A systematic comparison of VBM pipelines and their application to age prediction

Georgios Antonopoulos[1, 2], Shammi More[1, 2], Federico Raimondo[1, 2], Simon B. Eickhoff[1, 2], Felix Hoffstaedter[1, 2], and Kaustubh R. Patil[1, 2]

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[2]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany

Abstract

Voxel-based morphometry (VBM) analysis is commonly used for localized quantification of gray matter volume (GMV). Several alternatives exist to implement a VBM pipeline. However, how these alternatives compare and their utility in applications, such as the estimation of aging effects, remain largely unclear. This leaves researchers wondering which VBM pipeline they should use for their project. In this study, we took a user-centric perspective and systematically compared five VBM pipelines, together with registration to either a general or a study-specific template, utilizing three large datasets (n>500 each). Considering the known effect of aging on GMV, we first compared the pipelines in their ability of individual-level age prediction and found markedly varied results. To examine whether these results arise from systematic differences between the pipelines, we classified them based on their GMVs, resulting in near-perfect accuracy. To gain deeper insights, we examined the impact of different VBM steps using the region-wise similarity between pipelines. The results revealed marked differences, largely driven by segmentation and registration steps. We observed large variability in subject-identification accuracies, highlighting the interpipeline differences in individual-level quantification of GMV. As a biologically meaningful criterion we correlated regional GMV with age. The results were in line with the age-prediction analysis, and two pipelines, CAT and the combination of fMRIPrep for tissue characterization with FSL for registration, reflected age information better.

1 Introduction

Analysis of brain structure has provided important insights regarding its organization in health and disease. T1-weighted (T1w) images obtained using magnetic resonance imaging (MRI) are commonly used for this purpose. However, raw T1w images cannot be compared directly due to their semiquantitative nature and inter- and intrasubject variability [1]. Volumetric analysis of T1w images using voxel-based morphometry (VBM) [2, 3] allows the investigation of the volumetric composition of brain tissues across subjects. It estimates tissue volume in each voxel and brings individual brains in a common reference space permitting comparison. VBM analysis has provided a plethora of valuable insights, for instance, in neurodegenerative diseases [4–8] and psychiatric disorders [9].

VBM has been successfully applied to study aging [10–12]. Recently, prediction of individuals' age based on VBM-derived information has proven to be a validated proxy for brain integrity and overall health [13–15], and promising for individualized clinical applications [14, 16–19]. Brain-age prediction is an important and widely studied topic that aims to estimate the trajectory of healthy brain aging [20, 21].

To estimate the GVM from T1w images, some specific steps must be performed. The main steps of a VBM pipeline are as follows: i) **Segmentation** creates probability maps where each voxel is assigned a probability of belonging to specific brain tissues, usually gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). **Brain extraction**, which is the process of removing the skull from an image and leaving only actual brain tissues and CSF, is also a segmentation process but in some cases is performed prior to segmentation of GM, WM and CSF.

ii) **Spatial registration/normalization** to a reference brain space is performed so that anatomical regions are aligned. The reference space can be either a general template (e.g., MNI-152) or a study-/data-specific template (henceforth referred to as data-template) [22–24]. Data-templates are mainly used when comparing healthy subjects to patients to avoid bias due to general templates constructed from healthy populations. Several ways exist to create a data-template, and they are often created to match a standard space, such as the MNI space. Most VBM pipelines come with a general template.

iii) **Modulation** of the normalized tissue estimates aims at preserving the original amounts of tissue after spatial registration. To do so, normalized images are adjusted by the amount of local volume changes.

Since the introduction of VBM in 1995 [2], several alternatives and a multitude of options for each of the steps have been proposed. Even though various VBM pipelines utilize the same steps, the order of the steps may vary, and each step might use a different algorithm with several configurable options. Moreover, the pipelines can use those steps in a different order or perform some of them simultaneously and/or iteratively. It is also possible to create hybrid pipelines by combining the steps from different tools. Furthermore, optional steps, for example, whether to create a data template or use a general template provided by a

tool, add to the already vast number of choices. Consequently, even if a user chooses an *off-the-shelf* VBM pipeline is not completely absolved of further choices. How the outputs of VBM pipelines compare and their utility in different applications remain poorly studied, which can lead to suboptimal choices [25–27].

Previous work comparing VBM pipelines indeed provides evidence for differences. A comprehensive comparison between Computational Anatomy Toolbox (CAT) [28] version 12.7, two FSL-based and a hybrid (still FSL [29] dependent) pipelines has shown that the choice of preprocessing pipeline has an impact both in age prediction and sex classification [30]. The same study showed that regions driving the results are pipeline dependent, while the choice of the templates used for registration, general or data-template, has little or no impact. FSL and SPM [31] yield different outcomes, especially for cortical regions [32]. A comparison focusing on registration and segmentation steps of SPM and FSL concluded that these preprocessing steps drive the regions identified in multiple amyotrophic lateral sclerosis [26]. Segmentation and registration as implemented in SPM8 newseg, SPM8 DARTEL [33], and FSLVBM were found to have substantial influence on GMV estimates and their relationship to age [34]. This study additionally concluded that pipelines with limited degrees of freedom for local deformations might overestimate between-group differences. Finally, the selection of tissue probability maps (TPMs) as priors for segmentation systematically impacts the segmentation outcome and, in turn, affects the statistical estimates [35]. The CAT12 VBM pipeline was found to perform better in the detection of volumetric alterations in temporal lobe epilepsy compared to the VBM8 toolbox [36, 37].

Several studies have investigated the effects of individual VBM steps and their parametrization. A comparison of 14 deformation algorithms used for registration found that SyN [38] from the Advance Normalization Toolkit (ANTs) [39] and DARTEL (CAT) were among those with the best performance, with SyN exhibiting the highest consistency across subjects [40] as well as being among the most robust to noise, partial volume effects and magnetic field inhomogeneities [41]. Segmentation algorithms from SPM, ANTs and FSL showed relatively small differences in controls, but significant differences appeared when comparing brains with atrophies, suggesting that the segmentation algorithm should be selected according to the brain characteristics of the study-population [42]. Dadar and colleagues compared six segmentation tools and confirmed significant differences between the tools as well as within-tool differences based on interscanner analysis [43]. For brain extraction, although FSL-BET has been reported to have low performance [42], it does not influence subsequent segmentation [44]. A comparison of SPM12, SPM8 and FreeSurfer5.3 [45] showed that SPM12 estimates of total intracranial volume (TIV) align better with manual segmentation [46]. SPM-based estimates in autism spectrum disorder and typically developing controls were closest to manual segmentation in terms of TIV, followed by FreeSurfer, while FSL appeared to underestimate TIV [47].

Taken together, different VBM pipelines produce different outcomes. The disagreement in VBM pipelines hinders precise localization and valid interpretation of tissue volume in the downstream analysis, e.g., atrophy in patients with multiple sclerosis [48–50]. To date, there

is no standard method to calculate GMV or guidelines on which implementation of VBM is appropriate for a study at hand, e.g., age prediction. Additionally, the interaction of different algorithms and parameters in each step of VBM for estimating GMV and their effect on age estimates across the adult life-span, has not been thoroughly investigated. Moreover, the utility of a data-template created from healthy subjects and how it compares with a general template, especially in cross-site studies, remains unanswered. Here, to fill this gap, utilizing three large datasets (each n>500), we compared and evaluated five VBM pipelines including two *off-the-shelf* workflows and three modularly constructed pipelines utilizing commonly used neuroimaging tools. Each pipeline was implemented in two versions, one using a general template and one using a data-template, resulting in a total of 10 VBM pipelines. To remain consistent with our user-centric approach and developer guidelines, we adopted the default parameters unless there were specific recommendations from the developers [51]. First, we investigated whether different VBM pipelines produce GMV estimates that lead to different results in machine-learning-based predictions of individuals' chronological age. We also calculated regional correlation to age, as GMV is known to decrease with age in healthy subjects. This extrinsic evaluation provides a more objective and utilitarian proxy for comparison [19, 20, 52, 53] and a criterion based on biological factors. Additionally, we showed that the pipelines indeed produce distinct patterns of GMV using machine-learning-based classification. Specifically, we address the following questions:

- How do the pipelines differ at the *region-* and the *subject-level*?
- What impact do *brain extraction*, *segmentation* and *registration* have on GMV?
- What is the effect of using a *data-template* compared to a *general template*?
- How do the pipeline outcomes compare in *univariate* and *multivariate* analyses?
- Which pipeline better reflects *brain aging* and performs best in *brain-age prediction*?

With this comprehensive and systematic comparative analysis of VBM pipelines, we aim to provide essential information and recommendations to researchers to help them select the VBM pipeline that best matches their research goals.

2 Materials and Methods

2.1 Datasets

We analyzed T1w images of healthy individuals from three large datasets covering the adult lifespan,

eNKI [54]: population based sample of n=953 subjects, of which 573 had no psychiatric or neurological disorders or medication at the time of the scan (48.1±17.2 years, 630 female). CamCAN [55, 56]: n=634 aging individuals without serious psychiatric conditions or cognitive impairment (54.8±=18.4 years, 320 female). IXI [57]: multisite sample of n=582 normal and healthy subjects (49.4±16.7 years, 324 female). (Table S.1 in Supplementary Material)

2.2 Pipelines

CAT [28], a popularly used off-the-shelf VBM tool, is a successor of the first VBM pipeline implemented in SPM [3]. Here, we used the latest version CAT12.8 (r1813). Several general-purpose neuroimaging tools also provide functionality that can be used to create VBM pipelines. FSLVBM [58] uses tools from FSL [29] and is also widely used. ANTs [39] provides broad image processing and image analysis functionality, including all functions needed to perform VBM. Hybrid VBM pipelines that combine the functionality of different tools can be constructed, e.g., using fMRIPrep [59], which performs brain extraction using ANTs and then performs the rest of the steps using FSL.

We devised five VBM pipelines following the recommended steps and settings in the literature [39]: ANTs, ANTs-FSL, fMRIPrep-FSL, FSLVBM, and CAT. These pipelines were selected to reflect the choices that are common practice and easy to use. We used each pipeline with a standard template (the default templates for CAT and FSLVBM) irrespective of the dataset (general template) and with a dataset-specific template that was created and used for registration (data-template). Together, this resulted in ten pipelines.

2.2.1 ANTs

We used ANTs version 2.2.0. First, each scan was corrected using the N4 bias field correction [60] and then segmented to select intracranial tissues using Atropos-based brain extraction [61]. Next, Atropos segmentation initialized with K-means was applied to segment the images into GM, WM and CSF. The GM-map images were registered to a template (general or data-specific) using a sequence of transformations. First, rigid body and affine transformations were applied, followed by a nonlinear BsplineSyN transform with the parameters set as in [62]. The Jacobian matrix from the spatial transformation was used to modulate the segmented GM. Data-specific templates were created using the ANTs build template method with default values. To create the template images, the transformations were averaged and used iteratively [39, 63]. To keep the template shape stable over multiple iterations of template building, the inverse average warp was calculated and applied to the template image.

To facilitate the analysis, the data-template process was initialized using a general MNI template. Therefore, the final data-template was also in the MNI space. For all processes

requiring tissue masks and templates as well as for the registration to MNI, we used the ICBM 152 Non-linear Asymmetrical template version 2009a and corresponding tissue probability maps [64, 65].

2.2.2 FSLVBM

We used FSL version 6.0. The images were prepared by automatically reorienting and then cropping part of the neck and lower head. Then, BET was used to extract the intracranial part of the brain, which was then segmented into GM, WM and CSF using FAST. Data-specific templates were created following FSLVBM's process utilizing all GM images from a given dataset. GM segmented images were affinely registered to the ICBM-152 GM template, concatenated and averaged. This averaged image was then flipped along the x-axis, and the two mirror images were then reaveraged to obtain a first-pass, study-specific $affine$ GM template. Second, GM images were reregistered to this $affine$ GM template using nonlinear registration, averaged and flipped along the x-axis. Both mirror images were then averaged to create the final symmetric, study-specific, $non-linear$ GM template. The resulting data-template was in the MNI space. The GM images were then nonlinearly registered to the template (either general or data-specific) and modulated. As the general template, we used the FSL-provided template (see Table 1).

2.2.3 fMRIPrep-FSL

The reportedly poor quality of BET in brain extraction might lead to spurious results [42]; thus, we decided to test a pipeline that uses a better brain extraction as provided by ANTs followed by FSL for the rest of VBM processing. As fMRIPrep has been well validated and is gaining popularity, we chose to use the output of the fMRIPrep's structural processing. In this hybrid pipeline for image preparation and segmentation, we used fMRIPrep version stable 20.0.6 [59], which uses ANTs version 2.1.0. Each T1w volume was corrected for intensity nonuniformity (INU) using N4BiasFieldCorrection [60] and skull-stripped using 'antsBrainExtraction.sh' (using the OASIS template). Brain tissue segmentation into CSF, WM and GM was then performed using FSL FAST [66] (as used by the fMRIPrep FSL v5.0.9). This FAST parametrization diverges from the one in FSLVBM in the following parameters: (i) the Markov random field (MRF) beta value for the main segmentation phase was set to H=0.2, while the default value in FSLVBM was 0.1, and (ii) the MRF beta value for mixeltype was R=0.2, while the default in FSLVBM was 0.3. Template creation, spatial normalization, and modulation were identical to the FSLVBM pipeline.

2.2.4 ANTs-FSL

The exact same processing, as mentioned above in the ANTs pipeline, was used to prepare the images, correct bias field noise, perform brain extraction and finally perform tissue segmentation using ANTs' Atropos. The creation of a data-specific template, registration and modulation were implemented as in the FSLVBM pipeline. Note that the difference between this pipeline and the fMRIPrep-FSL pipeline is the tissue segmentation tool used.

2.2.5 CAT

CAT12.8 was used based on SPM12 (v7771) using MATLAB (R2017b) and compiled for containerization in Singularity (2.6.1). CAT provides a complete VBM pipeline including denoising with spatial-adaptive nonlocal means, bias-correction, skull-stripping, and linear and nonlinear spatial registration. Images are segmented by an adaptive maximum a-posteriori approach [67] with partial volume model [68]. For nonlinear transformation, the geodesic shooting algorithm [69] is used. As the default template, an IXI-based template transformed to MNI152NLin2009cAsym is provided. For the data-template, initially, all structural T1 images are segmented into GM, WM, and CSF and spatially coregistered to the MNI standard template using affine registration. The affine tissue segments were used to create the new sample-specific geodesic shooting template that consists of four iterative nonlinear normalization steps.

Table 1 summarizes the VBM steps of each pipeline we utilized in our analyses.

Pipeline	Skull stripping	Segmentation	Template (general/data-specific)	Registration/ Modulation
ANTs	ANTs Brain Extraction	Atropos	ICBM MNI152Nlin2009a	ANTsRegistration
			AntsBuildtemplate	
ANTs-FSL	ANTs Brain Extraction	Atropos	ICBM MNI152Nlin6th generation	FNIRT
			fslvbm_2_template	
fMRIPrep-FSL	ANTs Brain Extraction	FAST	ICBM MNI152Nlin6th generation	FNIRT
			fslvbm_2_template	
FSLVBM	BET	FAST	ICBM MNI152Nlin6th generation	FNIRT
			fslvbm_2_template	
CAT	CAT	CAT	ICBM MNI152Nlin2009c based	CAT
			CAT	

Table 1: Software/algorithm used for the main VBM steps in our analysis pipelines.

2.3 Parcellation scheme and quality control

To decrease the dimensionality of the data and thereby facilitate informative comparison and the use of machine-learning approaches, we extracted region-level averages. However, to preserve good spatial resolution, we selected a high granularity parcellation scheme. A combination of three atlases covering the whole brain and together constituting 1073 regions of interest (ROIs) was used: 1000 cortical regions from the Schaefer atlas [70], 36 subcortical regions from the Brainnetome Atlas [71] and 37 cerebellar regions [72]. Regional GMV values were calculated as the average of nonzero voxels within each region.

ANTs segmentation (Atropos), which was initiated with k-means, in some cases returned tissues in a different order, resulting in selecting the WM instead of the GM for further analysis. Therefore, we employed the following quality check to ensure that selected tissue represented GM. First, we discarded individuals who had a ratio of the mean of GM voxels over the mean of WM and CSF voxels of less than 1.5. Furthermore, images that were close to the 1.5 threshold as well as randomly sampled images were visually inspected for quality of segmentation. Because developing a thorough quality check or tackling this issue

inside Atropos is out of the scope of this work, the threshold for the ratio of mean GM over WM and CSF was experimentally identified. Although CAT has an internal quality control method, for consistency, we applied our test to all pipelines. We retained only subjects who passed the quality checks across all the pipelines.

2.4 Age prediction

We performed machine-learning-based analysis to predict the age of each subject using regional GMVs from each pipeline as features. We chose this as a suitable test given that age is reliably associated with GMV [19, 20, 52, 53] and because of the increasing importance of brain-age as a proxy for overall brain health [52, 73–75]. All features were standardized by removing the mean and scaling to unit variance in a cross-validation (CV)-consistent manner [76]. We utilized four machine-learning algorithms: relevance vector regression (RVR) [77], Gaussian process regression (GPR) [78], least absolute shrinkage and selection operator (LASSO) [79, 80], and kernel ridge regression (KRR) [81], in a nested 5-fold CV scheme repeated 5 times [82]. The age prediction performance was evaluated using the mean absolute error (MAE). To ensure that differences were not driven by factors other than the pipelines, we used the same data (subjects and regions) and models for each pipeline.

The evaluation was performed in two set ups, intradataset, and interdataset. In the interdataset evaluation, the models were trained using two datasets and then used to predict the third hold-out dataset. This analysis was performed for each pipeline separately.

2.5 Classification of pipelines

To confirm the existence of systematic differences in the outcomes of the pipelines, we performed machine-learning-based predictive analysis based on the multivariate patterns of regional GMV. The idea behind this analysis is that if a model can classify the pipeline producing a GMV image with a high accuracy, that would indicate that the model learned systematic differences between the VBM pipelines. We performed 10-class classification with subjects' regional GMVs as features and the pipelines as class labels. The features were standardized by removing the mean and scaling to unit variance in a CV-consistent manner [76] in two ways: i) within each feature and ii) within each subject. The former is standard preprocessing, while we implemented the latter to guard against trivial biases such as magnitude shifts. We used a linear support vector machine (SVM) with the default cost parameter of C=1 in a 5-fold CV scheme repeated 5 times.

2.6 Individual-level identification

We examined the within-subject consistency of GMV patterns when processed by different pipelines. To do so, we identified subjects across pipelines using a nearest neighbor search. Using each pipeline as a reference (query), we tried to match each subject with all the subjects of each other pipeline (database). As an identification metric, we used Pearson's correlation between two subjects' regional GMVs [83, 84]. Each subject was matched with the subject from another pipeline with the highest correlation coefficient. The identification performance between two pipelines was calculated using the differential identifiability (Idiff)

metric [84].

2.7 Region-level comparison

To obtain a better understanding of regions driving the differences between pipelines, we assessed the similarity in regional GMV estimates from different pipelines using univariate statistical analysis. These analyses were performed for subjects from all datasets combined as well as separately for each dataset. We estimated similarity in regional GMVs across subjects using Pearson's correlation coefficient for all possible pipeline pairs (in total 45). To investigate whether the size of parcels affects the regional similarities, we calculated for each ROI the median of correlation coefficients across the pairs of pipelines and correlated it with the number of voxels per region (see Figure S.6 in the Supplementary Material).

For all arithmetic operations on Pearson's r values, first Fisher's z transform was applied, and then the result was transformed back to Pearson's r value.

2.8 Extrinsic evaluation of similarity between pipelines

The pipeline comparisons described above are intrinsic in nature. Thus, although they provide important information regarding differences between the pipelines, they do not provide information regarding the correctness of the pipelines in estimating the GMV. Such a correctness assessment, although desirable, cannot currently be achieved due to a lack of ground truth data. Instead, we compared the pipelines based on their utility in capturing age-related information.

We first tested to what degree regional GMV estimates from each pipeline reflect subjects' age using univariate statistical analysis. To do so, we computed Pearson's r between the regional GMVs and subjects' ages for each pipeline separately. The resulting p values were corrected to control for the familywise error rate [85] due to multiple comparisons, again for all data combined as well as separately for each pipeline. We then performed an analysis of variance (ANOVA) to test whether the means of the correlation coefficients were significantly different.

Machine-learning-based analyses were performed using scikit-learn [86].

3 Results

3.1 Preprocessing and data-templates

For CAT and fMRIPrep, less than 0.4% of all subjects failed the preprocessing. For CAT, all outcomes passed our quality check. For FSLVBM, less than 2% of the subjects failed the QC. For fMRIPrep-FSL, there were slightly fewer subjects who failed QC than for FSLVBM. A considerable number of subjects failed ANTs segmentation (13% for eNKI, 5% for CamCAN and 12% for IXI). The QC results for the hybrid ANTs-FSL pipeline were similar to those of ANTs. The final number of subjects who qualified for further analyses was n=741 for eNKI, 593 for CamCAN and 418 for IXI (total n=1752).

The data-templates created by CAT and ANTs were sharper and more similar to general templates than those created by FSLVBM (templates are demonstrated in the Supplementary Material in Figures S.1, S.2, S.3).

3.2 VBM pipelines produce different results

3.2.1 Brain age prediction

We first performed individual-level prediction of chronological age using regional GMVs as features using four machine-learning algorithms (Figure 1). Within-dataset CV performance considerably varied among pipelines (Figure 1 (a)). The average performance across the learning algorithms and datasets was highest for the fMRIPrep-FSL general template ($MAE = 5.83$), followed by the FSLVBM general template ($MAE = 6.17$) and fMRIPrep-FSL data-template ($MAE = 6.18$). CAT with the data-template and with the general template showed similar performance of $MAE = 6.37$ and 6.39, respectively. The best average performance across datasets was achieved by the fMRIPrep-FSL general template with KRR ($MAE = 5.59$). ANTs performed the worst on average. All four learning algorithms generally showed similar performance for each pipeline (Supplementary Material Table S.2).

For cross-dataset predictions (Figure 1 (b)), the best performance averaged across datasets and models was again achieved by the fMRIPrep-FSL pipelines, with the data-template ($MAE = 6.21$) performing slightly better than the general template ($MAE = 6.26$) closely followed by CAT general template ($MAE = 6.45$). Here, the best overall predictions were again provided by the KRR algorithm. For the fMRIPrep-FSL data-template and general-template MAE was 6.06 and 6.13, respectively. For CAT, $MAE = 6.32$ and 6.42 with the general template and data-template, respectively. ANTs-FSL-derived GMVs performed the worst on average (Supplementary Material Table S.3).

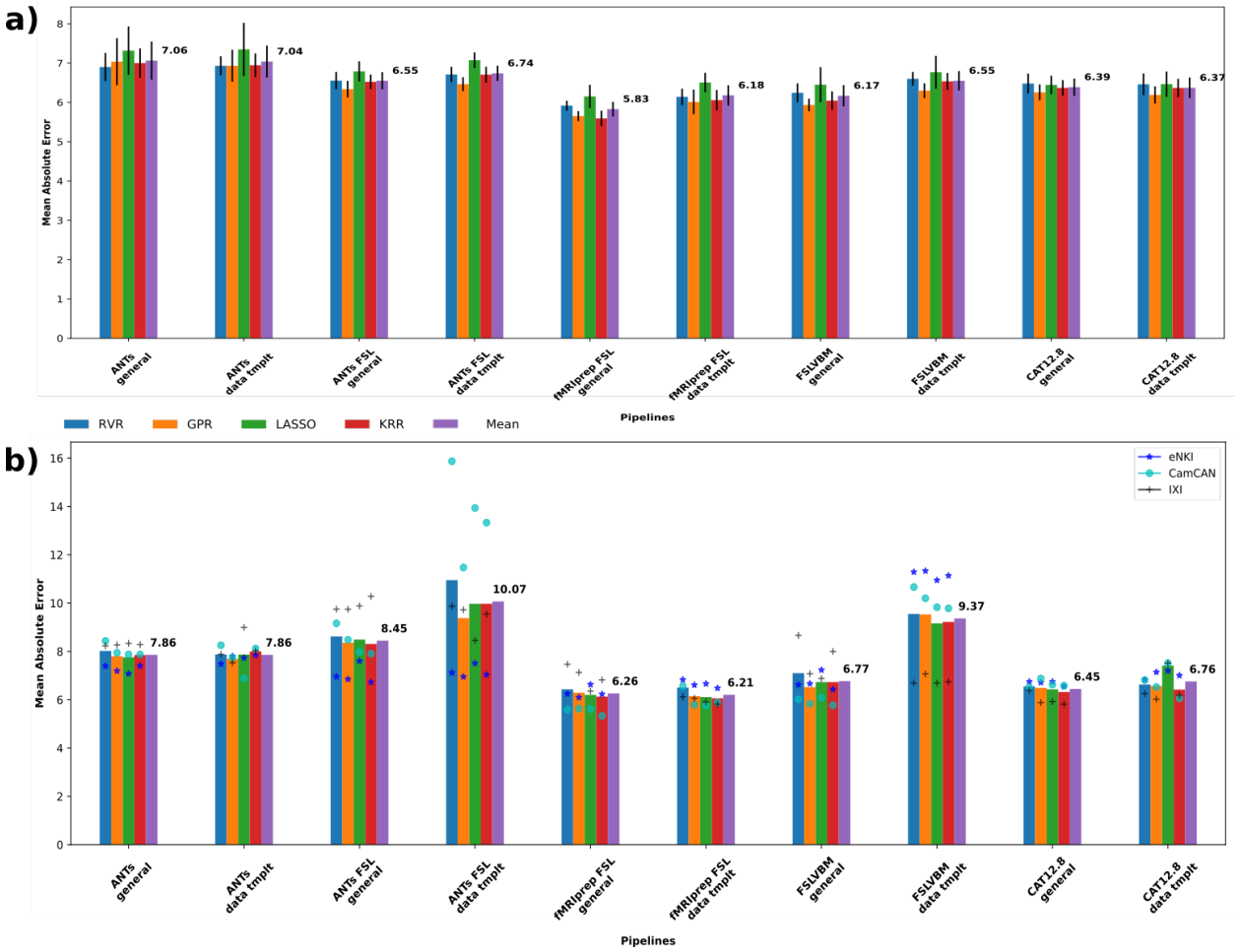

Figure 1: Age prediction for each pipeline. Blue, orange, green and red bars represent the averaged results of the three datasets per machine-learning algorithm, and the purple bars show the mean across models and datasets. a) Models trained and tested in the same dataset. Four models were tested using the three datasets in a nested K-fold cross-validation scheme. b) Age prediction for each pipeline when trained with two of the datasets and tested in the left-out one. Blue stars show the prediction performances on eNKI data, light blue circles the performances on CamCAN data, and black crosses on IXI data.

3.2.2 Machine-learning analysis confirms distinct GMV patterns

The machine-learning approach classified the pipelines with a near-perfect accuracy close to 100%. To rule out the possibility that this high accuracy was driven by systematic differences, that is, some pipelines over- or underestimating the GMV overall (which is indeed the case, see Supplementary Material Figure S.7), we performed an additional analysis where each subject's feature vector was z-scored independently, in effect removing the overall differences in GMV estimates. This analysis also resulted in high classification accuracy for

all the datasets, close to 100%. Detailed results are provided in the Supplementary Material (Figure S.4).

3.2.3 Identification shows individual-level differences

Pipelines differing only in the template showed high differential identifiability 43>Idiff>29. fMRIPrep-FSL and FSLVBM, both with data-template, had the highest Idiff= 45, followed by the two ANTs pipelines (Idiff= 43). The two CAT pipelines had the lowest mean Idiff values, with the data-template pipeline being the lowest. FSLVBM with data-template had the highest mean Idiff. Pipelines using FSL for registration and modulation, with a general template, had a mean Idiff= 33.7. The same pipelines with a data-template showed mean Idiff= 37.7. ANTs-FSL and fMRIPrep-FSL, when both using a general template had Idiff= 35 and when using a data-template Idiff= 34. Finally, ANTs and ANTs-FSL, which differ in registration (and modulation), had Idiff= 29 when both used general templates and Idiff= 30 for data-templates (Figure 2).

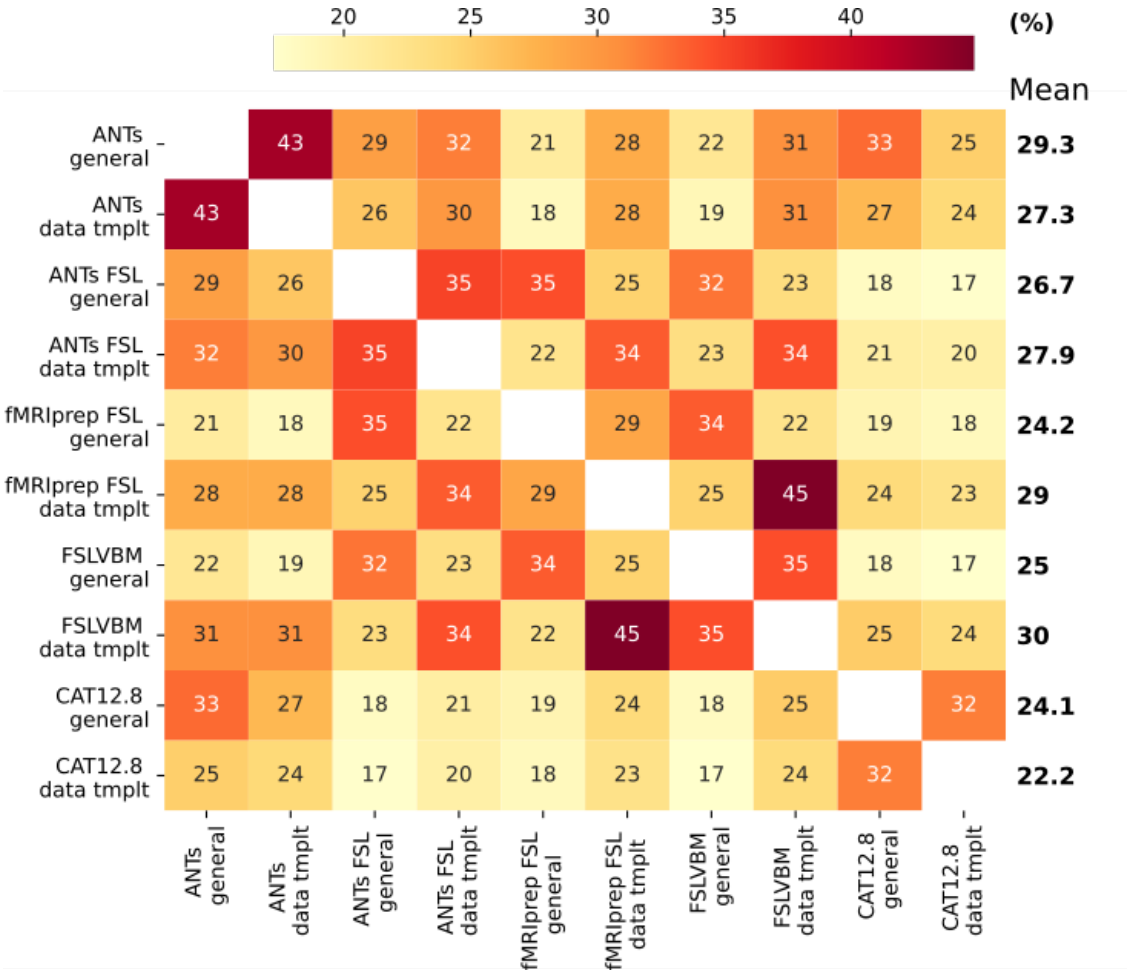

Figure 2: Identification performance in terms of differential identifiability. We used Pearson's coefficient to calculate similarity between subjects. The highest mean Idiff was found for FSLVBM data-template followed by ANTs general template. The two CAT pipelines showed the lowest mean Idiff values.

3.2.4 Univariate analysis and region-wise similarity

To better understand whether some VBM steps drive differences in the GMV estimates more than others, as well as to identify the regions showing significant differences, we performed several univariate statistical analyses. Some of the pipelines differ only in a single step; therefore, by examining the similarity between them, insightful conclusions can be extracted about the effect of this specific VBM step. We observed that the overall agreement between the pipelines, based on the median of the pairwise correlation values, varied across the regions, while most of the regions showed only low-to-moderate agreement (Figure 3). Only the regions close to the cingulum, temporal lobes and fusiform area showed relatively high agreement across the pipelines (median $r > 0.6$). Most of the subcortical regions showed low

agreement (median $r < 0.4$), except the caudate (median $r > 0.6$). In the cerebellum, all regions showed a median $r < 0.6$. Overall, these results indicate a low agreement across the pipelines.

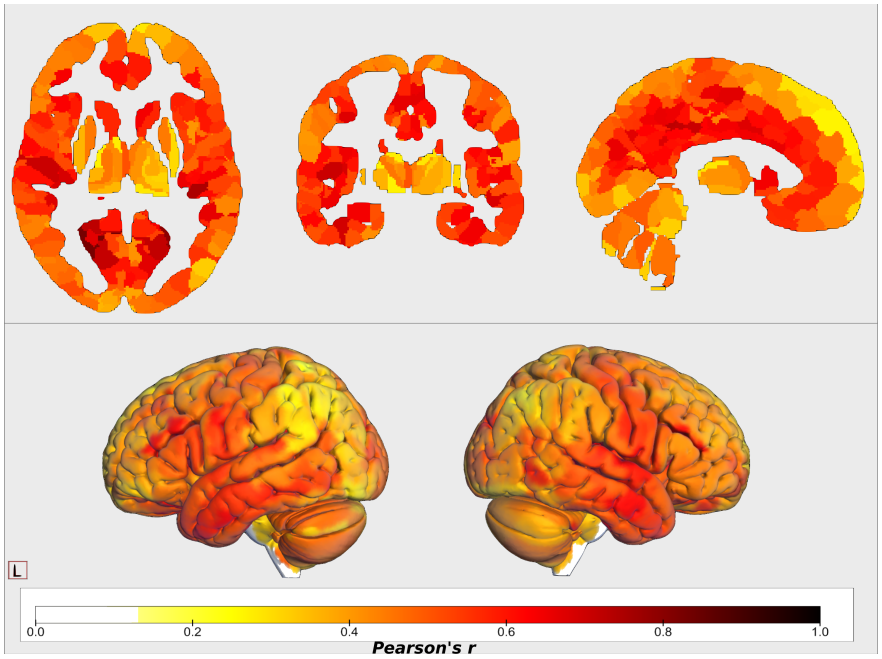

Figure 3: **Median values** of regional correlations calculated across subjects of all pairwise combinations of pipelines. The frontal lobe, subcortical regions and cerebellum showed lower similarity. First, correlations between regional GMVs across subjects were calculated for each pipeline pair. The median of these 45 values was then calculated as an overall agreement among the pipelines for each region.

The regionwise similarity between pairs of pipelines differed substantially. While ignoring pipeline pairs that differ only in the template (which are expected to be similar), maximum similarity was observed between fMRIPrep and FSLVBM both using a data-specific template (average $r = 0.76$), while the minimum similarity was between ANTs-FSL using the general template and CAT with both templates (average $r = 0.306$) (Figure 4).

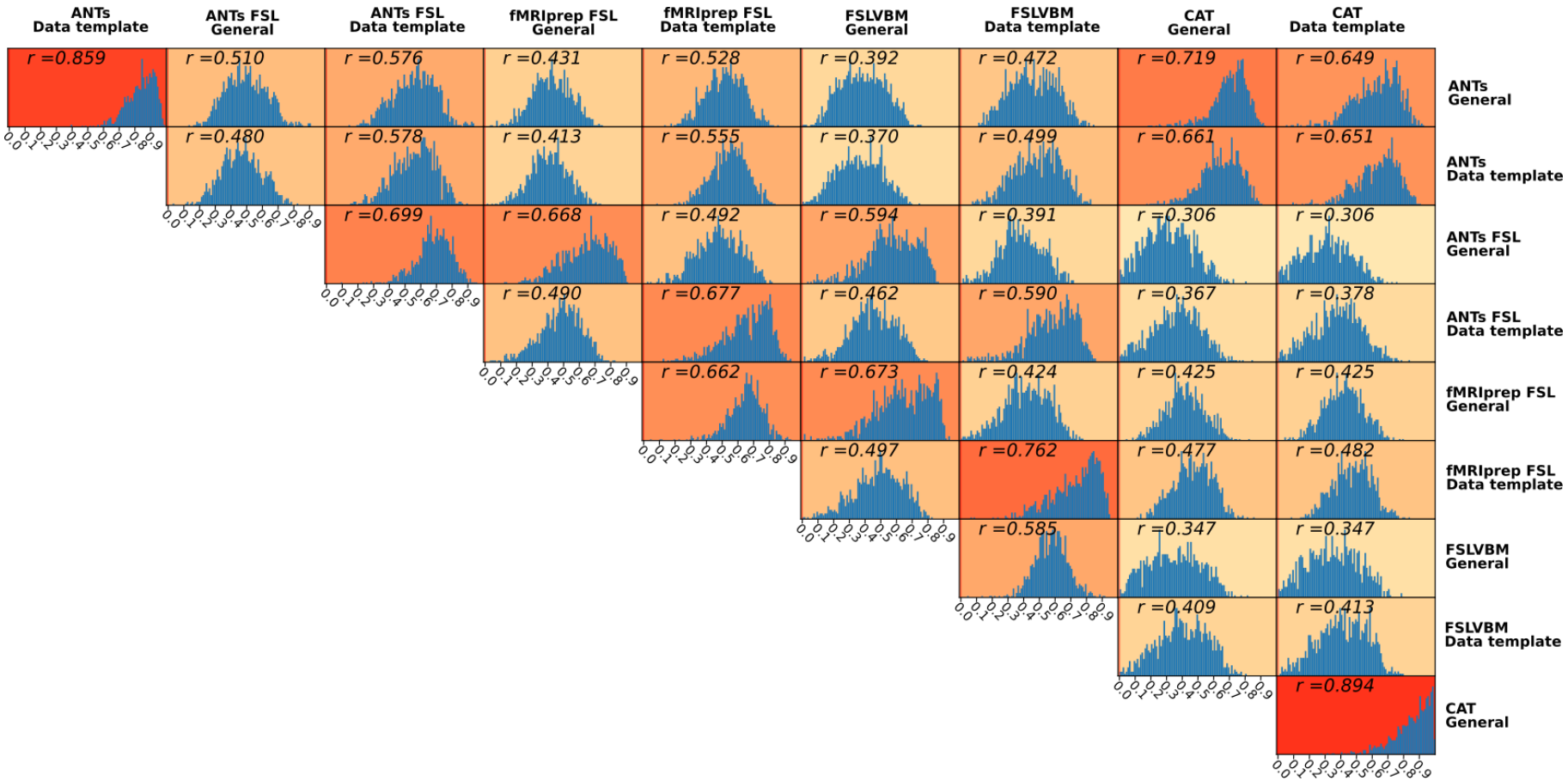

Figure 4: Histograms of regional interpipeline similarity for all pairs of pipelines. For each pair, we calculated Pearson's r coefficient for each region across all subjects. We used the Holm-Bonferroni method to correct for multiple comparisons. The histograms shown consist of those regions that survived the multiple comparison ($p < 0.05$).

3.2.5 Comparison between ANTs and CAT

High similarities were observed between the CAT and ANTs pipelines, despite differences in the steps, the order of the steps and the algorithms for each step. The highest similarity was observed when using the general templates (which themselves are different, as shown in Table 1) with $r = 0.72$ followed by $r = 0.66$ between the ANTs data-template and the CAT general template. A slightly lower similarity, of $r = 0.65$ was estimated when both pipelines used the data-templates as well as between the ANTs general template and the CAT data-template.

3.2.6 Effect of Registration, Segmentation, and Brain extraction

In the subsequent analyses, we compared pipelines differing in specific VBM steps to assess their specific impact.

Regionwise similarity between ANTs and ANTs-FSL that differed only in **registration** (and therefore in modulation) using the general template was moderate to low, average $r = 0.51$. When using data-specific templates, the similarity was higher for all data (0.58) but also for each of the three datasets (Figure 5 (a)).

ANTs-FSL and fMRIPrep-FSL share the same steps besides **segmentation**. When using the general template, the average region-wise similarity was $0,67$, and for the data-specific templates, the corresponding value was 0.68 (Figure 5 (b)).

FSLVBM and fMRIPrep-FSL differ in the **brain extraction** step. When both pipelines utilized the default FSL template, they had a similarity of 0.67. When the registration was performed using their respective data-specific template, the similarity increased to 0.76 (Figure 5 (c)).

Overall, similarities were higher when data-templates were used.

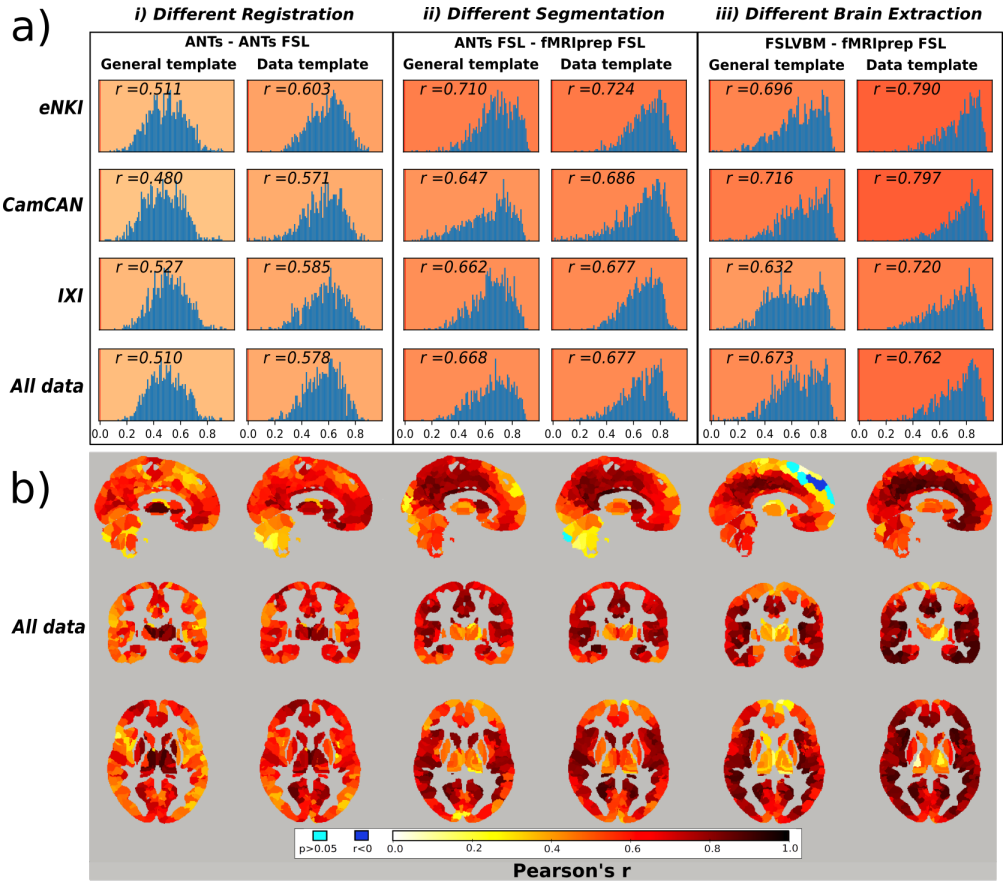

Figure 5: a) Histograms of regionwise correlation values between selected pairs of pipelines for all datasets. The r value represents the average correlation of all regions (that survived the Holm-Bonferroni correction) after transforming them to Fisher's z and then reverse transformed to r. The pipeline pairs are categorized according to the template they use in the registration step. *i*) Correlation between ANTs and ANTs-FSL, which differ only in the registration step. *ii*) ANTs compared to fMRIPrep-FSL. These two pipelines differ only in the segmentation step, as fMRIPrep utilizes FSL-based segmentation. Segmentation imposes fewer differences than registration, *iii*) FSLVBM and fMRIPrep-FSL only differ in the brain extraction step. This step has a similar effect to segmentation when a general template is used and higher similarity when a data-template is used. The data-specific template comparisons are also provided here for convenience reasons, although it should be noted that the template creation steps may differ for the pipeline pairs, resulting in the usage of different data-specific templates. b) Brain maps with regional similarity of selected pairs of pipelines calculated using all data. Similarity values are expressed in Pearson's r and were corrected using the Holm-Bonferroni method. Light blue represents regions without a significant association (p> 0.05) and blue represents regions with a negative correlation ($r < 0$). *i*) High similarity in subcortical areas and increased differences in cortical areas, especially when using a general template. *ii*) Different segmentations seem to have affected the cerebellum, subcortical areas and the posterior and anterior areas of the same axial level for both templates. *iii*) Brain extraction when using a general template caused more differences in the subcortical areas, superior frontal and the upper part of the cerebellum. It is noteworthy that negative values appear in the superior frontal lobe.

For ANTs compared to ANTs-FSL, the highest similarity values were in subcortical areas, and the lowest similarity values were in the ventrolateral and dorsolateral prefrontal cortices, especially when using a general template (Figure 5 $b(i)$). ANTs-FSL and fMRIPrep-FSL showed the least similarities in subcortical areas, the occipital lobe and prefrontal cortex (Figure 5 $b(ii)$). Finally, FSLVBM and fMRIPrep-FSL had the lowest similarity values in the subcortical areas, and the highest values were in the temporal lobes, medial prefrontal cortex and cingulate gyrus (Figure 5 $b(iii)$).

For each of the three datasets, similar figures separately with histograms of regional correlation values and Nifti files with all regional correlation values for the other pairs of pipelines can be found in the Supplementary Material.

3.2.7 Pipelines with the same registration

ANTs-FSL and FSLVBM, which share only the registration step, had a similarity of 0.59 for all data when using either the FSL default or the data-specific template. The similarity for the eNKI dataset was 0.65 for both templates; for the CamCAN dataset, the similarity was 0.60 for the general template and 0.63 for the data-template and 0.56 and 0.58 for IXI dataset, respectively.

3.2.8 General template versus data-specific template

The pipelines differing in the template, i.e., either general or a data-template, showed varying degrees of similarity (Table 2). The highest similarity was for CAT ($r > 0.9$), followed by ANTs (> 0.86) in all three datasets. The similarity was low to moderate for the three pipelines using FSL for registration and template creation steps (ANTs-FSL, FSLVBM, and fMRIPrep-FSL). Specifically, ANTs-FSL had a mean similarity across the three datasets of $r = 0.71$, fMRIPrep-FSL 0.66 and FSLVBM 0.59.

General template compared to the data-specific template

	ANTs	ANTs-FSL	fMRIPrep-FSL	FSLVBM	CAT
eNKI	0.879	0.718	0.646	0.573	0.908
CamCAN	0.876	0.694	0.678	0.596	0.910
IXI	0.864	0.713	0.668	0.605	0.916
Mean	0.873	0.708	0.664	0.591	0.911
All data	0.859	0.699	0.662	0.585	0.894

Table 2: The average values of regionwise correlation calculated across subjects for each pipeline when using a general template and a data-template. The *mean* across datasets is also presented, as well as the values from the same analysis performed with data from all datasets. It is noteworthy that when all data were combined, there was not an overall template created, but subjects were registered to the corresponding dataset template.

Univariate analysis is in line with the identification Idiff results. Pearson's r between the Idiff values and the regionwise correlations of pairs of pipelines was high, $r = 0.841, p < 0.05$ (more details in Supplementary Material Figure S.12).

3.3 Association with age

3.3.1 Correlation between age and regional GMV

We performed univariate analysis to assess how regional GMVs capture aging-related information. CAT showed the highest average correlation magnitude between regional GMVs and age irrespective of the template used for all datasets, followed by fMRIPrep-FSL with the general template. For CAT, the mean correlation across datasets was $r = -0.410$ and -0.406 with a general template and data-specific template, respectively (Table 3). The distribution of regional GMV-age correlation values was more narrowly distributed for CAT and ANTs, while they were more broadly distributed for pipelines using FSL (Figure 6 (a)). Overall, the regional GMV-age correlation was markedly different between the pipelines (Figure 6).

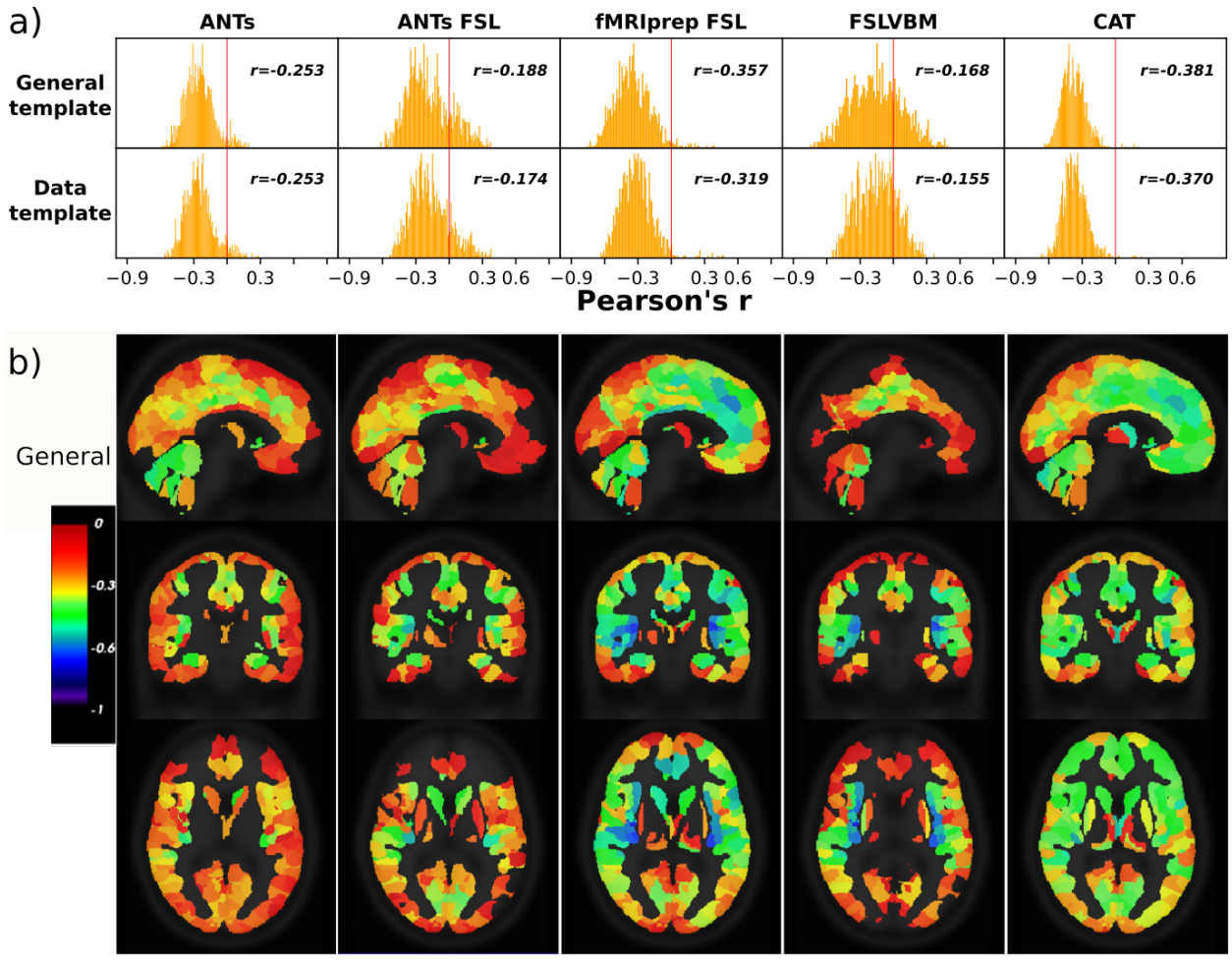

Figure 6: Correlation between regional GMV and age across subjects for the eNKI dataset. CAT had the fewest regions with a positive correlation with age (n=6 for the general template and 7 for the data-template). A few more regions with positive correlations had ANTs (n=27, n=31) and fMRIPrep-FSL (n=29 and 31). ANTs-FSL and FSLVBM have significantly higher numbers of regions with positive correlations as well as regions with non-significant correlations ($p>0.05$). Regions with positive or nonsignificant correlations appear transparent in the brain images. For ANTs, the cerebellar regions and regions of cingulate gyri and limbic lobes. ANTs-FSL and FSLVBM demonstrated the most regions with a positive correlation with age. The cerebellum in FSLVBM shows a very small association with age, while in ANTs-FSL, cerebellar regions have more medium to high r values. Finally, fMRIPrep-FSL and CAT have small r values in the superior parietal and occipital lobes and medium to high r values in the frontal parts of the brain.

One-way ANOVA revealed a statistically significant difference in the average r-coefficients of regional GMV and age between at least two pipelines for all datasets (Supplementary

General templates

	ANTs	ANTs-FSL	fMRIPrep-FSL	FSLVBM	CAT
eNKI	-0.258	-0.182	-0.324	-0.155	-0.388
CamCAN	-0.264	-0.197	-0.411	-0.224	-0.425
IXI	-0.274	-0.163	-0.337	-0.151	-0.416
Mean	-0.265	-0.181	-0.357	-0.177	-0.410
All data	-0.253	-0.188	-0.357	-0.168	-0.381

Data-specific template

	ANTS	ANTs-FSL	fMRIPrep-FSL	FSLVBM	CAT 12
eNKI	-0.262	-0.188	-0.291	-0.145	-0.385
CamCAN	-0.260	-0.193	-0.365	-0.202	-0.421
IXI	-0.270	-0.157	-0.298	-0.140	-0.413
Mean	-0.264	-0.179	-0.318	-0.162	-0.406
All data	-0.253	-0.174	-0.319	-0.155	-0.370

Table 3: Pearson's r-values were calculated between age and all regional GMVs across subjects. r-values were transformed to Fischer's z averaged and transformed back to r-values. CAT with the general template and with the data-template appears to preserve age-related information better than the other pipelines, followed by fMRIPrep-FSL and ANTs. There is high consistency between datasets, with CamCAN showing a higher relation to age for those pipelines that use FSL for registration and CAT.

Material Table S.5).

3.3.2 Comparison of regional age information between pipelines

The regional GMV-age correlation values not only differed but also showed opposing effects (Figure 7). In other words, some regions showed a positive correlation with age in one pipeline but a negative correlation in another pipeline (see Supplementary Material Figures S.16, S.17 and S.18). In particular, this was the case for FSLVBM and ANTs-FSL, which contained many regions with a positive correlation with age. Strikingly, the same two pipelines also exhibited a large number of regions with opposing correlations with age when using a different template.

When using all data, CAT had $n\_rois = 6$ ROIs with a positive correlation to age when using either template. fMRIPrep-FSL had n_rois =27 with the general template and 22 with the data-template, and ANTs had $n\_rois = 56$ for both templates. ANTs-FSL and FSLVBM had $n\_rois = 218$ and 280 regions positively correlated to age when using a general template and 184 and 226 regions when using a data-template, respectively. Two regions in the thalamus showed a positive correlation with age for all pipelines. In general, the regions with a positive correlation with age for all pipelines were mostly subcortical.

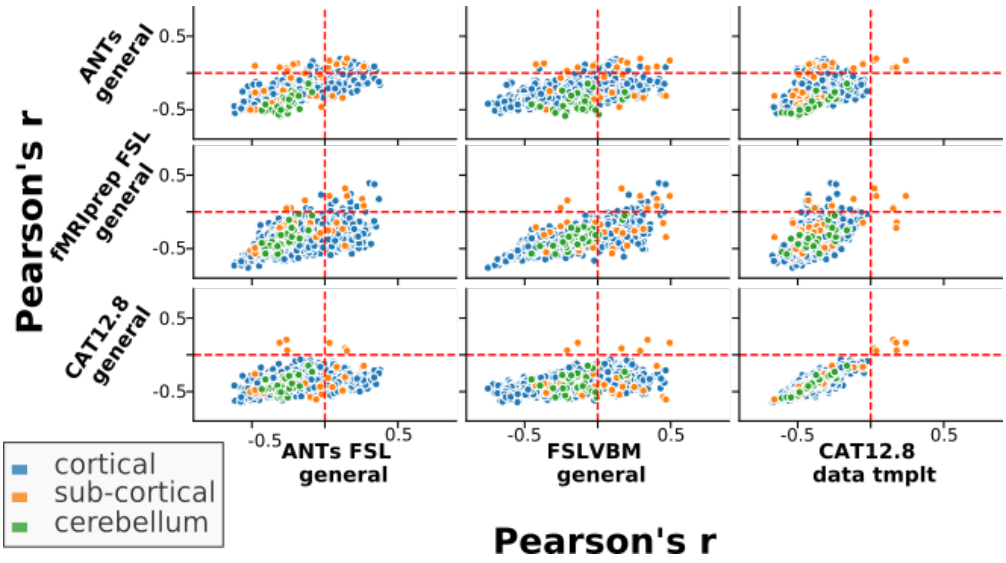

Figure 7: Pearson's r values between regional GMV and age calculated across subjects for selected pipelines plotted against the same measurements for other pipelines. The upper left and lower right quadrants of each subplot contain those regions that have correlations to age with opposite signs/directions between the two pipelines. ANTs-FSL and FSLVBM have the most ROIs with positive correlations to age. Here, we selected a few pipelines that cover the spectrum of the main tools we used and better illustrate how the same regions in different pipelines can have opposite relations to age. All pipeline combinations can be seen in Figure S.15 in the Supplementary Material.

3.3.3 Effect of Parcel size

We examined whether parcel size was associated with the agreement among the pipelines and with the agreement between ROIs and age. We observed no or marginal association between the overall similarity among the pipelines (calculated as the median of agreement between pipeline pairs) and parcel sizes (Pearson's correlation, all data: $r = -0.08, p = 0.006$, eNKI: $r = -0.02, p = 0.51$, CamCAN: $r = -0.11, p = 0.0002$, IXI: $r = 0.07, p = 0.022$) (Supplementary Material Figure S.19).

Correlation values between parcel size and the corresponding regional correlation values to age for each pipeline varied between pipelines as well as between datasets. The highest correlation was for CAT, with $r = -0.145$ when using the general template and $r = -0.134$ with the data-template (both $p < 0.05$). ANTs showed the next closest relation between parcel size and regional association with age, with $r = -0.105$ when using a general template and $r = -0.101$ when using a data-template (both $p < 0.05$). Those marginal negative correlations indicate that the fewer voxels are in an ROI, the better the relation of this ROI to age. All other correlation values were rather small, indicating that overall, the parcel sizes did not impact our results (Supplementary Material, for all data combined Figure S.23, eNKI Figure S.20, CamCAN Figure S.21 and IXI Figure S.22).

4 Discussion

"Which tool shall I use to perform my VBM analysis?", this is one of the very first questions that a researcher asks before starting a VBM study. The choice is often based on the literature or familiarity or recommendations. The current lack of an in depth comparison between VBM pipelines, the impact of the main steps on the outcome, and their utility precludes informative choice. Sparked by that, we compared 10 VBM pipelines derived from widely used tools on three large datasets covering the adult lifespan, acquired in different scanners and protocols. Two of the pipelines consisted of VBM steps from different tools. Our experiments were designed to facilitate a user-centric and systematic evaluation, which allows us to derive robust conclusions. Moreover, it permitted the examination of the effect of template use, i.e., general and data-template, as well as the effect of individual VBM steps.

Overall, we made the following observations based on analysis of the GMV estimates from different perspectives. The differences in individuals' brain-age predictions confirmed that different VBM pipelines produce different GMVs (Figure 1, Tables S.2 & S.3). The systematic differences between the pipelines were further confirmed by the high accuracy when predicting the pipelines using their GMVs (Figure S.4). A detailed univariate analysis of across-subject correlation (Figures 4) and identification using the subject-specific multivariate GMV pattern (Figures 2) showed that the individual steps of the VBM process as well as the choice of the template lead to the differences in the GMV estimates (see also Figure 5 and Table 2). Differences in GMV in turn impact the way age is reflected as we saw in univariate analysis correlating regional GMV with age (Figure 6 and Table 3).

First, we sought to establish whether the pipelines indeed lead to different results in applications. To this end, we performed predictive analysis using regional GMV as features and four machine-learning models commonly used in brain-age prediction. Individual-level age prediction showed variability in prediction accuracy (Figure 1), similar to what has been previously reported for voxel-level analysis and using CAT and FSL-based pipelines [30]. Our age-prediction accuracy for CAT and fMRIPrep-FSL are comparable to previous reports, considering our dataset size and the wide age range [87, 88]. To establish whether the differences in the pipelines are systematic, we performed classification analysis. The near-perfect classification performance in the prediction of pipelines (Figure S.4) provides evidence for systematically distinct outcomes of the pipelines, which could be learned by the machine-learning algorithm and is in line with previous research [26, 32, 34]. Importantly, removing overall GMV differences by standardizing each feature vector also provided similarly high accuracy. Based on these results, even though the pipelines differ in seemingly trivial ways, such as using different templates or segmentation algorithm, we can conclude that they produce diverging GMV patterns.

Taken together, these results suggest that combining data processed with different pipelines might not be fruitful. Data harmonization methods [89, 90], although designed for tackling cross-site differences, can also be explored to eliminate cross-pipeline differences. To this

end, we performed two preliminary analyses. First, we harmonized data across all the 10 pipelines and performed pipeline prediction analysis similar to 2.5. The pipelines could not be predicted with high accuracy after harmonization, however we also observed a bias towards specific pipelines (Supplementary Material Figure S.5). Second, we harmonized the three datasets processed with three different pipelines and performed leave-one-site-out age prediction analysis similar to section 2.4. This resulted in a higher MAE (MAE=8.5 using a GPR model, Supplementary Material Table S.4) compared to when using a single preprocessing pipeline (MAE=6.29-8.36 using a GPR model, Table S.3). In addition, we would like to note that harmonization can perform better when the biological variance of interest is explicitly preserved, such as age as the target in age prediction analysis. However, this means that the target value must be also available for the test data. This setup leads to data leakage when performing CV and cannot be applied on real test data, considering also that data from the test site or pipeline is needed for learning a harmonization model (in our analysis we harmonized all the data together). Thus, in its current form this approach is not suitable for ML applications. These results suggest that applying data harmonization methods in this context is challenging and needs further investigation.

The low to moderate identification performance and its variability across pipelines suggest that individual-level characteristics are, to a certain degree, captured differently by different pipelines (Figure 2). This result has important implications for data sharing and privacy issues [91]. As we show, with regionwise GMV data it is difficult to identify subjects when processed with different pipelines. Thus, when sharing such data, for instance, to perform multicenter analysis, it is important to keep the VBM pipeline consistent, including the template used.

Univariate analysis showed limited ROI-level similarity across pipelines, with an average regional similarity of $r = 0.51$ for pipelines using a general template. FSLVBM (using BET) and fMRIPrep-FSL (using ANTs brain extraction) showed high similarity, especially when a data-template was used (average $r = 0.76$) (Figure 5 (c)). When using the general template, the average similarity decreased but remained relatively high ($r = 0.67$). This suggests that differences in brain extraction are overshadowed by the subsequent steps. ANTs-FSL and fMRIPrep-FSL pipelines that differ mainly in segmentation (and the a priori template in brain extraction) showed relatively high agreement ($r = 0.67$ general template; $r = 0.68$ data-template), although slightly lower than what we show for brain extraction (Figure 5 (b)).

Differences between registration algorithms have been reported [41]. Our results are in line with this previous report. The registration step, evaluated as a comparison between ANTs and ANTs-FSL, had medium-to-high impact, with average agreement between these pipelines ranging across datasets, from $r = 0.48$ to $r = 0.53$ and $r = 0.57$ to $r = 0.6$ for general and data-template, respectively (Figure 5 (a)).

The impact of using different registration templates, general template versus data-template, was examined using pipelines that differ only in the template. This resulted in a wide-ranging

agreement from $r = 0.59$ to $r = 0.92$ (Table 2). ANTs and CAT create data-templates that are very similar to their respective general templates – likely due to their exhaustive registration algorithms and the iterative processes together with the fact that their template creation processes are initialized with a general template. Overall, the differences in data-template creation algorithms and the ensuing data-templates led to substantial differences across the tools. This is in agreement with previous research reporting a small impact of the template when using CAT [35]. Effectively, using a data-template imposes higher similarity between the subjects' images, which we also observed for some pipelines (Figure 4). Despite this high similarity, machine-learning-based analysis could reliably distinguish the pipelines. Univariate analysis of regionwise GMV-age correlations as well as age prediction were in favor of using a general template. Using subjects' data to create a data-template and then registering the same subjects to it is a circular process unless an independent subset is used for template creation; however, given the limited data, this is often hard to implement in practice. The latter, in combination with the high computational demands of the template-creation process, are in favor of using a general template.

Although ANTs and CAT share no common modules, they showed medium to high similarity (for all data sets ranged from $r = 0.65$ to $r = 0.72$; maximum was for $r = 0.74$ for the eNKI). According to the impact of individual steps in the final GMV, as shown in our pipeline comparison, CAT and ANTs are expected to yield differing GMV estimates unless there are similarities in their internal algorithmic mechanism, which seems to be the case. In fact, exhaustive registration to similar templates can lead to similar outcomes. ANTs-FSL with the general template and CAT (both templates) showed the lowest regionwise similarity across datasets. However, in our opinion, the low similarity between CAT, with either template, and FSLVBM using a general template needs special attention (Figure 4 and Supplementary Material, eNKI Figure S.8, CamCAN Figure S.9 and IXI Figure S.10). The reason is that they are both *off-the-shelf* pipelines and widely used in VBM projects. Regionally, the highest differences were present in the frontal lobe, superior parietal lobule and subcortical regions, specifically with regards to their association to age (Supplementary Material Figures S.15, S.16, S.17, S.18) Such differences enhance the risk of emanating different or even sometimes contradictory conclusions. From the projection of similarities between pipelines in the brain (Supplementary Material nifti files), it appears that high correlation values are not located in specific regions, nor is a specific pattern formed. However, segmentation and brain extraction seem to affect stronger subcortical and cerebellar areas and the superior frontal and occipital lobes. When comparing the registrations of ANTs and FNIRT, widespread differences appear in cortical areas and in the cerebellum (Figure 5 (b)).

The identification results (Figure 2) were very similar to the pairwise similarity estimated using Pearson's correlation (Figure 4). The agreement between the two methods was high (Pearson's correlation between pairwise similarity and Idiff, $r = 0.84$), and when using general templates, identification and univariate analysis were almost the same ($r = 0.955$, Supplementary Material Figure S.12). This agreement between two different methods to assess similarity between the pipelines provides confirmatory validity to our findings.

It is important to note that, mostly for brain extraction but also for segmentation and registration algorithms, there are important differences between the datasets (Figure 5). This indicates that properties such as the intensity range of the images can influence the results in different ways, e.g., the quality of segmentation varies across different scanning parameters [92–94].

By using three large datasets, we aimed to cover a wide range of MRI vendors as well as scanning parameters and settings. Different scanners were used not only across datasets but also within the same dataset, strengthening our results and conclusions independent of the datasets' idiosyncrasies.

The fMRIPrep-FSL combination showed the second highest correlation with age and the best brain-age predictions. This is not surprising given the nonexhaustive registration of FSL, which together with deep neural networks provides accurate brain-age prediction [25]. It is noteworthy that we used all subjects from the eNKI sample without separating the healthy part of the cohort as is usually done. When inspecting the age predictions of only healthy subjects, in intrasite predictions, and a mix of healthy and nonhealthy subjects, cross-site, separately, we did not observe a significant difference (see Supplementary Material Table S.2 and Table S.3). This can be explained by the fact that the nonlinear transformations wipe-out small differences compared to linear registration but also by the fact that the templates we used are based on healthy populations. In the age-prediction CAT showed performance similar to fMRIPrep-FSL but lower than what has been previously reported [17]. However, this difference can be driven by the machine-learning algorithms and the feature space employed. These results are in line with the univariate analysis we performed, where the same two pipelines had the highest (anti-) correlation with age (Figure 6). In addition, fewer ROIs showed a positive correlation with age for CAT and fMRIPrep-FSL than for other pipelines, which is in line with known GM atrophy with age [95–97]. Taken together, our results are in favor of CAT and fMRIPrep-FSL in regard to aging-related studies. Although some recent brain-age applications have shown that linear registration is preferable [16, 25], we decided to compare the whole VBM process using nonlinear registration. This choice was made so that we could approach the topic via a common space, permit the use of a parcellation atlas and facilitate the interpretability of the results.

The user-centric approach we followed in this project does not allow for an extensive evaluation of the potentials of the tools we used. CAT, ANTs, but to a certain degree also FSLVBM potentially can be tuned to provide more accurate brain-age predictions or regional associations to age. However, such an investigation is out of the scope of this work.

To summarize, our results show that all steps of a VBM pipeline have a considerable impact on the GMV estimates, and therefore, different pipelines produce different results. These differences in GMV estimates are reflected in univariate as well as multivariate analyses. The choice of registration has the highest impact, followed by segmentation and brain extraction algorithm. In the specific case of age-prediction, we recommend the combination of ANTs for brain extraction and FSL for segmentation (as implemented in fMRIPrep) and FSL nonlinear

registration or CAT 12.8, with the latter having the advantage of being available as an off-the-shelf pipeline. The option of using a general template is preferred for age-related studies and likely other studies with a similar set up, especially when analyzing scans from multiple datasets.

Ethics statement

Ethical approval and informed consent were obtained locally for each study covering both participation and subsequent data sharing. The ethics proposals for the use and retrospective analyses of the datasets were approved by the Ethics Committee of the Medical Faculty at the Heinrich-Heine-University Düsseldorf.

Data/code availability statement

The codes used for preprocessing, feature extraction and model training are available at `https://jugit.fz-juelich.de/g.antonopoulos/vbm_comparison_codes`.

Disclosure of competing interests

The authors report no competing interests.

Acknowledgments

This study is supported by Deutsche Forschungsgemeinschaft (DFG, PA 3634/1-1 and EI 816/21-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modelling for the Human Brain" and the European Union's Horizon 2020 Research and Innovation Program grant agreement 945539 (HBP SGA3).

References

[1] Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Deborah Blacker, Paul Maguire, Diana Rosas, Nikos Makris, Randy Gollub, Anders Dale, Bradford C. Dickerson, and Bruce Fischl. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *NeuroImage*, 46(1):177–192, May 2009.

[2] I. C. Wright, P. K. McGuire, J. B. Poline, J. M. Travere, R. M. Murray, C. D. Frith, R. S. J. Frackowiak, and K. J. Friston. A Voxel-Based Method for the Statistical Analysis of Gray and White Matter Density Applied to Schizophrenia. *NeuroImage*, 2(4):244–252, December 1995.

[3] John Ashburner and Karl J. Friston. Voxel-Based Morphometry—The Methods. *NeuroImage*, 11(6):805–821, June 2000.

[4] Hiroshi Matsuda. Voxel-based Morphometry of Brain MRI in Normal Aging and Alzheimer's Disease. *Aging and Disease*, 4(1):29, February 2013.

[5] Ching-Hung Lin, Chun-Ming Chen, Ming-Kuei Lu, Chon-Haw Tsai, Jin-Chern Chiou, Jan-Ray Liao, and Jeng-Ren Duann. VBM Reveals Brain Volume Differences between Parkinson's Disease and Essential Tremor Patients. *Frontiers in Human Neuroscience*, 7, 2013.

[6] Bijen Khagi, Kun Ho Lee, Kyu Yeong Choi, Jang Jae Lee, Goo-Rak Kwon, and Hee-Deok Yang. VBM-Based Alzheimer's Disease Detection from the Region of Interest of T1 MRI with Supportive Gaussian Smoothing and a Bayesian Regularized Neural Network. *Applied Sciences*, 11(13):6175, January 2021.

[7] Sean. J. Colloby, John. T. O'Brien, and John-Paul Taylor. Patterns of cerebellar volume loss in dementia with Lewy bodies and Alzheimer's disease: A VBM-DARTEL study. *Psychiatry Research: Neuroimaging*, 223(3):187–191, September 2014.

[8] J. B. Brewer. Fully-Automated Volumetric MRI with Normative Ranges: Translation to Clinical Practice. *Behavioural Neurology*, 21(1-2):21–28, 2009.

[9] Hosam Abozaid Yousef, Yasser Mohamed Bader-Eldein ElSerogy, Sherif Mohamed Abdelal, and Shaza Ragab Abdel-Rahman. Voxel-based morphometry in patients with mood disorder bipolar I mania in comparison to normal controls. *Egyptian Journal of Radiology and Nuclear Medicine*, 51(1):9, January 2020.

[10] Catriona D. Good, Ingrid S. Johnsrude, John Ashburner, Richard N. A. Henson, Karl J. Friston, and Richard S. J. Frackowiak. A Voxel-Based Morphometric Study of Ageing in 465 Normal Adult Human Brains. *NeuroImage*, 14(1):21–36, July 2001.

[11] Danielle J. Tisserand, Martin P.J. van Boxtel, Jens C. Pruessner, Paul Hofman, Alan C. Evans, and Jelle Jolles. A Voxel-based Morphometric Study to Determine Individual Differences in Gray Matter Density Associated with Age and Cognitive Change Over Time. *Cerebral Cortex*, 14(9):966–973, September 2004.

[12] Ali K Bourisly, Ahmed El-Beltagi, Jigi Cherian, Grace Gejo, Abrar Al-Jazzaf, and Mohammad Ismail. A voxel-based morphometric magnetic resonance imaging study of the brain detects age-related gray matter volume changes in healthy subjects of 21–45 years old. *The Neuroradiology Journal*, 28(5):450–459, October 2015. Publisher: SAGE Publications Ltd.

[13] M. Habes, D. Janowitz, G. Erus, J. B. Toledo, S. M. Resnick, J. Doshi, S. Van der Auwera, K. Wittfeld, K. Hegenscheid, N. Hosten, R. Biffar, G. Homuth, H. Völzke, H. J. Grabe, W. Hoffmann, and C. Da-

vatzikos. Advanced brain aging: relationship with epidemiologic and genetic risk factors, and overlap with Alzheimer disease atrophy patterns. *Translational Psychiatry*, 6(4):e775–e775, April 2016.

[14] Nikolaos Koutsouleris, Christos Davatzikos, Stefan Borgwardt, Christian Gaser, Ronald Bottlender, Thomas Frodl, Peter Falkai, Anita Riecher-Rössler, Hans-Jürgen Möller, Maximilian Reiser, Christos Pantelis, and Eva Meisenzahl. Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. *Schizophrenia Bulletin*, 40(5):1140–1153.

[15] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. C. Valdés Hernández, S. Muñoz Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, N. R. Wray, P. Redmond, R. E. Marioni, J. M. Starr, S. R. Cox, J. M. Wardlaw, D. J. Sharp, and I. J. Deary. Brain age predicts mortality. *Molecular Psychiatry*, 23(5):1385–1392, May 2018.

[16] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892, April 2010.

[17] B. A. Jonsson, G. Bjornsdottir, T. E. Thorgeirsson, L. M. Ellingsen, G. Bragi Walters, D. F. Gudbjartsson, H. Stefansson, K. Stefansson, and M. O. Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, 10(1):5409, November 2019.

[18] Longfei Su, Lubin Wang, Hui Shen, and Dewen Hu. Age-related Classification and Prediction Based on MRI: A Sparse Representation Method. *Procedia Environmental Sciences*, 8:645–652, January 2011.

[19] Deepthi P. Varikuti, Sarah Genon, Aristeidis Sotiras, Holger Schwender, Felix Hoffstaedter, Kaustubh R. Patil, Christiane Jockwitz, Svenja Caspers, Susanne Moebus, Katrin Amunts, Christos Davatzikos, and Simon B. Eickhoff. Evaluation of non-negative matrix factorization of grey matter in age prediction. *NeuroImage*, 173:394–410, June 2018.

[20] Katja Franke and Christian Gaser. Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained? *Frontiers in Neurology*, 10:789, 2019.

[21] Lea Baecker, Rafael Garcia-Dias, Sandra Vieira, Cristina Scarpazza, and Andrea Mechelli. Machine learning for brain age prediction: Introduction to methods and clinical applications. *eBioMedicine*, 72, October 2021.

[22] Ting Su, Pei-Wen Zhu, Biao Li, Wen-Qing Shi, Qi Lin, Qing Yuan, Nan Jiang, Chong-Gang Pei, and Yi Shao. Gray matter volume alterations in patients with strabismus and amblyopia: voxel-based morphometry study. *Scientific Reports*, 12(1):458, January 2022.

[23] Yin-Nan Zhang, Hui Li, Zhi-Wei Shen, Chang Xu, Yue-Jun Huang, and Ren-Hua Wu. Healthy individuals vs patients with bipolar or unipolar depression in gray matter volume. *World Journal of Clinical Cases*, 9(6):1304–1317, February 2021.

[24] Meng Li, Jianhao Yan, Shumei Li, Tianyue Wang, Hua Wen, Yi Yin, Shishun Fu, Luxian Zeng, Junzhang Tian, and Guihua Jiang. Altered gray matter volume in primary insomnia patients: a DARTEL-VBM study. *Brain Imaging and Behavior*, 12(6):1759–1767, December 2018.

[25] Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871, February 2021.

[26] Venkateswaran Rajagopalan and Erik P. Pioro. Disparate voxel based morphometry (VBM) results between SPM and FSL softwares in ALS patients with frontotemporal dementia: which VBM results to consider? 15:32.

[27] Nicola K. Dinsdale, Emma Bluemke, Stephen M. Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana I. L. Namburete. Learning patterns of the ageing brain in MRI using deep convolutional networks. *NeuroImage*, 224:117401, January 2021.

[28] Christian Gaser and R. Dahnke. CAT-a computational anatomy toolbox for the analysis of structural MRI data.

[29] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy E J Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney, Rami K Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J Michael Brady, and Paul M Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23 Suppl 1:S208–S219, 2004.

[30] Xinqi Zhou, Renjing Wu, Yixu Zeng, Ziyu Qi, Stefania Ferraro, Lei Xu, Xiaoxiao Zheng, Jialin Li, Meina Fu, Shuxia Yao, Keith M. Kendrick, and Benjamin Becker. Choice of Voxel-based Morphometry processing pipeline drives variability in the location of neuroanatomical brain markers. *Communications Biology*, 5(1):1–12, September 2022.

[31] Statistical Parametric Mapping: The Analysis of Functional Brain Images - 1st Edition.

[32] Veronica Popescu, Menno M. Schoonheim, Adriaan Versteeg, Nimisha Chaturvedi, Marianne Jonker, Renee Xavier de Menezes, Francisca Gallindo Garre, Bernard M. J. Uitdehaag, Frederik Barkhof, and Hugo Vrenken. Grey Matter Atrophy in Multiple Sclerosis: Clinical Interpretation Depends on Choice of Analysis Method. *PLOS ONE*, 11(1):e0143942, January 2016.

[33] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, October 2007.

[34] Dorothée V. Callaert, Annemie Ribbens, Frederik Maes, Stephan P. Swinnen, and Nicole Wenderoth. Assessing age-related gray matter decline with voxel-based morphometry depends significantly on segmentation and normalization procedures. *Frontiers in Aging Neuroscience*, 6, 2014.

[35] Logan Haynes, Amanda Ip, Ivy Y.K. Cho, Dennis Dimond, Christiane S. Rohr, Mercedes Bagshawe, Deborah Dewey, Catherine Lebel, and Signe Bray. Grey and white matter volumes in early childhood: A comparison of voxel-based morphometry pipelines. *Developmental Cognitive Neuroscience*, 46, October 2020.

[36] H. Matsuda, S. Mizumura, K. Nemoto, F. Yamashita, E. Imabayashi, N. Sato, and T. Asada. Automatic Voxel-Based Morphometry of Structural MRI by SPM8 plus Diffeomorphic Anatomic Registration Through Exponentiated Lie Algebra Improves the Diagnosis of Probable Alzheimer Disease. *AJNR: American Journal of Neuroradiology*, 33(6):1109–1114, 2012.

[37] Farnaz Farokhian, Iman Beheshti, Daichi Sone, and Hiroshi Matsuda. Comparing CAT12 and VBM8 for Detecting Brain Morphological Abnormalities in Temporal Lobe Epilepsy. *Frontiers in Neurology*, 8:428, August 2017.

[38] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. 12(1):26–41.

[39] Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, February 2011.

[40] Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E. Christensen, D. Louis Collins, James Gee, Pierre Hellier, Joo Hyun Song, Mark Jenkinson,

Claude Lepage, Daniel Rueckert, Paul Thompson, Tom Vercauteren, Roger P. Woods, J. John Mann, and Ramin V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802, July 2009.

[41] Yangming Ou, Hamed Akbari, Michel Bilello, Xiao Da, and Christos Davatzikos. Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. *IEEE transactions on medical imaging*, 33(10):2039–2065, October 2014.

[42] Eileanoir B. Johnson, Sarah Gregory, Hans J. Johnson, Alexandra Durr, Blair R. Leavitt, Raymund A. Roos, Geraint Rees, Sarah J. Tabrizi, and Rachael I. Scahill. Recommendations for the Use of Automated Gray Matter Segmentation Tools: Evidence from Huntington's Disease. *Frontiers in Neurology*, 8:519, 2017.

[43] Mahsa Dadar and Simon Duchesne. Reliability assessment of tissue classification algorithms for multi-center and multi-scanner data. *NeuroImage*, 217:116928, August 2020.

[44] Frederick Klauschen, Aaron Goldman, Vincent Barra, Andreas Meyer-Lindenberg, and Arvid Lundervold. Evaluation of automated brain MR image segmentation and volumetry methods. *Human Brain Mapping*, 30(4):1310–1327, June 2008.

[45] A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, February 1999.

[46] Ian B. Malone, Kelvin K. Leung, Shona Clegg, Josephine Barnes, Jennifer L. Whitwell, John Ashburner, Nick C. Fox, and Gerard R. Ridgway. Accurate automatic estimation of total intracranial volume: A nuisance variable with less nuisance. *Neuroimage*, 104:366–372, January 2015.

[47] Gajendra J. Katuwal, Stefi A. Baum, Nathan D. Cahill, Chase C. Dougherty, Eli Evans, David W. Evans, Gregory J. Moore, and Andrew M. Michael. Inter-Method Discrepancies in Brain Volume Estimation May Drive Inconsistent Findings in Autism. *Frontiers in Neuroscience*, 10:439, 2016.

[48] Jorge Sepulcre, Jaume Sastre-Garriga, Mara Cercignani, Gordon T. Ingle, David H. Miller, and Alan J. Thompson. Regional Gray Matter Atrophy in Early Primary Progressive Multiple Sclerosis: A Voxel-Based Morphometry Study. *Archives of Neurology*, 63(8):1175–1180, August 2006.

[49] Antonia Ceccarelli, Maria A. Rocca, Elisabetta Pagani, Bruno Colombo, Vittorio Martinelli, Giancarlo Comi, and Massimo Filippi. A voxel-based morphometry study of grey matter loss in MS patients with different clinical phenotypes. *NeuroImage*, 42(1):315–322, August 2008.

[50] Marco Battaglini, Antonio Giorgio, Maria L. Stromillo, Maria L. Bartolozzi, Leonello Guidi, Antonio Federico, and Nicola De Stefano. Voxel-wise assessment of progression of regional brain atrophy in relapsing-remitting multiple sclerosis. *Journal of the Neurological Sciences*, 282(1-2):55–60, July 2009.

[51] Nicholas J. Tustison, Hans J. Johnson, Torsten Rohlfing, Arno Klein, Satrajit S. Ghosh, Luis Ibanez, and Brian B. Avants. Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences. 7.

[52] James H. Cole, Rudra P. K. Poudel, Dimosthenis Tsagkrasoulis, Matthan W. A. Caan, Claire Steves, Tim D. Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. 163:115–124.

[53] Elizabeth R. Sowell, Bradley S. Peterson, Paul M. Thompson, Suzanne E. Welcome, Amy L. Henkenius, and Arthur W. Toga. Mapping cortical change across the human life span. *Nature Neuroscience*, 6(3):309–315, March 2003.

[54] Kate Brody Nooner, Stanley J. Colcombe, Russell H. Tobe, Maarten Mennes, Melissa M. Benedict, Alexis L. Moreno, Laura J. Panek, Shaquanna Brown, Stephen T. Zavitz, Qingyang Li, Sharad Sikka,

David Gutman, Saroja Bangaru, Rochelle Tziona Schlachter, Stephanie M. Kamiel, Ayesha R. Anwar, Caitlin M. Hinz, Michelle S. Kaplan, Anna B. Rachlin, Samantha Adelsberg, Brian Cheung, Ranjit Khanuja, Chaogan Yan, Cameron C. Craddock, Vincent Calhoun, William Courtney, Margaret King, Dylan Wood, Christine L. Cox, A. M. Clare Kelly, Adriana Di Martino, Eva Petkova, Philip T. Reiss, Nancy Duan, Dawn Thomsen, Bharat Biswal, Barbara Coffey, Matthew J. Hoptman, Daniel C. Javitt, Nunzio Pomara, John J. Sidtis, Harold S. Koplewicz, Francisco Xavier Castellanos, Bennett L. Leventhal, and Michael P. Milham. The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Frontiers in Neuroscience*, 6:152, 2012.

[55] Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, null Cam-Can, and Richard N. Henson. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144(Pt B):262–269, January 2017.

[56] Meredith A Shafto, Lorraine K Tyler, Marie Dixon, Jason R Taylor, James B Rowe, Rhodri Cusack, Andrew J Calder, William D Marslen-Wilson, John Duncan, Tim Dalgleish, Richard N Henson, Carol Brayne, and Fiona E Matthews. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 14:204, October 2014.

[57] IXI Dataset – Brain Development, https://brain-development.org/ixi-dataset/.

[58] Gwenaëlle Douaud, Stephen Smith, Mark Jenkinson, Timothy Behrens, Heidi Johansen-Berg, John Vickers, Susan James, Natalie Voets, Kate Watkins, Paul M. Matthews, and Anthony James. Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain: A Journal of Neurology*, 130(Pt 9):2375–2386, September 2007.

[59] Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. FMRIPrep: a robust preprocessing pipeline for functional MRI. 16(1):111–116.

[60] Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, June 2010.

[61] Brian B. Avants, Nicholas J. Tustison, Jue Wu, Philip A. Cook, and James C. Gee. An open source multivariate framework for n-tissue segmentation with evaluation on public data. 9(4):381–400.

[62] Nicholas James Tustison and Brian avants Avants. Explicit b-spline regularization in diffeomorphic image registration. *Frontiers in Neuroinformatics*, 7.

[63] Avants Bb, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, and Gee Jc. The optimal template effect in hippocampus studies of diseased populations. *NeuroImage*, 49(3), February 2010.

[64] VS Fonov, AC Evans, RC McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, July 2009.

[65] Vladimir Fonov, Alan C. Evans, Kelly Botteron, C. Robert Almli, Robert C. McKinstry, and D. Louis Collins. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327, January 2011.

[66] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, January 2001.

[67] J. C. Rajapakse, J. N. Giedd, and J. L. Rapoport. Statistical approach to segmentation of single-channel cerebral MR images. *IEEE transactions on medical imaging*, 16(2):176–186, April 1997.

[68] Jussi Tohka, Alex Zijdenbos, and Alan Evans. Fast and robust parameter estimation for statistical partial volume models in brain MRI. *NeuroImage*, 23(1):84–97, September 2004.

[69] John Ashburner and Karl J. Friston. Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. *NeuroImage*, 55(3):954–967, April 2011.

[70] Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, 28(9):3095–3114, September 2018.

[71] Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R. Laird, Peter T. Fox, Simon B. Eickhoff, Chunshui Yu, and Tianzi Jiang. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex*, 26(8):3508–3526, August 2016.

[72] Randy L. Buckner, Fenna M. Krienen, Angela Castellanos, Julio C. Diaz, and B. T. Thomas Yeo. The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(5):2322–2345, November 2011.

[73] James H. Cole and Katja Franke. Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends in Neurosciences*, 40(12):681–690.

[74] Ji Hye Won, Mansu Kim, Jinyoung Youn, and Hyunjin Park. Prediction of age at onset in parkinson's disease using objective specific neuroimaging genetics based on a sparse canonical correlation analysis. *Nature*, 10(1):11662.

[75] Shammi More, Georgios Antonopoulos, Felix Hoffstaedter, Julian Caspers, Simon B. Eickhoff, Kaustubh R. Patil, and the Alzheimer's Disease Neuroimaging Initiative. Brain-age prediction: a systematic comparison of machine learning workflows, November 2022. Pages: 2022.11.16.515405 Section: New Results.

[76] Shammi More, Simon B. Eickhoff, Julian Caspers, and Kaustubh R. Patil. Confound Removal and Normalization in Practice: A Neuroimaging Based Sex Prediction Case Study. In Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke, editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, Lecture Notes in Computer Science, pages 3–18, Cham, 2021. Springer International Publishing.

[77] Michael E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1(Jun):211–244, 2001.

[78] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, USA, November 2005.

[79] Fadil Santosa and William W. Symes. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, October 1986.

[80] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[81] Vladimir Vovk. Kernel ridge regression. In *Empirical inference*, pages 105–116. Springer, 2013.

[82] Russell A. Poldrack, Grace Huckins, and Gael Varoquaux. Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*, 77(5):534–540, May 2020.

[83] Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664–1671, November 2015.

[84] Enrico Amico and Joaquín Goñi. The quest for identifiability in human functional connectomes. *Scientific Reports*, 8(1):8254, May 2018.

[85] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[86] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[87] Claudia R Eickhoff, Felix Hoffstaedter, Julian Caspers, Kathrin Reetz, Christian Mathys, Imis Dogan, Katrin Amunts, Alfons Schnitzler, and Simon B Eickhoff. Advanced brain ageing in Parkinson's disease is related to disease duration and individual impairment. *Brain Communications*, 3(3):fcab191, July 2021.

[88] James H. Cole, Tiina Annus, Liam R. Wilson, Ridhaa Remtulla, Young T. Hong, Tim D. Fryer, Julio Acosta-Cabronero, Arturo Cardenas-Blanco, Robert Smith, David K. Menon, Shahid H. Zaman, Peter J. Nestor, and Anthony J. Holland. Brain-predicted age in Down syndrome is associated with beta amyloid deposition and cognitive decline. *Neurobiology of Aging*, 56:41–49, August 2017.

[89] Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M. Nasrallah, Theodore D. Satterthwaite, Yong Fan, Lenore J. Launer, Colin L. Masters, Paul Maruff, Chuanjun Zhuo, Henry Völzke, Sterling C. Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Daniel H. Wolf, Raquel Gur, Ruben Gur, John Morris, Marilyn S. Albert, Hans J. Grabe, Susan M. Resnick, R. Nick Bryan, David A. Wolk, Russell T. Shinohara, Haochang Shou, and Christos Davatzikos. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450, March 2020.

[90] Joaquim Radua, Eduard Vieta, Russell Shinohara, Peter Kochunov, Yann Quidé, Melissa J. Green, Cynthia S. Weickert, Thomas Weickert, Jason Bruggemann, Tilo Kircher, Igor Nenadić, Murray J. Cairns, Marc Seal, Ulrich Schall, Frans Henskens, Janice M. Fullerton, Bryan Mowry, Christos Pantelis, Rhoshel Lenroot, Vanessa Cropley, Carmel Loughland, Rodney Scott, Daniel Wolf, Theodore D. Satterthwaite, Yunlong Tan, Kang Sim, Fabrizio Piras, Gianfranco Spalletta, Nerisa Banaj, Edith Pomarol-Clotet, Aleix Solanes, Anton Albajes-Eizagirre, Erick J. Canales-Rodríguez, Salvador Sarro, Annabella Di Giorgio, Alessandro Bertolino, Michael Stäblein, Viola Oertel, Christian Knöchel, Stefan Borgwardt, Stefan du Plessis, Je-Yeon Yun, Jun Soo Kwon, Udo Dannlowski, Tim Hahn, Dominik Grotegerd, Clara Alloza, Celso Arango, Joost Janssen, Covadonga Díaz-Caneja, Wenhao Jiang, Vince Calhoun, Stefan Ehrlich, Kun Yang, Nicola G. Cascella, Yoichiro Takayanagi, Akira Sawa, Alexander Tomyshev, Irina Lebedeva, Vasily Kaleda, Matthias Kirschner, Cyril Hoschl, David Tomecek, Antonin Skoch, Therese van Amelsvoort, Geor Bakker, Anthony James, Adrian Preda, Andrea Weideman, Dan J. Stein, Fleur Howells, Anne Uhlmann, Henk Temmingh, Carlos López-Jaramillo, Ana Díaz-Zuluaga, Lydia Fortea, Eloy Martinez-Heras, Elisabeth Solana, Sara Llufriu, Neda Jahanshad, Paul Thompson, Jessica Turner, Theo van Erp, David Glahn, Godfrey Pearlson, Elliot Hong, Axel Krug, Vaughan Carr, Paul Tooney, Gavin Cooper, Paul Rasser, Patricia Michie, Stanley Catts, Raquel Gur, Ruben Gur, Fude Yang, Fengmei Fan, Jingxu Chen, Hua Guo, Shuping Tan, Zhiren Wang, Hong Xiang, Federica Piras, Francesca

Assogna, Raymond Salvador, Peter McKenna, Aurora Bonvino, Margaret King, Stefan Kaiser, Dana Nguyen, and Julian Pineda-Zapata. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage*, 218:116956, September 2020.

[91] Tonya White, Elisabet Blok, and Vince D. Calhoun. Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed. *Human Brain Mapping*, 43(1):278–291, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.25120.

[92] Vishwanatha M. Rao, Zihan Wan, David J. Ma, Pin-Yu Lee, Ye Tian, Andrew F. Laine, and Jia Guo. Improving Across-Dataset Brain Tissue Segmentation Using Transformer. *arXiv:2201.08741 [cs, eess]*, January 2022.

[93] Frithjof Kruggel, Jessica Turner, and L. Tugan Muftuler. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. *NeuroImage*, 49(3):2123–2133, February 2010.

[94] Sergi Valverde, Arnau Oliver, Mariano Cabezas, Eloy Roura, and Xavier Lladó. Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. *Journal of Magnetic Resonance Imaging*, 41(1):93–101, 2015.

[95] Farnaz Farokhian, Chunlan Yang, Iman Beheshti, Hiroshi Matsuda, and Shuicai Wu. Age-Related Gray and White Matter Changes in Normal Adult Brains. *Aging and disease*, 8(6):899, 2017.

[96] Efstathios D. Gennatas, Brian B. Avants, Daniel H. Wolf, Theodore D. Satterthwaite, Kosha Ruparel, Rastko Ciric, Hakon Hakonarson, Raquel E. Gur, and Ruben C. Gur. Age-Related Effects and Sex Differences in Gray Matter Density, Volume, Mass, and Cortical Thickness from Childhood to Young Adulthood. *Journal of Neuroscience*, 37(20):5065–5073, 2017.

[97] Elouise A. Koops, Emile de Kleine, and Pim van Dijk. Gray matter declines with age and hearing loss, but is partially maintained in tinnitus. *Scientific Reports*, 10(1):21801, December 2020. Number: 1 Publisher: Nature Publishing Group.

[98] Rafael Garcia-Dias, Cristina Scarpazza, Lea Baecker, Sandra Vieira, Walter H.L. Pinaya, Aiden Corvin, Alberto Redolfi, Barnaby Nelson, Benedicto Crespo-Facorro, Colm McDonald, Diana Tordesillas-Gutiérrez, Dara Cannon, David Mothersill, Dennis Hernaus, Derek Morris, Esther Setien-Suero, Gary Donohoe, Giovanni Frisoni, Giulia Tronchin, João Sato, Machteld Marcelis, Matthew Kempton, Neeltje E.M. van Haren, Oliver Gruber, Patrick McGorry, Paul Amminger, Philip McGuire, Qiyong Gong, René S. Kahn, Rosa Ayesa-Arriola, Therese van Amelsvoort, Victor Ortiz-García de la Foz, Vince Calhoun, Wiepke Cahn, and Andrea Mechelli. Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. *Neuroimage*, 220:117127, October 2020.

Supplementary Material

- MRI acquisition details

Sample	**Scanner**	**Sequence**	**Tesla**	**Slices**	**Voxel size (mm)**	**Time parameters (TR / TE / TI [ms])**	**Other parameters (FA/FOV [°/mm])**
eNKI	Siemens Magnetom TrioTim	3D MP-RAGE	3T	176	1 x 1 x 1	1900/2.52/900	9/250x250
Cam-CAN	Tim Trio Siemens	3D MP-RAGE	3.0T	192	1 x 1 x 1	2,250/2.99/900	9/256 x 256
IXI-Guys	Phillips	3D MP-RAGE	1.5T	192	1.2 x 0.94 x 0.94	9.813/4.603/	8/256 x 256
IXI-HH	Discovery GE	3D FSPGR	3T	176	1.2 x 0.94 x 0.94	9.6/4.6/450	N/A /256 x 256
IXI-IOP	GE	N/A	1.5	N/A	1.2 x 0.94 x 0.94	N/A	N/A /256 x 256

Table S.1: MRI acquisition details for the three datasets.

- Preprocessing

All pipelines were run in a high-throughput compute cluster except CAT12.8, which was run in a high-performance computing cluster.

- Study specific templates of each pipeline

The templates created by CAT and ANTs appear to be sharper than those created by FSLVBM. For ANTs, the template was 197x233x189; for CAT 175x199x175; and for FSL-based pipelines, 91x109x91. Templates built with eNKI are presented in Figure S.1, templates built based on CamCAN subjects can be found in Figure S.2 and for IXI in Figure S.3.

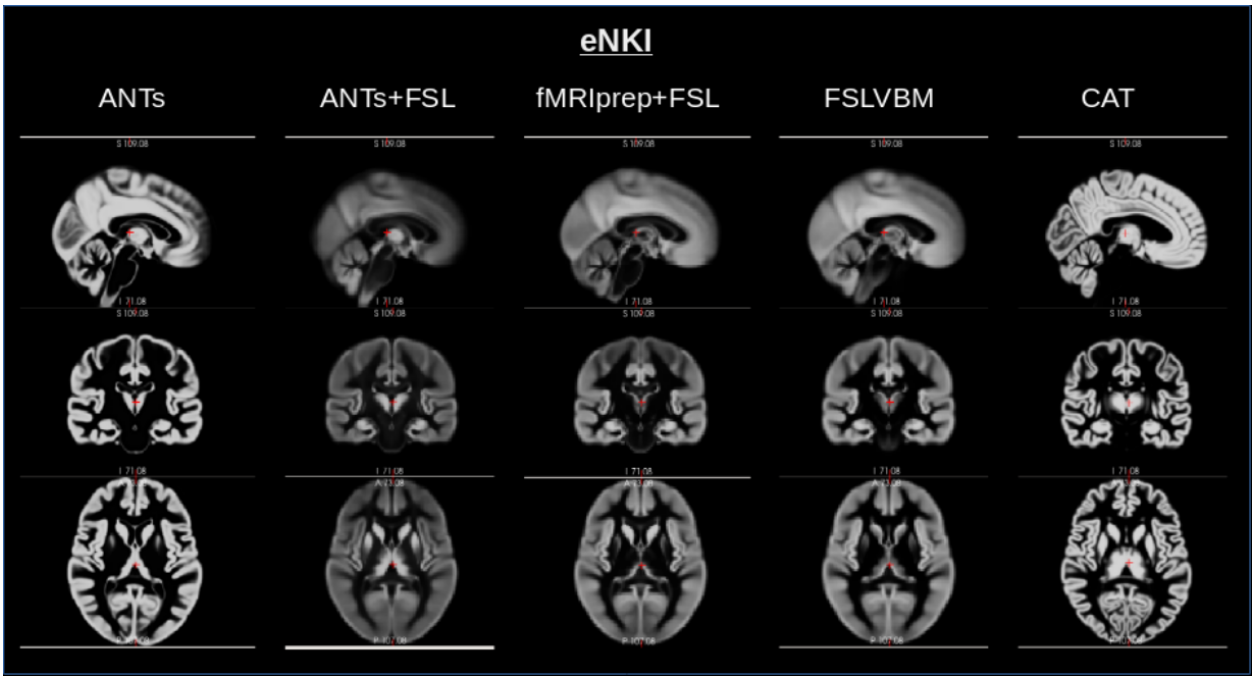

Figure S.1: Templates created for each pipeline. Default CAT and ANTs processes created sharper templates.

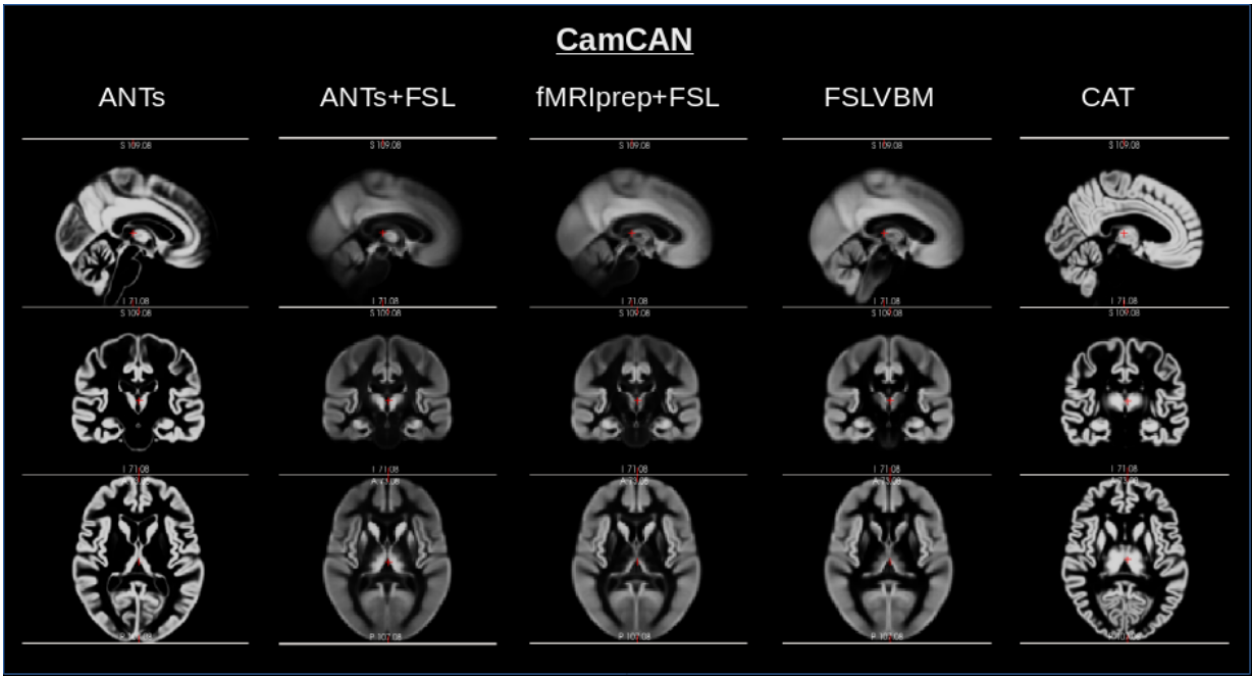

Figure S.2: Templates created for each pipeline CamCAN dataset.

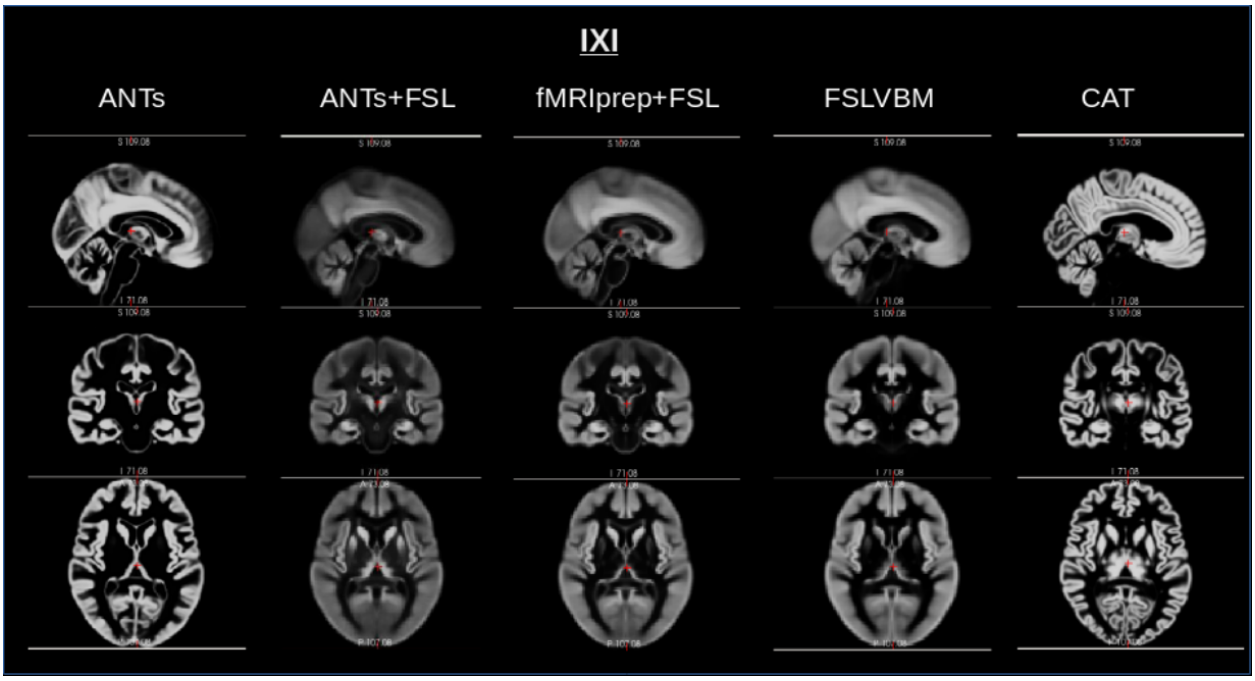

Figure S.3: Templates created for each pipeline for the IXI dataset

- Initialization of ANTs Atropos

It is worth mentioning that to follow our user-centric perspective, we used K-means clustering to initialize ANTs Atropos segmentation instead of some a-priori probability map. This resulted in relatively many subjects failing the preprocessing, as tissue samples were assigned unexpected labels by the K-means algorithm, resulting in confusion between white matter and gray matter. Although this could be approached by adjusting Atropos, such effort is out of scope of this project; instead, we performed custom quality control to detect such failed preprocessing.

- Detailed results of age predictions

Table S.2 shows the analytical results of age prediction when models are trained and tested within each site.

Pipelines	Models	MAE per model all datasets	MAE all models all datasets
ANTs General template	RVR	6.9	**7.06**
	GPR	7.04	
	LASSO	7.32	
	KRR	7.00	
ANTs Data template	RVR	6.93	**7.04**
	GPR	6.93	
	LASSO	7.35	
	KRR	6.95	
ANTs FSLVBM General template	RVR	6.56	**6.55**
	GPR	6.34	
	LASSO	6.79	
	KRR	6.52	
ANTs FSLVBM Data template	RVR	6.71	**6.74**
	GPR	6.46	
	LASSO	7.08	
	KRR	6.71	
fMRIPrep FSL General template	RVR	5.92	**5.83**
	GPR	5.65	
	LASSO	6.15	
	KRR	5.59	
fMRIPrep FSL Data template	RVR	6.14	**6.18**
	GPR	6.01	
	LASSO	6.51	
	KRR	6.06	
FSLVBM General template	RVR	6.25	**6.17**
	GPR	5.93	
	LASSO	6.45	
	KRR	6.05	
FSLVBM Data template	RVR	6.60	**6.55**
	GPR	6.30	
	LASSO	6.77	
	KRR	6.54	
CAT 12 General template	RVR	6.48	**6.39**
	GPR	6.26	
	LASSO	6.45	
	KRR	6.37	
CAT 12 Data template	RVR	6.46	**6.37**
	GPR	6.19	
	LASSO	6.47	
	KRR	6.37	

Table S.2: Results of age prediction using a multivariate approach. Four models were tested in the three datasets in a nested K-fold scheme. The third column contains the averaged results of the three datasets per model. The last column shows the average of all datasets and all models for each pipeline.

Table S.3 shows the analytical results of age prediction when the models are trained with

two of the datasets and tested with the leftout dataset.

Pipeline	**Models**	**Test eNKI**	**Test CamCAN**	**Test IXI**	**Mean test (datasets)**	**Mean test (datasets & Pipelines)**
ANTs General template	RVR	7.39	8.43	8.23	8.02	7.86
	GPR	7.19	7.94	8.27	7.80	
	LASSO	7.08	7.88	8.33	7.76	
	KRR	7.40	7.87	8.28	7.85	
ANTs Data template	RVR	7.48	8.26	7.88	7.87	7.86
	GPR	7.80	7.74	7.53	7.69	
	LASSO	7.73	6.88	8.99	7.87	
	KRR	7.84	8.12	8.03	8.00	
ANTs-FSL General template	RVR	6.96	9.16	9.75	8.62	8.44
	GPR	6.85	8.48	9.74	8.36	
	LASSO	7.60	7.97	9.88	8.49	
	KRR	6.73	7.91	10.28	8.31	
ANTs-FSL Data template	RVR	7.12	15.87	9.87	10.95	10.07
	GPR	6.95	11.47	9.72	9.38	
	LASSO	7.51	13.94	8.45	9.97	
	KRR	7.04	13.33	9.54	9.97	
FMRIprep-FSL General template	RVR	6.25	5.58	7.47	6.43	6.26
	GPR	6.10	5.63	7.13	6.29	
	LASSO	6.63	5.62	6.36	6.20	
	KRR	6.23	5.33	6.82	6.13	
FMRIprep-FSL Data template	RVR	6.83	6.56	6.12	6.50	6.21
	GPR	6.61	5.79	6.06	6.15	
	LASSO	6.66	5.76	5.91	6.11	
	KRR	6.48	5.89	5.82	6.06	
FSLVBM General template	RVR	6.62	6.02	8.66	7.10	6.77
	GPR	6.67	5.83	7.08	6.52	
	LASSO	7.23	6.09	6.88	6.73	
	KRR	6.43	5.77	8.00	6.73	
FSLVBM Data template	RVR	11.28	10.67	7.63	9.92	9.36
	GPR	11.42	10.66	6.70	9.55	
	LASSO	10.95	9.82	6.69	9.16	
	KRR	11.14	9.78	6.74	9.22	
CAT12.8 General template	RVR	6.75	6.53	6.39	6.55	6.45
	GPR	6.71	6.88	5.88	6.49	
	LASSO	6.75	6.62	5.92	6.43	
	KRR	6.55	6.58	5.82	6.32	
CAT12.8 Data template	RVR	6.83	6.81	6.25	6.63	6.76
	GPR	7.14	6.53	6.02	6.57	
	LASSO	7.20	7.51	7.52	7.41	
	KRR	7.01	6.06	6.19	6.42	

Table S.3: Cross-dataset age prediction results. For each pipeline we trained four models using two of the datasets and predicted the age of the subjects on the third, left-out dataset. The optimal parameters for each model were selected using a 5-fold cross validation scheme in the training datasets. The second to last column contains the mean of each model across all combinations of the datasets for training and testing for each pipeline. The last column contains mean across models and dataset combinations for each pipeline.

- Classifying subjects' images based on the preprocessing pipeline

We used two methods to scale the features prior to classification, using a linear SVM: i) within each feature and ii) within each subject, (both with standard scaling). The classification results were close to perfect using both methods (Figure S.4).

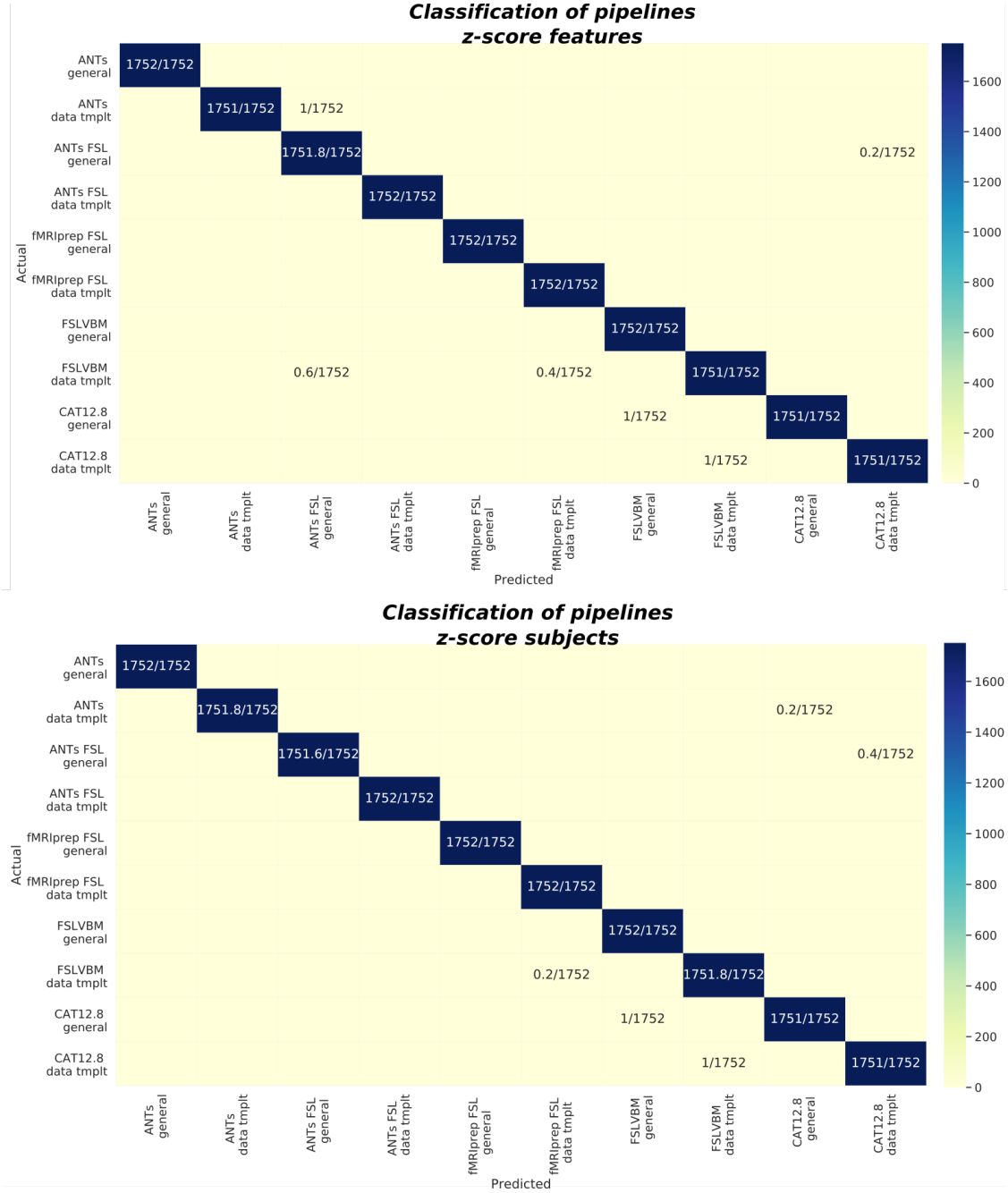

Figure S.4: Confusion matrices of multiclass classification of preprocessed subjects from all pipelines using the preprocessing pipelines as labels. We tried two methods for scaling the features aiming at ruling out that the overall intensity differences drive the classification. We used standard scaling, which standardizes features by removing the mean and scaling to unit variance, and MinMax scaling, which scales and translates each feature individually such that it is in the given range on the training set, here between zero and one.

- Harmonization across pipelines

The impact of harmonization across pipelines was also tested. We performed age-prediction and pipeline-prediction as we did for non-harmonized data in 2.4 and 2.5. Harmonization was performed using Neuroharmonize [98]. For age-prediction we selected eNKI processed by fMRIprep-FSL, IXI processed by FSLVBM and CamCAN processed by CAT 12.8 all with a general template. Table S.4 shows the age-prediction results when each of the three dataset is processed by different pipeline and then all data are harmonized.

model	Train: eNKI-IXI Test: CamCAN	Train: eNKI-CamCAN Test: IXI	Train: CamCAN-IXI Test: eNKI	Average
RVR	MAE=10.7	MAE=8.11	MAE=7.9	MAE=8.9
GPR	MAE=10.3	MAE=7.6	MAE=7.5	MAE=8.5

Table S.4: Brain age prediction results for harmonized data. The brain age prediction process was performed using eNKI processed by fMRIprep-FSL, IXI processed by FSLVBM and CamCAN processed by CAT 12.8 all with a general template. Data were harmonized and then used for individuals age prediction with GPR and RVR as in 2.4.

The results of the pipeline classification with harmonized data are shown in Figure S.5.

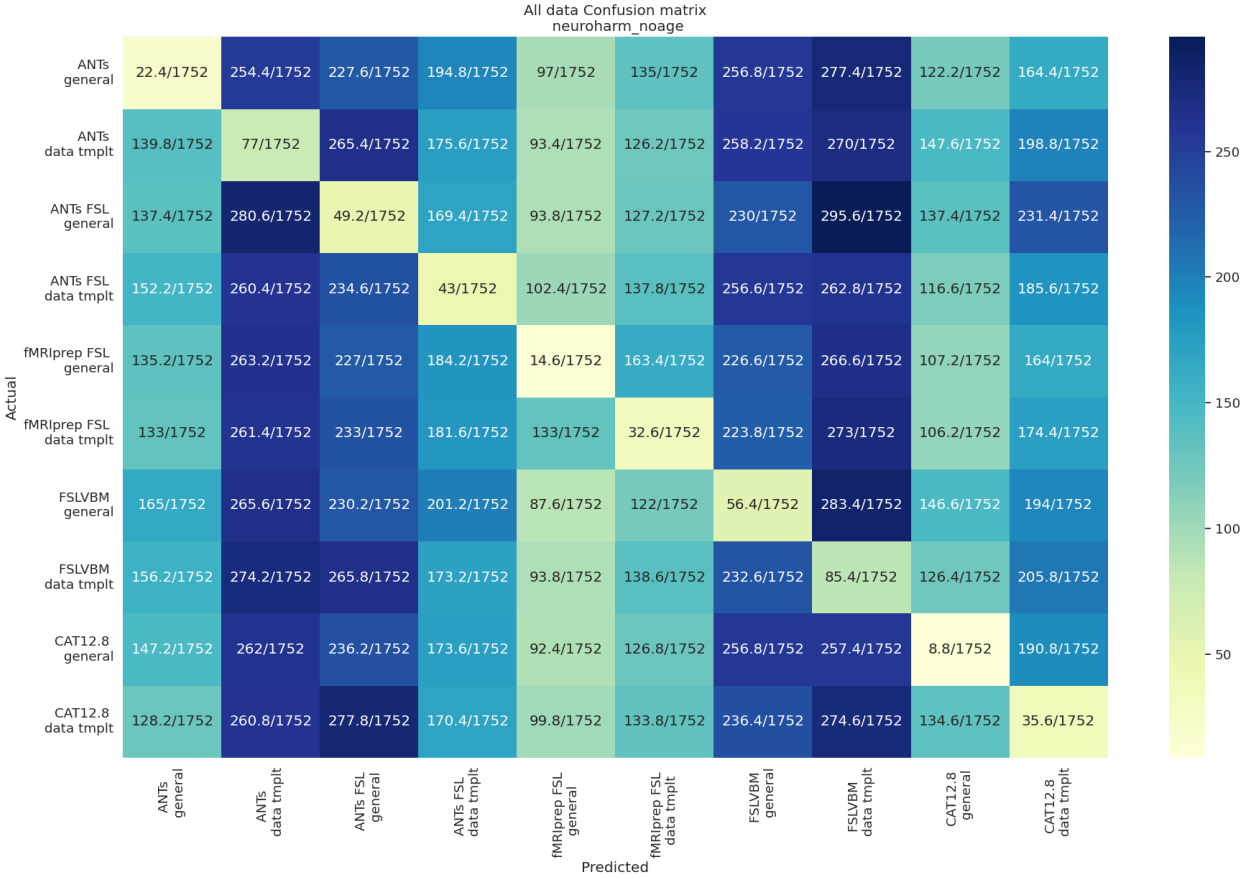

Figure S.5: Classification of pipelines based on harmonized GMV. Harmonization was performed across pipelines and the rest of the process was as in 2.5.

- Univariate analysis chart

Figure S.6 illustrates the univariate analysis we followed.

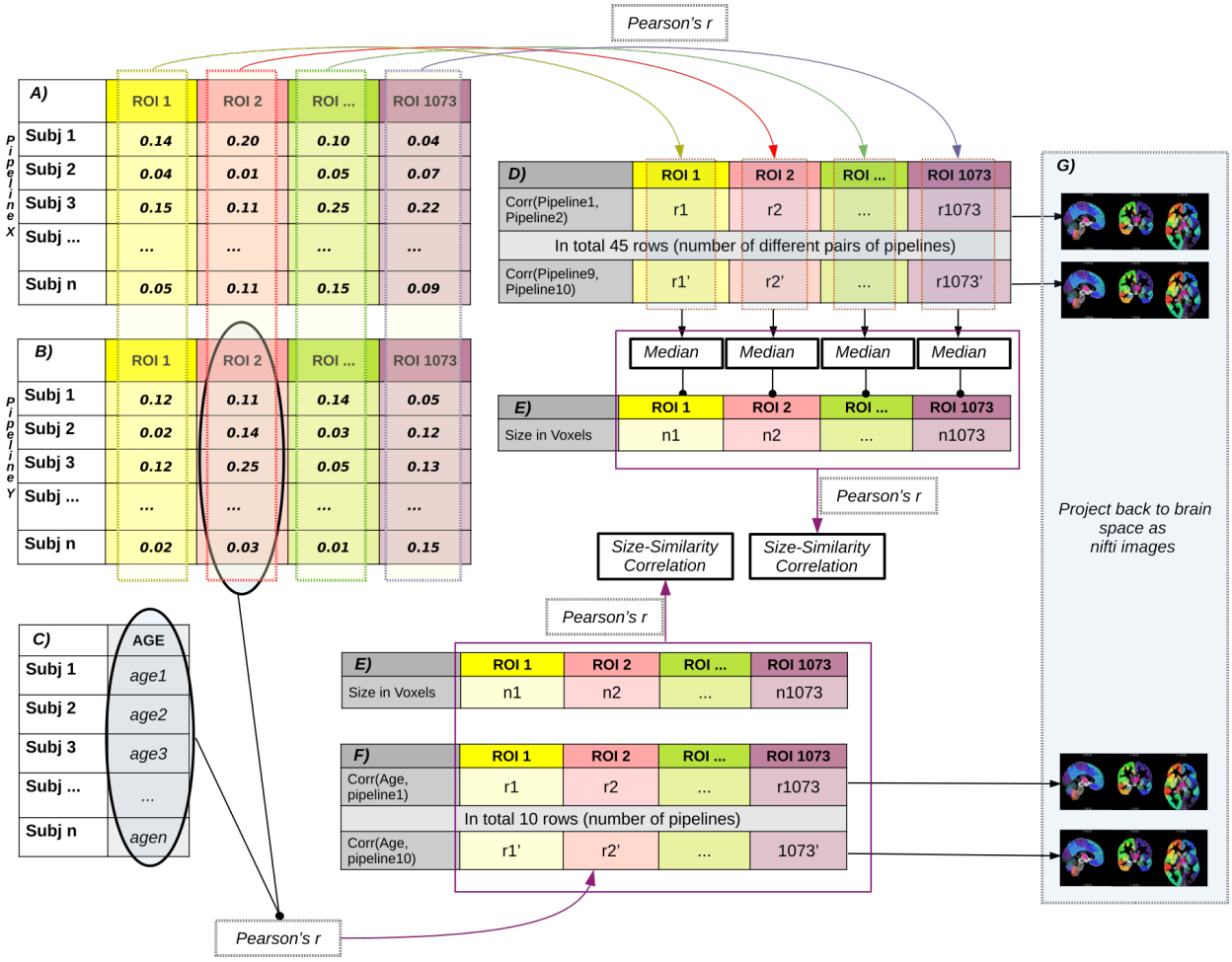

Figure S.6: Depiction of the process to calculate regional correlations across subjects for a pipeline pair. For a given pair of pipelines (panels A and B in the figure) for each region, we calculate the correlation across subjects. By performing this process for all pairs of pipelines and all regions, we obtain a regional correlation matrix (panel D). The overall agreement between the pipelines was calculated as the median for each region across all pipeline pairs, which was then used to correlate with the size of the parcels (panel E). For each pipeline and each region, we calculated Pearson's correlation across subjects between regional GMV (shown here for panel B) and age (panel C). Regional correlation values between pipelines (panel D) or with age (panel F) were projected on the brain for visualization purposes.

- Total GMV plots

The following image S.7 presents the total GMV of all subjects for each pipeline and each dataset.

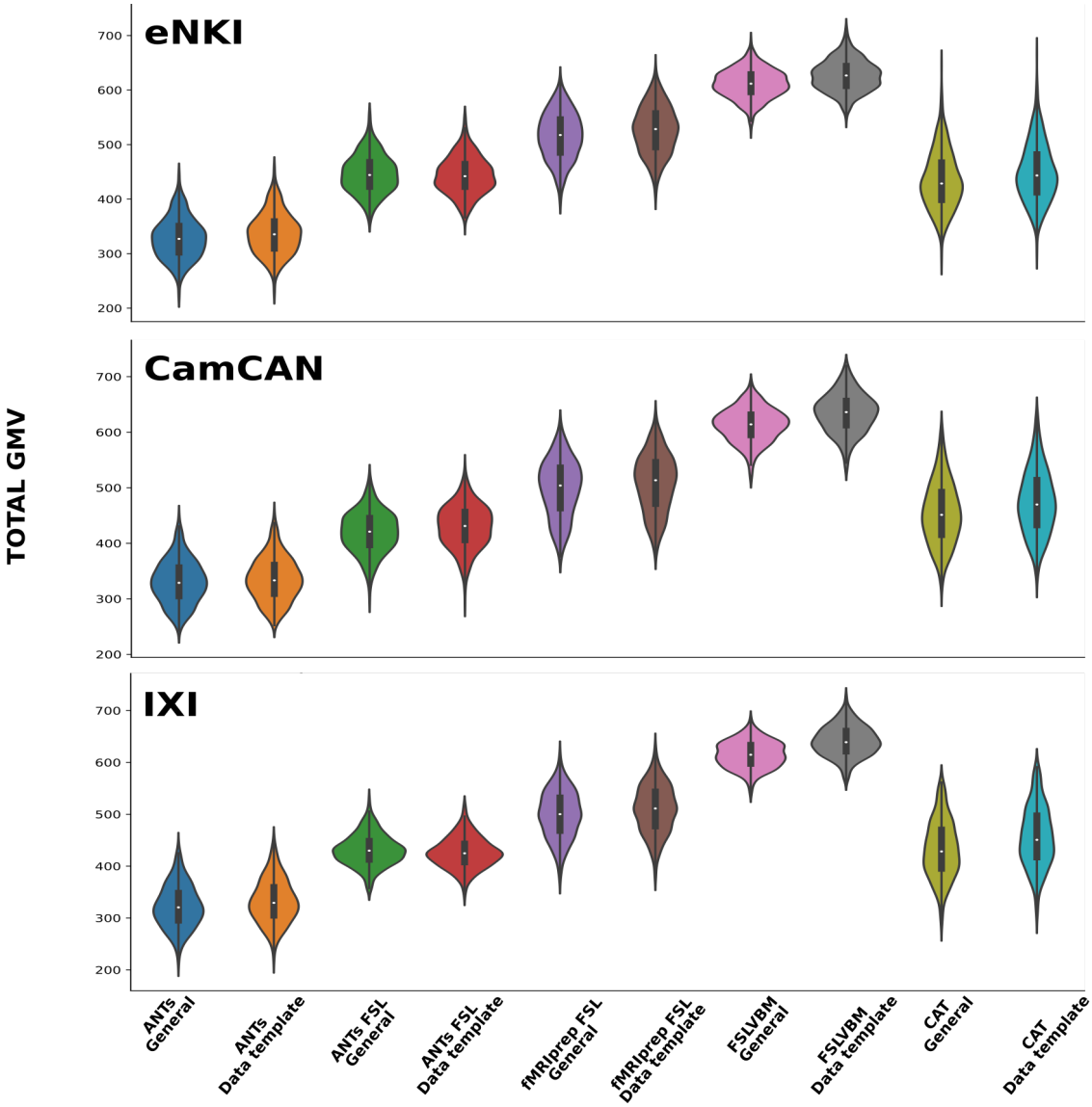

Figure S.7: Total GMV of all subjects for all pipelines and all datasets. Important differences in the total intensities between all pipelines. Only ANTs-FSL and CAT have similar means. Not surprisingly, the template appears to have no impact on the total GMV of subjects. High consistency is observed for the same pipelines across datasets.

- **Similarity between pipelines** as expressed by the regionwise Pearson's correlation across subjects for each pair of pipelines. For CamCAN in Figure S.9 and for IXI in S.10. From the figures of the three datasets, we see that similarities are consistent across datasets. However, some lack of variability is still identified, most likely due to differences in the quality of the images among datasets.

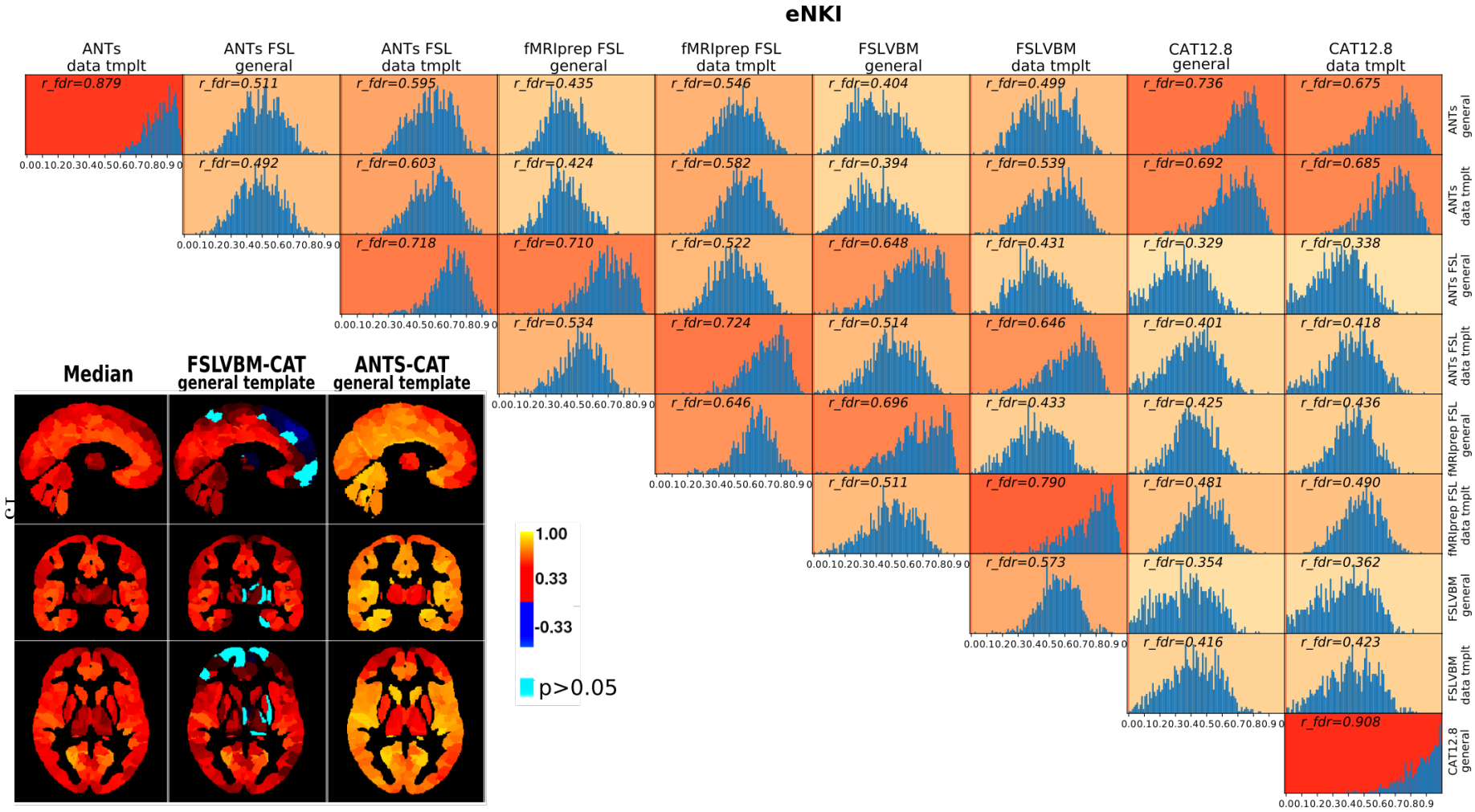

Figure S.8: eNKI: Histograms of Pearson's R values for all regions across subjects and for all combinations of pipelines. Niftis represent R values in the brain for the pipelines with Max mean, Min mean and the comparison of the two pipelines with the highest correlation to age.

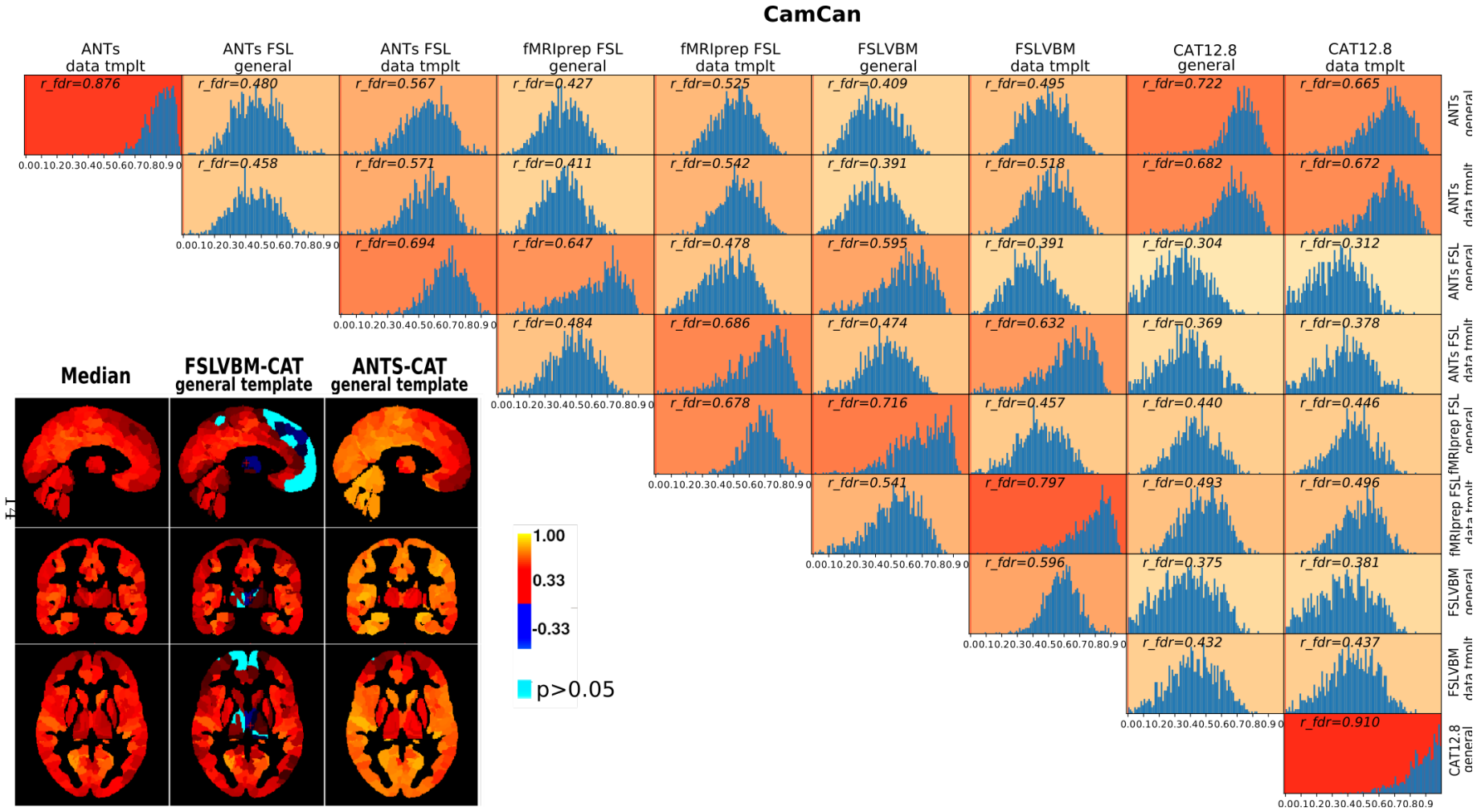

Figure S.9: CamCAN: Histograms of Pearson's R values for all regions across subjects and for all combinations of pipelines. Niftis represent R values in the brain for the pipelines with Max mean, Min mean and the comparison of the two pipelines with the highest correlation to age.

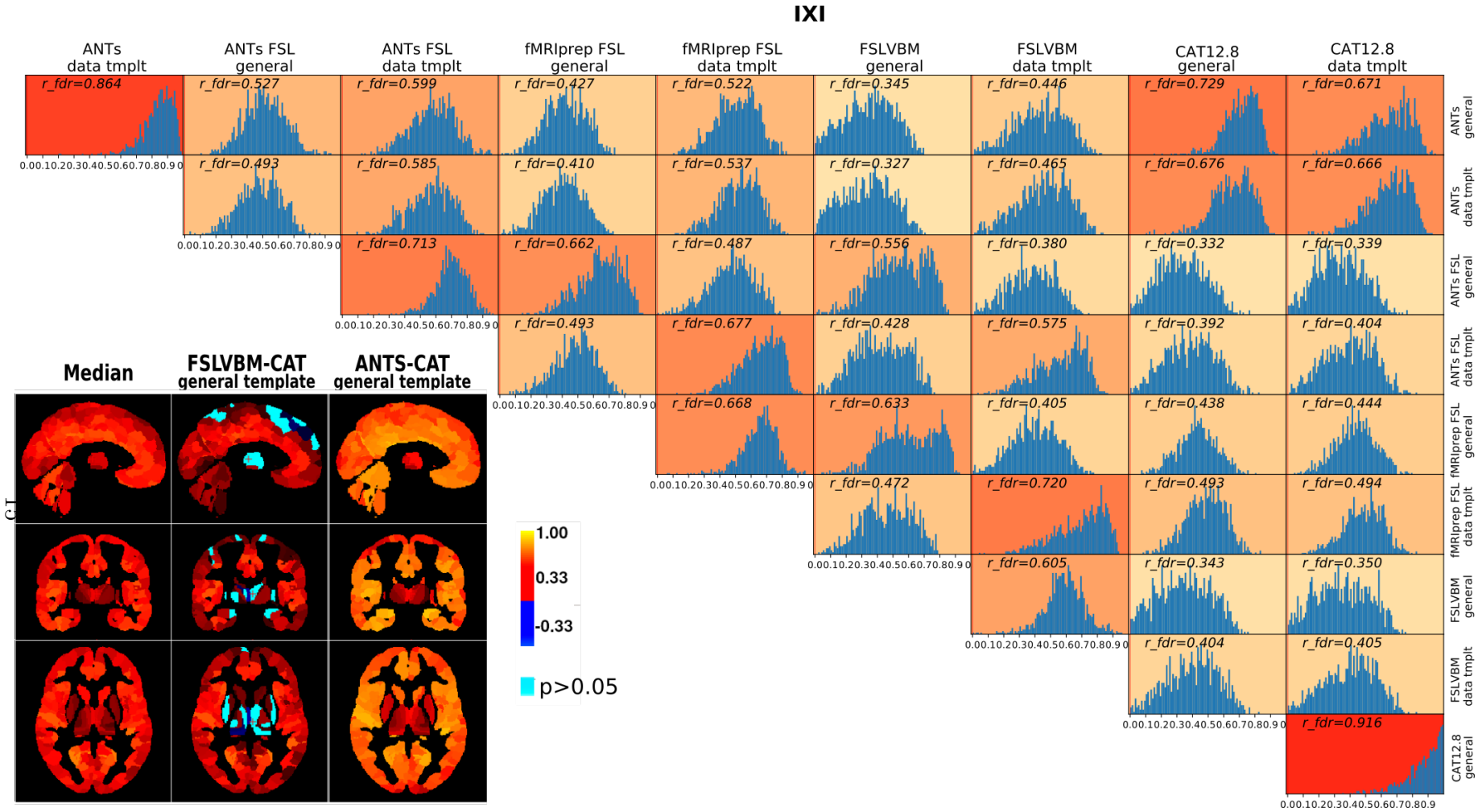

Figure S.10: IXI: Histograms of Pearson's R values for all regions across subjects and for all combinations of pipelines. Niftis represent R values in the brain for the pipelines with Max mean, Min mean and the comparison of the two pipelines with the highest correlation to age.

Correlations of pipelines that differ only in the template for the three datasets S.11

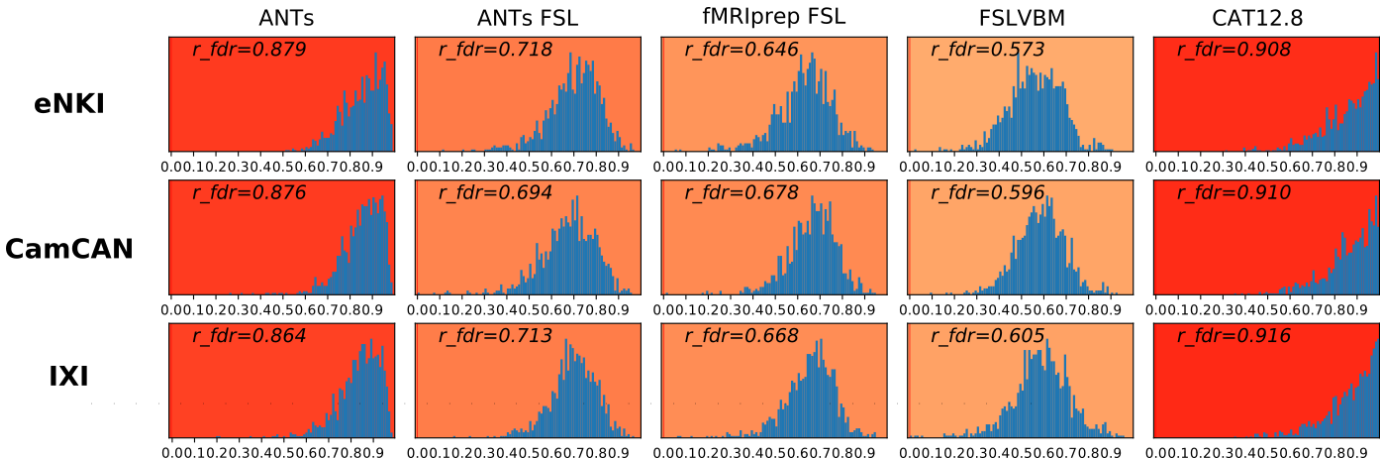

Figure S.11: Mean correlation of regions across subjects for all datasets between pipelines that only differ in the template used for spatial normalization.

The correlation between differential identifiability and Pearson's correlations calculated between pairs of pipelines was examined to assess the agreement between the two methods. The results can be seen in Figure S.12

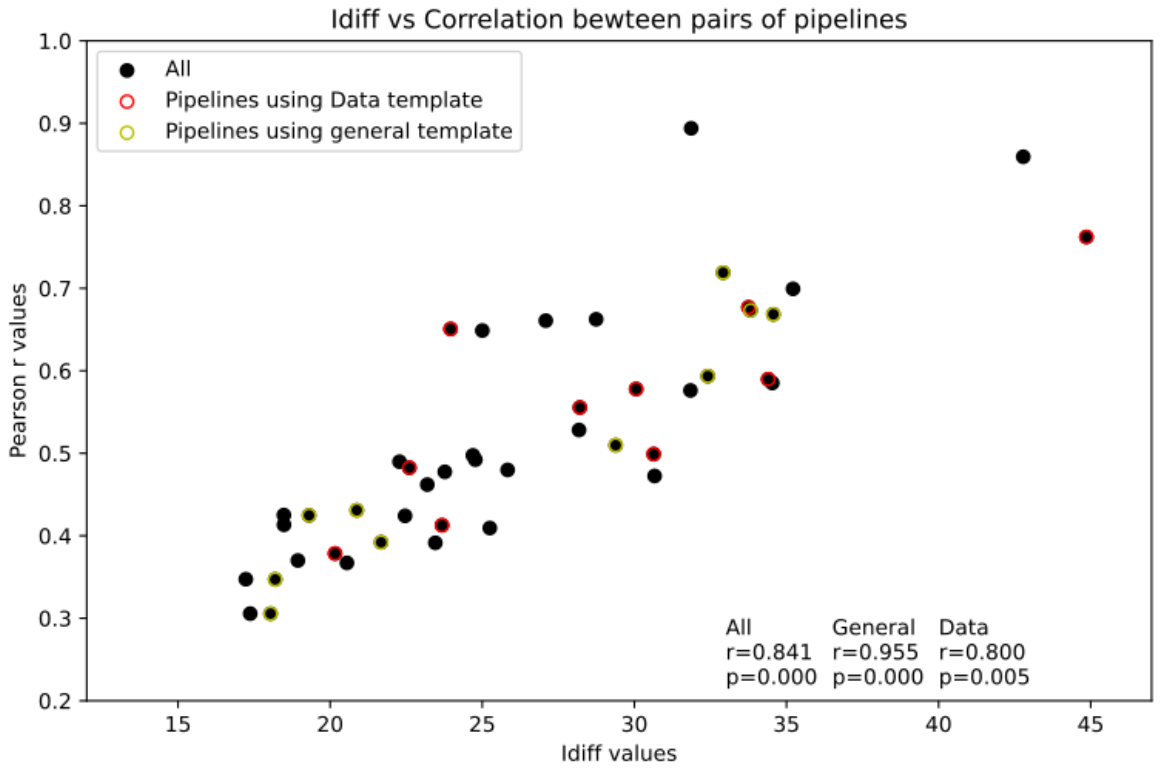

Figure S.12: Strong correlations were observed between the two methods we used to assess similarity between pipelines, the univariate analysis and identification. Especially between the pipelines using the general template, the correlation was $r = 0.955$. For pipelines using data-templates, the correlation was $r = 0.8$. The correlation for all pipeline pairs was $r = 0.841$. All correlations had $p < 0.05$.

Age-ROIs correlations for all pipelines in Figure S.13

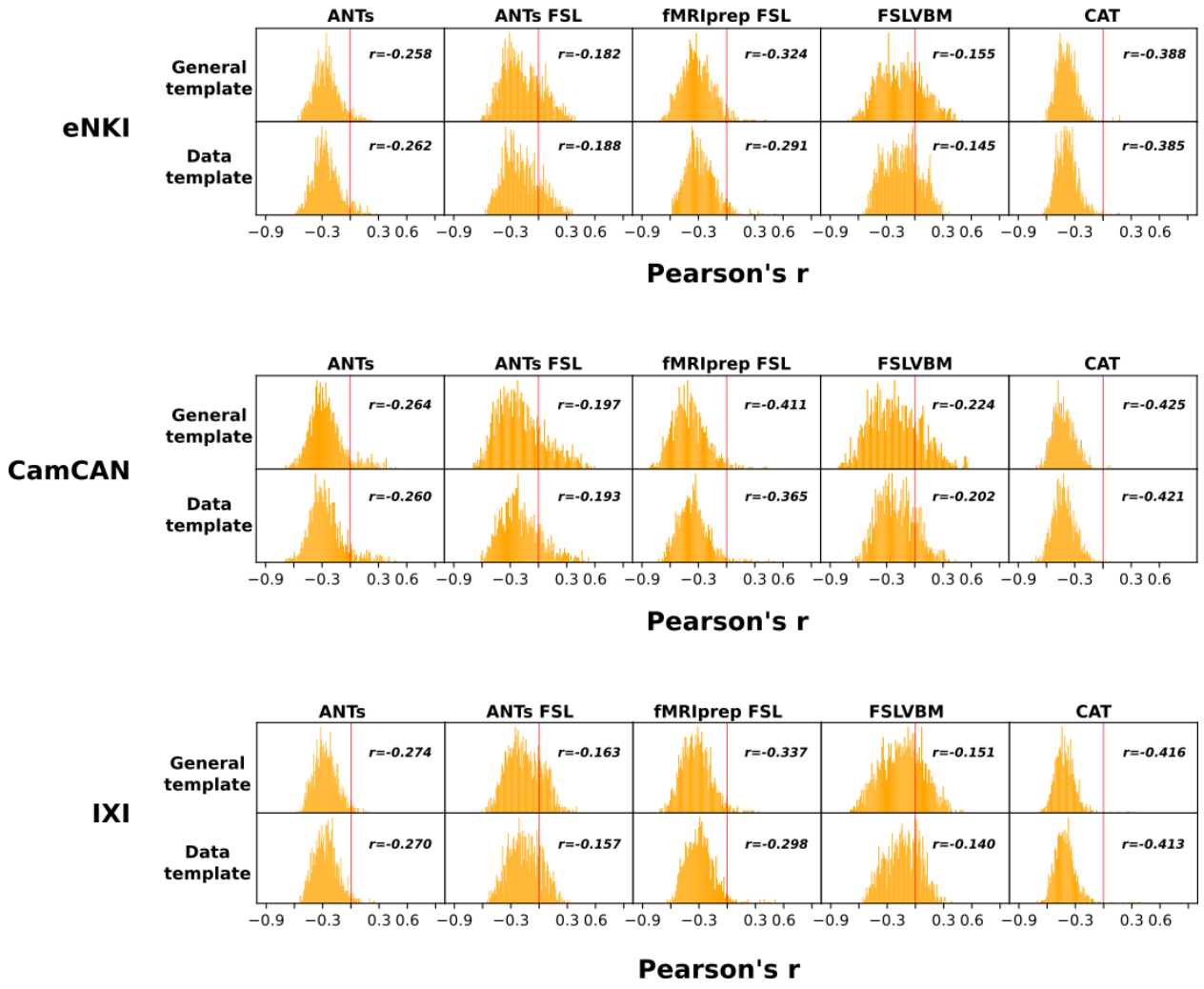

Figure S.13: Correlation between regions and age of subjects for all pipelines.

Table S.5 ANOVA results for correlation values between ROIs and age for all pipelines across subjects of all datasets. Figure S.14 depicts the same correlation values between all regions and age calculated across subjects of all datasets, per pipeline.

Dataset	F score	p-value
eNKI	509.99	4.18E-193 <0.05
CamCAN	400.45	5.11E-156 <0.05
IXI	637.771	5.18E-234 <0.05
All data	324.468	5.18E-234 <0.05

Table S.5: One-way ANOVA for the three datasets was performed for the three pipelines that used general templates and had the highest overall correlation to age.

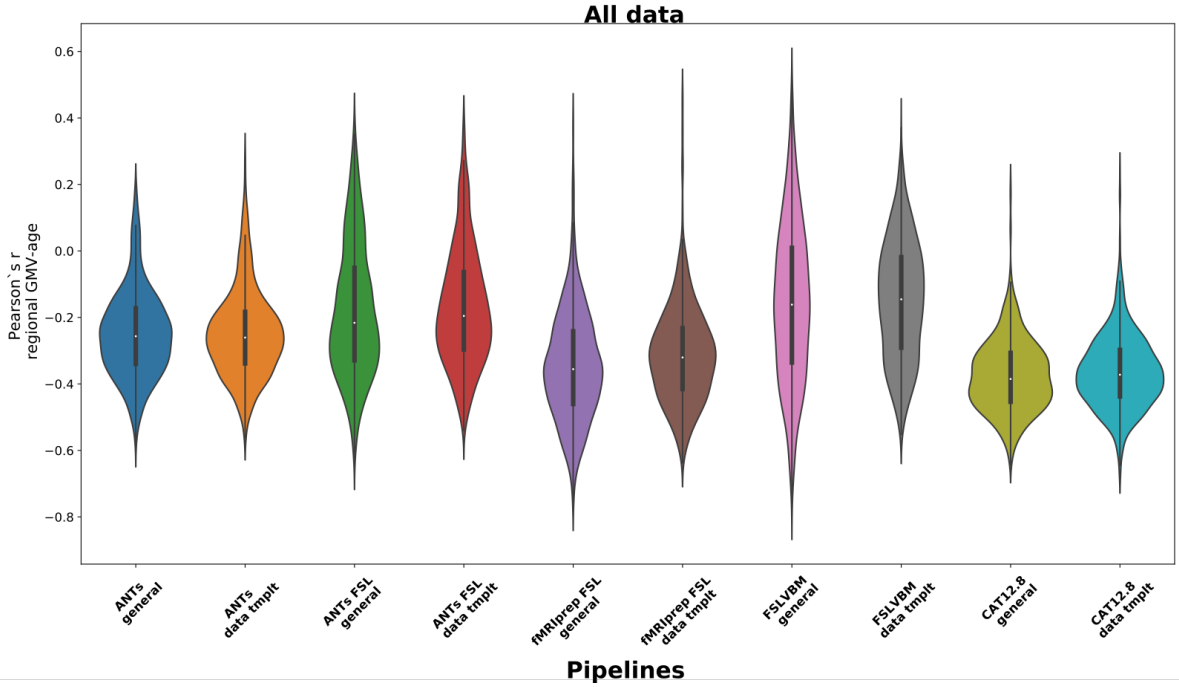

Figure S.14: ROI-age correlations for all pipelines and all datasets. One-way ANOVA showed that there were significant differences between pipelines in ROI-Age correlations.

Scatter plots for each pipeline demonstrate the size of each ROI on the x-axis and the Pearson's r value between ROI and age calculated across subjects. The first figure (Figure S.21) is for the CamCAN dataset, and Figure S.22 is for IXI.

Paired comparisons of ROI-Age correlation values between pipelines

Figure S.15 shows pairplots of age-region correlations across subjects for all data, and Figures S.16, S.17 and S.18 show for eNKI, CamCAM and IXI, respectively.

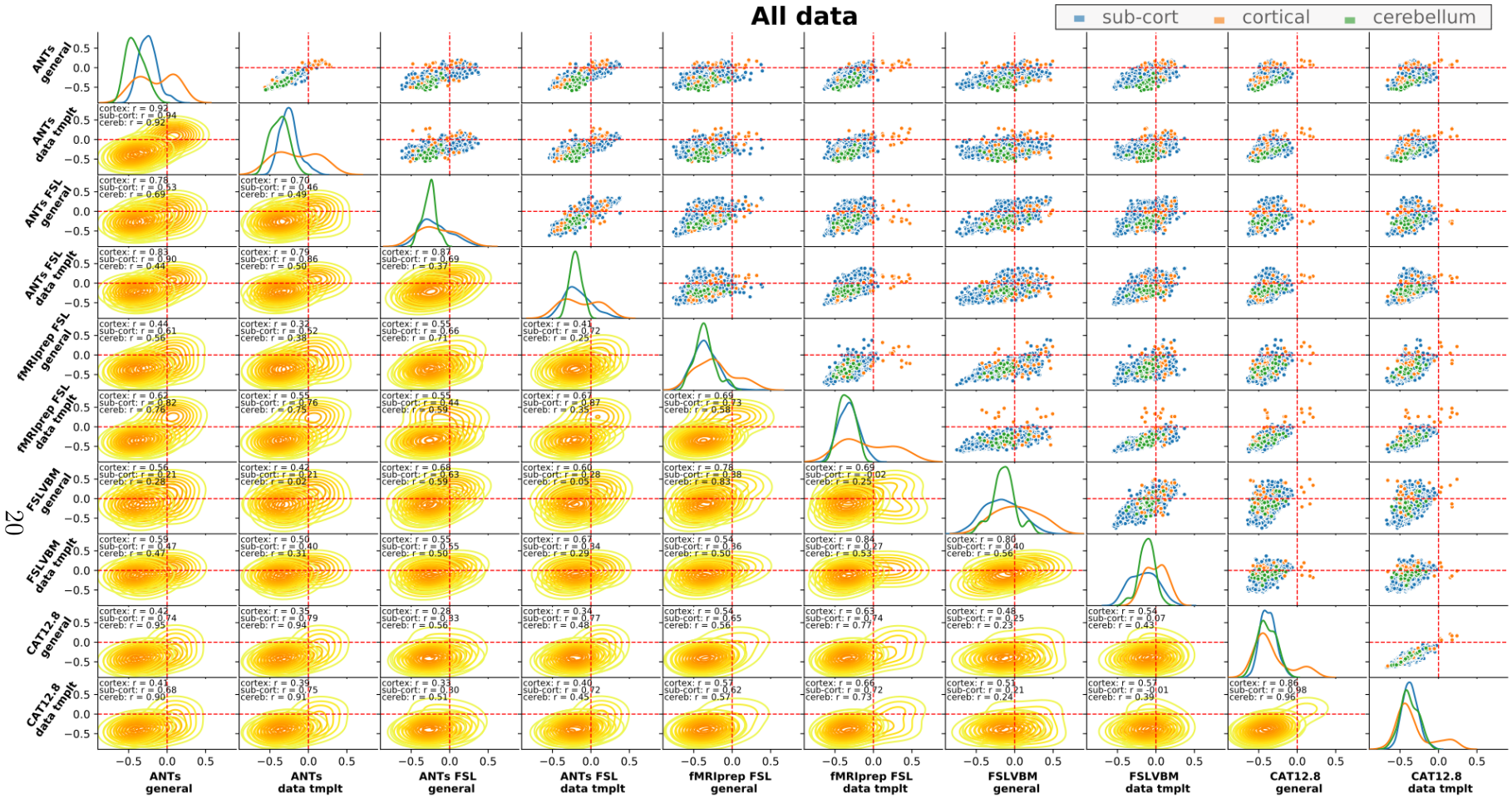
All data
sub-cort
cortical
cerebellum
ANTs general
ANTs data tmplt
ANTs FSL general
ANTs FSL data tmplt
fMRIprep FSL general
fMRIprep FSL data tmplt
FSLVBM general
FSLVBM data tmplt
CAT12.8 general
CAT12.8 data tmplt

Figure S.15:

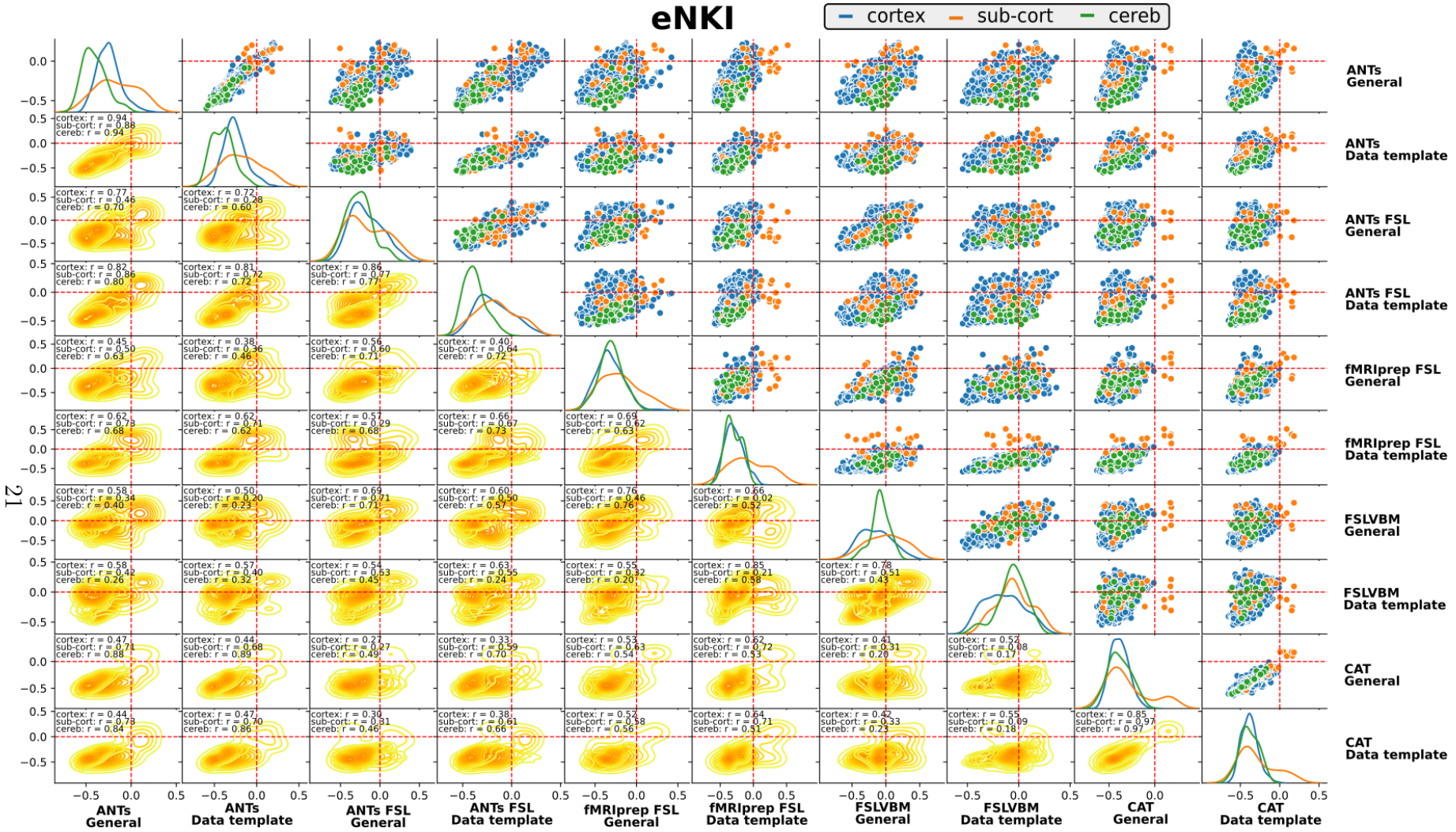
eNKI
cortex
sub-cort
cereb
ANTs General
ANTs Data template
ANTs FSL General
ANTs FSL Data template
fMRIprep FSL General
fMRIprep FSL Data template
FSLVBM General
FSLVBM Data template
CAT General
CAT Data template
21

Figure S.16:

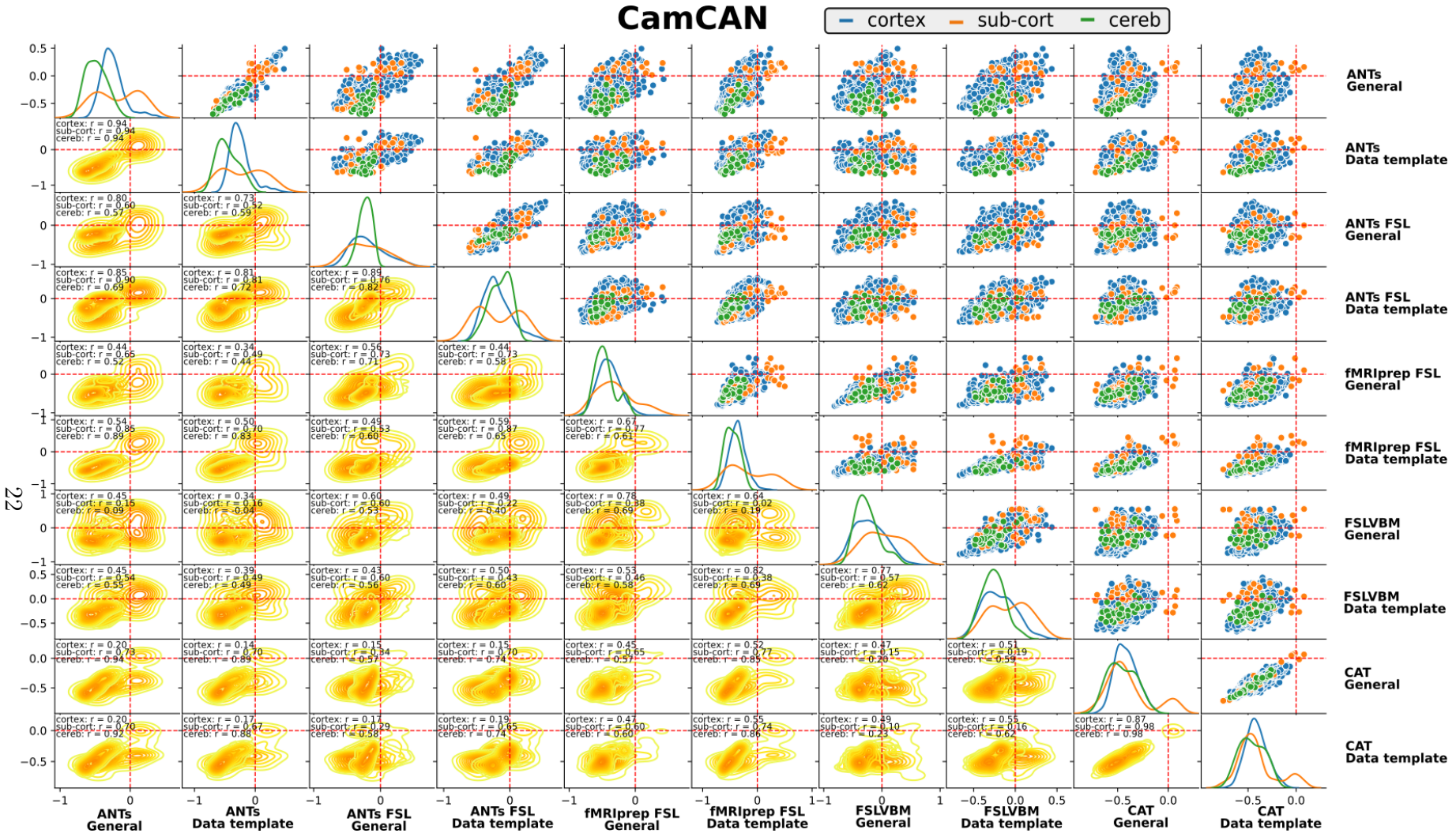
CamCAN
cortex
sub-cort
cereb
ANTs General
ANTs Data template
ANTs FSL General
ANTs FSL Data template
fMRIprep FSL General
fMRIprep FSL Data template
FSLVBM General
FSLVBM Data template
CAT General
CAT Data template

Figure S.17:

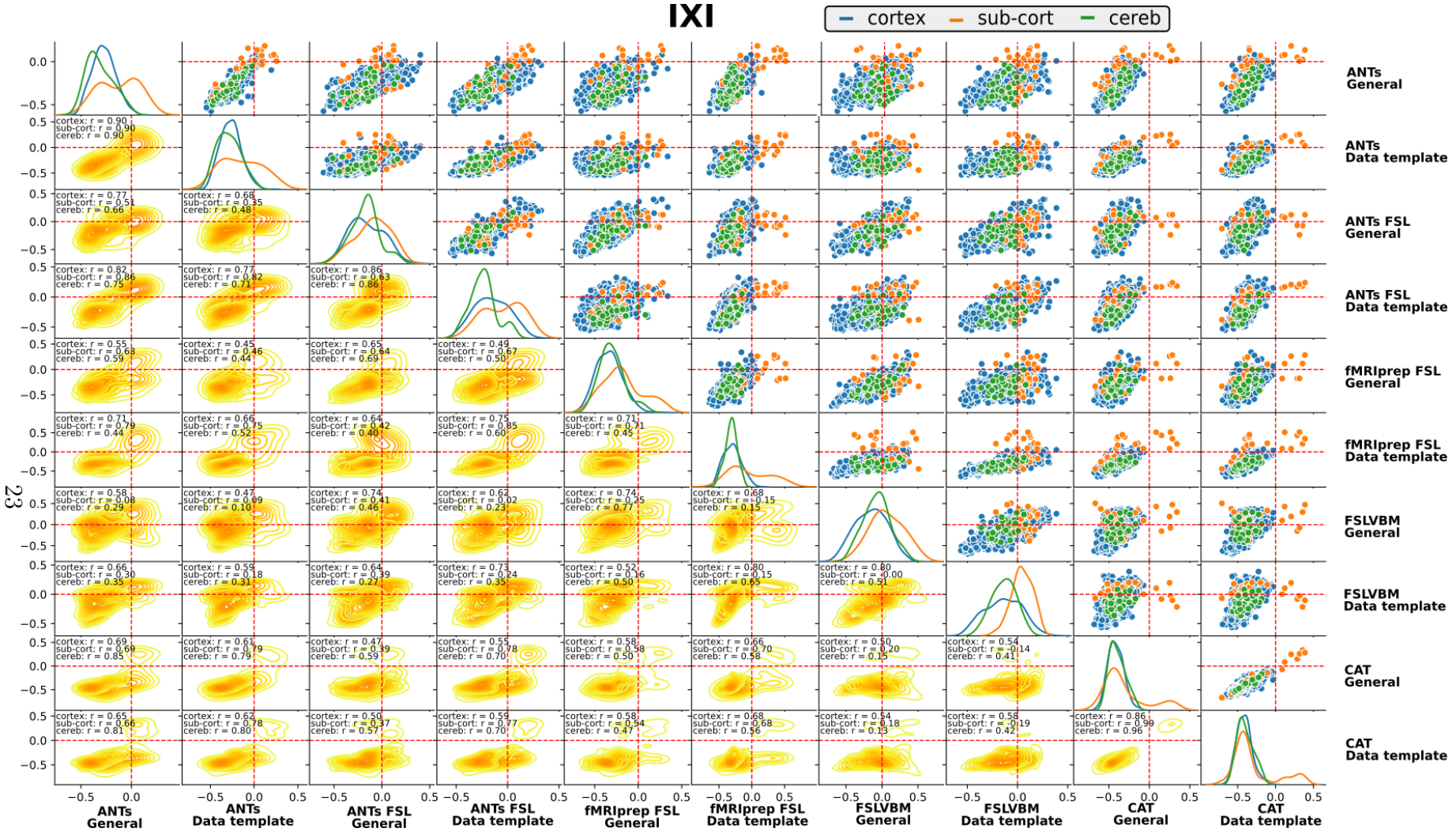
IXI
cortex
sub-cort
cereb
ANTs General
ANTs Data template
ANTs FSL General
ANTs FSL Data template
fMRIprep FSL General
fMRIprep FSL Data template
FSLVBM General
FSLVBM Data template
CAT General
CAT Data template

Figure S.18:

The effect of region size

- Association between the overall similarity among the pipelines (calculated as the median of agreement between pairs) and parcel sizes

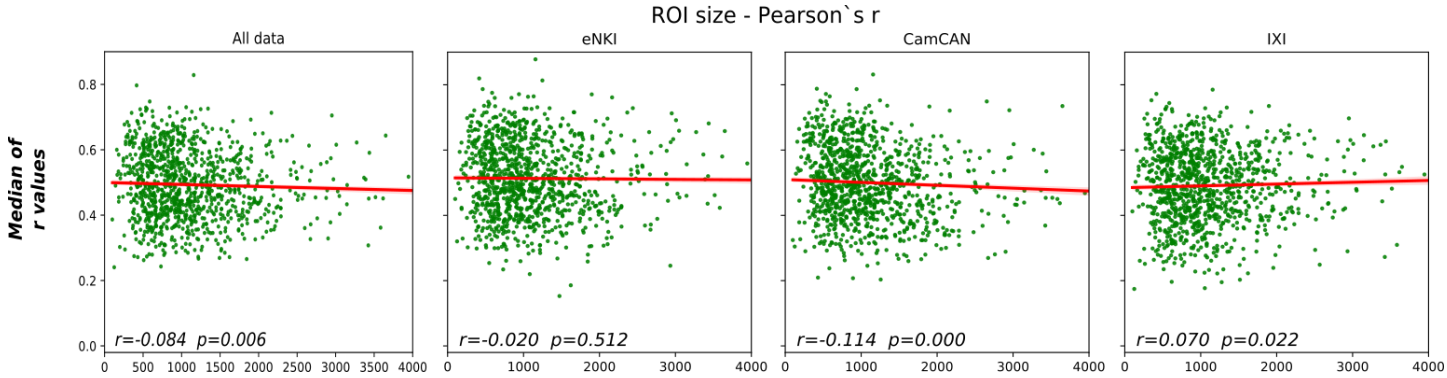

Figure S.19: Median values from all pairs of pipelines of Pearson's r correlations across subjects for all regions plotted against the size of the regions. A nonsignificant correlation was found for the eNKI dataset, and very low correlations were found for the other two datasets.

-The association between the size of regions and the corresponding ROI-age correlation values. CAT appears to have a higher association between the size of each ROI and the correspondence correlation value with age for eNKI (r=-0.128 for both templates) and Cam-CAN. ANTs had similar values but only for the eNKI dataset (=r=-0.121 for general template and -0.125 for data template). For the IXI dataset, FSLVBM with the general template had the highest values (r=0.102). Notably, FSLVBM had a positive correlation when ANTs and CAT had negative values. Figure S.23 provides the same analysis when data from all the datasets are combined.

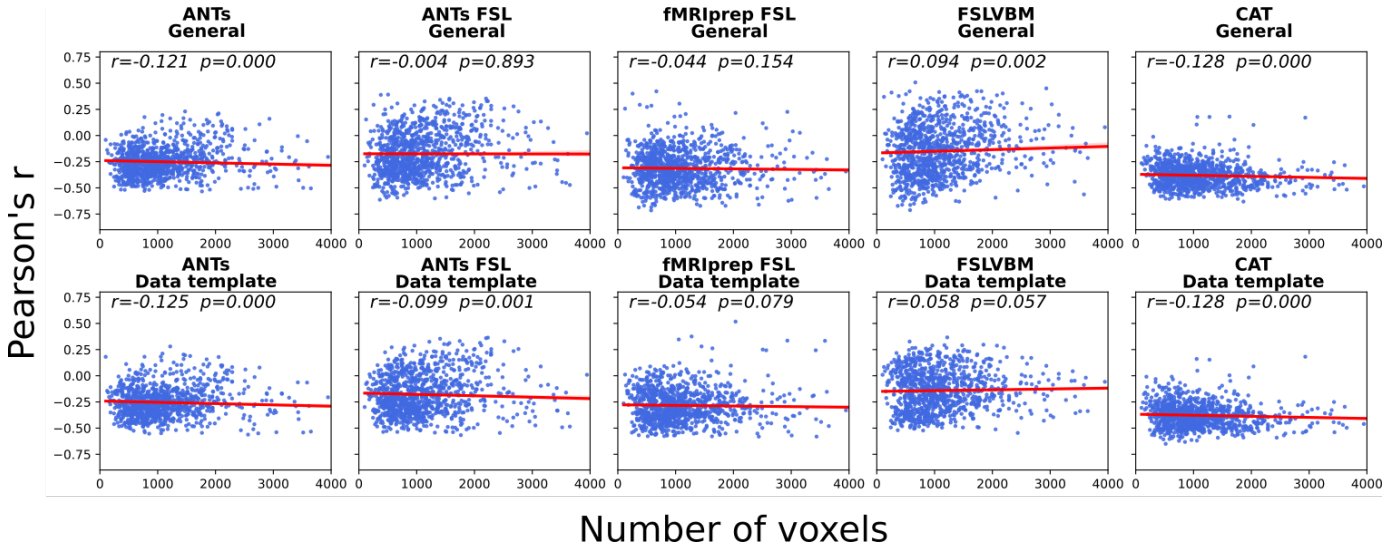

Figure S.20: Scatter plots with the y-axis representing regional correlation to age and the x-axis representing the size of the corresponding ROI, for all pipelines estimated in all datasets. For each pipeline, we estimated the Pearson's r and p values. Red lines represent the linear regression line.

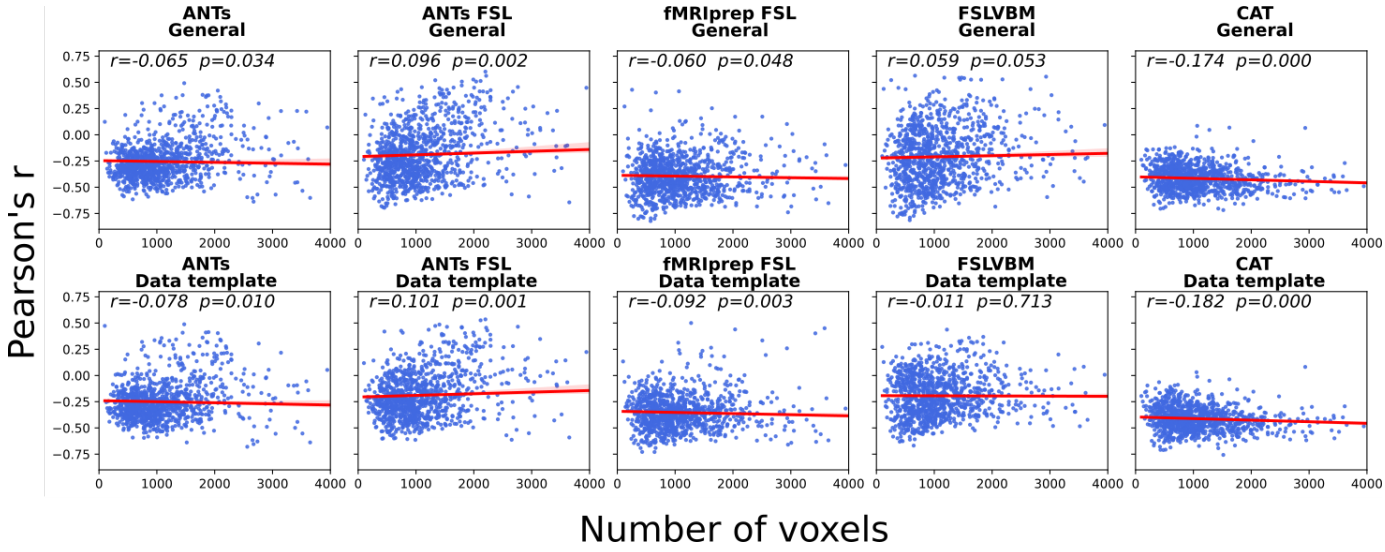

Figure S.21: Scatter plots with the y-axis representing regional correlation to age and the x-axis representing the size of the corresponding ROI for all pipelines estimated in the eNKI dataset. For each pipeline, we estimated the Pearson's r and p values. Red lines represent the linear regression line.

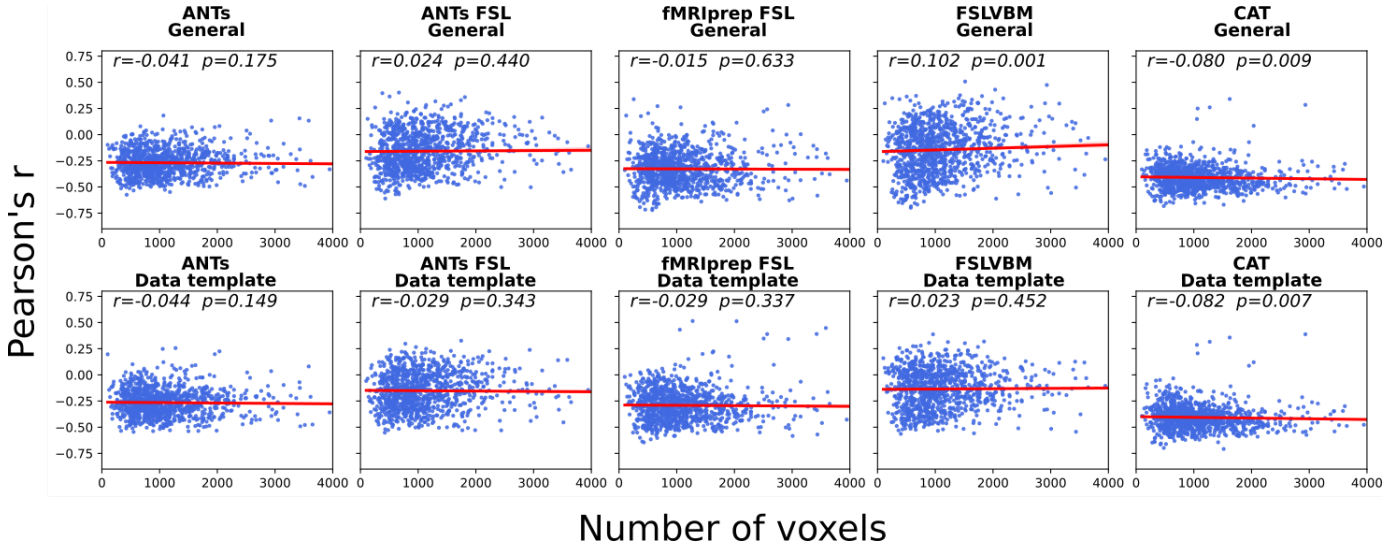

Figure S.22: Scatter plots with the y-axis representing regional correlation to age and the x-axis representing the size of the corresponding ROI for all pipelines estimated in the Cam-CAN dataset. For each pipeline, we estimated the Pearson's r and p values. Red lines represent the linear regression line.

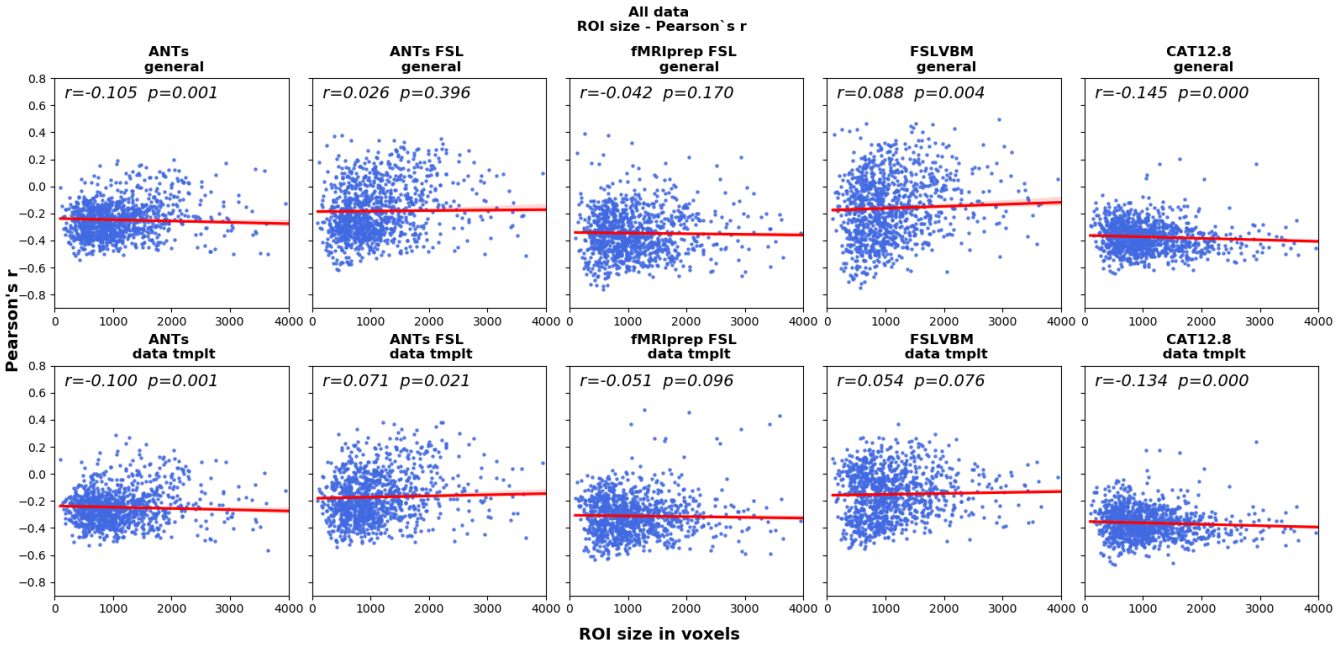

Figure S.23: Scatter plots with the y-axis representing regional correlation to age and the x-axis representing the size of the corresponding ROI for all pipelines estimated in the IXI dataset. For each pipeline, we estimated the Pearson's r and p values. Red lines represent the linear regression line.