frontiers | Frontiers in Neurology

TYPE Original Research
PUBLISHED 19 December 2022
DOI 10.3389/fneur.2022.960760

OPEN ACCESS

EDITED BY
Pedro Gomez-Vilda,
Polytechnic University of Madrid, Spain

REVIEWED BY
Lauren Jean O'Donnell,
Harvard Medical School, United States
Karmele López-de-Ipiña,
University of the Basque Country, Spain
Jiri Mekyska,
Brno University of Technology, Czechia
Rafael Martinez Olalla,
Polytechnic University of Madrid, Spain

*CORRESPONDENCE
Kenneth A. Weber II
kenweber@stanford.edu

SPECIALTY SECTION
This article was submitted to Neurological Biomarkers, a section of the journal Frontiers in Neurology

RECEIVED 03 June 2022
ACCEPTED 25 November 2022
PUBLISHED 19 December 2022

CITATION
Weber KA II, Teplin ZM, Wager TD, Law CSW, Prabhakar NK, Ashar YK, Gilam G, Banerjee S, Delp SL, Glover GH, Hastie TJ and Mackey S (2022) Confounds in neuroimaging: A clear case of sex as a confound in brain-based prediction. *Front. Neurol.* 13:960760. doi: 10.3389/fneur.2022.960760

Confounds in neuroimaging: A clear case of sex as a confound in brain-based prediction

Kenneth A. Weber II[1]*, Zachary M. Teplin[1], Tor D. Wager[2], Christine S. W. Law[1], Nitin K. Prabhakar[3], Yoni K. Ashar[4], Gadi Gilam[1,5], Suchandrima Banerjee[6], Scott L. Delp[7], Gary H. Glover[8], Trevor J. Hastie[9] and Sean Mackey[1]

[1]Systems Neuroscience and Pain Lab, Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Palo Alto, CA, United States, [2]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, United States, [3]Division of Physical Medicine and Rehabilitation, Department of Orthopaedic Surgery, Stanford University School of Medicine, Palo Alto, CA, United States, [4]Department of Psychiatry, Weill Cornell Medicine, New York, NY, United States, [5]The Institute of Biomedical and Oral Research, Faculty of Dental Medicine, Hebrew University of Jerusalem, Jerusalem, Israel, [6]General Electric Healthcare, Chicago, IL, United States, [7]Department of Bioengineering and Mechanical Engineering, Stanford University, Palo Alto, CA, United States, [8]Radiological Sciences Laboratory, Department of Radiology, Stanford University School of Medicine, Palo Alto, CA, United States, [9]Department of Statistics, Stanford University, Palo Alto, CA, United States

Muscle weakness is common in many neurological, neuromuscular, and musculoskeletal conditions. Muscle size only partially explains muscle strength as adaptions within the nervous system also contribute to strength. Brain-based biomarkers of neuromuscular function could provide diagnostic, prognostic, and predictive value in treating these disorders. Therefore, we sought to characterize and quantify the brain's contribution to strength by developing multimodal MRI pipelines to predict grip strength. However, the prediction of strength was not straightforward, and we present a case of sex being a clear confound in brain decoding analyses. While each MRI modality—structural MRI (i.e., gray matter morphometry), diffusion MRI (i.e., white matter fractional anisotropy), resting state functional MRI (i.e., functional connectivity), and task-evoked functional MRI (i.e., left or right hand motor task activation)—and a multimodal prediction pipeline demonstrated significant predictive power for strength ($R^2 = 0.108–0.536$, $p \leq 0.001$), after correcting for sex, the predictive power was substantially reduced ($R^2 = -0.038–0.075$). Next, we flipped the analysis and demonstrated that each MRI modality and a multimodal prediction pipeline could significantly predict sex (accuracy = 68.0%–93.3%, AUC = 0.780–0.982, $p < 0.001$). However, correcting the brain features for strength reduced the accuracy for predicting sex (accuracy = 57.3%–69.3%, AUC = 0.615–0.780). Here we demonstrate the effects of sex-correlated confounds in brain-based predictive models across multiple brain MRI modalities for both regression and classification models. We discuss

implications of confounds in predictive modeling and the development of brain-based MRI biomarkers, as well as possible strategies to overcome these barriers.

KEYWORDS

neuroimaging, magnetic resonance imaging, brain, muscle strength, sex, machine learning, confounding variables, biomarkers

Introduction

The neuromuscular system is highly adaptable. Given appropriate training, muscle strength can increase to meet the physical demands placed on the body. Morphometric adaptations associated with strength training include changes in the contractile elements of the muscles and the non-contractile tissues—the most visible being the increase in muscle size, muscle hypertrophy, which typically accompanies gains in strength (1). Strength, however, is only partially explained by the structural and architectural properties of the muscles. Adaptions within the central nervous system are understood to be a factor in force generation (2). Early gains in strength can precede any evidence of hypertrophy, and strength training in one limb can increase strength in the contralateral limb (3). Neural adaptations may increase force generation through improved intra- and inter-muscular coordination or more complete activation of the motoneuron pool. Thus, in addition to the muscles, the brain plays a role in strength as well (2).

Magnetic resonance imaging (MRI) has become a key tool for the non-invasive mapping of the human brain. With MRI, we can quantitatively characterize both structural and functional properties of the brain. From T_1-weighted and T_2-weighted structural imaging, the morphometry of the cortical and subcortical gray matter can be measured. Diffusion-weighted imaging can probe the integrity of the white matter tracts connecting gray matter regions. With functional MRI, dynamic fluctuations in neural signaling can be extracted to assess network-level neural processing across the brain and identify brain regions activated during experimental tasks. Together these complementary measures provide a valuable multimodal macroscale representation of the human brain.

Multivariate predictive modeling and machine-learning techniques are increasingly being adopted in the brain MRI field to develop and implement brain-based biomarkers of health and disease. These models are beginning to show promise in extracting patterns of information from the high-dimensional set of brain features to make predictions across individuals. A biomarker, broadly defined, is an indicator of normal biological processes, pathogenic processes, or responses to therapeutic intervention (4). Muscle weakness is a common finding in many neurological, neuromuscular, and musculoskeletal diseases. The pattern of weakness can help localize the site of pathology or injury and inform care; however, when weakness is identified, the contributing component, muscular or neural, may not be clear. A valid brain-based MRI biomarker of strength could be helpful for clinical and research communities in several ways: (1) prognosis (i.e., for indicating the likely progression of health or disease; (2) identifying patients likely to respond to a particular treatment (i.e., prediction); (3) identifying a specific disorder based on the pattern of brain pathology (i.e., diagnosis); and (4) identifying targets for therapeutic intervention (4).

However, the interpretation of multivariate prediction models can be problematic in the presence of confounds—variables that are not of direct interest but correlated with the predicted variable (5–7). Confounds create ambiguity in the source of the information driving the prediction. While a model may accurately predict a measure, the neurobiological information driving the prediction could be related to the confound and not the target measure itself. We sought to use multimodal brain MRI and grip strength measures to characterize and quantify the brain's contribution to strength through multivariate predictive modeling. While each MRI modality and a multimodal model significantly predicted strength in an independent testing dataset, the prediction of strength was not straightforward. Here we present a case of sex being a clear confound in the brain decoding analyses, obfuscating the relationship between the brain features and strength and creating ambiguity in the interpretation of the measure being predicted: strength or sex.

Materials and methods

Dataset

Multimodal 3T brain MRI datasets were acquired from the Washington University, University of Minnesota, and Oxford University (WU-Minn) Human Connectome Project (HCP) 1,200 subjects release, which contains structural MRI (T_1-weighted and T_2-weighted), diffusion MRI, resting state fMRI, and task fMRI in healthy young adults (8). For the 3T datasets, imaging was performed supine in 1,113 participants (average age $\pm$ one standard deviation $=$ 28.8 $\pm$ 3.7 years, 606 females)

using a Siemens 3T Connectome Skyra magnetic resonance scanner (location Washington University, St. Louis, MO, USA) equipped with a standard 32-channel Siemens receive head coil and a specially designed Siemens body transmission coil to accommodate the specialized gradients of the connectome scanners (56 cm bore, maximum gradient strength 100 mT/m). Healthy was defined broadly in the HCP to be representative of the variability in behavior, ethnicity, and socioeconomic status of the population. The exclusion criteria were limited to diabetes, hypertension, severe neurodevelopmental disorders, and diagnosed neuropsychiatric disorders and neurological conditions. The HCP has released preprocessed datasets that follow the HCP standard processing pipelines (9). The HCP preprocessed images were used as inputs into the analyses. Any processing steps in addition to the standard HCP pipeline are described in the corresponding sections.

Gray matter features

The HCP preprocessing pipeline for the T_1-weighted and T_2-weighted structural images included averaging of the T_1-weighted or T_2-weighted images if multiple images were collected; distortion correction; alignment with MNI152 space; bias field correction; and spatial normalization to 0.7 mm^3 MNI152 space. The processed T_1-weighted image is considered the native volume space for each subject. In preprocessing, the T_2-weighted image and processed diffusion MRI and fMRI datasets were registered to the T_1-weighted space prior to normalization to MNI152 space. Cortical (i.e., pial and white matter surfaces) and subcortical regions were automatically segmented with FreeSurfer (version = 5.2) using both the T_1-weighted and T_2-weighted images to estimate the cerebral cortical ribbon (10). The extracted and tabulated FreeSurfer measures included in the HCP 1,200 subjects release were used as the features in the gray matter prediction pipelines. The features included 68 cortical thickness and 68 cortical area measures from 34 regions (17 left and 17 right regions) based on the Desikan-Killiany cortical atlas and subcortical volumes from 19 regions (nine left and nine right subcortical regions and one midline brainstem region) totaling 155 gray matter features (Supplementary Table 1) (11, 12).

White matter features

The HCP preprocessing pipeline for diffusion MRI included b_0 image intensity normalization, distortion correction, eddy-current correction, motion correction, gradient non-linearity distortion correction, and registration to the T_1-weighted image. The spatial transformation from the T_1-weighted registration was also applied to the diffusion vectors and gradient field tensors. The Oxford Center for fMRI of the Brain's (FMRIB) Software Library (FSL) was used to calculate the diffusion metrics and perform normalization to MNI152 space (13, 14). Fractional anisotropy (FA) maps were generated by fitting diffusion tensors to the processed diffusion MRI dataset registered to the T_1-weighted structural images using's FSL's dtifit with correction for gradient non-linearities. FSL's tract-based spatial statistics (TBSS) was then used to non-linearly normalize the FA images (FNIRT) to the FMRIB58 FA template in 1 mm^3 MNI152 space. The white matter features were extracted using Nilearn (version = 0.5.2), an open-source Python module for statistical learning on neuroimaging data (15). Nilearn's NiftiLabelsMasker (resampling target = data) was used to extract the mean fractional anisotropy within 48 white matter regions from The Johns Hopkins University-International Consortium for Brain Mapping's white-matter labels 1 mm^3 atlas (JHU-ICBM-DTI-81) included with FSL (16), which were used as features in the white matter prediction pipelines (Supplementary Table 2).

Resting state features

Resting state timeseries were acquired in two sessions with eyes open and relaxed fixation on a crosshair. Within each session, two runs (14 min and 24 s each) were completed with the images acquired with alternated phase encoding directions (i.e., left-right and right-left). If resting state data from two sessions were present, only the data from the first complete session were analyzed. The HCP preprocessing pipeline for resting state fMRI consisted of distortion correction, motion correction, denoising using FMRIB's ICA-based X-noiseifier (FIX), registration to the T_1-weighted image, and spatial normalization to 2 mm^3 MNI152 space (17). Nilearn's NiftiLabelsMasker (resampling target = data, spatial smoothing = none, band-pass temporal filter = 0.008–0.100 Hz) was used to extract the mean time series from the regions of the asymmetric bootstrap analysis of stable clusters (BASC) 122 region brain parcellation from the preprocessed resting state images in MNI152 space for each phase encoding direction run (18). We chose the 122 region BASC parcellation because it performed the best of the predefined atlases and almost as well as the best data-driven brain parcellation methods (i.e., Dictionary Learning ℓ_1) in the recent study by Dadi et al. (19), which completed an exhaustive comparison of 240 resting state fMRI classification prediction pipelines across multiple datasets and conditions. We chose a predefined atlas to increase the efficiency of the analyses as well as the interpretability of the functional connectivity measures, which otherwise vary depending on the sampling (e.g., when using a data driven parcellation method). Tangent pairwise functional connectivity was calculated using Nilearn's ConnectivityMeasure for each phase encoding direction, and then averaged resulting in 7,381 resting state features per connectivity measure (20, 21).

Motor task features

The task-evoked fMRI experiments followed the resting state fMRI sessions. The motor task was adapted from Buckner et al. and designed to map brain motor areas (22, 23). Each run consisted of alternating 12 s blocks of tapping the left or right fingers, squeezing of the left or right toes, and movement of the tongue. Each movement block was performed twice and preceded by 3 s visual cues. Each run also contained three 15 s fixation blocks for a total of 3 min and 24 s per run. Two motor task runs were performed with the images acquired using alternated phase encoding directions (i.e., left-right and right-left). The HCP preprocessing pipeline for task fMRI consisted of distortion correction, motion correction, registration to the T_1-weighted image, and spatial normalization to 2 mm^3 MNI152 space. Statistical maps of the preprocessed images in MNI152 space were generated for each run using FSL's Improved Linear Model (FILM) with prewhitening (24, 25). The functional images were spatially smoothed with a 4 mm^3 full width half maximum (FWHM) Gaussian smoothing kernel and then high-pass temporally filtered (cutoff = 200 s). The task was modeled using the hemodynamic response function (double-gamma, phase 0 s) convolved vectors for the visual cue and the left fingers, right fingers, left toes, right toes, and tongue movements as explanatory variables. We included the temporal derivatives of the task blocks as covariates of no interest. We then generated average activation maps across the two runs in a second-level fixed effects analysis. The contrast parameter estimates (COPE) for the movement blocks relative to the fixation blocks were extracted voxelwise from the second-level analyses for the left and right finger tapping movements using Nilearn's NiftiMasker and a gray matter mask. The gray matter mask included the cortical gray matter, subcortical nuclei, brainstem, and cerebellum resulting in 194,807 features for each finger tapping movement (i.e., left and right hands), which were used in the motor task prediction pipelines. The left and right finger tapping movements were each assessed separately in their own prediction pipelines.

Strength

Grip strength from the dominant hand was used as the measure of strength. Strength testing was performed using the National Institutes of Health (NIH) Toolbox Grip Strength Test and measured seated using a Jamar Plus Digital dynamometer with the feet on the ground, the elbow bent at 90°, the arm against the trunk, and the wrist in neutral (26). For each hand, the participant performed a less than full force practice trial followed by a full force trial, in which the participant squeezed as hard as possible for a count of three. We used the raw grip strength values normalized to the entire NIH Toolbox Normative Sample (age ≥18 years) without adjusting for sex or age. A score of 100 indicates performance that was at the national average, and a score of 115 or 85 indicates performance one standard deviation above or below the national average, respectively (27). The grip strength testing protocol has been shown to have good to excellent test-retest reliability and good validity compared to a Biodex System 3 Isokinetic Dynamometer (Biodex Medical Systems, Inc. Shirley, New York, USA).

Training and testing datasets

From the HCP 1,200 subjects release 3T data, 1,047 participants had complete MRI datasets. A complete dataset included at least one T_1-weighted structural image, one T_2-weighted structural image, completion of 50% of the diffusion MRI acquisition, and completion of one set of the left-right and right-left phase encoding runs for both the motor task and resting state fMRI acquisitions. Of the complete MRI datasets, one participant's data were excluded for a missing grip strength score, and another participant was excluded for having an outlier grip strength score of 55.3, which was more than five standard deviations below the HCP's average grip strength score. The final dataset consisted of 1,045 participants (age = 28.7 ± 3.7, 514 females) with an average grip strength score of 116.7 ± 11.2. The final dataset was then split into training and testing datasets. The training dataset was used to train the prediction pipeline, and the testing dataset was used as an independent, holdout dataset to provide an unbiased estimate of the prediction pipeline performance. The HCP 1,200 subjects release 3T dataset contains 143 pairs of monozygotic and 85 pairs of dizygotic twins confirmed by genotyping as well as non-twin siblings. To preserve independence of the training and testing datasets from genetic and environmental factors, the 75 non-related participants from unique families, based on the mother, father, and family identifiers, were assigned to the testing dataset (age = 28.5 ± 3.7 years, 38 females, grip strength score = 117.5 ± 10.4). The remaining 970 participants, which included twin and sibling participants, were assigned to the training dataset (age = 28.8 ± 3.7 years, 525 females, grip strength score = 116.6 ± 11.2). The training and testing datasets did not significantly differ on age ($t = 0.677$, $p = 0.499$), sex ($X^2 = 0.335$, $p = 0.563$), handedness ($X^2 = 2.909$, $p = 0.088$), or grip strength ($U = 34{,}177.5$, $p = 0.383$; Table 1).

Prediction pipelines

Analyses were performed using Scikit-learn (version = 0.21.2), an open-source python package for machine-learning (28). For the first-level modeling, we used linear regression (LinearRegression), which is a statistical method that determines the best linear function for all points (X, y) that minimizes the sum of the squared errors *via* ordinary least squares.

TABLE 1 Summary of training and testing datasets.

	Training ($n = 970$)	Testing ($n = 75$)	p-Value
Age (years)	28.8 ± 3.7	28.5 ± 3.7	0.499
Female (n)	525	38	0.563
Right-handed (n)	885	64	0.088
Grip strength score	116.6 ± 11.2	117.5 ± 10.4	0.383

Because of the small size of the dataset in comparison to the high-dimensional feature space, especially for the resting state and motor task prediction pipelines, the regression model is prone to overfitting. Therefore, we used non-sparse ($\ell 2$) models to penalize excessive model complexity and encourage generalizability. As the number of features (p) for the resting state and motor task fMRI greatly exceeded the number of participants (n) (i.e, $p >> n$), dimensionality reduction was performed using principal components analysis (PCA). Using PCA, we transformed the features to the set of uncorrelated principal components (n – 1), each a linear combination of the original features using singular value decomposition. The features were winsorized (average ± 3 SD) to limit extreme values, and then each feature was mean centered and scaled to unit variance. Winsorizing and scaling were performed prior to PCA, based on the training data, and then used to transform the testing data, which kept the processing independent of the testing data. This was similarly done for PCA and the calculation of the tangent functional connectivity measure, which requires the calculation of a group average covariance matrix. Hyperparameter tuning of the $\ell 2$ regularization parameter ($C = 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1{,}000, 10{,}000, 100{,}000$) was performed *via* nested five-fold cross-validation using grid search and the mean squared error as the performance metric. We then repeated the grid search again using a finer range of regularization parameters.

Multimodal model

A stacked ensemble of prediction pipelines was used to combine the first-level prediction pipelines for the gray matter, white matter, resting state, and motor task features. The left and right motor task prediction pipelines were included separately. In a standard stacking ensemble, the training dataset is used to fit a first-level model, and then the training dataset predictions from the first-level models are used to train a second-level model. The use of the same training dataset to fit and generate the regressors for the second-level model can lead to overfitting, especially for datasets that have a large number of features relative to the number of samples, which is the case for the resting state and motor tasks. To overcome this barrier, we used a 10-fold cross-validation framework for each first-level prediction pipeline and passed the out-of-fold predictions as regressors to the second-level model (i.e., pre-validation) (29). Linear regression without regularization was used for the second-level multimodal modeling.

Testing performance

We assessed the performance of the prediction pipelines for each of the first-level models as well as the second-level multimodal model in the independent, holdout testing dataset to provide an unbiased estimate of model of performance and generalizability. Testing performance was assessed using the mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R^2) calculated with scikit-learn. The following equation was used for R^2:

$$R^2 = 1 - \frac{\sum_i (y_i - p_i)^2}{\sum_i (y_i - \bar{y})^2}$$

In this equation, y_i = the measured grip strength scores, p_i = the predicted grip strength scores, and $\bar{y}$ = the average of the measured grip strength scores. When using a model fit to a training dataset to make a prediction on a testing dataset, R^2 can be negative if the model performs worse than a baseline model that always predicts $\bar{y}$. Permutation testing was used to validate the unbiased nature of the prediction pipelines in which the first-level training dataset grip strength scores were randomly permuted prior to entering the prediction pipeline (30). We recorded the MAE, RMSE, and R^2 of the testing dataset predictions for 10,000 permutations of the training dataset to create a null distribution of performance metrics. From the null distribution, we calculated a p-value (one-tailed) for the performance of the prediction pipeline fit to the unpermuted training dataset to provide an estimate of statistical significance relative to chance.

Sex as a confound

Strength was significantly greater in males than females (Figure 2 and Table 3), making sex a potential confound in the analyses, so we then investigated the performance of the prediction pipelines after correcting the brain features and grip strength scores for sex. Sex correction was performed by demeaning each feature and the strength scores using the mean values of the corresponding sex. Demeaning as applied is equivalent to using linear regression to correct for a binary variable. To further investigate strength and sex in brain prediction, we flipped the analyses and explored the power of the brain features to predict sex by using a logistic regression classification model with $\ell 2$ penalization (LogisticRegression).

We performed the feature selection and model training similar to the strength regression prediction pipelines for the first-level models as well as the multimodal classification model (i.e., 10-fold cross-validated stacked classification ensemble). The sex prediction pipeline performance was compared with and without correcting the brain features for grip strength. Strength correction was performed using linear regression and then training on the residuals. For sex prediction, testing performance was assessed using percent accuracy and the area under the receiver operating characteristic curve (AUC) calculated with scikit-learn. The AUC was calculated by plotting the true positive rate (TPR) as a function of the false positive rate (FPR) at varying threshold settings. The strength and sex correction parameters were calculated on the training dataset and then applied to the testing dataset. Finally, we assessed the ability to predict sex using the grip strength measures alone and logistical regression.

Visualization of first-level prediction pipeline features

To visualize and compare the brain features that made reliable contributions to the prediction of strength and sex, bootstrapping was performed. For each first-level model, p-values for the model coefficients were calculated by fitting the models to 10,000 bootstrapped samples from the training dataset (hyperparameter tuning same as that determined from model fit to entire training dataset), then converting the model coefficient to Z-values (bootstrapped mean of each coefficient / bootstrapped standard deviation of each coefficient), and finally transforming the Z-values to p-values using a two-tailed normal distribution. The coefficient p-values were corrected with a false discovery rate (FDR) of $q < 0.05$ to correct for multiple comparisons (31). When PCA was employed, the model coefficients for these pipelines were first inverse transformed back to the original feature space before conversion to p-values. The significant model coefficients (FDR-corrected) were visualized to assess the contributions of the first-level features to the prediction of strength and sex. We performed the FDR correction using the MNE Python package (version = 0.21.0).

The labeling of the 122 regions in the BASC parcellation, used to calculate functional connectivity, is unordered and does not correspond to any particular resting state functional brain network. To assist in interpretation of the resting state functional networks underlying prediction, each of the 122 regions was assigned to a resting state functional network based on the maximum percent overlap between the region and binary maps of the visual, somatomotor, dorsal attention, ventral attention, limbic, frontoparietal, and default mode networks as well as subcortical and cerebellar networks. The cortical networks were defined using the seven network parcellation developed by Yeo et al. (23). The subcortical network was defined used the Harvard-Oxford subcortical atlas and contained the thalamus, caudate, putamen, globus pallidus, hippocampus, amygdala, and brainstem (11, 32–34). The cerebellar network was defined using the probabilistic human cerebellum atlas developed by Diedrichsen et al. (35). Both the subcortical and cerebellar atlases are included with FSL. The FDR-corrected model coefficients weights for the resting state prediction pipeline were visualized using circular graphs plotting both the positive and negative connections within and between the resting state functional networks using Circos (36). The FDR-corrected white matter and task model coefficients were visualized overlaid the MNI152 T_1-weighted brain template. Finally, we used Spearman correlations to directly compare the non-corrected model coefficients and quantify the similarity between the strength and sex prediction models.

Statistical testing

Student t-tests and correlation analyses were performed using the SciPy Python package (version = 1.2.1). An $\alpha < 0.05$ was considered statistically significant. Correlation and AUC plots were generated using the Matplotlib Python package (version = 3.1.0).

Results

We assessed the performance of each modality on the independent testing dataset ($n = 75$), providing an unbiased estimate of performance. Each modality had predictive power for strength that exceeded chance accuracy ($p \leq 0.001$) with the resting state prediction pipeline having the highest performance of the individual modalities, explaining more than 45% of the variance in strength ($R^2 = 0.452$). The gray matter and white matter prediction pipelines performed comparably to each other, both explaining more than 30% of the variance in strength. The left and right hand prediction pipelines had the lowest performance, each explaining < 15% of the variance in strength (Figure 1 and Table 2). The multimodal prediction pipeline outperformed the individual modality prediction pipelines explaining more than 50% of the variance in strength ($R^2 = 0.535$, MAE = 5.95, and RMSE = 7.09; Figure 1 and Table 2).

Grip strength, however, was significantly higher in males than in females in both the training ($p < 0.001$) and testing ($p < 0.001$) datasets, and sex was identified as a potential confound (Figure 2 and Table 3). After correcting for sex, the predictive power was substantially reduced for each first-level model (Figure 1 and Table 2). While the resting state prediction pipeline continued to have the highest performance of the individual modalities, resting state functional connectivity only explained slightly more than 7% of the variance in strength after correcting for sex ($R^2 = 0.073$, $p < 0.001$). The performance

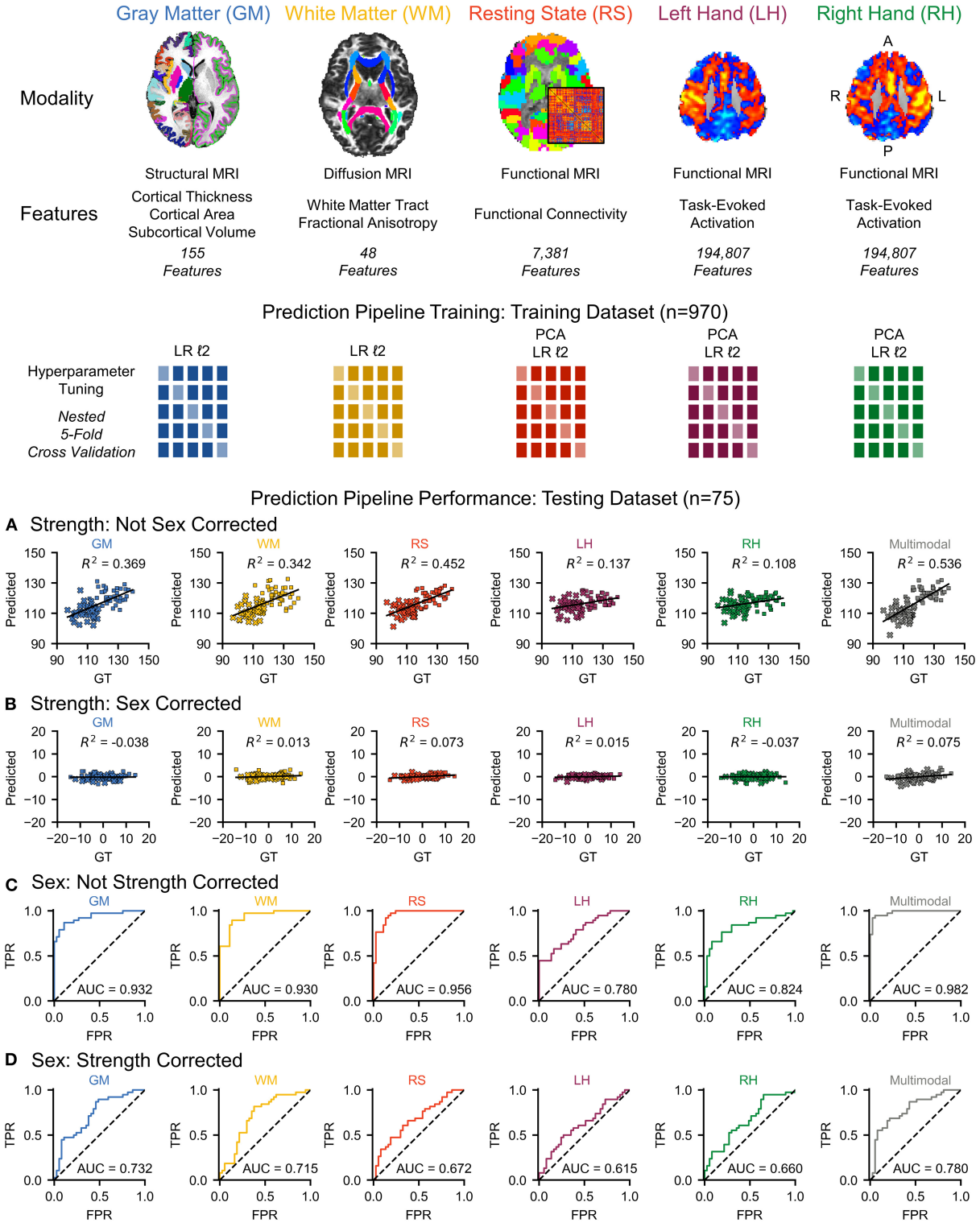

FIGURE 1

We used multimodal brain MRI and grip strength scores to characterize and quantify the brain's contribution to strength through multivariate predictive modeling. Multimodal brain MRI provides measures on gray matter (GM) morphometry, white matter (WM) fractional anisotropy, resting state (RS) functional connectivity, and left hand (LH) and right hand (RH) motor task activation. We trained a prediction pipeline for each MRI modality on the training dataset (n = 970) using non-sparse ($\ell 2$) linear regression (LR) models with principal component analysis (PCA) for dimensionality reduction, when applied, and nested five-fold cross-validation for hyperparameter tuning. We then used stacked ensembles of the prediction pipelines to combine the first-level models into second-level multimodal models. **(A)** Each MRI modality and a multimodal model

(Continued)

FIGURE 1 (Continued)
significantly predicted strength in an independent testing dataset ($n = 75$). **(B)** However, sex was identified as a potential confound, and correcting for sex substantially reduced the predictive power for each MRI modality and the multimodal model (female = ×, male = ■). **(C)** Next, we flipped the analysis and investigated the use of multimodal brain MRI to predict sex. Each of the MRI modalities and a multimodal model significantly predicted sex. **(D)** Correcting the brain features for strength, however, reduced the accuracy of each MRI modality and a multimodal model for predicting sex. GT, ground truth; TPR, true positive rate; FPR, false positive rate; AUC, area under the curve. L, left; R, right; A, anterior; P, posterior.

TABLE 2 Grip strength prediction testing performance ($n = 75$).

Pipeline	MAE	*p*-Value	RMSE	*p*-Value	R^2	*p*-Value
Not sex corrected						
Gray matter	6.92	<0.001	8.27	<0.001	0.369	<0.001
White matter	6.41	<0.001	8.44	<0.001	0.342	<0.001
Resting state	6.18	<0.001	7.71	<0.001	0.452	<0.001
Left hand	8.11	0.002	9.67	0.002	0.137	<0.001
Right hand	7.89	0.004	9.83	0.001	0.108	0.001
Multimodal	5.95	<0.001	7.09	<0.001	0.536	<0.001
Sex corrected*						
Gray matter	5.20	0.126	6.04	0.014	−0.038	0.014
White matter	4.89	0.087	5.89	0.066	0.013	0.069
Resting state	4.82	0.005	5.71	<0.001	0.073	<0.001
Left hand	4.99	0.026	5.89	0.003	0.015	0.003
Right hand	5.07	0.071	6.04	0.018	−0.037	0.018
Multimodal	4.78	0.001	5.71	<0.001	0.075	<0.001

*Testing performance assessed using grip strength scores corrected for sex.

of the second-level multimodal prediction pipeline was likewise reduced with similar performance as the resting state prediction pipeline ($R^2 = 0.075$, $p < 0.001$).

Next, we flipped the analysis and investigated the use of multimodal brain MRI to predict sex. Each of the MRI modalities significantly predicted sex ($p < 0.001$). The resting state prediction pipeline had the highest performance of the individual modalities, predicting sex with 89.3% accuracy (AUC = 0.956). The gray matter and white matter prediction pipelines performed comparably to each other, both having accuracies > 80%. The left and right hand prediction pipelines had the lowest performance, each with accuracies lower than 80%. The second-level multimodal prediction pipeline had an accuracy of 93.3% (AUC = 0.982; Figure 1 and Table 4). Correcting the brain features for strength, however, reduced the accuracy of predicting sex for each modality, the resting state prediction pipeline had the lowest accuracy of the individual modalities (accuracy = 57.3%, AUC = 0.672). The multimodal prediction accuracy dropped to < 70% after correction for strength (AUC = 0.780; Figure 1 and Table 4). In comparison, using grip strength alone predicted sex with 94.7% accuracy (AUC = 0.985, $p < 0.001$).

Finally, we compared the coefficients between the strength prediction and sex prediction models without correction for sex or strength, respectively, to assess the similarity in the brain features driving the predictions. The FDR-corrected coefficient maps between the strength and sex prediction show substantial overlap for each modality, indicating that both models are using similar information in their predictions (Figures 3, 4). To quantify the level of similarity, we used Spearman correlations to directly compare the model coefficients between the strength and sex prediction models. The model coefficients were moderately to strongly correlations with the white matter models having the greatest correlation ($\rho = 0.762$, $p < 0.001$) and the resting state models having the lowest correlation ($\rho = 0.482$, $p < 0.001$), further demonstrating a substantial but not complete degree of similarity in the information underlying the prediction of strength and sex (Table 5).

Discussion

We initially sought to characterize and quantify the brain's contribution to strength by developing multimodal brain-based MRI pipelines to predict grip strength. Sex, however, was identified as a clear confound in the analyses, complicating the prediction of strength. Each MRI modality and a multimodal prediction pipeline could accurately predict strength; however,

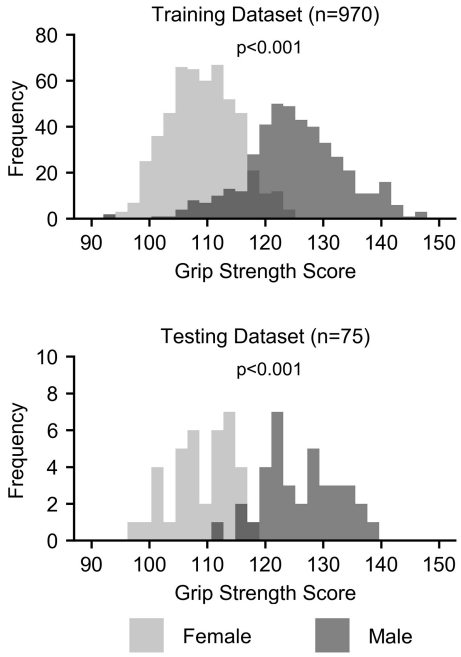

FIGURE 2
Histograms of grip strength scores by sex for the training and testing datasets.

correcting for sex substantially reduced the accuracy. When flipping the analyses and classifying sex from brain MRI, each MRI modality and a multimodal prediction pipeline could accurately predict sex; however, correcting the brain features for strength reduced the accuracy of predicting sex. Comparing the coefficients between the models, we identified substantial but not complete similarities between the brain features driving the prediction of strength and sex. As demonstrated by our results, the interpretation of brain-based prediction models can be problematic in the presence of confounds: are the models predicting strength or sex? In the following, we discuss confounds in brain decoding analyses, their implications in the development of brain-based biomarkers, and possible strategies to overcome these challenges and improve interpretation.

Confounds are a well-known issue in statistical modeling and machine-learning applications (5–7). Confounds create ambiguity in the source of the information driving the prediction and uncertainty in the measure being predicted. In other words, confounds affect the relationship between the features and the predicted measure. In the case of strength prediction, sex is a clear confound in this sample. While males in this sample have only ≈ 15% greater grip strength than females on average, the distribution of strength is clearly bimodal, with highly significant differences in the distribution of strength between the sexes (Cohen's $d > 2.0$). In fact, grip strength alone predicted sex with > 90% test accuracy. Considering the substantial reduction in strength prediction after correcting for sex, the ability to predict sex with high accuracy using the multimodal brain features, and the similarity between the strength and sex prediction models, the strength prediction model, at least in part, is likely learning patterns of features associated with sex. If the accuracy of strength prediction was the only aim, then sex being a confound is less of a concern. However, the initial goal of this study was to uncover the brain's contribution to strength to develop models that may be used as brain-based biomarkers in conditions affecting strength. We are not interested in using features for strength prediction that are related to sex but invariant to strength. Therefore, for the intended application, the interpretation of the models becomes as important as accuracy.

Sex is a recognized confound in neuroimaging. Males on average have larger total brain size than females (37). In addition to global differences in brain size, localized alterations in cortical thickness and subcortical volumes have also been identified between sexes, and gray matter morphometry has been used to predict sex as well as white matter characteristics and both resting state and task-evoked fMRI measures (38–43). Recently, sex differences in brain-based biomarkers of intelligence have been uncovered, suggesting that the generation of intelligence in males and females may utilize distinct brain networks (44). Debate exists regarding the interpretation, meaning, and importance of sex differences in the brain (45–48). A recent quantitative synthesis by Eliot et al. (49) challenges the importance of sex differences in the human brain, arguing that many structural and functional sex differences identified with MRI are largely negligible after correcting for brain size. A more recent large-scale brain-based MRI study from the UK Biobank with more than 40,000 participants uncovered sex differences in two-thirds of the gray matter brain measures investigated, which were independent of brain size when accounting for non-linear relationships between local brain region morphometry and brain size (50). While the effect of sex was small after correction, structural and functional differences in the brain, even if small, could have meaningful consequences on brain function, especially in aggregate (51, 52). Overall, the meaning of these sex differences and their role in human behavior, mental health, and the brain in health and disease remain to be uncovered. When taken together and with respect to the present findings, identifying and interpreting sex differences in the brain is complex, especially when considering sex-correlated confounds such as strength and brain size.

Age is another important confound in neuroimaging. The brain is not a static organ but changes over the lifespan (50). Similarly to sex prediction, gray matter, white matter, and resting state and task-evoked fMRI brain measures can predict age (53–58). The aging, cognition, and dementia fields have made substantial progress in developing biomarkers of normal and pathological age-related changes in the brain. The difference between predicted and actual brain age, the brain age gap, is

TABLE 3 Grip strength scores by sex.

	Female	Male	Cohen's *d*	*p*-Value
Training ($n = 970$)	109.1 ± 6.0	125.5 ± 9.1	2.1	<0.001
Testing ($n = 75$)	109.0 ± 5.2	126.1 ± 6.6	2.9	<0.001

TABLE 4 Sex prediction testing performance ($n = 75$).

Pipeline	Accuracy (%)	*p*-Value	AUC	*p*-Value
Not strength corrected				
Gray matter	84.0	<0.001	0.932	<0.001
White matter	85.3	<0.001	0.930	<0.001
Resting state	89.3	<0.001	0.956	<0.001
Left hand	68.0	0.005	0.780	<0.001
Right hand	78.7	<0.001	0.824	<0.001
Multimodal	93.3	<0.001	0.982	<0.001
Strength corrected				
Gray matter	68.0	0.001	0.732	<0.001
White matter	69.3	<0.001	0.715	<0.001
Resting state	57.3	0.121	0.672	0.005
Left hand	58.7	0.057	0.615	0.051
Right hand	61.3	0.043	0.660	0.015
Multimodal	69.3	<0.001	0.780	<0.001

the most commonly used measure, providing a single metric to assess brain health with respect to normal aging (59). Brain age biomarkers are already showing potential diagnostic and prognostic value in several conditions including schizophrenia and Alzheimer's disease (60–65). Disease-related processes as well as genetic, lifestyle, and health factors are thought to affect brain aging, but the impact of these factors are still unclear (See Wrigglesworth et al. (66) for review). Sex differences in the brain that accompany aging have also been identified (67–69). A recent study demonstrated differences in the predicted age gap between males and females with a family history of Alzheimer's disease. Sex, *APOE* genotype status, and physical activity demonstrated a significant interaction. Males with +*APOE4* genotype who engage in physical activity had younger predicted brain age than the corresponding females, suggesting that physical activity influences brain aging differentially in +*APOE4* males and females (70). In the prediction of age, sex is not likely a significant confound that impacts interpretability as long as sex is balanced across the age groups. However, if the older participants were disproportionately male and the younger participants were disproportionately female, sex would be a confound, similar to strength in the present study. Sex differences in brain structure and function over the lifespan could impact the accuracy of the age prediction models and substantiate correcting the features for sex or modeling each sex separately. In regards to age as a confound in the prediction of strength, participants in the HCP were healthy young adults (age = 22–37 years), so we were not concerned about age being an important confound. Strength, however, on average declines with age (71–73), and even within the HCP dataset, grip strength was weakly negatively correlated with age ($n = 1{,}045$, $r = -0.104$, $p < 0.001$). If we were developing predictive models of strength across a broader age range, age would become a confound in the analyses making it difficult to determine whether we are predicting strength or age.

So far, we have discussed participant-related confounds of sex and age and their potential confounding effects in brain-based predictions. Additional participant-related confounds include anatomical variability (e.g., brain size and shape), arousal level (e.g., sleep and caffeine use), mental and emotional state, head movement (i.e., ability to stay still), and overall health status (e.g. presence of medical conditions, medication use, and comorbidities) (74). Procedure-related confounds occur from factors within the study design that can influence the participant measures and include experimental instructions (e.g., eyes open or closed in resting state fMRI), time of day, auditory and visual noise, and image acquisition (e.g., imaging parameters, software, and hardware) and processing methods (e.g., spatial normalization methods) (75–77). Brain shape and size are known to vary across populations, change with age, and be affected by diseases (78–82). For example, ventricular enlargement and cortical atrophy often present in

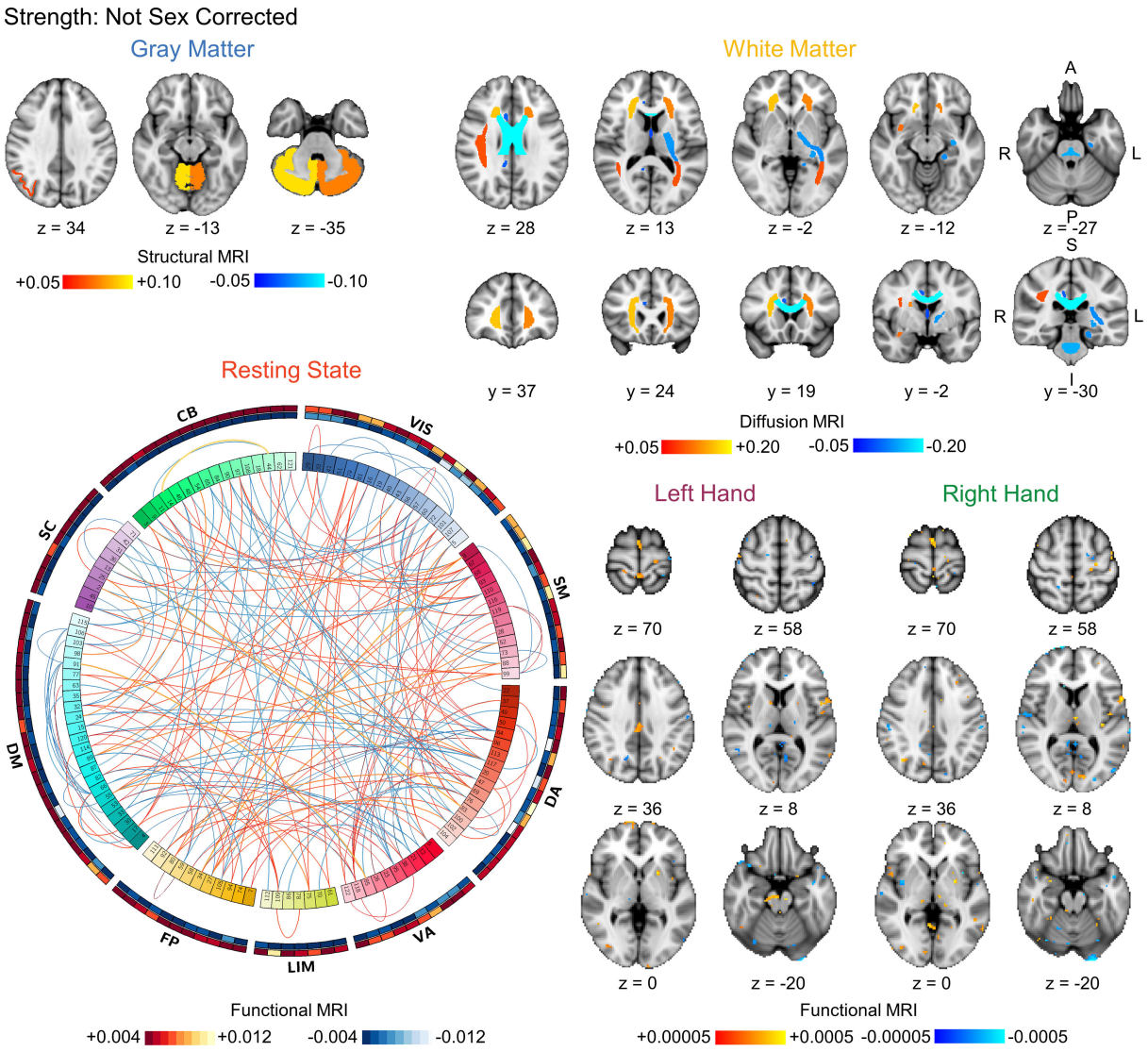

FIGURE 3
Visualization of the bootstrapped FDR-corrected ($q < 0.05$) first-level model coefficients of each MRI modality for strength prediction without correction for sex. Note the similarity between these models and the sex prediction models without correction for strength shown in Figure 4. The outer bands in the resting state connectivity graph show the sum of the positive and negative connections passing FDR-correction. VIS, visual; SM, somatomotor; DA, dorsal attention; VA, ventral attention; LIM, limbic; FP, frontoparietal; DM, default mode; SC, subcortical; CB, cerebellar. L, left; R, right; A, anterior; P, posterior; S, superior; I, inferior.

elderly patients, and standard spatial normalization algorithms may not be adequate for this population (83). This could lead to age-related biases in the measurements of cortical structures. A model that predicts age using the measurement error vs. actual age-related neurophysiological changes may accurately predict age, but the model would not likely be a strong candidate biomarker. Additionally, even when developing a brain-based diagnostic marker between participants with a specific condition and an age- and sex-matched group of healthy controls, other confounds may still be present. For example, in chronic pain, comorbidities such as anxiety and depression are often present, and the patients may be taking medications for their conditions (84). The presence of comorbidities and medication use could be considered confounds in the development of chronic pain markers as each of the three factors (chronic pain, comorbidities, and medication use) may differ from the control group. Similar issues can arise when combining datasets across multiple sites, where site could be a confound (77, 85). Overall, if any factors vary with respect to the predicted measure, these factors may be potential confounds. The degree of concern regarding these factors will depend on the goal of the biomarker and may require additional interrogation to interpret the features driving the prediction and further validate the model (86). As an example, Liem et al. (53) were concerned that head motion may confound

TABLE 5 Correlation between strength and sex prediction model coefficients*.

Pipeline	ρ	p-Value
Gray matter	0.653	<0.001
White matter	0.762	<0.001
Resting state	0.482	<0.001
Left hand	0.686	<0.001
Right hand	0.689	<0.001

*Models not corrected for sex or strength, respectively.

age prediction because head motion affects brain imaging and motion may be higher in older participants. Therefore, the authors included additional analyses to be confident that motion was not driving the prediction of age (53).

In conventional univariate analyses, controlling for confounds can be done statistically by including them as covariates in the model. For example, motion regressors are commonly modeled as covariates when generating subject-level task-evoked fMRI activation maps to account for signal correlated with subject motion, and for group-level fMRI analyses, sex and age can be included to remove their effects when comparing activity between groups. Including confounds as features in decoding models, however, would be counterproductive because confounds contain information that by definition we do not want to use in prediction, so alternative strategies need to be employed. With sampling-based approaches you train models on subsamples of the dataset in which the sampling removes the correlation between the confound and predicted measure and the overall effect of the confound (7). For example, you could match males and females on strength, and only include those matched pairs in the training dataset. This would eliminate the sex differences in strength. However, in the case of strength with strong sex differences, a large portion of the data at the high and low strength levels would be discarded using sampling-based approaches, reducing power and also limiting the domain and range on which the model was trained. Sampling-based approaches can be performed within a cross-validated framework using a stratification factor (e.g., StratifiedKFold). While typically performed *post hoc*, controlling for confounds can be done *a priori as* done in prospective trials where groups are commonly matched on age and sex at enrollment in a study.

While sampling-based approaches may be feasible for one and maybe two confounds, accounting for multiple confounds becomes increasingly complicated and costly (i.e., dropping of unmatched data) and is largely not feasible. Regression-based approaches statistically remove the variance attributed to the confounds from the dataset by fitting a model containing the confound variables to the dataset (7). The modeling is then performed on the residuals, in which the confounding signal has been removed. Regression-based approaches were used in the current study to correct for sex and grip strength. After correcting for sex, only the resting state and the multimodal strength prediction models significantly predicted strength better than a random model. A drawback to regression-based approaches is that they increase the complexity of the prediction pipeline in that the regression-based confound correction model is another step within the prediction pipeline with its own parameters. Another recent alternative proposed by Zhao et al. (87) uses a deep-learning approach with generative adversarial networks that models confounding effect in the feature-learning process such that the model learns patterns of features that are invariant to confounds. The authors validate their method with classification and regression models with continuous and categorical confounds, and the code is openly available on GitHub (https://github.com/qingyuzhao/br-net) (87).

Inherent to our goal of developing brain-based biomarkers of strength is that we need the models to be based on features that underlie the production of force as we want the predictions to be sensitive to changes in strength over time. Sex is a non-modifiable factor, and a strength model based on sex-correlated features invariant of strength would not likely be sensitive to changes in strength. As described and demonstrated, sex is a major confound in the prediction of strength. When accounting for sex, the accuracy for predicting strength from multimodal brain MRI is largely lost. As discussed, males have larger brains than females on average, and the magnitude of the sex differences decreases after correcting for brain size. Sex and brain size are not the only factors that covary. Strength and muscle mass also strongly covary with sex and brain size (88). Strength, muscle mass, and brain size are known to decrease in older age (71–73, 89, 90). Also, physical exercise can decrease age-related losses in the brain's gray and white matter (91–93). While less studied than in the brain and human studies are lacking, sex differences may also present in the spinal cord. Male rats have larger spinal cords by weight and a greater number of motoneurons (94) as well as differences in skeletal muscle fiber morphometry, spindle density, and innervation (95, 96). When taken together, it is not too great of a leap to suggest that a component of the neuromuscular system, such as skeletal muscle mass, could mediate the relationship between sex and brain size. The larger brain size in males could be to support the larger skeletal muscle mass (97).

The cross-sectional nature of the HCP limits the current study to (repeat imaging is only available in a small subset of the dataset, $n < 50$) associations between the brain imaging features and strength. Longitudinal studies with a strength training exercise intervention could determine whether the strength prediction pipelines track increases in strength. If so, this would indicate that the strength prediction models are relying on some information related to strength and provide evidence that the brain features predictive of strength are causally linked to force production. If not, then the models may be relying more on sex-correlated features for the prediction of strength. Longitudinal

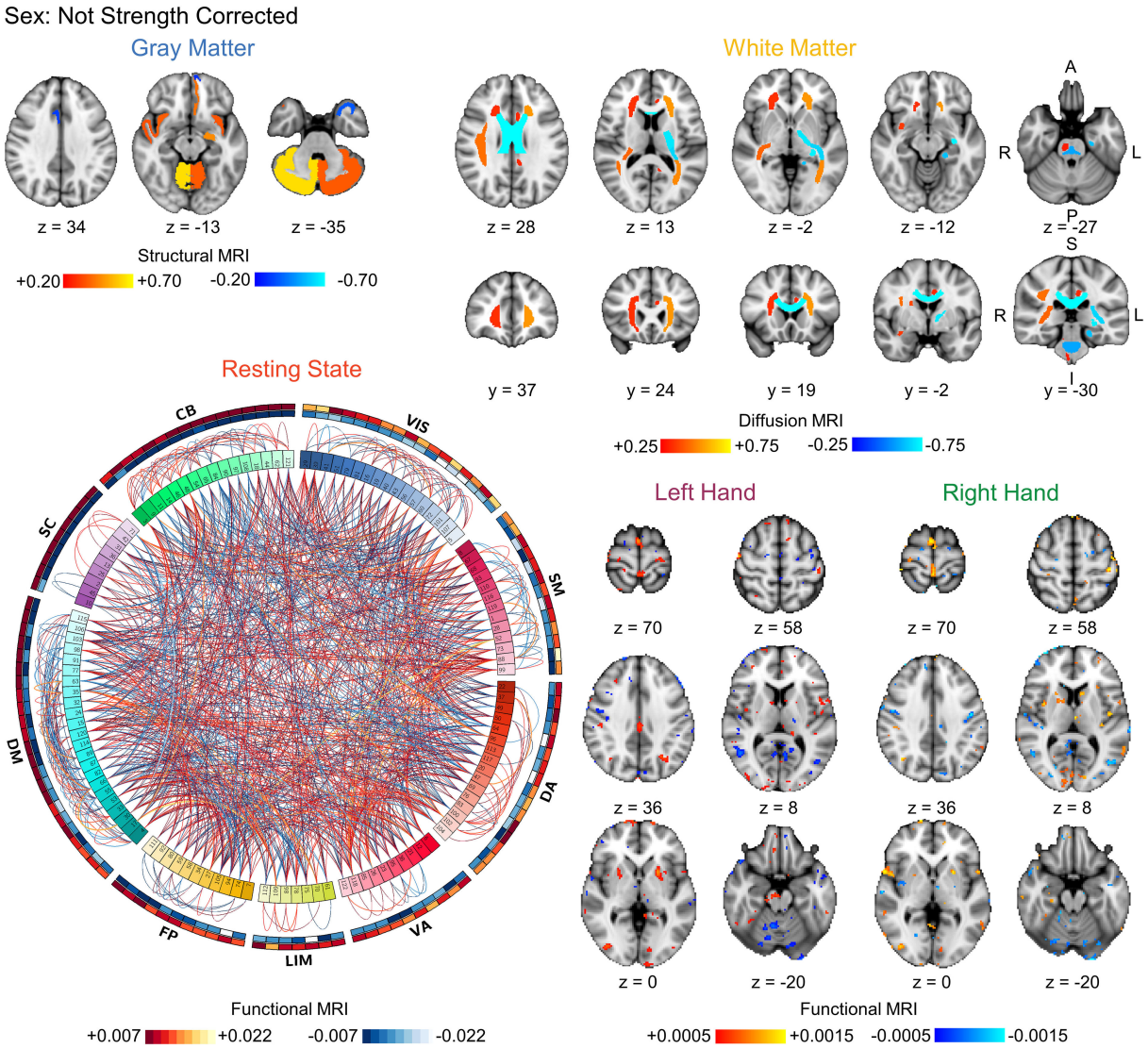

FIGURE 4
Visualization of the bootstrapped FDR-corrected ($q < 0.05$) first-level model coefficients of each MRI modality for sex prediction without correction for strength. Note the similarity between these models and the strength prediction models without correction for sex shown in Figure 3. The outer bands in the resting state connectivity graph show the sum of the positive and negative connections passing FDR-correction. VIS, visual; SM, somatomotor; DA, dorsal attention; VA, ventral attention; LIM, limbic; FP, frontoparietal; DM, default mode; SC, subcortical; CB, cerebellar. L, left; R, right; A, anterior; P, posterior; S, superior; I, inferior.

studies employing within subject models can also help address the issue of confounds allowing us to develop models that predict change in strength from the changes in brain features. A similar argument supporting the use of longitudinal studies has been made by Vidal-Pineiro et al. (98) in the age-prediction field where models developed from cross-sectional studies did not predict longitudinal changes in brain age. The authors suggest that brain age may relate more to early-life factors than longitudinal brain changes, leading to the recommendation that future studies use longitudinal designs when predicting individual changes in brain age is the goal.

Additional improvements can also be made from the current study. The fMRI experiment used to map task-evoked left- and right-hand motor activity was a finger tapping task and not designed to capture signals related to strength. An experiment tuned to force generation may have greater predictive power (99). Similarly, a more diverse sample with greater variation in strength should improve our ability to predict strength. While the use of HCP dataset provided a large, homogenous dataset for this study, future work should include data across sites with varying imaging equipment, imaging parameters, and sample characteristics to generate

models that generalize across sites and to the population at large.

The application of multivariate predictive modeling to neuroimaging is increasing our ability to extract clinically relevant information from the brain and make predictions across individuals. While brain-based predictive models have great potential as biomarkers for neurological, neuromuscular, and musculoskeletal conditions, the presence of confounds creates ambiguity in the source of the information driving the prediction and the interpretation of the measure being predicted, decreasing their potential clinical utility. Here we focused on the effects of sex-correlated confounds in brain-based predictive modeling across multiple MRI modalities for both regression and classification models. In addition to sex, other patient-related and procedural confounds are well-known in neuroimaging. Methods to better assess the influence of these confounds on the predicted measure and the development of strategies to mitigate the effects of confounds will increase the interpretability and validity of brain-based biomarkers and further promote their translation to clinical practice.

Data availability statement

WU-Minn Human Connectome Project (HCP) 1200 subjects release dataset is available from the HCP (www.humanconnectome.org). The use of family tree structure, exact age, and handedness required access to the restricted HCP dataset, which includes potentially identifiable information, and acceptance of the HCP's restricted data use terms. The codes for the HCP preprocessing pipeline and preprocessed datasets are available through HCP. The analysis and prediction pipelines were developed using open-source, commercially usable Python packages (Nilearn, sci-kit Learn, MNE, and SciPy).

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

KW: conceptualization. KW, CL, and SM: data curation. KW, TW, CL, NP, YA, GG, SD, GHG, TH, and SM: methodology. KW, ZT, TW, NP, YA, GG, SD, GHG, TH, and SM: investigation. KW and ZT: formal analysis. KW, ZT, and SM: writing—original draft. KW, ZT, TW, CL, NP, YA, GG, SD, GHG, TH, and SM: writing—review and editing. KW and SM: funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grants from the National Institute of Neurological Disorders and Stroke (K23NS104211, L30NS108301, K24NS126781, and R61NS118651) and the National Center for Advancing Translational Sciences (TL1TR002386) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgments

Portions of the computing for this project was performed on the Sherlock cluster (www.sherlock.stanford.edu). We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support. We want to thank Dr. Nirao Shah at Stanford University for reviewing the manuscript and providing feedback.

Conflict of interest

Author SB was employed by General Electric Healthcare.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2022.960760/full#supplementary-material

References

1. Vigotsky AD, Schoenfeld BJ, Than C, Brown JM. Methods matter: the relationship between strength and hypertrophy depends on methods of measurement and analysis. *PeerJ.* (2018) 6:e5071. doi: 10.7717/peerj.5071

2. Enoka RM. Muscle strength and its development. New perspectives. *Sports Med.* (1988) 6:146–68. doi: 10.2165/00007256-198806030-00003

3. Moritani T, deVries HA. Neural factors versus hypertrophy in the time course of muscle strength gain. *Am J Phys Med.* (1979) 58:115–30.

4. FDA-NIH Biomarker Working Group. *Best (Biomarkers, Endpoints, and Other Tools) Resource*. Silver Spring, MD: Food and Drug Administration (2016).

5. Alfaro-Almagro F, McCarthy P, Afyouni S, Andersson JLR, Bastiani M, Miller KL, et al. Confound modelling in UK biobank brain imaging. *Neuroimage.* (2021) 224:117002. doi: 10.1016/j.neuroimage.2020.117002

6. Smith SM, Nichols TE. Statistical challenges in "big data" human neuroimaging. *Neuron.* (2018) 97:263–8. doi: 10.1016/j.neuron.2017.12.018

7. Snoek L, Miletić S, Scholte HS. How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage.* (2019) 184:741–60. doi: 10.1016/j.neuroimage.2018.09.074

8. Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K. The Wu-Minn human connectome project: an overview. *Neuroimage.* (2013) 80:62–79. doi: 10.1016/j.neuroimage.2013.05.041

9. Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage.* (2013) 80:105–24. doi: 10.1016/j.neuroimage.2013.04.127

10. Fischl B. Freesurfer. *Neuroimage.* (2012) 62:774–81. doi: 10.1016/j.neuroimage.2012.01.021

11. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage.* (2006) 31:968–80.

12. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron.* (2002) 33:341–55. doi: 10.1016/S0896-6273(02)00569-X

13. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage.* (2012) 62:782–90. doi: 10.1016/j.neuroimage.2011.09.015

14. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, et al. Advances in functional and structural mr image analysis and implementation as FSL. *Neuroimage.* (2004) 23:S208–19. doi: 10.1016/j.neuroimage.2004.07.051

15. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform.* (2014) 8:14. doi: 10.3389/fninf.2014.00014

16. Mori S, Oishi K, Jiang H, Jiang L, Li X, Akhter K, et al. Stereotaxic white matter atlas based on diffusion tensor imaging in an Icbm template. *Neuroimage.* (2008) 40:570–82. doi: 10.1016/j.neuroimage.2007.12.035

17. Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage.* (2014) 90:449–68. doi: 10.1016/j.neuroimage.2013.11.046

18. Bellec P, Rosa-Neto P, Lyttelton OC, Benali H, Evans AC. Multi-level bootstrap analysis of stable clusters in resting-state f MRI. *Neuroimage.* (2010) 51:1126–39. doi: 10.1016/j.neuroimage.2010.02.082

19. Dadi K, Rahim M, Abraham A, Chyzhyk D, Milham M, Thirion B, et al. Benchmarking functional connectome-based predictive models for resting-state f MRI. *Neuroimage.* (2019) 192:115–34. doi: 10.1016/j.neuroimage.2019.02.062

20. Varoquaux G, Baronnet F, Kleinschmidt A, Fillard P, Thirion B. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. *Med Image Comput Comput Assist Interv.* (2010) 13(Pt 1):200–8. doi: 10.1007/978-3-642-15705-9_25

21. Varoquaux G, Craddock RC. Learning and comparing functional connectomes across subjects. *Neuroimage.* (2013) 80:405–15. doi: 10.1016/j.neuroimage.2013.04.007

22. Buckner RL, Krienen FM, Castellanos A, Diaz JC, Yeo BT. The organization of the human cerebellum estimated by intrinsic functional connectivity. *J Neurophysiol.* (2011) 106:2322–45. doi: 10.1152/jn.00339.2011

23. Yeo BT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol.* (2011) 106:1125–65. doi: 10.1152/jn.00338.2011

24. Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of f MRI data. *Neuroimage.* (2001) 14:1370–86. doi: 10.1006/nimg.2001.0931

25. Worsley KJ. Statistical analysis of activation images. In: Jezzard P, Matthews PM, Smith SM, editors. *Functional MRI: An Introduction to Methods*. New York, NY: Oxford University Press (2001). doi: 10.1093/acprof:oso/9780192630711.003.0014

26. Reuben DB, Magasi S, McCreath HE, Bohannon RW, Wang YC, Bubela DJ, et al. Motor assessment using the NIH toolbox. *Neurology.* (2013) 80(11 Suppl 3):S65–75. doi: 10.1212/WNL.0b013e3182872e01

27. Beaumont JL, Havlik R, Cook KF, Hays RD, Wallner-Allen K, Korper SP, et al. Norming plans for the NIH toolbox. *Neurology.* (2013) 80(11 Suppl 3):S87–92. doi: 10.1212/WNL.0b013e3182872e70

28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* (2011) 12:2825–30. Available online at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

29. Tibshirani RJ, Efron B. Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol.* (2002) 1:1. doi: 10.2202/1544-6115.1000

30. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E. An fMRI-based neurologic signature of physical pain. *N Engl J Med.* (2013) 368:1388–97. doi: 10.1056/NEJMoa1204471

31. Kohoutová L, Heo J, Cha S, Lee S, Moon T, Wager TD, et al. Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat Protoc.* (2020) 15:1399–435. doi: 10.1038/s41596-019-0289-5

32. Frazier JA, Chiu S, Breeze JL, Makris N, Lange N, Kennedy DN, et al. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *Am J Psychiatry.* (2005) 162:1256–65. doi: 10.1176/appi.ajp.162.7.1256

33. Goldstein JM, Seidman LJ, Makris N, Ahern T, O'Brien LM, Caviness VS Jr, et al. Hypothalamic abnormalities in schizophrenia: sex effects and genetic vulnerability *Biol Psychiatry.* (2007) 61:935–45. doi: 10.1016/j.biopsych.2006.06.027

34. Makris N, Goldstein JM, Kennedy D, Hodge SM, Caviness VS, Faraone SV, et al. Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophr Res.* (2006) 83:155–71. doi: 10.1016/j.schres.2005.11.020

35. Diedrichsen J, Balsters JH, Flavell J, Cussans E, Ramnani N. A probabilistic MR atlas of the human cerebellum. *Neuroimage.* (2009) 46:39–46. doi: 10.1016/j.neuroimage.2009.01.045

36. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* (2009) 19:1639–45. doi: 10.1101/gr.092759.109

37. Ruigrok AN, Salimi-Khorshidi G, Lai MC, Baron-Cohen S, Lombardo MV, Tait RJ, et al. a meta-analysis of sex differences in human brain structure. *Neurosci Biobehav Rev.* (2014) 39:34–50. doi: 10.1016/j.neubiorev.2013.12.004

38. Sen B, Parhi KK. Predicting male vs. female from task- fMRI brain connectivity. *Annu Int Conf IEEE Eng Med Biol Soc.* (2019) 2019:4089–92. doi: 10.1109/EMBC.2019.8857236

39. Duarte-Carvajalino JM, Jahanshad N, Lenglet C, McMahon KL, de Zubicaray GI, Martin NG, et al. Hierarchical topological network analysis of anatomical human brain connectivity and differences related to sex and kinship. *Neuroimage.* (2012) 59:3784–804. doi: 10.1016/j.neuroimage.2011.10.096

40. Weis S, Patil KR, Hoffstaedter F, Nostro A, Yeo BTT, Eickhoff SB. Sex classification by resting state brain connectivity. *Cereb Cortex.* (2020) 30:824–35. doi: 10.1093/cercor/bhz129

41. Ritchie SJ, Cox SR, Shen X, Lombardo MV, Reus LM, Alloza C, et al. Sex differences in the adult human brain: evidence from 5216 UK biobank participants. *Cereb Cortex.* (2018) 28:2959–75. doi: 10.1093/cercor/bhy109

42. Leming M, Suckling J. Deep learning for sex classification in resting-state and task functional brain networks from the UK biobank. *Neuroimage.* (2021) 241:118409. doi: 10.1016/j.neuroimage.2021.118409

43. Anderson NE, Harenski KA, Harenski CL, Koenigs MR, Decety J, Calhoun VD, et al. Machine learning of brain gray matter differentiates sex in a large forensic sample. *Hum Brain Mapp.* (2019) 40:1496–506. doi: 10.1002/hbm.24462

44. Jiang R, Calhoun VD, Cui Y, Qi S, Zhuo C, Li J, et al. Multimodal data revealed different neurobiological correlates of intelligence between males and females. *Brain Imaging Behav.* (2020) 14:1979–93. doi: 10.1007/s11682-019-00146-z

45. Eliot L. The trouble with sex differences. *Neuron.* (2011) 72:895–8. doi: 10.1016/j.neuron.2011.12.001

46. Joel D, Berman Z, Tavor I, Wexler N, Gaber O, Stein Y, et al. Sex beyond the genitalia: the human brain mosaic. *Proc Natl Acad Sci U S A*. (2015) 112:15468–73. doi: 10.1073/pnas.1509654112

47. Del Giudice M, Lippa RA, Puts DA, Bailey DH, Bailey JM. Schmitt DP. Joel et al.'s method systematically fails to detect large, consistent sex differences. *Proc Natl Acad Sci U S A*. (2016) 113:E1965. doi: 10.1073/pnas.1525534113

48. Sanchis-Segura C, Aguirre N, Cruz-Gómez ÁJ, Félix S, Forn C. Beyond "sex prediction": estimating and interpreting multivariate sex differences and similarities in the brain. *Neuroimage*. (2022) 257:119343. doi: 10.1016/j.neuroimage.2022.119343

49. Eliot L, Ahmed A, Khan H, Patel J. Dump the "dimorphism": comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neurosci Biobehav Rev*. (2021) 125:667–97. doi: 10.1016/j.neubiorev.2021.02.026

50. Williams CM, Peyre H, Toro R, Ramus F. Neuroanatomical norms in the UK biobank: the impact of allometric scaling, sex, and age. *Hum Brain Mapp*. (2021) 42:4623–42. doi: 10.1002/hbm.25572

51. Hirnstein M, Hausmann M. Sex/gender differences in the brain are not trivial-a commentary on Eliot et al. (2021). *Neurosci Biobehav Rev*. (2021) 130:408–9. doi: 10.1016/j.neubiorev.2021.09.012

52. Williams CM, Peyre H, Toro R, Ramus F. Sex differences in the brain are not reduced to differences in body size. *Neurosci Biobehav Rev*. (2021) 130:509–11. doi: 10.1016/j.neubiorev.2021.09.015

53. Liem F, Varoquaux G, Kynast J, Beyer F, Kharabian Masouleh S, Huntenburg JM, et al. Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage*. (2017) 148:179–88. doi: 10.1016/j.neuroimage.2016.11.005

54. Franke K, Ziegler G, Klöppel S, Gaser C. Estimating the age of healthy subjects from T1-weighted MRI scans using Kernel methods: exploring the influence of various parameters. *Neuroimage*. (2010) 50:883–92. doi: 10.1016/j.neuroimage.2010.01.005

55. Dosenbach NU, Nardos B, Cohen AL, Fair DA, Power JD, Church JA, et al. Prediction of individual brain maturity using f MRI. *Science*. (2010) 329:1358–61. doi: 10.1126/science.1194144

56. Brown TT, Kuperman JM, Chung Y, Erhart M, McCabe C, Hagler DJ Jr, et al. Neuroanatomical assessment of biological maturity. *Curr Biol*. (2012) 22:1693–8. doi: 10.1016/j.cub.2012.07.002

57. Erus G, Battapady H, Satterthwaite TD, Hakonarson H, Gur RE, Davatzikos C, et al. Imaging patterns of brain development and their relationship to cognition. *Cereb Cortex*. (2015) 25:1676–84. doi: 10.1093/cercor/bht425

58. Peng H, Gong W, Beckmann CF, Vedaldi A, Smith SM. Accurate brain age prediction with lightweight deep neural networks. *Med Image Anal*. (2021) 68:101871. doi: 10.1016/j.media.2020.101871

59. Franke K, Gaser C. Ten years of brainage as a neuroimaging biomarker of brain aging: what insights have we gained? *Front Neurol*. (2019) 10:789. doi: 10.3389/fneur.2019.00789

60. Clausen AN, Fercho KA, Monsour M, Disner S, Salminen L, Haswell CC, et al. Assessment of brain age in posttraumatic stress disorder: findings from the enigma ptsd and brain age working groups. *Brain Behav*. (2021) 12:e2413. doi: 10.1002/brb3.2413

61. Wrigglesworth J, Yaacob N, Ward P, Woods RL, McNeil J, Storey E, et al. Brain-predicted age difference is associated with cognitive processing in later-life. *Neurobiol Aging*. (2022) 109:195–203. doi: 10.1016/j.neurobiolaging.2021.10.007

62. Huang W, Li X, Li H, Wang W, Chen K, Xu K, et al. Accelerated brain aging in amnestic mild cognitive impairment: relationships with individual cognitive decline, risk factors for alzheimer disease, and clinical progression. *Radiol Artif Intell*. (2021) 3:e200171. doi: 10.1148/ryai.2021200171

63. Chung Y, Addington J, Bearden CE, Cadenhead K, Cornblatt B, Mathalon DH, et al. Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. *JAMA Psychiatry*. (2018) 75:960–8. doi: 10.1001/jamapsychiatry.2018.1543

64. Löwe LC, Gaser C, Franke K. The effect of the apoe genotype on individual brainage in normal aging, mild cognitive impairment, and alzheimer's disease. *PLoS ONE*. (2016) 11:e0157514. doi: 10.1371/journal.pone.0157514

65. Shahab S, Mulsant BH, Levesque ML, Calarco N, Nazeri A, Wheeler AL, et al. Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls. *Neuropsychopharmacology*. (2019) 44:898–906. doi: 10.1038/s41386-018-0298-z

66. Wrigglesworth J, Ward P, Harding IH, Nilaweera D, Wu Z, Woods RL, et al. Factors associated with brain ageing - a systematic review. *BMC Neurol*. (2021) 21:312. doi: 10.1186/s12883-021-02331-4

67. Király A, Szabó N, Tóth E, Csete G, Faragó P, Kocsis K, et al. Male brain ages faster: the age and gender dependence of subcortical volumes. *Brain Imaging Behav*. (2016) 10:901–10. doi: 10.1007/s11682-015-9468-3

68. Peng F, Wang L, Geng Z, Zhu Q, Song Z. A cross-sectional voxel-based morphometric study of age- and sex-related changes in gray matter volume in the normal aging brain. *J Comput Assist Tomogr*. (2016) 40:307–15. doi: 10.1097/RCT.0000000000000351

69. Coffey CE, Lucke JF, Saxton JA, Ratcliff G, Unitas LJ, Billig B, et al. Sex differences in brain aging: a quantitative magnetic resonance imaging study. *Arch Neurol*. (1998) 55:169–79. doi: 10.1001/archneur.55.2.169

70. Subramaniapillai S, Rajagopal S, Snytte J, Otto AR, PREVENT-AD Research Group, Einstein G, et al. Sex differences in brain aging among adults with family history of Alzheimer's disease and APOE4 genetic risk. *Neuroimage Clin*. (2021) 30:102620. doi: 10.1016/j.nicl.2021.102620

71. Frontera WR, Hughes VA, Fielding RA, Fiatarone MA, Evans WJ, Roubenoff R. Aging of skeletal muscle: a 12-yr longitudinal study. *J Appl Physiol*. (2000) 88:1321–6. doi: 10.1152/jappl.2000.88.4.1321

72. Goodpaster BH, Park SW, Harris TB, Kritchevsky SB, Nevitt M, Schwartz AV, et al. The loss of skeletal muscle strength, mass, and quality in older adults: the health, aging and body composition study. *J Gerontol A Biol Sci Med Sci*. (2006) 61:1059–64. doi: 10.1093/gerona/61.10.1059

73. Marcus R. Relationship of age-related decreases in muscle mass and strength to skeletal status. *J Gerontol A Biol Sci Med Sci*. (1995) 50:86–7. doi: 10.1093/gerona/50A.Special_Issue.86

74. Duncan NW, Northoff G. Overview of potential procedural and participant-related confounds for neuroimaging of the resting state. *J Psychiatry Neurosci*. (2013) 38:84–96. doi: 10.1503/jpn.120059

75. Logothetis NK, Murayama Y, Augath M, Steffen T, Werner J, Oeltermann A. How not to study spontaneous activity. *Neuroimage*. (2009) 45:1080–9. doi: 10.1016/j.neuroimage.2009.01.010

76. Agcaoglu O, Wilson TW, Wang YP, Stephen J, Calhoun VD. Resting state connectivity differences in eyes open versus eyes closed conditions. *Hum Brain Mapp*. (2019) 40:2488–98. doi: 10.1002/hbm.24539

77. Kostro D, Abdulkadir A, Durr A, Roos R, Leavitt BR, Johnson H, et al. Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *Neuroimage*. (2014) 98:405–15. doi: 10.1016/j.neuroimage.2014.04.057

78. Chen HS, Kumar VA, Johnson JM, Chen MM, Noll KR, Hou P, et al. Effect of brain normalization methods on the construction of functional connectomes from resting-state functional MRI in patients with gliomas. *Magn Reson Med*. (2021) 86:487–98. doi: 10.1002/mrm.28690

79. Forsberg L, Sigurdsson S, Fredriksson J, Egilsdottir A, Oskarsdottir B, Kjartansson O, et al. The ages-reykjavik study atlases: non-linear multi-spectral template and atlases for studies of the ageing brain. *Med Image Anal*. (2017) 39:133–44. doi: 10.1016/j.media.2017.04.009

80. Pai PP, Mandal PK, Punjabi K, Shukla D, Goel A, Joon S, et al. Brahma: population specific T1, T2, and flair weighted brain templates and their impact in structural and functional imaging studies. *Magn Reson Imaging*. (2020) 70:5–21. doi: 10.1016/j.MRI.2019.12.009

81. Rao NP, Jeelani H, Achalia R, Achalia G, Jacob A, Bharath RD, et al. Population differences in brain morphology: need for population specific brain template. *Psychiatry Res Neuroimaging*. (2017) 265:1–8. doi: 10.1016/j.pscychresns.2017.03.018

82. Yang G, Zhou S, Bozek J, Dong HM, Han M, Zuo XN, et al. Sample sizes and population differences in brain template construction. *Neuroimage*. (2020) 206:116318. doi: 10.1016/j.neuroimage.2019.116318

83. Ganzetti M, Liu Q, Mantini D. A spatial registration toolbox for structural mr imaging of the aging brain. *Neuroinformatics*. (2018) 16:167–79. doi: 10.1007/s12021-018-9355-3

84. Gureje O, Von Korff M, Simon GE, Gater R. Persistent Pain and well-being: a world health organization study in primary care. *JAMA*. (1998) 280:147–51. doi: 10.1001/jama.280.2.147

85. Carré A, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Sci Rep*. (2020) 10:12340. doi: 10.1038/s41598-020-69298-z

86. Lee JJ, Kim HJ, Ceko M, Park BY, Lee SA, Park H, et al. A neuroimaging biomarker for sustained experimental and clinical pain. *Nat Med*. (2021) 27:174–82. doi: 10.1038/s41591-020-1142-7

87. Zhao Q, Adeli E, Pohl KM. Training confounder-free deep learning models for medical applications. *Nat Commun*. (2020) 11:6010. doi: 10.1038/s41467-020-19784-9

88. Kilgour AH, Todd OM, Starr JM. A systematic review of the evidence that brain structure is related to muscle structure and their relationship to brain and muscle function in humans over the lifecourse. *BMC Geriatr.* (2014) 14:85. doi: 10.1186/1471-2318-14-85

89. Voorbij AI, Steenbekkers LP. The composition of a graph on the decline of total body strength with age based on pushing, pulling, twisting and gripping force. *Appl Ergon.* (2001) 32:287–92. doi: 10.1016/S0003-6870(00)00068-5

90. Caspi Y, Brouwer RM, Schnack HG, van de Nieuwenhuijzen ME, Cahn W, Kahn RS, et al. Changes in the intracranial volume from early adulthood to the sixth decade of life: a longitudinal study. *Neuroimage.* (2020) 220:116842. doi: 10.1016/j.neuroimage.2020.116842

91. Erickson KI, Voss MW, Prakash RS, Basak C, Szabo A, Chaddock L, et al. Exercise training increases size of hippocampus and improves memory. *Proc Natl Acad Sci U S A.* (2011) 108:3017–22. doi: 10.1073/pnas.1015950108

92. Herold F, Törpel A, Schega L, Müller NG. Functional and/or structural brain changes in response to resistance exercises and resistance training lead to cognitive improvements - a systematic review. *Eur Rev Aging Phys Act.* (2019) 16:10. doi: 10.1186/s11556-019-0217-2

93. Won J, Callow DD, Pena GS, Gogniat MA, Kommula Y, Arnold-Nedimala NA, et al. Evidence for exercise-related plasticity in functional and structural neural network connectivity. *Neurosci Biobehav Rev.* (2021) 131:923–40. doi: 10.1016/j.neubiorev.2021.10.013

94. Mierzejewska-Krzyzowska B, Bukowska D, Taborowska M, Celichowski J. Sex differences in the number and size of motoneurons innervating rat medial gastrocnemius muscle. *Anat Histol Embryol.* (2014) 43:182–9. doi: 10.1111/ahe.12060

95. Mierzejewska-Krzyzowska B, Drzymała-Celichowska H, Celichowski J. Gender differences in the morphometric properties of muscle fibres and the innervation ratio of motor units in rat medial gastrocnemius muscle. *Anat Histol Embryol.* (2011) 40:249–55. doi: 10.1111/j.1439-0264.2011.01066.x

96. Gartych M, Jackowiak H, Bukowska D, Celichowski J. Evaluating sexual dimorphism of the muscle spindles and intrafusal muscle fibers in the medial gastrocnemius of male and female rats. *Front Neuroanat.* (2021) 15:734555. doi: 10.3389/fnana.2021.734555

97. Raichlen DA, Gordon AD. Relationship between exercise capacity and brain size in mammals. *PLoS ONE.* (2011) 6:e20601. doi: 10.1371/journal.pone.0020601

98. Vidal-Pineiro D, Wang Y, Krogsrud SK, Amlien IK, Baaré WF, Bartres-Faz D, et al. Individual variations in 'brain age' relate to early-life factors more than to longitudinal brain change. *Elife.* (2021) 10:e69995. doi: 10.7554/eLife.69995

99. Cramer SC, Weisskoff RM, Schaechter JD, Nelles G, Foley M, Finklestein SP, et al. Motor cortex activation is related to force of squeezing. *Hum Brain Mapp.* (2002) 16:197–205. doi: 10.1002/hbm.10040