ORIGINAL RESEARCH

Cross-Scanner Harmonization of Neuromelanin-Sensitive MRI for Multisite Studies

Kenneth Wengler, PhD,[1*] Clifford Cassidy, PhD,[2] Marieke van der Pluijm, MS,[3,4]
Jodi J. Weinstein, MD,[1,5] Anissa Abi-Dargham, MD,[5] Elsmarieke van de Giessen, MD, PhD,[3]
and Guillermo Horga, MD, PhD[1*]

Background: Neuromelanin-sensitive magnetic resonance imaging (NM-MRI) is a validated measure of neuromelanin concentration in the substantia nigra–ventral tegmental area (SN–VTA) complex and is a proxy measure of dopaminergic function with potential as a noninvasive biomarker. The development of generalizable biomarkers requires large-scale samples necessitating harmonization approaches to combine data collected across sites.
Purpose: To develop a method to harmonize NM-MRI across scanners and sites.
Study Type: Prospective.
Population: A total of 128 healthy subjects (18–73 years old; 45% female) from three sites and five MRI scanners.
Field Strength/Sequence: 3.0 T; NM-MRI two-dimensional gradient-recalled echo with magnetization-transfer pulse and three-dimensional T1-weighted images.
Assessment: NM-MRI contrast (contrast-to-noise ratio [CNR]) maps were calculated and CNR values within the SN–VTA (defined previously by manual tracing on a standardized NM-MRI template) were determined before harmonization (raw CNR) and after ComBat harmonization (harmonized CNR). Scanner differences were assessed by calculating the classification accuracy of a support vector machine (SVM). To assess the effect of harmonization on biological variability, support vector regression (SVR) was used to predict age and the difference in goodness-of-fit (Δr) was calculated as the correlation (between actual and predicted ages) for the harmonized CNR minus the correlation for the raw CNR.
Statistical Tests: Permutation tests were used to determine if SVM classification accuracy was above chance level and if SVR Δr was significant. A P-value <0.05 was considered significant.
Results: In the raw CNR, SVM MRI scanner classification was above chance level (accuracy = 86.5%). In the harmonized CNR, the accuracy of the SVM was at chance level (accuracy = 29.5%; $P = 0.8542$). There was no significant difference in age prediction using the raw or harmonized CNR ($\Delta r = -0.06$; $P = 0.7304$).
Data Conclusion: ComBat harmonization removes differences in SN–VTA CNR across scanners while preserving biologically meaningful variability associated with age.
Level of Evidence: 2
Technical Efficacy: 1

J. MAGN. RESON. IMAGING 2021;54:1189–1199.

The search for biomarkers in neuropsychiatric disorders is a major research agenda for the field and aligns with ongoing precision medicine initiatives.[1–4] Neuromelanin-sensitive magnetic resonance imaging (NM-MRI) is a noninvasive imaging technique that exploits the paramagnetic nature of neuromelanin–iron complexes that form as the result of dopamine metabolism in dopaminergic neurons of the midbrain.[5–7] NM-MRI has been validated as a marker of both dopaminergic function and dopaminergic neurodegeneration,[8,9] indicating potential for NM-MRI as a biomarker of the

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27679

Received Feb 22, 2021, Accepted for publication Apr 19, 2021.

*Address reprint requests to: K.W., Columbia University, New York, NY 10032, USA. E-mail: kenneth.wengler@nyspi.columbia.edu; G.H., Columbia University, New York, NY 10032, USA. E-mail: horgag@nyspi.columbia.edu

From the [1]Department of Psychiatry, New York State Psychiatric Institute, Columbia University, New York, USA; [2]University of Ottawa Institute of Mental Health Research, affiliated with The Royal, Ottawa, Ontario, Canada; [3]Department of Radiology and Nuclear Medicine, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; [4]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; and [5]Department of Psychiatry, Stony Brook University, Stony Brook, New York, USA

dopaminergic system. Furthermore, NM-MRI has several features positioning it as an ideal biomarker candidate: Its noninvasive nature and lack of ionizing radiation allowing for repeated and longitudinal measurement, its ease of implementation, and its excellent test–retest reliability.[10–12] The development and characterization of biomarkers demands large sample sizes to facilitate training and testing of statistical models and to provide generalizability.[1] Multisite studies facilitate obtaining such large sample sizes, but a method for combining NM-MRI data from multiple sites has yet to be developed.

Combining MRI data from multiple sites can introduce unwanted, nonbiological variability into the data from differences in hardware (eg, MRI scanner or head coil) or software (eg, pulse sequence parameters).[13–15] Including covariates in analyses (i.e. statistically controlling for site) does not sufficiently remove this variance and may perform no better than models which ignore the confounds.[16] Several methods for MRI data harmonization[16,17]—that is, the explicit removal of nonbiological variability (eg, hardware and software related effects)—have been developed for neuroimaging data, but ComBat[18] has emerged as a literature standard. ComBat is an empirical Bayes method for harmonizing both the mean and variance of a measure (eg, cortical thickness) across batches and is particularly robust for small sample sizes. This approach has been previously applied to MRI measurement of cortical thickness,[17,19] regional brain volumes[19,20] and cortical surface area,[19] diffusion tensor imaging,[21,22] resting-state functional MRI,[23,24]and task-based functional MRI.[23]

For a harmonization method to be useful, it must maintain biologically meaningful variability (eg, differences associated with diagnostic status). Because NM accumulates in dopaminergic midbrain neurons over the lifespan,[25–28] the NM-MRI signal should increase with age,[29] thus providing biological variability that can be used to test a NM-MRI harmonization approach in healthy subjects. A possible added benefit of harmonization is the ability to improve both the reproducibility and statistical power of downstream analyses due to the removal of unwanted, nonbiological variability.[17,19] This would be particularly beneficial for biomarker development where greater reproducibility and statistical power could lead to more generalizable biomarkers.

Thus, the aims of this study were to introduce a method for harmonization of contrast-to-noise ratio (CNR) maps calculated from NM-MRI data across multiple sites, scanner vendors, and acquisition parameters, to assess the ability of this method to harmonize the data while maintaining biologically meaningful variability associated with age, and its ability to improve both the reproducibility and statistical power of the expected positive relationship between the NM-MRI signal and age.[25–29]

Methods and Materials

Participants

All subjects provided written informed consent and institutional review board approval was obtained from the three institutes. Inclusion criteria were: age ≥ 18 and no MRI contraindications. Exclusion criteria were: history of neurological or psychiatric diseases, pregnancy or nursing, and inability to provide written consent. Two-dimensional gradient-recalled echo with magnetization-transfer pulse (GRE-MT) NM-MRI and anatomical three-dimensional (3D) T1-weighted (T1w) images were collected from healthy subjects at three sites: Site 1 (3 T MR750 and 3 T Signa Premier, GE, Milwaukee, WI), Site 2 (3 T Prisma, Siemens, Erlangen, Germany), and Site 3 (3 T Ingenia and 3 T Ingenia Elition, Philips, Best, The Netherlands). Due to hardware and/or software differences, the NM-MRI sequence parameters differed across MRI scanners; a detailed description of sequence parameters is given in Table 1. Sequence parameters for the 3D T1w acquisitions are listed in Table 2.

Data Processing

NM-MRI data were preprocessed using a pipeline combining statistical parametric mapping (SPM) and advanced normalization tools (ANTs) previously shown to achieve excellent test–retest reliability.[12] This pipeline consisted of the following steps: 1) if averages were acquired separately (i.e. offline averaging), realignment to correct for motion using "SPM-Realign" and averaging of the realigned images using "SPM-ImCalc"; 2) brain extraction of the T1w image using "antsBrainExtraction.sh"; 3) spatial normalization of the brain-extracted T1w image to the MNI152NLin2009cAsym template space using "antsRegistrationSyN.sh" (rigid + affine + deformable syn); 4) coregistration of the NM-MRI image to the T1w image using "antsRegistrationSyN.sh" (rigid); 5) spatial normalization of the NM-MRI images to template space by a single-step transformation combining the transformations estimated in steps 3 and 4 using "antsApplyTransforms"; and 6) spatial smoothing of the spatially normalized NM-MRI image with a 1 mm full-width-at-half-maximum Gaussian kernel using "SPM-Smooth."

The preprocessed NM-MRI images in the template space were then used to calculate NM-MRI CNR maps. NM-MRI CNR at each voxel (CNR_v) was calculated as the percent signal difference in NM-MRI signal intensity at a given voxel (I_v) from the signal intensity in the crus cerebri (I_{CC})—a region of white matter known to have minimal NM content—as:

$$\text{CNR}_v = \{[I_v - \text{mode}(I_{\text{CC}})]/\text{mode}(I_{\text{CC}})\} \times 100$$

where $\text{mode}(I_{\text{CC}})$ is calculated for each participant from a kernel-smoothing-function fit to a histogram of all voxels in

TABLE 1. NM-MRI Acquisition Parameters

	Site 1		Site 2	Site 3	
Parameter	GE MR750	GE Signa Premier	Siemens Prisma	Philips Ingenia	Philips Elition
RO FOV (mm)	220	220	220	199	199
PE FOV (mm)	165	165	165	162	162
SS FOV (mm)	30	30	30	22	22
RO resolution (mm)	0.43	0.43	0.43	0.39	0.39
PE resolution (mm)	0.43	0.43	0.43	0.39	0.39
Slice thickness (mm)	3	1.5	3	2.5	2.5
Acquisition matrix (RO × PE)	512 × 320	512 × 320	512 × 320	512 × 416	512 × 416
Number of slices	10	20	10	8	8
Slice gap (mm)	0	0	0	0.25	0.25
Slice orientation	// AC-PC line	// AC-PC line	// AC-PC line	⊥ 4th ventricle	⊥ 4th ventricle
TE (msec)	3.9	4.8	3.9	3.9	3.9
TR (msec)	250	500	273	260	260
FA (°)	40	40	40	40	40
NEX	8	5	10	2	2
Averaging mode	Online	Online	Offline	Online	Online
BW (Hz/Px)	122	122	315	75	75
MT offset (Hz)	1200	1200	1200	1200	1200
MT duration (msec)	10	10	10	15.6	15.6
Acquisition time (minutes:seconds)	8:04	10:04	11:02	13:20	13:20
Receive coil	32-channel head coil	48-channel head coil	20-channel head–neck coil	32-channel head coil	32-channel head coil

RO = read-out direction; FOV = field-of-view; PE = phase-encoding direction; SS = slice-selection direction; TE = echo time; TR = repetition time; FA = flip angle; NEX = number of averages; BW = bandwidth; MT = magnetization transfer pulse; AC-PC = anterior cingulate-posterior cingulate; // = parallel to; ⊥ = perpendicular to.

the crus cerebri (CC) mask following prior work.[8] Masks of both the CC and substantia nigra–ventral tegmental area (SN–VTA) complex were taken from a previous study where they were manually drawn on an NM-MRI template averaged from 40 subjects[8] and have been subsequently used in other NM-MRI studies.[12,30]

CNR values of each voxel in the SN–VTA mask were then harmonized using the ComBat harmonization model[18,21] to remove nonbiological variability while maintaining biological variability associated with age and sex. This model can be written as:

$$y_{ijv} = \alpha_v + \mathrm{X}_{ij}^T \boldsymbol{\beta}_v + \gamma_{iv} + \delta_{iv}\varepsilon_{ijv}$$

where y_{ijv} is the CNR value of MRI scanner i ($i \in \{1, \ldots, 5\}$), subject j ($j \in \{1, \ldots, N\}$; where N is the number of subjects), and SN–VTA voxel v ($v \in \{1, \ldots, 1807\}$); α_v is the average CNR value over subjects for SN–VTA voxel v; $\mathbf{X}$ is a design matrix for the covariates of interest (age and sex); $\boldsymbol{\beta}_v$ is a vector of regression coefficients corresponding to $\mathbf{X}$ for SN–VTA voxel v; γ_{iv} and δ_{iv} are the additive and multiplicative effects of MRI scanner i for SN–VTA voxel v, respectively; and ε_{ijv} are the error terms that are assumed

TABLE 2. Three-dimensional T1w Acquisition Parameters

	Site 1		Site 2	Site 3	
Parameter	**GE MR750**	**GE Signa Premier**	**Siemens Prisma**	**Philips Ingenia**	**Philips Elition**
RO FOV (mm)	240	240	166	284	284
PE FOV (mm)	240	240	240	284	284
SS FOV (mm)	176	176	166	170	170
RO resolution (mm)	0.8	0.8	0.8	0.9	0.9
PE resolution (mm)	0.8	0.8	0.8	0.9	0.9
Slice thickness (mm)	0.8	0.8	0.8	0.9	0.9
TE (msec)	3.1	3.4	564	4.1	4.1
TR (msec)	7.9	2500	3200	9.0	9.0
FA (°)	12	8	8	8	8

RO = read-out direction; FOV = field-of-view; PE = phase-encoding direction; SS = slice-selection direction; TE = echo time; TR = repetition time; FA = flip angle.

to follow a normal distribution with mean 0 and variance σ_v^2.

The ComBat-harmonized CNR values are defined as:

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - \widehat{\alpha}_v - \mathrm{X}_{ij}\widehat{\boldsymbol{\beta}}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \widehat{\alpha}_v + \mathrm{X}_{ij}\widehat{\boldsymbol{\beta}}_v$$

where γ_{iv}^* and δ_{iv}^* are the empirical Bayes estimates of γ_{iv} and δ_{iv}, respectively. Harmonization of the SN–VTA CNR was performed using a publicly available MATLAB (MathWorks, Natick, MA) package hosted at https://github.com/Jfortin1/ComBatHarmonization/tree/master/Matlab.

Scanner Effects

To visualize potential differences in NM-MRI SN–VTA CNR across scanners, we calculated the mean SN–VTA CNR distribution for each scanner. Specifically, for each SN–VTA voxel we calculated the mean over subjects (within a scanner) and fit a kernel distribution to the histogram of all SN–VTA voxels for visualization purposes. To quantify potential differences in SN–VTA CNR we calculated the within-subject median CNR over voxels.

Classification of MRI Scanner

Following a previous application of ComBat to MRI data,[17] we evaluated the performance of the proposed NM-MRI harmonization method using support vector machines (SVM) to classify the MRI scanner from the pattern of SN–VTA CNR.[23] Specifically, the input features for the SVM were either the raw or harmonized CNR values of the 1807 voxels within the SN–VTA mask. Mean classification accuracy was calculated using 5-fold cross-validation (1000 repetitions) of a one-vs.-one error correcting output code linear SVM (MATLAB, *fitcecoc* with hyperparameter *C* optimized in an inner 3-fold cross-validation loop)—this method was previously shown to improve performance over other multiclass classifiers.[31] To account for imbalanced sample sizes across scanners, misclassification costs were set to be inversely proportional to class frequencies for each one-vs.-one classifier.[32]

Multivariate Prediction of Age

To determine the ability of the proposed NM-MRI harmonization method to maintain biologically meaningful variability, we used linear epsilon-insensitive support vector regressions (SVR) to predict age from the voxelwise pattern of SN–VTA CNR. Specifically, the input features for the SVR were the raw or harmonized CNR values of the 1807 voxels within the SN–VTA mask. Root-mean-square error (RMSE) and Pearson correlation coefficient (*r*) between the actual and predicted ages were calculated using 10-fold cross-validation (1000 repetitions) linear SVR (MATLAB, *fitrlinear* with hyperparameter *C* optimized in an inner 5-fold cross-validation loop). Due to significant side effects observed in the raw CNR, site was regressed out for each feature prior to input in the SVR (for the raw CNR only).[33]

To evaluate differences in the ability to predict age using SN–VTA CNR, the differences in RMSE (ΔRMSE) and *r (*Δ*r)* between the SVRs using the raw and harmonized CNR was calculated for every combination of the 1000 repetitions.

To determine the reproducibility of the age-prediction results, 1000 random subsets of subjects of varying sample sizes (32, 64, and 96) were selected and SVR was used to predict age from the voxelwise pattern of SN–VTA CNR. In a SVR, the weight of a predictor variable or feature (eg, CNR in voxel v) on a predicted outcome variable (eg, age) is reflected in the β-coefficient of each voxel such that the effect of interest is the multivoxel (spatial) pattern of β-coefficients. To investigate the reliability of the pattern of β-coefficients for predicting age, the similarity of the pattern was evaluated by calculating the spatial correlation (Pearson r) between SVR β-coefficients from a given subset of subjects and every other random subset with the same sample size.[34]

Standard Univariate Voxelwise Analysis of Age Effects

To determine the statistical power of the age effect (significant positive relationship with age), we performed a standard voxelwise analysis via robust linear regressions (MATLAB, *fitlm* with “RobustOpts,” “on”) that predicted CNR at every voxel v within the SN–VTA as follows:

$$\mathrm{CNR}_v = \beta_0 + \beta_1 \cdot \mathrm{age} + \sum_{i=2}^{n+1} \beta_i \cdot \text{nuisance covariate} + \varepsilon$$

where $n = 5$ for the raw CNR (sex and four dummy variables for the five MRI scanners as nuisance covariates) and $n = 1$ for the harmonized CNR (sex as a nuisance covariate). To correct for multiple comparisons, we calculated the spatial extent of an effect as the number of voxels, k, exhibiting a significant correlation with age (voxel-level height threshold for t test of regression coefficient β_1 of $P < 0.05$, one-sided).

A leave-one-out analysis (i.e. 128 folds) was performed to estimate the unbiased correlation between age and the mean CNR in *age-effect voxels* (those voxels showing a significant relationship with age), where the age-effect voxels used to read out the mean CNR in a given subject were determined in a sample excluding that subject (n = 127).[8,35]

The variability in the location of age-effect voxels was estimated using the results from the leave-one-out analysis. The overlap in the spatial location of the age-effect voxels was evaluated by calculating the Dice similarity coefficient between the age-effect voxels determined from one of the 128 folds and those determined by the other 127 folds.

To further investigate the effect of harmonization on statistical power, we compared the number of age-effect voxels for each fold of the leave-one-out analysis between the raw and harmonized CNR.

Left–Right Asymmetries

Left–right asymmetries (referred to here as laterality effect) were investigated by comparing the difference in left and right hemisphere SN–VTA CNR. For this we calculated the difference in the median CNR over voxels between the left hemisphere and the right hemisphere for each subject.

Statistical Analysis

One-way ANOVAs were used to test for significant differences in SN–VTA CNR across scanners by comparing the means (over subjects within a scanner) of the median (over voxels within a subject) CNR (five levels) and to test for a significant difference in age across scanners (5 levels). Bonferroni post-hoc tests were used to test for significant differences in age for each of the 5 MRI scanners.

Permutation tests were used to determine if the mean (over cross-validation repetitions) SVM classification accuracy was significantly greater than expected by chance (one-tailed test). This was done by comparing the mean classification accuracy to a null distribution of classification accuracies generated by running the SVM 10,000 times with MRI scanner labels randomly shuffled each time.

Permutation tests were used to determine if the mean (over cross-validation repetitions) SVR performance (RMSE and r) was significantly better than expected by chance (one-tailed tests). This was done by comparing the mean performance to a null distribution of age prediction performance generated by running the SVR 10,000 times with age values randomly shuffled each time. These same null distributions were subsequently used to determine if the mean ΔRMSE and Δr were greater or less than expected by chance (two-tailed tests), but here the new null distributions were generated by taking the difference between every combination of the previous null distributions for the raw and harmonized CNR.

Permutation tests were used to determine if the number of age-effect voxels in the standard univariate voxelwise analyses was greater than expected by chance (one-tailed tests). This was done by comparing the number of age-effect voxels (positive relationship with age determined by voxel-level height threshold for t-test of regression coefficient $P < 0.05$, one-sided) to a null distribution of the number for age-effect voxels generated by running the standard univariate voxelwise analysis 10,000 times with age values randomly shuffled each time. These same null distributions were subsequently used to determine if the number of age-effect voxels was significantly different for the raw and harmonized CNR (two-tailed test), but here the new null distributions were generated by taking the difference between every combination of the previous null distributions for the raw and harmonized CNR.

Partial Pearson correlations (r_{partial}) were used to calculate the unbiased correlation between age and the mean CNR in age-effect voxels (from leave-one-out analysis) while controlling for the nuisance covariates included in the robust linear regression (two-tailed tests). Because the correlations are overlapping, the Meng, Rosenthal, and Rubin’s z-test was used to compare the unbiased correlation coefficients between the raw and harmonized CNR.[36]

A two-way analysis of variance (ANOVA) was used to test for significant differences in the mean (over random subsamples within a sample size or over cross-validation repetitions for the full sample) spatial correlation of SVR β-coefficients between sample sizes (four levels) and the raw and harmonized CNR (two levels). Bonferroni post-hoc tests were used to test for significant differences in the mean spatial correlations from the raw and harmonized CNR for each of the four sample sizes (n = 32, 64, 96, and 128).

Permutation tests were used to determine if there was a significant difference in the median (over leave-one-out folds) Dice coefficient for the age-effect voxels from the raw and harmonized CNR (two-tailed test).

Multiple linear regression models were used to test for significant laterality effects for each scanner with the laterality effect of each subject as the dependent variable and each scanner as categorical independent variables, and t tests were performed on the regression coefficients for each scanner. One-way ANOVAs were also used to test for significant differences in the laterality effect across scanners (five levels).

For all tests, $P < 0.05$ was considered significant.

Results

Demographics

NM-MRI and anatomical T1w MRI data were collected on 128 healthy subjects from studies at Site 1, Site 2, and Site 3. At Site 1, 51 subjects were collected on a 3 T GE MR750 (mean ± standard deviation age: 34.5 ± 14.6 years; 22 female and 29 male) and 29 subjects on a 3 T GE Signa Premier (age: 29.0 ± 7.4 years; 15 female and 14 male); at Site 2, 24 subjects were collected on a 3 T Siemens Prisma (age: 27.8 ± 8.9 years; 12 female and 12 male); and at Site 3, 12 subjects were collected on a 3 T Philips Ingenia (age: 23.9 ± 4.1 years; four female and eight male) and 12 subjects on a 3 T Philips Ingenia Elition (hereafter referred to as Philips Elition; age: 24.3 ± 2.1 years; five female and seven male). Age was significantly different across scanners ($F_{4,123} = 9.28$), with this difference due to subjects acquired on the GE MR750 being significantly older than those acquired on any of the other scanners. No significant differences were observed for any comparisons within the other four scanners (all $P > 0.6506$).

Effect of NM-MRI Harmonization on Scanner Effects

Substantial differences in the mean SN–VTA CNR distributions were visually apparent across the five MRI scanners (Figure 1a) and were seemingly removed with ComBat harmonization (Figure 1b). For the raw CNR values, the median over SN–VTA voxels differed significantly between scanners ($F_{4,123} = 72.45$; Figure 1c). For the harmonized CNR values, no differences between scanners were present ($F_{4,123} = 0.13$, $P = 0.969$; Figure 1d). A SVM multivoxel pattern analysis (MVPA) using the raw CNR showed significant, above chance level accuracy for MRI scanner classification (mean ± standard deviation classification accuracy = 86.5 ± 1.8%; Figure 1e). Using the harmonized CNR, the accuracy of the SVM classifier was not above chance level (classification accuracy = 29.5 ± 3.0%, $P = 0.8542$; Figure 1e). A control analysis to rule out that site effects were driven by subjects scanned in the GE MR750 scanner being older showed that excluding subjects above 35 years old (without which the difference in age across scanners was no longer significant: $F_{4,123} = 1.97$, $P = 0.1051$) had little effect on site classification, which remained above chance level for the raw CNR (classification accuracy = 84.5 ± 2.2%).

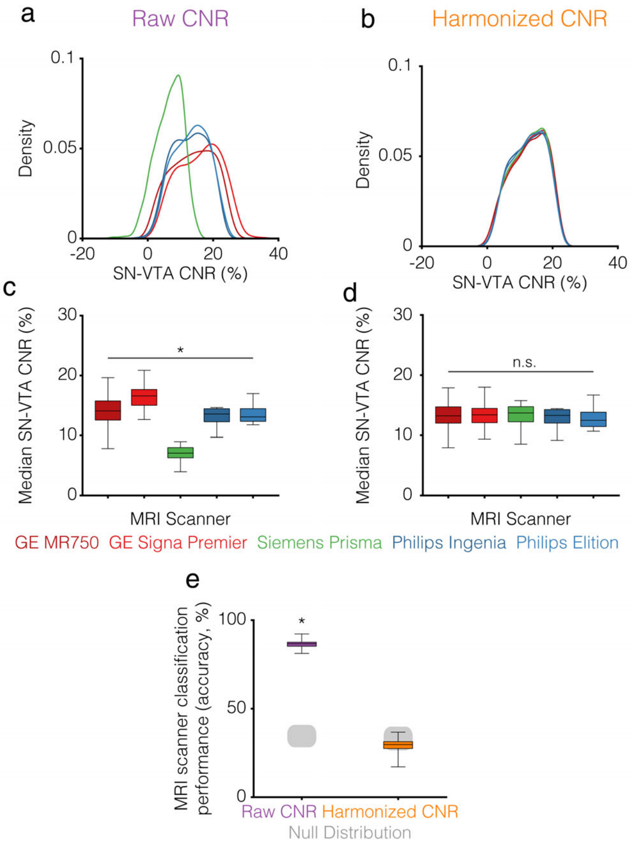

FIGURE 1: ComBat harmonization of neuromelanin-sensitive MRI (NM-MRI) substantia nigra–ventral tegmental area (SN–VTA) contrast-to-noise ratio (CNR). Kernel distributions of the mean SN–VTA CNR from all voxels for each of the five MRI scanners are shown for before (a) and after (b) harmonization. Boxplots show the distribution of median SN–VTA CNR of the raw (c) and the harmonized (d) CNR from subjects for each of the five MRI scanners. The asterisk (*) denotes $P < 0.05$ for an ANOVA comparing the median SN–VTA CNR across MRI scanners, and results that were not significant are labeled "n.s." (e) Boxplot showing the mean performance for classification of MRI scanner using 5-fold cross-validated linear support vector machine (SVM) across 1000 random splits of the data. The null distributions indicate empirical chance-level performance (5th–95th percentile shown) determined by randomly shuffling each subject's MRI scanner label 10,000 times. The asterisk (*) denotes $P < 0.05$ for the permutation test comparing the mean classification accuracy over the 1000 random splits to the null distribution. For all boxplots, the minimum, 25th percentile, 50th percentile (median), 75th percentile, and maximum are shown.

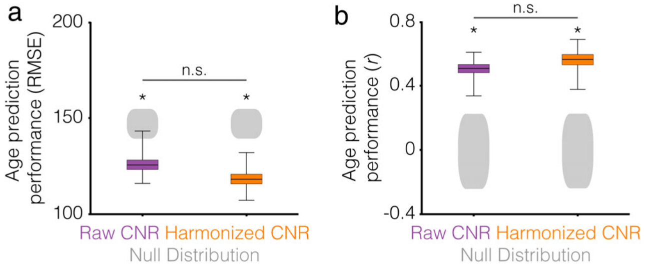

FIGURE 2: Effects of ComBat harmonization of neuromelanin-sensitive MRI (NM-MRI) substantia nigra–ventral tegmental area (SN–VTA) contrast-to-noise ratio (CNR) on biologically meaningful variability. Boxplots showing the mean performance for age prediction for two performance metrics: root-mean square error (RMSE; a) and Pearson correlation coefficient (*r*; b). The null distributions (gray regions) indicate empirical chance-level performance (5th–95th percentile shown) determined by randomly shuffling each subject's age 10,000 times. Asterisks (*) denote $P < 0.05$ and "n.s." denotes non-significance for the permutation test comparing the mean age prediction performance over the 1000 random splits to the null distribution. For all boxplots, the minimum, 25th percentile, 50th percentile (median), 75th percentile, and maximum are shown.

Effect of NM-MRI Harmonization on Biologically Meaningful Variability

A SVR using the raw CNR could predict age significantly above chance level (mean ± standard deviation RMSE = 125.91 ± 3.83 months; $r = 0.51 \pm 0.04$; Figure 2). Performance was also significantly above chance level using the harmonized CNR (RMSE = 118.35 ± 3.99 months, $r = 0.56 \pm 0.05$; Figure 2). There was no significant difference in age prediction between the SVRs using the raw or harmonized CNR (ΔRMSE = 7.56 ± 5.53; $P = 0.1551$; $\Delta r = -0.06 \pm 0.06$; $P = 0.7304$).

Effect of NM-MRI Harmonization on Statistical Power

In the raw CNR, the number of age-effect voxels in the SN–VTA was within chance level (241 of 1807 voxels, $P = 0.072$; Figure 3a). In the harmonized CNR, the number of age-effect voxels was significantly above chance level (324 of 1807 voxels; Figure 3b). The unbiased correlation between SN–VTA CNR and age was significant for both the raw CNR ($r_{\text{partial}} = 0.19$) and the harmonized CNR ($r_{\text{partial}} = 0.25$); although the harmonized CNR correlation was numerically greater than the raw CNR, this effect was not statistically significant ($z = 0.90$, $P = 0.37$). We also observed significantly fewer age-effect voxels in the raw CNR (mean ± standard deviation: 239.9 ± 16.4 voxels) compared to the harmonized CNR (323.9 ± 12.5 voxels) (Figure 3c).

Effect of NM-MRI Harmonization on Reproducibility

For the spatial correlation of SVR β-coefficients, we found a significant interaction between sample size (n = 32, 64, 96, or 128) and CNR type (raw or harmonized) suggesting a significant improvement in reproducibility for the harmonized CNR ($F_{3,7991992} = 9{,}288.75$; Figure 4a). Bonferroni post-hoc tests revealed increases in reproducibility for the harmonized CNR that were significant for sample sizes of 64 (raw CNR: $r = 0.53 \pm 0.08$, harmonized CNR: $r = 0.55 \pm 0.07$), 96 (raw CNR: $r = 0.73 \pm 0.05$, harmonized CNR: $r = 0.75 \pm 0.04$), and 128 (raw CNR: $r = 0.99 \pm 0.01$, harmonized CNR: $r = 0.99 \pm 0.00$) but not significant for a sample size of 32 (raw CNR: $r = 0.34 \pm 0.11$, harmonized CNR: $r = 0.34 \pm 0.12$; $P = 0.817$). We also found significantly less overlap of age-effect voxels for the raw CNR (median ± interquartile range Dice coefficient = 0.956 ± 0.046) compared to the harmonized CNR (Dice coefficient = 0.973 ± 0.022) (Figure 4b).

Left–Right Asymmetries

In the raw CNR, significant laterality effects were found for the GE MR750 ($t_{123} = 3.74$), GE Premier ($t_{123} = -3.13$), and Siemens Prisma ($t_{123} = -2.68$), but not for the Philips Ingenia ($t_{123} = 0.41$, $P = 0.6836$) or Philips Elition ($t_{123} = 1.22$, $P = 0.2234$), with positive *t*-statistics indicating greater CNR in the left hemisphere than the right hemisphere. Additionally, significant differences in laterality effects were observed across scanners ($F_{4,123} = 8.15$). In the harmonized CNR, no significant laterality effects were found for any scanner (GE MR750: $t_{123} = -0.50$, $P = 0.6180$; GE Premier: $t_{123} = -0.49$, $P = 0.6267$; Siemens Prisma: $t_{123} = -0.84$, $P = 0.4012$; Philips Ingenia: $t_{123} = 0.23$, $P = 0.8222$; Philips Elition: $t_{123} = 0.43$, $P = 0.6709$) and laterality effects were not significantly different across scanners ($F_{4,123} = 0.23$, $P = 0.9209$).

Discussion

We have presented a method for harmonizing NM-MRI data across sites and scanners to remove nonbiological variability due to factors such as hardware and software differences. In addition to effectively removing nonbiological variability, the harmonization method maintained biologically relevant variability (here, age effects) while increasing both reproducibility and statistical power.

As previously seen when using ComBat to harmonize MRI measures of cortical thickness,[17] we found that ComBat successfully removes systematic biases associated with scanner across multiple sites in which acquisition protocols were not fully harmonized. Of note, we observed significant differences in SN–VTA CNR values between two GE scanners (MR750 and Signa Premier) both at the same institute (Site 1), but the MR750 data were acquired with a slice thickness of 3 mm while the Signa Premier data were acquired with a slice thickness of 1.5 mm. Furthermore, data from Sites 1 and 2 were acquired with slices oriented parallel to the AC-PC line while data from Site 3 were acquired with slices oriented perpendicular to the floor of the fourth ventricle, suggesting

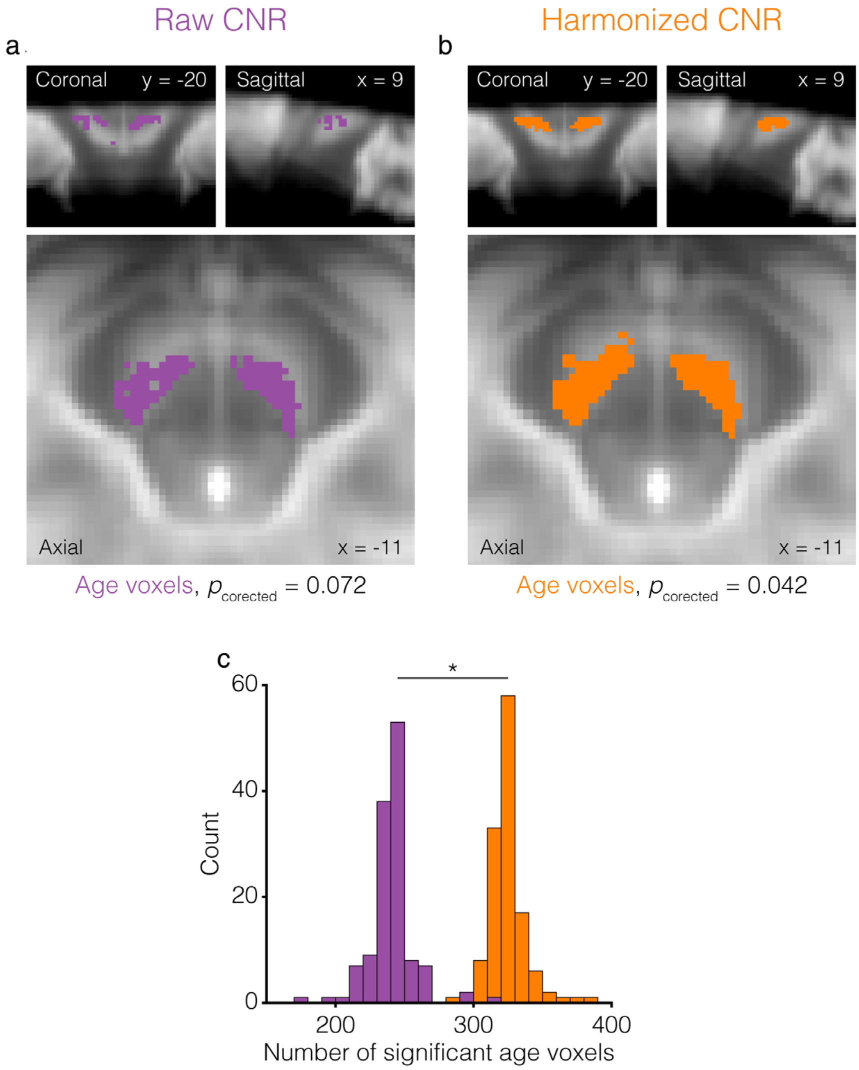

FIGURE 3: Effects of ComBat harmonization of neuromelanin-sensitive MRI (NM-MRI) substantia nigra–ventral tegmental area (SN–VTA) contrast-to-noise ratio (CNR) on statistical power. Maps of SN-VTA voxels where the raw (a) and harmonized (b) NM-MRI CNR were positively correlated with age (purple, thresholded at $P < 0.05$, one-sided, voxel level) overlaid on the average NM-MRI CNR map from all subjects (NM-MRI template). Histograms showing the number of age-effect voxels over each of the 128 leave-one-out folds for the raw and harmonized CNR (c). The asterisk (*) denotes $P < 0.05$ for the permutation test comparing the mean number of age-effect voxels from the raw and harmonized CNR to the null distribution.

that the proposed NM-MRI harmonization method can account for variability associated with slice orientation. The ability of ComBat to harmonize these data suggests that the "batch" effects can envelop all nonbiological sources of variability and alleviates the need to have perfectly matched acquisition protocols, which may not be feasible due to software and hardware limitations. It is possible that the benefits from harmonization could be reduced in studies where acquisition protocols are matched, but a previous study employing ComBat harmonization on such a dataset showed it to remain beneficial.[23]

Of almost equal importance to removing nonbiological variability from NM-MRI data across sites, is maintaining biologically meaningful variability. A harmonization approach to facilitate multisite studies focused on biomarker development must maintain the relevant biological variability. In the case of an NM-MRI biomarker for diagnosis of Parkinson's disease, schizophrenia or any other neuropsychiatric condition, this would typically be the effect of diagnosis. Other biomarkers may instead focus on prediction of treatment response or other clinically relevant outcomes. Here, we provided a proof-of-concept demonstration that ComBat maintains biologically meaningful variability in NM-MRI by showing that it preserves variability associated with age among healthy individuals. This is similar to previous ComBat harmonization studies which successfully maintained biologically meaningful variability in structural and diffusion MRI.[17,21] Future studies should verify that biologically meaningful variability associated with diagnosis, treatment response, symptom severity, and others, is maintained in

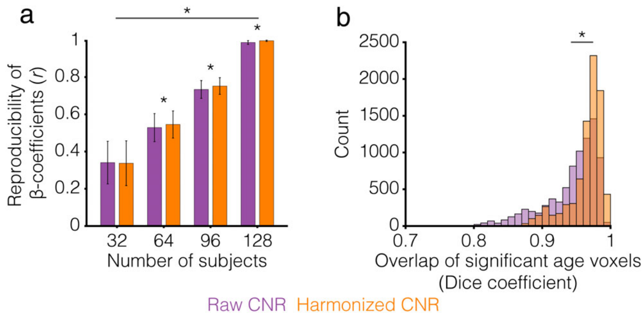

FIGURE 4: Effects of ComBat harmonization of neuromelanin-sensitive MRI (NM-MRI) substantia nigra–ventral tegmental area (SN–VTA) contrast-to-noise ratio (CNR) on the reproducibility of spatial patterns for predicting age. (a) A bar graph showing the mean (±SD) reproducibility of β-coefficients of 10-fold cross-validated linear SVR predicting age from 1000 random 10-fold splits of the data (n = 128) or 1000 random subsets of subjects (n = 32, 64, and 96). Asterisks (*) denote *P* < 0.05 for a two-way ANOVA comparing the mean correlation of β-coefficients between the raw and harmonized CNR for different sample sizes. (b) Histogram showing the overlap of age-effect voxels over each of the 128 leave-one-out folds for the raw and harmonized CNR. The asterisk (*) denotes *P* < 0.05 for the permutation test comparing the difference in median Dice coefficient for the raw and harmonized CNR to the null distribution.

clinical populations, although we would expect for this to be the case based on our current data and previous studies using ComBat.

We used SVM to successfully classify MRI scanner from SN–VTA CNR in the raw CNR and SVR to successfully predict age from SN–VTA CNR in both the raw and harmonized CNR. MVPA[37] has been widely applied in other neuroimaging modalities, such as functional MRI, and this work has led to the successful development of biomarkers of pain[37] and depression subtypes[38] among others. A potential benefit of MVPA analyses would be the ability to detect topographical patterns of effects without circularity or the need for repeated tests that are required for standard univariate voxelwise analyses. Future work should investigate the usefulness of MVPA for the development of clinically useful NM-MRI biomarkers.

Our finding of significantly fewer age-effect voxels in the raw CNR compared to the harmonized CNR suggests that the *effective* statistical power is increased in the harmonized CNR when using standard analysis procedures. This result should be interpreted with caution as it could be partially due to the difference in degrees of freedom in the regressions for the raw and harmonized CNR, the latter of which does not include site covariates. We use the term *effective* statistical power here, since, although the degrees of freedom differ, these are the two statistical tests that would be performed in actuality, and are *effectively* on equal experimental footing.

In this study we harmonized the CNR within the SN–VTA as opposed to the raw NM-MRI signal intensity in the SN–VTA or the raw signal for the entire image. This is in line with previous studies that harmonized the outcome measure (eg, regional brain volumes) instead of the raw MRI signal.[17,19–24] We also performed harmonization on the raw MRI signal (i.e. NM-MRI signal within the CC and the SN–VTA separately prior to calculating CNR) but this approach failed to remove the nonbiological variability (data not shown).

Previous work using NM-MRI observed left–right asymmetries in SN–VTA CNR, with healthy individuals showing higher CNR in the left hemisphere.[39] Here, we investigated if these laterality effects were present in each of the five MRI scanners and if the effects were consistent across the scanners. We showed that in the raw CNR, laterality effects were present in three out of the five MRI scanners; more importantly, laterality effects across these three scanners were inconsistent—one showed greater CNR in the left hemisphere, while the other two showed greater CNR in the right hemisphere—suggesting that left–right asymmetries are not caused by biologically relevant mechanisms.

Limitations

Due to the inclusion of only subjects without neurological or psychiatric disorders, the results of this study might not be generalizable to certain patient populations. This is particularly important for patient populations where differences in NM-MRI signal are known to occur as in Parkinson's disease.[8,39] Theoretically, any systematic differences between a patient population and controls should be maintained by ComBat (as was observed here with age), but this needs to be empirically demonstrated. Additionally, we did not present an exhaustive evaluation or comparison of methods to harmonize NM-MRI data but instead focused on a method previously shown to perform well in other MRI applications. As such, while ComBat achieved the goals of removing nonbiological variability and maintaining biologically meaningful variability, other harmonization methods may achieve better performance. NM-MRI has also been applied to the locus coeruleus (LC) to image the integrity of the noradrenergic system in health and neuropsychiatric disease.[40] Our study was limited by data using a field-of-view that did not provide full coverage for the LC, and we were thus unable to harmonize LC CNR. Given our current findings, we expect a similar harmonization approach to be beneficial for LC NM-MRI, but a study designed to test this is required. Lastly, we did not acquire neurocognitive tests on out participants, which limited our ability to investigate differences in cognitive characteristics across subsamples. If present, differences in SN–VTA CNR across scanners caused by neurocognitive differences would have been considered as nonbiological variability and removed during the harmonization process. Future studies should investigate the impact of cognitive characteristics and determine if the proposed harmonization approach maintains biological variability associated with them.

Conclusion

We have presented a method using ComBat to harmonize NM-MRI data that effectively removed nonbiological variability while maintaining biologically relevant variability (associated with age) and produced modest improvements in statistical power and reproducibility. Our results suggest that harmonization is unlikely to be harmful or obscure the biological effects under investigation, at least as long as they are identified a priori. This approach paves the way for combining NM-MRI data across sites and scanners, facilitating large-scale multisite studies to develop NM-MRI-based biomarkers.

Acknowledgments

J.J.W. received support from the NIMH (K23-MH115291); A.A.D. received support from the NIMH (R01-MH109635); E.v.d.G. received support from the ZonMw (Veni grant 91618075); G.H. received support from the NIMH (R01-MH114965, R01-MH117323).

References

1. Abi-Dargham A, Horga G. The search for imaging biomarkers in psychiatric disorders. Nat Med 2016;22:1248-1255.
2. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372(9):793-795.
3. Insel TR. The NIMH research domain criteria (RDoC) project: Precision medicine for psychiatry. Am J Psychiatry 2014;171(4):395-397.
4. Singh I, Rose N. Biomarkers in psychiatry. Nature 2009;460(7252): 202-207.
5. Sulzer D, Cassidy C, Horga G, et al. Neuromelanin detection by magnetic resonance imaging (MRI) and its promise as a biomarker for Parkinson's disease. NPJ Parkinsons Dis 2018;4(1):11.
6. Trujillo P, Summers PE, Ferrari E, et al. Contrast mechanisms associated with neuromelanin-MRI. Magn Reson Med 2017;78(5):1790-1800.
7. Zecca L, Bellei C, Costi P, et al. New melanic pigments in the human brain that accumulate in aging and block environmental toxic metals. Proc Natl Acad Sci 2008;105(45):17567-17572.
8. Cassidy CM, Zucca FA, Girgis RR, et al. Neuromelanin-sensitive MRI as a noninvasive proxy measure of dopamine function in the human brain. Proc Natl Acad Sci 2019;116(11):5108-5117.
9. Watanabe Y, Tanaka H, Tsukabe A, et al. Neuromelanin magnetic resonance imaging reveals increased dopaminergic neuron activity in the substantia nigra of patients with schizophrenia. PLoS One 2014;9(8): e104619.
10. Langley J, Huddleston DE, Liu CJ, Hu X. Reproducibility of locus coeruleus and substantia nigra imaging with neuromelanin sensitive MRI. Magn Reson Mater Phys Biol Med 2017;30(2):121-125.
11. Pluijm M, Cassidy C, Zandstra M, et al. Reliability and reproducibility of neuromelanin-sensitive imaging of the substantia nigra: A comparison of three different sequences. J Magn Reson Imaging 2021;53(3):712–721. http://doi.org/10.1002/jmri.27384.
12. Wengler K, He X, Abi-Dargham A, Horga G. Reproducibility assessment of neuromelanin-sensitive magnetic resonance imaging protocols for region-of-interest and voxelwise analyses. Neuroimage 2020;208: 116457.
13. Han X, Jovicich J, Salat D, et al. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. Neuroimage 2006;32(1): 180-194.
14. Takao H, Hayashi N, Ohtomo K. Effect of scanner in longitudinal studies of brain volume changes. J Magn Reson Imaging 2011;34(2): 438-444.
15. Zhu T, Hu R, Qiu X, et al. Quantification of accuracy and precision of multi-center DTI measurements: A diffusion phantom and human brain study. Neuroimage 2011;56(3):1398-1411.
16. Rao A, Monteiro JM, Mourao-Miranda J, Alzheimer's Disease Initiative. Predictive modelling using neuroimaging data in the presence of confounds. Neuroimage 2017;150:23-49.
17. Fortin J-P, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 2018;167: 104-120.
18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8(1): 118-127.
19. Radua J, Vieta E, Shinohara R, et al. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 2020;218:116956.
20. Pomponio R, Erus G, Habes M, et al. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage 2020;208:116450.
21. Fortin J-P, Parker D, Tunç B, et al. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 2017;161:149-170.
22. Villalón-Reina JE, Martínez K, Qu X, et al. Altered white matter microstructure in 22q11. 2 deletion syndrome: A multisite diffusion tensor imaging study. Mol Psychiatry 2020;25(11):2818-2831.
23. Nielson DM, Pereira F, Zheng CY, et al. Detecting and harmonizing scanner differences in the ABCD study-annual release 1.0. BioRxiv 2018;309260.
24. Yu M, Linn KA, Cook PA, et al. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum Brain Mapp 2018;39(11):4213-4227.
25. Zecca L, Fariello R, Riederer P, Sulzer D, Gatti A, Tampellini D. The absolute concentration of nigral neuromelanin, assayed by a new sensitive method, increases throughout the life and is dramatically decreased in Parkinson's disease. FEBS Lett 2002;510(3):216-220.
26. Zecca L, Stroppolo A, Gatti A, et al. The role of iron and copper molecules in the neuronal vulnerability of locus coeruleus and substantia nigra during aging. Proc Natl Acad Sci 2004;101(26):9843-9848.
27. Zucca FA, Vanna R, Cupaioli FA, et al. Neuromelanin organelles are specialized autolysosomes that accumulate undegraded proteins and lipids in aging human brain and are likely involved in Parkinson's disease. NPJ Parkinsons Dis 2018;4(1):17.
28. Zecca L, Gallorini M, Schünemann V, et al. Iron, neuromelanin and ferritin content in the substantia nigra of normal subjects at different ages: Consequences for iron storage and neurodegenerative processes. J Neurochem 2001;76(6):1766-1773.
29. Xing Y, Sapuan A, Dineen RA, Auer DP. Life span pigmentation changes of the substantia nigra detected by neuromelanin-sensitive MRI. Mov Disord 2018;33(11):1792-1799.
30. Wengler K, Ashinoff BK, Pueraro E, Cassidy CM, Horga G, Rutherford BR. Association between neuromelanin-sensitive MRI signal and psychomotor slowing in late-life depression. Neuropsychopharmacology 2020;1-7.
31. Fürnkranz J. Round robin classification. J Mach Learn Res 2002;2(Mar): 721-747.

32. King G, Zeng L. Logistic regression in rare events data. Political Anal 2001;9(2):137-163.

33. Snoek L, Miletić S, Scholte HS. How to control for confounds in decoding analyses of neuroimaging data. Neuroimage 2019;184: 741-760.

34. Zhang Y, Kimberg DY, Coslett HB, Schwartz MF, Wang Z. Multivariate lesion-symptom mapping using support vector regression. Hum Brain Mapp 2014;35(12):5861-5876.

35. Cassidy CM, Carpenter KM, Konova AB, et al. Evidence for dopamine abnormalities in the substantia nigra in cocaine addiction revealed by neuromelanin-sensitive MRI. Am J Psychiatry 2020;177(11):1038-1047.

36. Meng X-L, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. Psychol Bull 1992;111(1):172, 175.

37. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo C-W, Kross E. An fMRI-based neurologic signature of physical pain. N Engl J Med 2013; 368(15):1388-1397.

38. Drysdale AT, Grosenick L, Downar J, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 2017;23(1):28-38.

39. Isaias IU, Trujillo P, Summers P, et al. Neuromelanin imaging and dopaminergic loss in Parkinson's disease. Front Aging Neurosci 2016; 8:196.

40. Betts MJ, Kirilina E, Otaduy MC, et al. Locus coeruleus imaging as a biomarker for noradrenergic dysfunction in neurodegenerative diseases. Brain 2019;142(9):2558-2571.