relied on usable T1w images (Alfaro-Almagro et al., 2018), the subjects with low image quality of their T1w images were excluded for further analyses. The quality of T1w images was evaluated quantitatively using the Computational Anatomy Toolbox (CAT12) (Dahnke, Yotter, & Gaser, 2013; Gaser & Dahnke, 2016), which generated a single aggregated metric on a 100-point scale for the overall quality of each T1w image, with 100 the best possible. The T1w images with scores below 75 were excluded (Gaser & Dahnke, 2016; Gilmore, Buser, & Hanson, 2021).

For FA and MD, two head motion parameters and one registration quality parameter were used for quality control. These two head motion parameters were each subject's mean and largest values of the volumetric movements between adjacent dMRI frames. The registration quality parameter was each subject's mean deformation of the TBSS nonlinear registration. For CTh, FreeSurfer's Euler number, which summarized surface reconstruction quality (Rosen et al., 2018), was used for quality control. In addition, because T1w/T2w ratio myelin maps were sensitive to subtle errors of registration or surface placement (Glasser et al., 2013), following the multidimensional outlier detection method described below in Section 2.4, an outlier score of myelin map was calculated per subject and was used for CTh quality control. For RSFC, two head motion parameters were used for quality control. These two head motion parameters were each subject's mean and largest values of the framewise displacement between adjacent EPI volumes. For the seven quality control metrics described above, data in the range above the upper inner fence of the distribution of that metric were excluded from further analyses. Here, the upper inner fence was the third quartile (*Q3*) plus 1.5 times the interquartile range (*IQR*) of the distribution, and the observations above it are commonly defined as mild (greater than $Q3 + 1.5 \times IQR$, but smaller than $Q3 + 3 \times IQR$) or extreme outliers (greater than $Q3 + 3 \times IQR$) in statistics (Tukey, 1977). This upper inner fence threshold was applied in the quality control of neuroimaging data (Monereo-Sanchez et al., 2021).

2.4 | Primary screening: Calculation of outlier scores

In primary screening, every subject that passed the quality control was parameterized with an outlier score per imaging phenotype. The outlier score quantified the degree of outlierness in that imaging phenotype, and extreme outliers were identified based on the outlier scores. In statistics, extreme outliers in distribution are defined as the observations above the *Q3* plus three times the *IQR* of that distribution (Tukey, 1977). For a unidimensional imaging phenotype (VV, WMLV), using VV as an example, the number of *IQR*s away from the *Q3* of the VV distribution was used to define VV outlier scores:

$$\text{Outlier score} = \frac{VV - Q3}{IQR} \tag{1}$$

In this way, the unit of outlier score is *IQR*, and an extreme outlier has an outlier score of greater than 3. WMLV outlier scores were calculated similarly.

For each multidimensional imaging phenotype (FA, MD, CTh, RSFC), an autoencoder was used to calculate the outlier scores (Hawkins, He, Williams, & Baxter, 2002). Setting the dimensionality of the imaging phenotype as M and the number of subjects in the UKB discovery group as N, the inputs to the autoencoder were the values of that imaging phenotype across the whole group ($M \times N$), and the autoencoder was trained to replicate this input at its output. By definition, outliers only comprised a small portion of a dataset; therefore, the trained autoencoder cannot replicate these outliers as good as the non-outliers. This resulted in larger replication errors for the outlying subjects. These replication errors (also known as "autoencoder reconstruction error") were measured by the root mean square errors between each input and the autoencoder-predicted output. Because these replication errors were unidimensional, similar to the calculation of outlier scores for unidimensional imaging phenotypes, the number of *IQR*s away from the *Q3* of the replication error distribution was used to define outlier scores:

$$\text{Outlier score} = \frac{\text{error} - Q3}{IQR} \tag{2}$$

Still, the unit of outlier score is *IQR*, and an extreme outlier has an outlier score of greater than 3.

In the above analyses, to control for the effects of covariates (age, brain volume, and the image quality metrics described in Section 2.3) on outlier detection, their correlations with VV, WMLV, and the autoencoder replication errors of multidimensional imaging phenotypes were evaluated (Figure S1). The covariates with correlation >0.3 were regressed out from VV, WMLV, or the replication errors before applying Equation (1) or (2). As a result, age, brain volume, and CAT12's T1w image quality metric were regressed out from VV. Age was regressed out from WMLV. Age was also regressed out from the autoencoder replication errors of MD. FreeSurfer's Euler number was regressed out from the autoencoder replication errors of CTh.

Each autoencoder used in the present study 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 the autoencoder. The autoencoders were implemented using the "trainAutoencoder" function in the MATLAB and were trained using a GPU cluster (https://hpc.nih.gov). When the input dataset was too large to fit into the GPU memory, multiple autoencoders were used. In these scenarios, the input data were split into four to five smaller subgroups in a stratified manner, preserving the ratio of age and sex in each subgroup. For each subgroup, an autoencoder was trained using the data of that subgroup as the input. The trained autoencoders were then applied to the full dataset and the output of the whole group was obtained by averaging the outputs from each of these autoencoders.

2.5 | Evaluation of reliability of outlier scores and elimination of less reliable imaging phenotype

A subgroup (1,391 subjects) of the UKB discovery group subjects had a repeat MRI session (also known as "retest") 2–3 years after the initial imaging visit (also known as "test"). The test and retest data of these subjects were used to evaluate long-term reliability of outlier scores. Unlike the primary screening, in the reliability analysis, the volume measurements of unidimensional imaging phenotypes or the autoencoder replication errors of multidimensional imaging phenotypes were no longer adjusted for covariates. For each unidimensional imaging phenotype, the *Q3* and *IQR* were calculated from the full test data and applied to calculate outlier scores for both test data and, for subjects who were scanned twice, retest data. For each multidimensional imaging phenotype, the autoencoders trained on the full test data were applied to the retest data. The reliability was quantified by intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979) between the outlier scores of the test and retest data using a one-way random effects model:

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

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 & Wong, 1996). Reliability was defined as excellent (ICC > 0.8), good (0.8 > ICC > 0.6), moderate (0.6 > ICC > 0.4), fair (0.4 > ICC > 0.2), or poor (ICC < 0.2) (Guo et al., 2012) in the present study.

Any imaging phenotype with moderate/fair/poor outlier score reliability was excluded from further analysis of individual outliers. This resulted in the exclusion of RSFC (for details, see Section 3.2).

2.6 | Secondary screening: Investigation of individual extreme outliers

The automatic quality control described in Section 2.3 excluded most data collection/processing errors. However, a small number of errors could remain in this large cohort. For example, potentially low quality T2w FLAIR images and potential segmentation errors of white matter lesions were not accounted for because of the lack of a well-established tool for automatic assessment of the quality of T2w FLAIR images or white matter lesion segmentation. To capture potential data collection/processing errors that may occur in extreme outliers, for each remaining imaging phenotype, the extreme outlier subjects were first manually inspected and the ones with the errors were eliminated. For each VV extreme outlier subject, ventricle segmentation was visually inspected by overlaying the border of the segmented ventricle mask on the T1w image. For each WMLV extreme outlier subject, white matter lesion segmentation was visually inspected by overlaying the border of the segmented lesion mask on the T2w FLAIR image. The FA or MD extreme outlier subjects were visually checked for registration and field of view (FOV) coverage. For CTh extreme outlier subjects, their white/pial surface segmentation was visually checked via HCP pipeline structural quality control scenes (https://github.com/Washington-University/StructuralQC; v1.4.0).

A subgroup (120 subjects) of the remaining non-artifactual extreme outlier subjects were radiologically reviewed. This subgroup included all top-ranked extreme outlier subjects and randomly sampled non-top extreme outlier subjects to ensure a wide coverage (Figure S2). T1w MPRAGE and T2w FLAIR images, as well as the ages of these subjects, were provided to a board-certified neuroradiologist (D. S. R.). The instructions to the neuroradiologist were to identify any major findings that might plausibly account for the extreme outlier score—not to identify subtle abnormalities that would have required dedicated review on clinical-grade display systems. When the neuroradiologist was uncertain of the diagnosis, UKB health outcomes data (UKB Category 1712) were used in an attempt to determine the diagnosis. These data recorded the first occurrence of various diseases, including neuropsychiatric and neurological disorders. Based on the radiological review results, the subjects in the subgroup were further divided into two subgroups: a subgroup of the extreme outlier subjects with radiological findings (117 subjects), and another subgroup of the extreme outlier subjects which appeared normal to the neuroradiologist (3 subjects). The cases from these two subgroups that would be interesting for follow-up were highlighted.

2.7 | Evaluation of the relationships between outlier scores of different imaging phenotypes

The relationships between outlier scores of different imaging phenotypes were quantified using Pearson cross-correlation coefficients in the UKB discovery group. Two representative relationships of outlier scores, WMLV versus VV, and WMLV versus FA, were also visualized using scatterplots. In each scatterplot, three zones were defined to categorize extreme outlier subjects. For WMLV versus VV, Zone I covered the subjects who were VV extreme outliers but with normal WMLV (WMLV outlier score < 1.5), Zone II covered the subjects who were both VV and WMLV extreme outliers, and Zone III covered the subjects who were WMLV extreme outliers but with normal VV (VV outlier 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:

$$\text{Density}_{\text{Zone I}} = \frac{\text{Number of subjects in Zone I}}{(1.5 - \min(\text{WMLV outlier score})) * (\max(\text{VV outlier score}) - 3)} \tag{4}$$

$$\text{Density}_{\text{Zone II}} = \frac{\text{Number of subjects in Zone II}}{(\max(\text{WMLV outlier score}) - 3) * (\max(\text{VV outlier score}) - 3)} \tag{5}$$

$$\text{Density}_{\text{Zone III}} = \frac{\text{Number of subjects in Zone III}}{(\max(\text{WMLV outlier score}) - 3) * (1.5 - \min(\text{VV outlier score}))} \tag{6}$$

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 recomputed. A one-way analysis of variance (ANOVA) was then performed to evaluate the differences across the zones using the bootstrap samples. Similar analyses were also carried out to evaluate the relationship between WMLV and FA outlier scores.

2.8 | Outlier detection and screening in the HCP dataset

Similar outlier detection procedures were carried out separately in the HCP dataset to identify interesting extreme outliers in this young adult cohort (for details, see Supplementary Methods). Briefly, 3 T MRI data from the 1,200 Subjects Release (1,113 subjects: 507 males and 606 females; age 22–37) were used (Glasser et al., 2016). Because of the lack of HCP T2w FLAIR data and poor WMLV segmentation accuracy when only using T1w images (Hotz et al., 2021), WMLV was excluded from the outlier detection of the HCP dataset. All the HCP extreme outliers (12 subjects) without data collection/processing errors were radiologically reviewed, and the cases that would be interesting for follow-up were highlighted.

3 | RESULTS

3.1 | Properties of outlier score distributions

The results presented throughout the rest of the manuscript were obtained using the UKB discovery group unless otherwise specified. The outlier score histogram of each imaging phenotype is shown in Figure 2. These distributions were all right-skewed and more leptokurtic than a standard normal distribution (see Table 1 for skewness and kurtosis values). The percentage of extreme outliers ranged from a lowest of 0.2% in RSFC, to a highest of 3.9% in WMLV (Table 1). These percentages are all much higher than a standard normal distribution predicts, because the criterion of $Q3 + 3 \times IQR$ for defining extreme outliers (referred to as "outlier" hereafter) in each distribution is equivalent to about 4.7 times the *SD* plus the mean in a standard normal distribution. One would predict only 0.0001% of the data above mean $+ 4.7 \times$ *SD* in a standard normal distribution. Taken together, the results suggest that the outlier score distributions were all more outlier-prone than a standard normal distribution.

3.2 | Long-term test–retest reliability of outlier scores

A subgroup of the discovery group subjects had a repeat MRI session 2–3 years after the initial visit. The outlier scores of test versus retest of each imaging phenotype are visualized in the scatterplots of Figure 3a–f, respectively. VV outlier scores had excellent test–retest reliability, as indicated by the close-to-one value of the ICC (ICC = 0.98) between test and retest outlier scores. The test–retest reliabilities of WMLV and FA outlier scores were lower than VV but still excellent (WMLV ICC = 0.82; FA ICC = 0.86). The test–retest reliabilities of MD and CTh outlier scores were lower than the former three but still in the range of good reliability (MD ICC = 0.72; CTh ICC = 0.64).

However, RSFC outlier scores had a low test–retest ICC (ICC = 0.40, Figure 3f). Because of this low reliability, among the subjects with available test–retest data, no subject had both test and retest RSFC identified consistently as an outlier. This change in test–retest outlier scores was found to be correlated with the change of global signal amplitude ($r = .43$, Figure 3g). Here, global signal amplitude was defined as the *SD* of the global signal (Wong, Olafsson, Tal, & Liu, 2013). Indeed, the RSFC outlier score itself was found to be correlated with global signal amplitude ($r = .51$, Figure 3h). This association was unlikely due to head motion, because the subjects with large head motion were excluded in the automatic quality control. The association between RSFC outlier score and global signal amplitude also persisted when using partial correlations to evaluate RSFC, although they became negatively correlated in this case ($r = -.69$, Figure S3a). Global signal regression reduced their association, but RSFC outlier score was still moderately correlated with global signal amplitude ($r = .42$, Figure S3b). Remarkably, when we carried out similar analyses on the HCP dataset, the results were very similar (Figure S3c–h). Thus, RSFC was eliminated for further individual-level outlier screening due to its low individual test–retest reliability.

3.3 | Summary of the screening results of individual outliers

The total number of outliers across all individual imaging phenotypes (excluding RSFC) was 1,258. Because there were subjects who were outliers in more than one imaging phenotype, there were 1,026 distinct subjects that made up these 1,258 outliers.

Through the screening of each outlier, 87 outliers were associated with data collection/processing errors. This was true despite the use

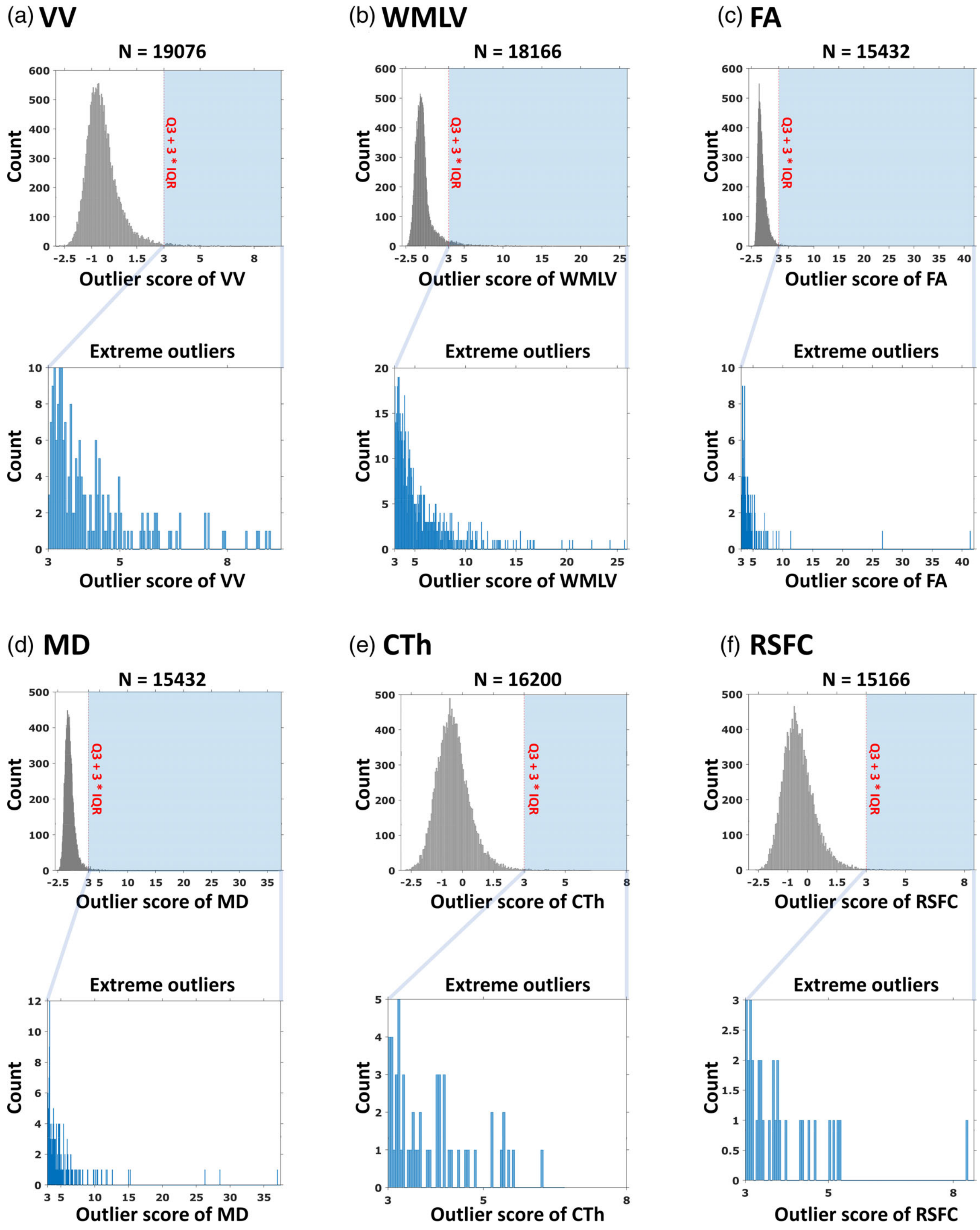


FIGURE 2 Outlier score histograms. (a) Ventricular volume (VV). (b) White matter lesion volume (WMLV). (c) Fractional anisotropy (FA). (d) Mean diffusivity (MD). (e) Cortical thickness (CTh). (f) Resting-state functional connectivity (RSFC). The zoom panels show the outlier score histograms of extreme outlier subjects

of automatic quality control to exclude poor quality data before outlier detection. Interestingly, none of the VV outliers were associated with data collection/processing errors. More frequent data collection/processing errors were found in the WMLV, FA, or MD outliers as compared to the VV outliers (Table 2). The errors were found to be most frequent (22.2%, 12/54) in the CTh outliers. Some of the errors occurred at the data acquisition stage, due to head motion artifacts (Figure S4a) or the selection of a wrong FOV (Figure S4b). Others occurred at the data processing stage, such as incorrect segmentation (Figure S4c) or incorrect registration (Figure S4d).

Of the remaining 1,171 outliers (954 distinct subjects) that did not have data collection/processing errors, 120 distinct subjects were reviewed by a board-certified neuroradiologist (Table 2). These 120 subjects included all top-ranked outliers and randomly sampled non-top-ranked outliers, and the outlier scores of this representative subgroup spanned almost the whole range above $Q3 + 3 \times IQR$ of

TABLE 1 Summary of outlier score distributions for the main dataset (UKB discovery group)

Phenotype	VV	WMLV	FA	MD	CTh	RSFC[a]
Number of subjects	19,076	18,166	15,432	15,432	16,200	15,166
Skewness	1.92	4.28	7.38	7.01	0.99	1.01
Kurtosis	11.47	36.56	227.61	156.27	6.20	6.01
Number of outliers	190 (1.0%)	706 (3.9%)	134 (0.9%)	174 (1.1%)	54 (0.3%)	33 (0.2%)

Abbreviations: CTh, cortical thickness; FA, fractional anisotropy; MD, mean diffusivity; RSFC, resting-state functional connectivity; UKB, UK Biobank; VV, ventricular volume; WMLV, White matter lesion volume.

[a] Full correlations without global signal regression.

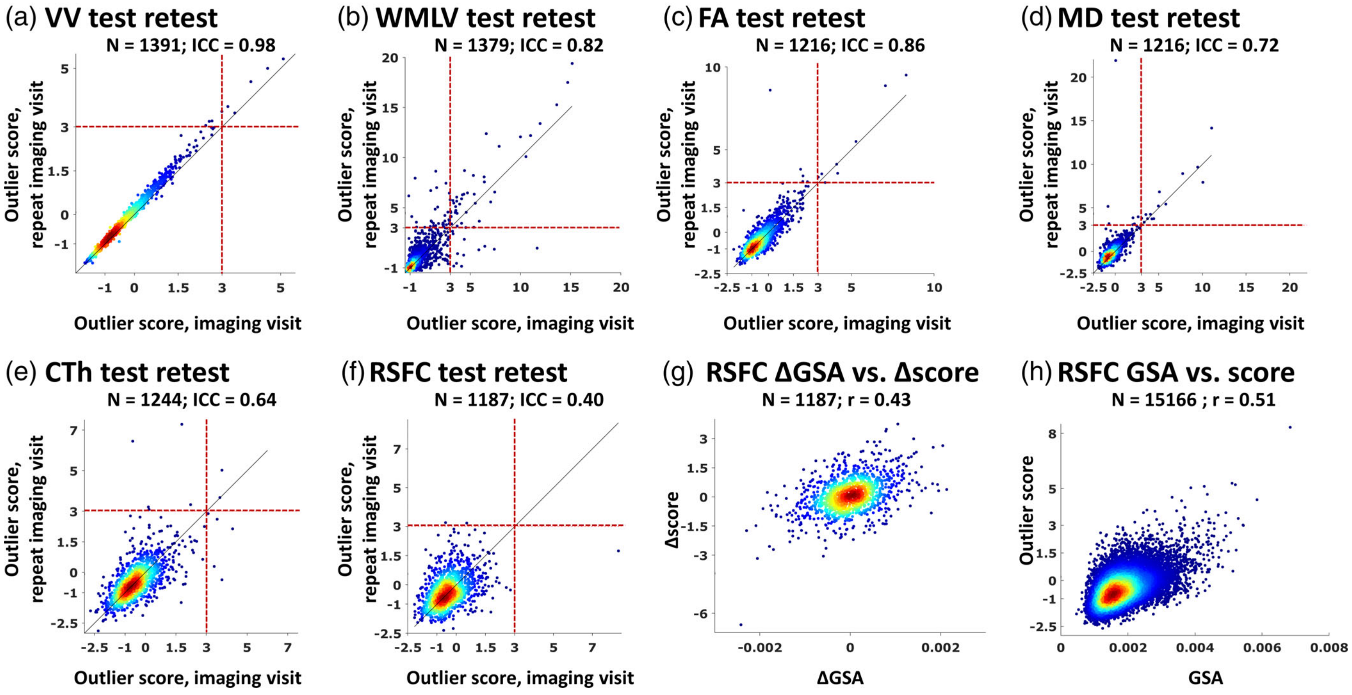


FIGURE 3 Long-term test–retest reliability of outlier scores. (a) Ventricular volume (VV). (b) White matter lesion volume (WMLV). (c) Fractional anisotropy (FA). (d) Mean diffusivity (MD). (e) Cortical thickness (CTh). (f) Resting-state functional connectivity (RSFC). For (a–f), in each scatterplot, each subject's outlier score of the initial imaging visit (also known as "test"; year 2014+) is plotted against the outlier score of the first repeat imaging visit (also known as "retest"; year 2019+). ICC: intraclass correlation between outlier scores of the two visits. Red dashed line: $Q3 + 3 \times IQR$. (g) The scatterplot of test–retest global signal amplitude (GSA) change versus test–retest RSFC outlier score change. For (a–g), only the UK Biobank (UKB) subjects that had both test and retest data available are shown in these scatterplots. (h) The scatterplot of GSA versus RSFC outlier score (RSFC calculated using full correlations)

the outlier score distribution in each imaging phenotype (see Figure S2 for details). In this subgroup, 117 subjects (97.5%, 117/120) were identified with radiological findings, and these findings covered a diverse category of phenotypes, such as large ventricles, masses, cysts, white matter lesions, infarcts, encephalomalacia, and prominent sulci. Representative individual outlier subjects are reported in the next few subsections per imaging phenotype.

3.4 | Individual outliers of VV

As an example, Figure 4a shows a VV outlier subject versus a normal subject. This subject had significantly enlarged lateral ventricles compared to a normal one (about 7.9 × *IQR* away between these two subjects in VV outlier score distribution). Forty-one of the VV outliers were reviewed by the neuroradiologist. Thirty-eight of the VV outliers being read were identified with radiological findings of large ventricles. Some of them had relatively clear etiology: A third ventricle mass (possibly a choroid plexus papilloma), a fourth ventricle mass (possibly an ependymoma), a colloid cyst, and a frontoparietal arachnoid cyst, all of which could cause obstructive hydrocephalus, were found in four VV outlier subjects, respectively (Figure 4b). The other major pathologies identified in the VV outliers were infarcts, nodules, agenesis of corpus callosum, and white matter lesions (Figure S5a).

In addition, a few VV outliers that were read would be potentially interesting for follow-up because they had large ventricles of unknown etiology and they did not present any other noticeable pathology (Figure 4c, left panel). Such VV outliers of unknown

TABLE 2 Summary of radiological review results of the outlier subjects in the main dataset

Phenotype		VV	WMLV	FA	MD	CTh
Outliers without data issue		190 (100%)	640 (90.7%)	129 (96.3%)	170 (97.7%)	42 (77.8%)
Outliers read by neuroradiologist		41	62	37	37	18
Radiological comments	Normal			2		1
	Large ventricles	38	18	9	8	5
	White matter lesions	27	62	29	31	9
	Mass	2	1			1
	Cyst	4	1	1	2	1
	Infarct	6	16	9	11	4
	Encephalomalacia			3	3	
	Prominent sulci	2	1	3	1	3
	Other findings	4	9	11	10	3

Note: Empty entries are zeros.

Abbreviations: CTh, cortical thickness; FA, fractional anisotropy; MD, mean diffusivity; RSFC, resting-state functional connectivity; VV, ventricular volume; WMLV, White matter lesion volume.

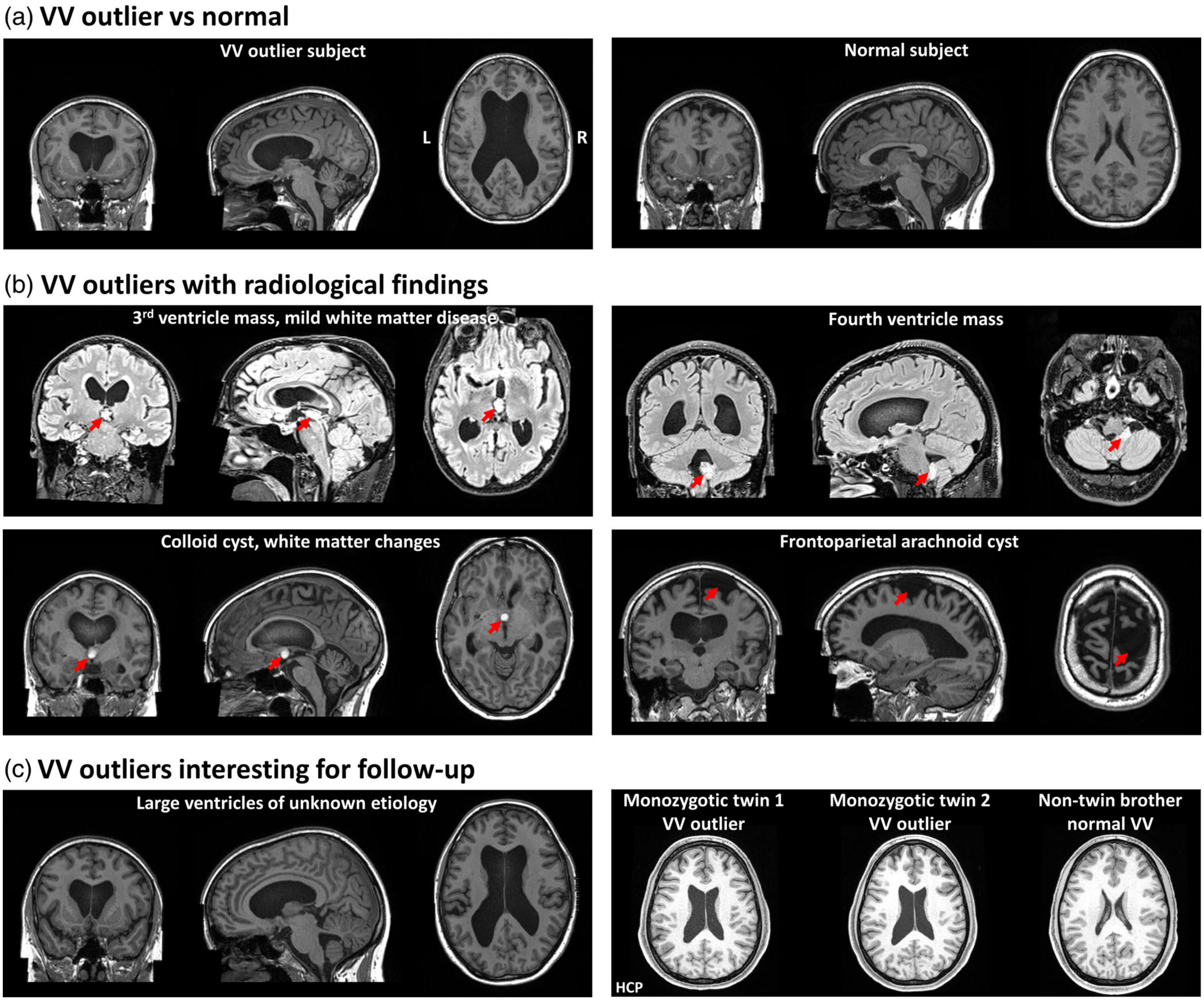

(a) VV outlier vs normal
VV outlier subject
L
R
Normal subject
(b) VV outliers with radiological findings
3rd ventricle mass, mild white matter disease
Fourth ventricle mass
Colloid cyst, white matter changes
Frontoparietal arachnoid cyst
(c) VV outliers interesting for follow-up
Large ventricles of unknown etiology
Monozygotic twin 1 VV outlier
Monozygotic twin 2 VV outlier
Non-twin brother normal VV
HCP

FIGURE 4 Individual outliers of ventricular volume (VV). (a) Structural images of an example of a VV outlier subject (left column) and an example of a normal VV subject (right column). (b) Structural images showing radiological findings in four representative VV outlier subjects. (c) Structural images of VV outlier subjects interesting for follow-up. Left column: A subject with large ventricles of unknown etiology. Right column: Structural images of a family in the Human Connectome Project (HCP) dataset (monozygotic twins and their non-twin brother). The twins had large ventricles of unknown etiology, but their non-twin brother had normal VV. The structural images in (a), (b), and the left column of (c) are reproduced by kind permission of UK Biobank ©

etiology were also present in the HCP dataset. In one family (Figure 4c, right panel), female monozygotic twins were both VV outliers, but their non-twin brother had normal VV; in another family (Figure S5b), one twin of a male monozygotic twin pair was a VV outlier, 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 outliers were associated with multiple different brain pathologies, and some of them had uncertain etiology requiring additional follow-up investigation.

3.5 | Individual outliers of white matter-based imaging phenotypes

Outlier detection of white matter-based imaging phenotypes was performed with WMLV, FA, and MD, respectively. As an example, Figure 5a shows a WMLV outlier subject versus a normal subject (about 26.7 × *IQR* away between these two subjects in WMLV outlier score distribution). The outlier subject had irregular periventricular white matter lesions extending into the deep white matter with large confluent areas, whereas the example normal subject had only tiny lesions on the periventricular caps. Figure 5b shows regional FA deviation maps of an FA outlier subject versus a normal subject (about 9.9 × *IQR* away between these two subjects in FA outlier score distribution). For this representative outlier subject, regional FA negatively deviated in all 27 white matter ROI used in this study, whereas the FA of the representative normal subject had almost no deviations. Figure S6a shows regional MD deviation maps of an MD outlier subject versus a normal subject (about 5.5 × *IQR* away between these two subjects in MD outlier score distribution), in which a large positive MD deviation was observed in the left superior longitudinal fasciculus of this outlier subject.

A proportion of the white matter outliers without any data acquisition or processing errors were reviewed by the neuroradiologist, and most of them were identified with radiological findings: This includes all the reviewed WMLV outliers, 94.6% (35/37) of the reviewed FA outliers, and all the reviewed MD outliers (Table 2). For instance, likely multiple sclerosis was identified in a subject who was an outlier in WMLV, FA, and MD (Figure 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 an outlier in all three white matter-based imaging phenotypes (Figure 5c). A parahippocampal cyst was identified in an MD outlier subject (Figure 5c). Encephalomalacia (Figure 5c) was identified in a subject who was an outlier in both FA and MD.

The etiology of the findings in some white matter outliers was uncertain. For example, the left panel of Figure 5d shows an outlier subject in WMLV, FA, and MD measures, who was read as having severe biparietal atypically distributed white matter lesions of unknown etiology. Other than the subjects with radiological findings, a small number of the white matter outlier subjects reviewed appeared normal to the neuroradiologist (Figure 5d, right panel, and Figure S6b). For example, the right panel of Figure 5d shows an FA outlier subject had an anomalously low FA value in the genu of corpus callosum specifically, but his T1w and T2w FLAIR images were normal-appearing. All these outliers of unknown etiology and normal-appearing outliers would be interesting for follow-up to determine the mechanism or whether they eventually present with specific clinical symptoms. Taken together, these results indicate that the non-artifactual outliers of white matter-based imaging phenotypes were associated with a large variety of different radiological findings. Normal-appearing outliers, each with unique FA or MD patterns, only constituted a small fraction of the white matter outliers.

3.6 | Individual outliers of CTh

We next examined the individuals with outlying CTh. As an example, Figure 6a shows regional CTh deviation maps of an outlier subject versus a normal subject (about 5.5 × *IQR* away between these two subjects in CTh outlier score distribution). Widespread negative CTh deviations, representing thinner cortices in these regions, were observed in this outlier subject. Among the non-artifactual CTh outlier subjects that were read, 94.4% (17/18) of them were identified with radiological findings, such as prominent sulci, atrophy, or an infarct affecting the nearby cortices (Figure 6b). Taken together, these results suggest that most non-artifactual CTh outliers were associated with radiological findings.

3.7 | Outlier score relationships across imaging phenotypes

The relationship of outlier scores across different imaging phenotypes was assessed via pairwise Pearson correlation coefficients (Figure 7a). The correlation between some white matter-based imaging phenotypes (WMLV vs. MD) was moderate ($.4 < r < .6$), indicating they can capture similar outlying patterns in the white matter. All the other correlations were weak ($.2 < r < .4$) or very weak ($r < .2$), indicating they were complementary and provided independent information.

To further illustrate these relationships, Figure 7b shows a scatterplot of WMLV versus VV outlier scores, which were weakly correlated ($r = .25$). Very few subjects were both VV and WMLV outliers, as evidenced by the sparser data in Zone II than Zone I or III (Figure 7b). Indeed, the density of Zone II was significantly lower than the other two zones ($p \approx 0$, one-way ANOVA of 100,000 bootstrap samples, Figure S7a). 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, Figure 7c shows a scatterplot of WMLV versus FA outlier scores ($r = .36$). The density of Zone II was significantly lower than Zone III ($p \approx 0$, one-way ANOVA of 100,000 bootstrap samples, Figure S7b) but was close to Zone I. Figure S7c shows two examples of the outliers in Zone II. The upper panel of Figure S7c shows a subject that was an outlier in both VV and WMLV. This subject, diagnosed with

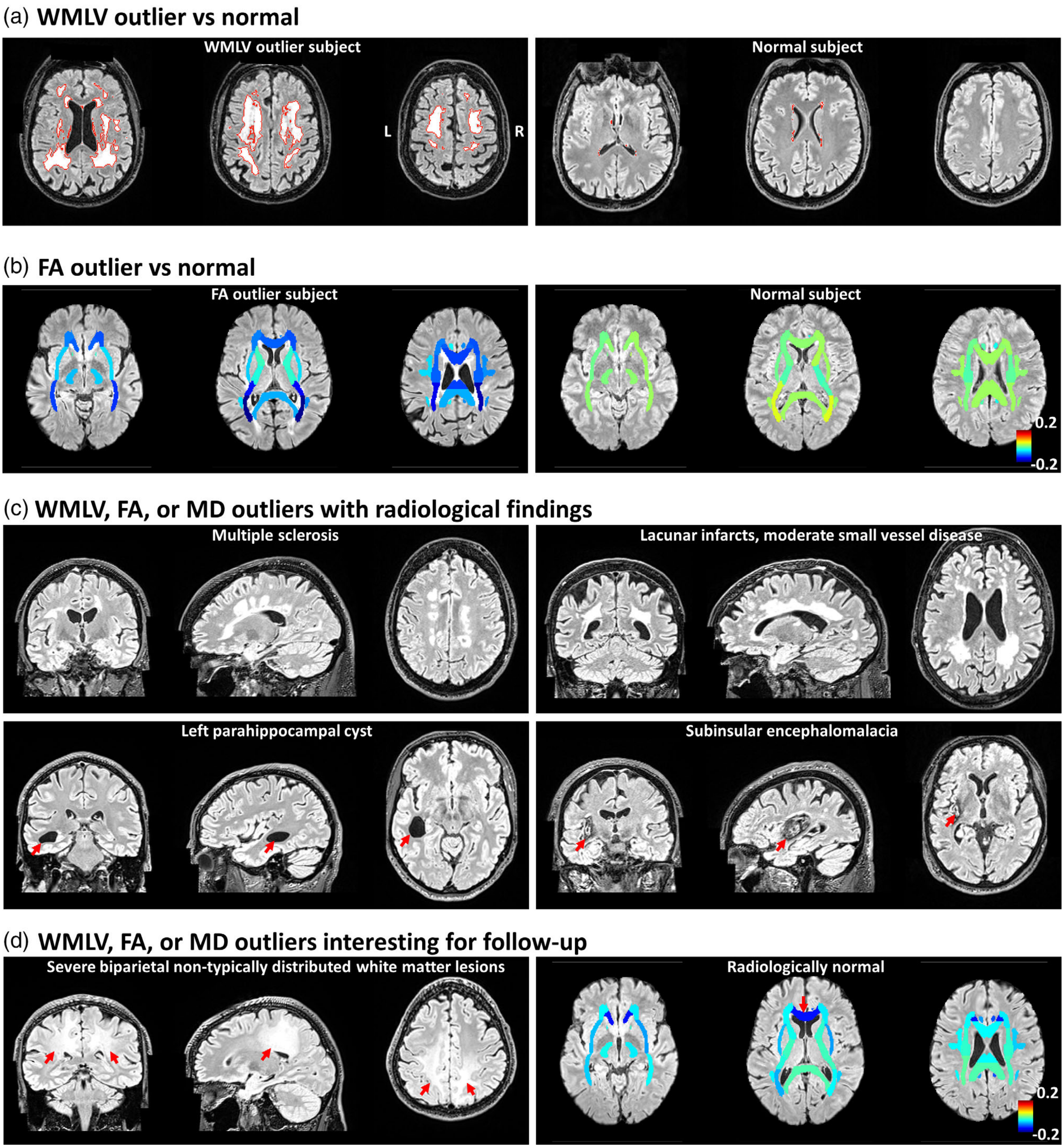

(a) WMLV outlier vs normal
WMLV outlier subject
L
R
Normal subject
(b) FA outlier vs normal
FA outlier subject
Normal subject
0.2
-0.2
(c) WMLV, FA, or MD outliers with radiological findings
Multiple sclerosis
Lacunar infarcts, moderate small vessel disease
Left parahippocampal cyst
Subinsular encephalomalacia
(d) WMLV, FA, or MD outliers interesting for follow-up
Severe biparietal non-typically distributed white matter lesions
Radiologically normal
0.2
-0.2

FIGURE 5 Individual outliers of white matter-based imaging phenotypes. (a) T2w FLAIR images of an example of a white matter lesion volume (WMLV) outlier subject (left column) and an example of a normal WMLV subject (right column). The red line represents the boundary of white matter lesions. (b) Regional fractional anisotropy (FA) deviation maps (overlaid on T2w FLAIR images) of an example of an FA outlier subject (left column) and an example of a normal FA subject (right column). (c) Structural images showing radiological findings in representative WMLV, FA, or mean diffusivity (MD) outlier subjects of multiple sclerosis (a WMLV, FA, and MD outlier), lacunar infarcts with moderate small vessel disease (a WMLV, FA, and MD outlier), cyst (an MD outlier), and encephalomalacia (an FA and MD outlier). (d) WMLV, FA, or MD outlier interesting for follow-up. Left column: T2w FLAIR images of an outlier subject with severe biparietal nontypical distributed white matter lesions of uncertain etiology (a WMLV, FA, and MD outlier). Right column: regional FA deviation map (overlaid on T2w FLAIR images) of an FA outlier subject that was radiologically normal. FA was anomalously low in the genu of corpus callosum and the cause was unknown. For the regional FA deviation maps in (b) and (d), each 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 region of interest (ROI) is displayed in its full size instead of only the Tract-Based Spatial Statistics (TBSS) skeleton. The structural images in this figure are reproduced by kind permission of UK Biobank ©

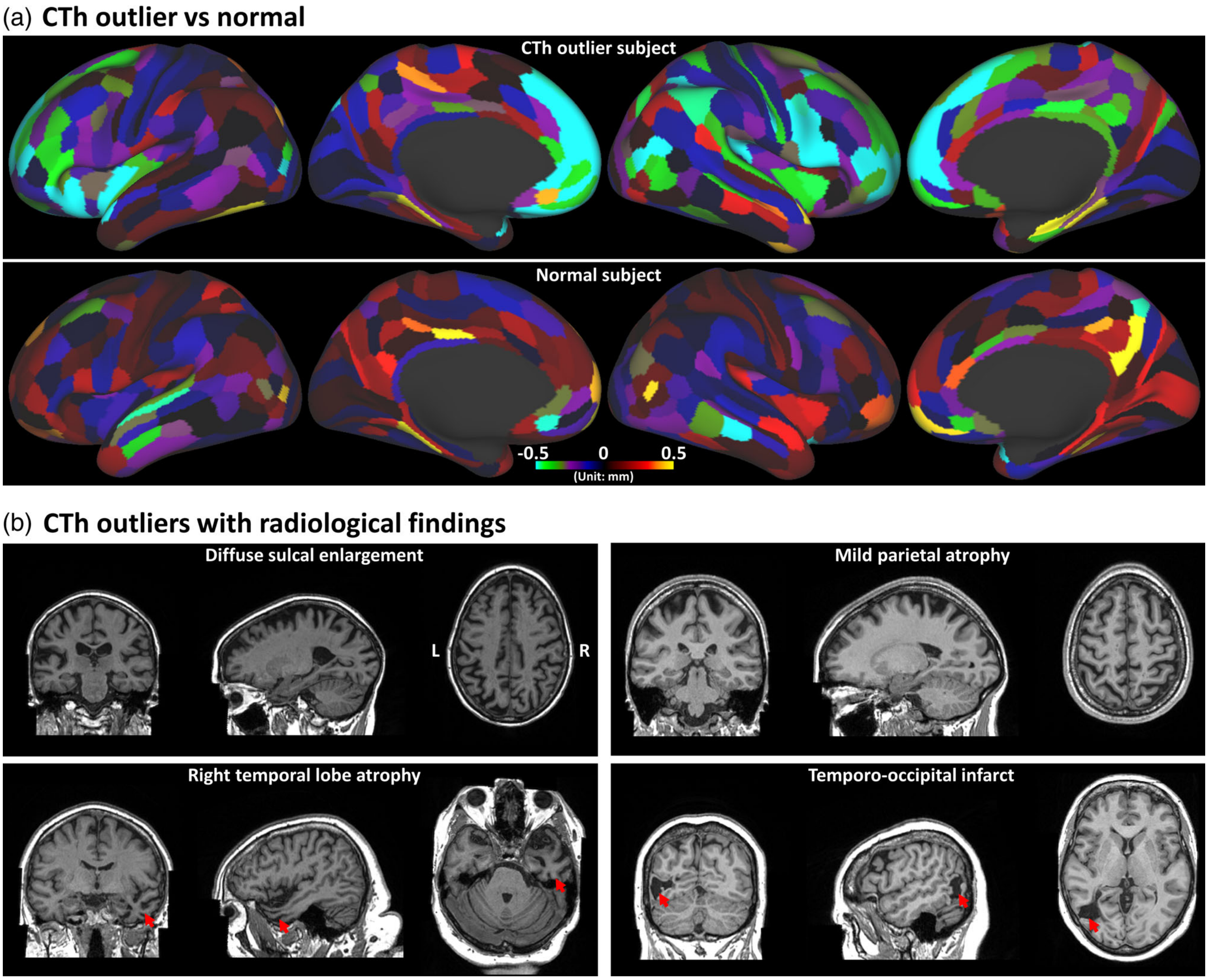

(a) CTh outlier vs normal
CTh outlier subject
Normal subject
-0.5
0
0.5
(Unit: mm)
(b) CTh outliers with radiological findings
Diffuse sulcal enlargement
L
R
Mild parietal atrophy
Right temporal lobe atrophy
Temporo-occipital infarct

FIGURE 6 Individual outliers of cortical thickness (CTh). (a) Regional CTh deviation maps (displayed on inflated cortical surfaces) of an example of a CTh outlier subject (first row) and an example of a normal CTh subject (second row). A regional CTh deviation map visualizes how the CTh values in a subject deviate from the autoencoder-predicted CTh values. (b) Structural images showing radiological findings in four representative CTh outlier subjects. These images are reproduced by kind permission of UK Biobank ©

ventriculomegaly and moderate white matter disease, had both periventricular and deep white matter lesions. The lower panel of Figure S7c shows a subject that was an outlier in all VV, WMLV, FA, and MD. The radiological read determined there was small vessel disease evidenced by the white matter lesions, and possible Alzheimer's disease evidenced by the parieto-temporal atrophy.

4 | DISCUSSION

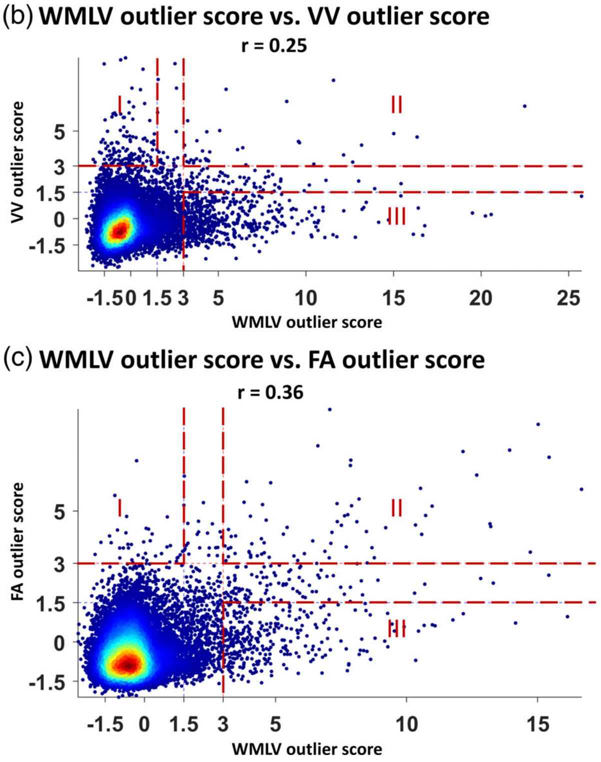
In this study, a semiautomated, two-level outlier detection and screening methodology was used to investigate outliers in the MRI phenotypes of VV, WMLV, FA, MD, CTh, and RSFC (Figure 1). Outlier score distributions were all more outlier-prone than a standard normal distribution (Figure 2). Except for RSFC, outlier scores of these imaging phenotypes had good-to-excellent reliability as assessed by test–retest ICC of outlier scores (Figure 3). Due to the low test–retest reliability of RSFC outlier scores, RSFC outliers were excluded from further individual-level outlier analyses (Figures 3 and S3). Through the screening of individual outliers, outliers of most imaging phenotypes were associated with no or very few data collection/processing errors, whereas the errors were found to be most frequent in CTh outliers (Table 2). Among the non-artifactual outliers being reviewed radiologically, most were associated with radiological findings (Table 2, Figures 4bc, 5c, 6b, and S5), though a small fraction appeared normal to the neuroradiologist (Figure 5d, right panel, and Figure S6b). The outlier scores of different imaging phenotypes were mostly independent, indicating that they each added information (Figure 7).

4.1 | Evaluation of unsupervised outlier detection

A common practice to evaluate an unsupervised outlier detection approach is to apply the method to a labeled dataset to calculate

(a) **Outlier score correlations across imaging phenotypes**

	WMLV	FA	MD	CTh
VV	0.25	0.19	0.30	0.12
WMLV		0.36	0.51	-0.00
FA			0.38	0.05
MD				0.06


(b) **WMLV outlier score vs. VV outlier score**

(c) **WMLV outlier score vs. FA outlier score**

FIGURE 7 Relationship between outlier scores of different imaging phenotypes. (a) Correlations between the outlier scores of different imaging phenotypes. The two representative relationships shown in (b) and (c) are encircled with red boxes. (b) White matter lesion volume (WMLV) outlier score plotted against ventricular volume (VV) outlier score. Zone I covered the VV outliers with normal WMLV. Zone II covered the subjects who were outliers in both VV and WMLV. Zone III covered the WMLV outliers with normal VV. (c) WMLV outlier score plotted against fractional anisotropy (FA) outlier score. Zone I covered the FA outliers with normal WMLV. Zone II covered the subjects who were outliers in both FA and WMLV. Zone III covered the WMLV outliers with normal FA

outlier scores without using the labels first, and these labels were used later as the ground truth when evaluating the performance of the unsupervised method (Aggarwal, 2017; Goldstein & Uchida, 2016). However, in the present study, the UKB data were unlabeled. Gibson et al. examined 1,000 subjects of this cohort radiologically (Gibson et al., 2017), but unfortunately, we were unsuccessful in obtaining their radiological annotations. It should be noted that the outliers defined in the present study were composed of not only the subjects with radiological findings, but also the subjects with data collection/processing errors that were not eliminated by the automatic quality control, as well as the radiologically normal-appearing outlier subjects, who still had large deviations from the group average (Figure 5d, right panel). These radiologically normal-appearing outlier subjects are interesting because they could be the ones at higher risk to develop noticeable pathologies (de Groot et al., 2013). Therefore, instead of using any existing labels, we opted to evaluate our approach by quantifying how well the outlier scores align with the amounts of deviations from the group averages. For unidimensional imaging phenotypes, the outlier scores were linearly transformed from unidimensional volume measurements, so they exactly quantified such deviations. For multidimensional imaging phenotypes, the amount of individual deviations can be measured by the correlation distance between each subject and the group average, and the autoencoder-derived outlier scores were found to be significantly correlated with such individual deviations (Figure S8). The outlier scores also showed larger dynamic ranges than the amounts of deviations from the group averages (slopes >1; Figure S8), indicating that they were more sensitive in distinguishing small differences of outlierness than using the amounts of deviations from the group averages. Taken together, these results confirm that the outlier scores reliably characterized the degree of individual deviations from the group averages in this unlabeled dataset.

4.2 | The approach to screen individual outliers in a large neuroimaging dataset

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. There has been much less work on studying outlying individuals and the associated imaging phenotypes in these large neuroimaging datasets (Marquand et al., 2016; Mourao-Miranda et al., 2011; Pinaya et al., 2019; van Hespen et al., 2021). To begin to fill this gap, we set out to investigate individual outliers from more than 15,000 UKB subjects.

Outlier detection was performed for all the commonly studied, well-established brain imaging phenotypes. The data of these imaging phenotypes were well-curated, as most acquisition and processing errors were excluded from the data using eight quantitative quality control metrics, and the image quality metrics that were correlated with the outlier scores were regressed out (Figure S9). However, a small fraction of the acquisition and processing errors were not accounted for in the automatic quality control. Our outlier detection approach was able to capture more data collection/processing errors (Figure S4) that were missed in the quality control. In addition, the outlier score assigned to each individual can be utilized as a useful summary index for assessing the effectiveness of different data denoising strategies at the individual level. For example, three common processing strategies regarding the global signal in RSFC were evaluated in this way (Figures 3g–h and S3). Taken together, our method is valuable for curating large neuroimaging datasets.

A neuroradiologist read the structural images of non-artifactual outliers. A large percentage (97.5%, 117/120) of them had radiological findings, such as large ventricles, 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 outlier subject (Figure 4b) was diagnosed with a colloid cyst causing hydrocephalus, and the neuroradiologist's read recommended this individual to see a neurosurgeon for follow-up. The aforementioned manual radiological screening study (Gibson et al., 2017) of the first 1,000 UKB subjects showed that only 1.8% of the UKB subjects screened via systematic radiologist review had radiological findings in their brain MRIs. The much higher percentage (97.5%) of the subjects identified with radiological findings among our outlier subjects indicates that our method can effectively identify a subgroup that is greatly enriched with radiological findings from a large dataset.

4.3 | Potential underlying mechanisms of outlier subjects with unknown etiology or were radiologically normal

Among the outliers identified with radiological findings, a few presented with unknown etiology. For example, eight UKB and five HCP VV outlier subjects had very large ventricles of uncertain etiology. The VV of these UKB subjects, ranging between 87.4 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 was between 45.4 and 56.2 ml, which were still much larger than the volumes of normal young healthy subjects. Interestingly, the data also showed unexplained variations of VV between two monozygotic twin pairs in these HCP VV outlier subjects. Two female individuals within the first monozygotic twin pair had anomalously large VV (Figure 4c, right panel), suggesting shared congenital, developmental, or environmental causes. In another monozygotic twin pair, only one twin had anomalous large VV (Figure S5b). This is probably due to environmental influences or a de novo mutation early in development.

Another interesting case of unknown etiology was a UKB subject who was an outlier for WMLV, FA, and MD (Figure 5d, left panel). In this subject, severe bilateral, confluent, and symmetrical white matter lesions were identified in the parietal white matter. Such lesion patterns were different from small vessel disease or multiple sclerosis, but were 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 outlier subject with unknown etiology.

Two UKB FA outliers, two HCP FA outliers, and one HCP MD outlier were not identified with any data collection/processing issues or radiological findings, which are potentially interesting for investigating the underlying mechanisms of their large FA or MD deviations. For these FA outliers, anomalously low FA values were found either in the corpus callosum, superior longitudinal fasciculus, cingulum, posterior thalamic radiation, or limbs of the internal capsule (Figure 5d, right panel, and Figure S6b). A previous study showed that low FA in normal-appearing white matter preceded the conversion of low FA regions into white matter lesions (de Groot et al., 2013). Therefore, these FA outlier subjects may be at risk to develop lesions later in the regions of anomalously low FA. For the MD outlier subject, many small perivascular spaces were found on his structural MRI image. These perivascular spaces were not abnormal but could be responsible for the increased MD. Taken together, all the outliers discussed above would benefit from follow-up assessments to study underlying mechanisms and to see if they progress to any known clinical phenotype.

4.4 | Generalizability of outlier detection to new UKB subjects

Since the UKB will ultimately enroll 100,000 subjects for brain imaging, it is important to verify that the outlier detection method used on the first more than 15,000 subjects in the present study can be applied to the rest of this population. Therefore, we made use of the second 20,000 UKB subjects released recently as a separate group for assessing the generalizability (referred to as *UKB replication group*; see Table S3 for detailed demographic information). These two groups were of comparable size and had no overlapping subjects. For each unidimensional imaging phenotype, generalizability was assessed by directly comparing the outlier score distribution obtained from the discovery group against the distribution obtained from the replication group, and no significant difference was found between these two distributions (Figure S10a,b. Two-sample Kolmogorov–Smirnov tests: for VV, $p = .41$; for WMLV, $p = .62$). For each multidimensional imaging phenotype, first, the generalizability of the discovery group subjects' outlier scores was evaluated by the ICC between two sets of their outlier scores calculated separately using two different autoencoders: one autoencoder trained using the discovery group itself, and another autoencoder trained using the replication group. The results showed that the ICC ranged from 0.87 to 0.99 (Figure S10c–e). Second, the generalizability of the replication group subjects' outlier scores was evaluated by the ICC between two sets of their outlier scores calculated separately using two different autoencoders: one autoencoder trained using the replication group itself, and another autoencoder trained using the discovery group. The results showed that the ICC ranged from 0.96 to 0.99 (Figure S10c–e). Taken together, these results indicate that the UKB replication group is consistent with the UKB discovery group, and the results suggest that our trained outlier detection models can be generalized to new UKB subjects.

4.5 | Impact of neuroimaging data processing software on outlier detection

The choice of neuroimaging data processing software can have a substantial impact on analysis results (Botvinik-Nezer et al., 2020; Velazquez, Mateos, Pasaye, Barrios, & Marquez-Flores, 2021). Using

FreeSurfer and CAT12 as examples, both software tools can quantify CTh from MRI data, but they presented systematic differences in CTh estimations (Righart et al., 2017; Seiger, Ganger, Kranz, Hahn, & Lanzenberger, 2018). Therefore, it is interesting to evaluate whether a different data processing software could change the outlier detection results. To this end, for each imaging phenotype, we chose an additional neuroimaging data processing software different from the one used in the main Methods for obtaining an additional set of preprocessed data (for processing details, see Supplementary Methods) and then carried out outlier detection on them. The results showed that outlier detection in most imaging phenotypes was not affected in a major way by using a different data processing software, as indicated by the strong positive correlations (r: .78–.95) between these two sets of outlier scores (Figure S11a–d), as well as the relatively high Dice similarity coefficients (DC: 0.71–0.83) between the two sets of outliers (Table S5). An exception was CTh, which only had a moderate positive correlation ($r = .46$) between FreeSurfer-based outlier scores versus CAT12-based outlier scores (Figure S11e) and a low Dice similarity coefficient ($\mathrm{DC} = 0.13$) between the outliers (Table S5). This could be explained by the aforementioned systematic CTh difference between the two software (Righart et al., 2017; Seiger et al., 2018). We opted to use FreeSurfer-based results in the present study because its surface-based approach is advantageous in alleviating partial volume effects (Velazquez et al., 2021).

4.6 | Technical considerations

There are a few technical considerations in the present study. First, for the subjects without usable T2w FLAIR images, ventricles were segmented using only T1w images. This choice should not affect the outlier detection of VV, because in our additional analysis of comparing T1w-only ventricle segmentation versus T1w and T2w combined ventricle segmentation in 18,031 UKB subjects, a close-to-one correlation ($r = .999$) was found between the VV values obtained from these two approaches. Second, spatial smoothing was not applied on the CTh data because more smoothing can degrade neuroanatomical features and individual variability of CTh (Zeighami & Evans, 2021), and it is generally suggested to avoid spatial smoothing in the HCP-style approach (Coalson, Van Essen, & Glasser, 2018). Third, we used the HCP multimodal parcellation atlas (Glasser, Coalson, et al., 2016) for parcellating CTh, because it is one of the most comprehensive atlases of human cortical areas. Since this choice may be arbitrary, we performed CTh outlier detection using another well-established parcellation atlas (Fan et al., 2016), and choosing this different atlas did not significantly affect the results: A strong positive correlation ($r = .77$) was found between these two sets of outlier scores derived based on different atlases (Figure S12), and the two sets of outliers also showed relatively high similarity ($\mathrm{DC} = 0.55$). Nevertheless, finding an optimal parcellation for outlier detection remains an open research topic in the field.

4.7 | Conclusions

The present study characterized individual outliers across multiple brain MRI phenotypes from more than 15,000 subjects. Every subject was parameterized with an outlier score per imaging phenotype to quantitate the outlierness. Outlier score distributions were all more outlier-prone than a standard normal distribution. The approach enabled the assessment of test–retest reliability via the outlier scores, which ranged from excellent reliability for VV, WMLV, and FA, to good reliability for MD and CTh. RSFC was excluded for individual-level outlier screening due to its low test–retest reliability. The individual-level analyses of the outliers revealed that a small number of outliers were due to data collection/processing errors, demonstrating the usefulness of outlier detection in curating large neuroimaging datasets. Most of the remaining non-artifactual outliers were due to different brain pathologies as determined by a neuroradiologist, indicating that our approach can effectively identify a subgroup that is greatly enriched with radiological findings from a large unlabeled cohort. Several outliers had unknown etiology or were normal-appearing, and these outliers are candidates for follow-up to determine the mechanism or whether they eventually progress to a clinical phenotype. Our analysis suggests that unsupervised outlier detection of large neuroimaging datasets is valuable for data curation, reliability assessment, and identification of individuals for medical follow-up or further study of novel mechanisms. Outlier detection methods should contribute to the effort of developing automatic processes to analyze and interpret brain imaging data in large population cohorts.

ACKNOWLEDGMENTS

This study was supported by NIH/NINDS Intramural Research Program (Project numbers: NS002989 and NS003119). The authors 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).

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

AUTHOR CONTRIBUTIONS

Zhiwei Ma, Daniel S. Reich, Jeff H. Duyn, and **Alan P. Koretsky**: Designed the study. **Zhiwei Ma**, **Daniel S. Reich**, and **Sarah Dembling**: Performed the analyses. **Zhiwei Ma**: Drafted the initial manuscript. **Zhiwei Ma, Daniel S. Reich, Sarah Dembling, Jeff H. Duyn, and Alan P. Koretsky:** Revised the manuscript.

DATA AVAILABILITY STATEMENT

Access to the UK Biobank data can be requested via Access Management System at https://bbams.ndph.ox.ac.uk/ams/. The Human Connectome Project data are publicly available for download upon registration at https://db.humanconnectome.org.

ORCID

Zhiwei Ma https://orcid.org/0000-0003-2928-402X
Daniel S. Reich https://orcid.org/0000-0002-2628-4334
Alan P. Koretsky https://orcid.org/0000-0002-8085-4756

REFERENCES

Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-3-319-47578-3

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., … Smith, S. M. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, *166*, 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., … Biobank, U. (2012). UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, *1*(3), 123–126. https://doi.org/10.1016/j.hlpt.2012.07.003

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. https://doi.org/10.1016/j.neuroimage.2015.10.019

Basser, P. J., Mattiello, J., & LeBihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance. Series B*, *103*(3), 247–254. https://doi.org/10.1006/jmrb.1994.1037

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Coalson, T. S., Van Essen, D. C., & Glasser, M. F. (2018). The impact of traditional neuroimaging methods on the spatial localization of cortical areas. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(27), E6356–E6365. https://doi.org/10.1073/pnas.1801582115

Dahnke, R., Yotter, R. A., & Gaser, C. (2013). Cortical thickness and central surface estimation. *NeuroImage*, *65*, 336–348. https://doi.org/10.1016/j.neuroimage.2012.09.050

de Groot, M., Verhaaren, B. F., de Boer, R., Klein, S., Hofman, A., van der Lugt, A., … Vernooij, M. W. (2013). Changes in normal-appearing white matter precede development of white matter lesions. *Stroke*, *44*(4), 1037–1042. https://doi.org/10.1161/STROKEAHA.112.680223

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Dickie, E. W., Anticevic, A., Smith, D. E., Coalson, T. S., Manogaran, M., Calarco, N., … Voineskos, A. N. (2019). Ciftify: A framework for surface-based analysis of legacy MR acquisitions. *NeuroImage*, *197*, 818–826. https://doi.org/10.1016/j.neuroimage.2019.04.078

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., … Jiang, T. (2016). The Human Brainnetome Atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, *26*(8), 3508–3526. https://doi.org/10.1093/cercor/bhw157

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Gaser, C., & Dahnke, R. (2016). CAT-A computational anatomy toolbox for the analysis of structural MRI data. Presented at the 22nd Annual Meeting of the Organization for Human Brain Mapping.

Geraldes, R., Ciccarelli, O., Barkhof, F., De Stefano, N., Enzinger, C., Filippi, M., … Jacqueline Palace on behalf of the MAGNIMS Study Group. (2018). The current role of MRI in differentiating multiple sclerosis from its imaging mimics. *Nature Reviews. Neurology*, *14*(4), 213. https://doi.org/10.1038/nrneurol.2018.39

Gibson, L. M., Littlejohns, T. J., Adamska, L., Garratt, S., Doherty, N., UK Biobank Imaging Working Group, … Sudlow, C. L. (2017). Impact of detecting potentially serious incidental findings during multi-modal imaging. *Wellcome Open Research*, *2*, 114. https://doi.org/10.12688/wellcomeopenres.13181.3

Gilmore, A. D., Buser, N. J., & Hanson, J. L. (2021). Variations in structural MRI quality significantly impact commonly used measures of brain anatomy. *Brain Informatics*, *8*(1), 7. https://doi.org/10.1186/s40708-021-00128-2

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., … Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L., Auerbach, E. J., Behrens, T. E., … Van Essen, D. C. (2016). The Human Connectome Project's neuroimaging approach. *Nature Neuroscience*, *19*(9), 1175–1187. https://doi.org/10.1038/nn.4361

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., … WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS One*, *11*(4), e0152173. https://doi.org/10.1371/journal.pone.0152173

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. *Cerebral Cortex*, *26*(1), 288–303. https://doi.org/10.1093/cercor/bhu239

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., … Smith, S. M. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage*, *95*, 232–247. https://doi.org/10.1016/j.neuroimage.2014.03.034

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

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*(4), 1471–1483. https://doi.org/10.1016/j.neuroimage.2012.03.027

Hagler, D. J., Jr., Hatton, S., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., … Dale, A. M. (2019). Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage*, *202*, 116091. https://doi.org/10.1016/j.neuroimage.2019.116091

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). *Outlier detection using replicator neural networks*. Berlin: Springer.

Holmes, A. J., Hollinshead, M. O., O'Keefe, T. M., Petrov, V. I., Fariello, G. R., Wald, L. L., … Buckner, R. L. (2015). Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. *Scientific Data*, *2*, 150031. https://doi.org/10.1038/sdata.2015.31

Hotz, I., Deschwanden, P. F., Liem, F., Mérillat, S., Kollias, S., & Jäncke, L. (2021). Performance of three freely available methods for extracting white matter hyperintensities: FreeSurfer, UBO detector and BIANCA. *bioRxiv*, 2020.2010.2017.343574. doi:https://doi.org/10.1101/2020.10.17.343574

Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., … Allen, N. E. (2020). The UK Biobank imaging enhancement of 100,000 participants: Rationale, data collection, management and future directions. *Nature Communications*, *11*(1), 2624. https://doi.org/10.1038/s41467-020-15948-9

Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies. *Biological Psychiatry*, *80*(7), 552–561. https://doi.org/10.1016/j.biopsych.2015.12.023

Maximov, I. I., van der Meer, D., de Lange, A. G., Kaufmann, T., Shadrin, A., Frei, O., … Westlye, L. T. (2021). Fast qualitY conTrol meThod foR derIved diffUsion Metrics (YTTRIUM) in big data analysis: U.K. Biobank 18,608 example. *Human Brain Mapping*. *42*(10), 3141–3155. https://doi.org/10.1002/hbm.25424

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., … Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393

Moller, M. F. (1993). A scaled conjugate-gradient algorithm for fast supervised learning. *Neural Networks*, *6*(4), 525–533. https://doi.org/10.1016/S0893-6080(05)80056-5

Monereo-Sanchez, J., de Jong, J. J. A., Drenthen, G. S., Beran, M., Backes, W. H., Stehouwer, C. D. A., … Jansen, J. F. A. (2021). Quality control strategies for brain MRI segmentation and parcellation: Practical approaches and recommendations - insights from the Maastricht study. *NeuroImage*, *237*, 118174. https://doi.org/10.1016/j.neuroimage.2021.118174

Mori, S., Oishi, K., Jiang, H., Jiang, L., Li, X., Akhter, K., … Mazziotta, J. (2008). Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. *NeuroImage*, *40*(2), 570–582. https://doi.org/10.1016/j.neuroimage.2007.12.035

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*(3), 793–804. https://doi.org/10.1016/j.neuroimage.2011.06.042

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage*, *211*, 116604. https://doi.org/10.1016/j.neuroimage.2020.116604

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. *Human Brain Mapping*, *40*(3), 944–954. https://doi.org/10.1002/hbm.24423

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. https://doi.org/10.1016/j.neuroimage.2014.10.044

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. W., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, *107*, 107–115. https://doi.org/10.1016/j.neuroimage.2014.12.006

Righart, R., Schmidt, P., Dahnke, R., Biberacher, V., Beer, A., Buck, D., … Muhlau, M. (2017). Volume versus surface-based cortical thickness measurements: A comparative study with healthy controls and multiple sclerosis patients. *PLoS One*, *12*(7), e0179590. https://doi.org/10.1371/journal.pone.0179590

Rosen, A. F. G., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., … Satterthwaite, T. D. (2018). Quantitative assessment of structural image quality. *NeuroImage*, *169*, 407–418. https://doi.org/10.1016/j.neuroimage.2017.12.059

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. https://doi.org/10.1016/j.neuroimage.2013.11.046

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*(1), 119–124. https://doi.org/10.1212/01.wnl.0000167542.89697.0f

Seiger, R., Ganger, S., Kranz, G. S., Hahn, A., & Lanzenberger, R. (2018). Cortical thickness estimations of FreeSurfer and the CAT12 toolbox in patients with Alzheimer's disease and healthy controls. *Journal of Neuroimaging*, *28*(5), 515–523. https://doi.org/10.1111/jon.12521

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., … Behrens, T. E. (2006). Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage*, *31*(4), 1487–1505. https://doi.org/10.1016/j.neuroimage.2006.02.024

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining* (1st ed.). Boston, MA: Pearson Addison Wesley.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

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

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. *Scientific Reports*, *11*(1), 7714. https://doi.org/10.1038/s41598-021-87013-4

Velazquez, J., Mateos, J., Pasaye, E. H., Barrios, F. A., & Marquez-Flores, J. A. (2021). Cortical thickness estimation: A comparison of FreeSurfer and three voxel-based methods in a test-retest analysis and a clinical application. *Brain Topography*, *34*(4), 430–441. https://doi.org/10.1007/s10548-021-00852-2

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. https://doi.org/10.1016/j.neuroimage.2013.07.057

Yendiki, A., Koldewyn, K., Kakunoori, S., Kanwisher, N., & Fischl, B. (2014). Spurious group differences due to head motion in a diffusion MRI study. *NeuroImage*, *88*, 79–90. https://doi.org/10.1016/j.neuroimage.2013.11.027

Zeighami, Y., & Evans, A. C. (2021). Association vs. prediction: The impact of cortical surface smoothing and parcellation on brain age. *Frontiers in Big Data*, *4*, 637724. https://doi.org/10.3389/fdata.2021.637724

SUPPORTING INFORMATION

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

How to cite this article: Ma, Z., Reich, D. S., Dembling, S., Duyn, J. H., & Koretsky, A. P. (2022). Outlier detection in multimodal MRI identifies rare individual phenotypes among more than 15,000 brains. *Human Brain Mapping*, *43*(5), 1766–1782. https://doi.org/10.1002/hbm.25756