Published in final edited form as:
Neuroimage. 2023 October 01; 279: 120292. doi:10.1016/j.neuroimage.2023.120292.

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

Georgios Antonopoulos,
Shammi More,
Federico Raimondo,
Simon B. Eickhoff,
Felix Hoffstaedter,
Kaustubh R. Patil[*]
Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany; 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

[*]Corresponding author at: Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany. k.patil@fz-juelich.de (K.R. Patil).

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.

CRediT authorship contribution statement
Georgios Antonopoulos: Formal analysis, Software, Validation, Visualization, Writing – original draft. **Shammi More:** Data curation, Writing – review & editing. **Federico Raimondo:** Software, Writing – review & editing. **Simon B. Eickhoff:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **Felix Hoffstaedter:** Data curation, Writing – review & editing. **Kaustubh R. Patil:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest
None

Appendix A. Supplementary data
Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neuroimage.2023.120292.

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 (Jovicich et al., 2009). Volumetric analysis of T1w images using voxel-based morphometry (VBM) (Wright et al., 1995; Ashburner and Friston, 2000) 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 (Matsuda, 2013; Lin et al., 2013; Khagi et al., 2021; Colloby et al., 2014; Brewer, 2009) and psychiatric disorders (Yousef et al., 2020).

VBM has been successfully applied to study aging (Good et al., 2001; Tisserand et al., 2004; Bourisly et al., 2015). Recently, prediction of individuals' age based on VBM-derived information has proven to be a validated proxy for brain integrity and overall health (Habes et al., 2016; Koutsouleris et al., 2014-09-01; Cole et al., 2018), and promising for individualized clinical applications (Franke et al., 2010; Jonsson et al., 2019; Koutsouleris et al., 2014-09-01; Su et al., 2011; Varikuti et al., 2018). Brain-age prediction is an important and widely studied topic that aims to estimate the trajectory of healthy brain aging (Franke and Gaser, 2019; Baecker et al., 2021).

To estimate the GMV 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) (Su et al., 2022; Zhang et al., 2021; Li et al., 2018). 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 (Wright et al., 1995), 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 (Peng et al., 2021; Rajagopalan and Pioro, 2015; Dinsdale et al., 2021).

Previous work comparing VBM pipelines indeed provides evidence for differences. A comprehensive comparison between Computational Anatomy Toolbox (CAT) (Gaser and Dahnke, 2016) version 12.7, two FSL-based and a hybrid (still FSL (Smith et al., 2004) dependent) pipelines has shown that the choice of preprocessing pipeline has an impact both in age prediction and sex classification (Zhou et al., 2022). 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 (Friston Karl et al., 2007) yield different outcomes, especially for cortical regions (Popescu et al., 2016). 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 (Rajagopalan and Pioro, 2015). Segmentation and registration as implemented in SPM8 newseg, SPM8 DARTEL (Ashburner, 2007), and FSLVBM were found to have substantial influence on GMV estimates and their relationship to age (Callaert et al., 2014). 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 (Haynes et al., 2020). The CAT12 VBM pipeline was found to perform better in the detection of volumetric alterations in temporal lobe epilepsy compared to the VBM8 toolbox (Matsuda et al., 2012; Farokhian et al., 2017a).

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 (Avants et al., 2008) from the Advance Normalization Toolkit (ANTs) (Avants et al., 2011a) and DARTEL (CAT) were among those with the best performance, with SyN exhibiting the highest consistency across subjects (Klein et al., 2009) as well as being among the most robust to noise, partial volume effects and magnetic field inhomogeneities (Ou et al., 2014). 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 (Johnson et al., 2017). Dadar and colleagues compared six segmentation tools and confirmed significant differences between the tools as well as within-tool differences based on inter-scanner analysis (Dadar and Duchesne, 2020). For brain extraction, although FSL-BET has been reported to have low performance (Johnson et al., 2017), it does not influence subsequent segmentation (Klauschen et al., 2008). A comparison of SPM12, SPM8 and FreeSurfer5.3 (Dale et al., 1999) showed that SPM12 estimates of total intracranial volume (TIV) align better with manual segmentation (Malone et al., 2015). 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 (Katuwal et al., 2016).

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 (Sepulcre et al., 2006; Ceccarelli et al., 2008; Battaglini et al., 2009). 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 (Tustison et al., 2013). 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 (Cole et al., 2017b; Franke and Gaser, 2019; Varikuti et al., 2018; Sowell et al., 2003) 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 (Nooner et al., 2012): 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 (Taylor et al., 2017; Shafto et al., 2014): n = 634 aging individuals without serious psychiatric conditions or cognitive impairment (54.8 ± = 18.4 years, 320 female). IXI (https://brain-development.org/ixi-dataset/): 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 (Gaser and Dahnke, 2016), a popularly used off-the-shelf VBM tool, is a successor of the first VBM pipeline implemented in SPM (Ashburner and Friston, 2000). 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 (Douaud et al., 2007) uses tools from FSL (Smith et al., 2004) and is also widely used. ANTs (Avants et al., 2011a) 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 (Esteban et al., 2019), 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 (Avants et al., 2011a): 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 (Tustison et al., 2010) and then segmented to select intracranial tissues using Atropos-based brain extraction (Avants et al., 2011b). 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 Tustison and Avants (2013). 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 (Avants et al., 2010, 2011a). 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 (Fonov et al., 2009, 2011).

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 (Johnson et al., 2017); 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 (Esteban et al., 2019), which uses ANTs version 2.1.0. Each T1w volume was corrected for intensity nonuniformity (INU) using N4BiasFieldCorrection (Tustison et al., 2010) and skull-stripped using 'antsBrainExtraction.sh' (using the OASIS template). Brain tissue segmentation into CSF, WM and GM was then performed using FSL FAST (Zhang et al., 2001) (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). GAT 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 (Rajapakse et al., 1997) with partial volume model (Tohka et al., 2004). For nonlinear transformation, the geodesic shooting algorithm (Ashburner and Friston, 2011) 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.

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 (Schaefer et al., 1991), 36 subcortical regions from the Brainnetome Atlas (Fan et al., 2016) and 37 cerebellar regions (Buckner et al., 2011). 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 (Cole et al., 2017b; Franke and Gaser, 2019; Varikuti et al., 2018; Sowell et al., 2003) and because of the increasing importance of brain-age as a proxy for overall brain health (Cole et al., 2017b; Cole and Franke, 2017; Won et al., 2020; More et al., 2022). All features were standardized by removing the mean and scaling to unit variance in a cross-validation (CV)-consistent manner (More et al., 2021). We utilized four machine-learning algorithms: relevance vector regression (RVR) (Tipping, 2001), Gaussian process regression (GPR) (Rasmussen and Williams, 2005), least absolute shrinkage and selection operator (LASSO) (Santosa and Symes, 1986; Tibshirani, 1996), and kernel ridge regression (KRR) (Vovk, 2013), in a nested 5-fold CV scheme repeated 5 times (Poldrack et al., 2020). 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 (More et al., 2021) 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 (Finn et al., 2015; Amico and Goñi, 2018). 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 (Amico and Goñi, 2018).

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 (Holm, 1979) 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 (Pedregosa et al., 2011).

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 (Fig. 1). Within-dataset CV performance considerably varied among pipelines (Fig. 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 (Fig. 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).

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 (Fig. 2).

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 (Fig. 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.

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 $r = 0.306$) (Fig. 4).

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 (Fig. 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 (Fig. 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 (Fig. 5(c)).

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

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 (Fig. 5b(i)). ANTs-FSL and fMRIPrep-FSL showed the least similarities in subcortical areas, the occipital lobe and prefrontal cortex (Fig. 5b(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 (Fig. 5b(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.

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 (Fig. 6(a)). Overall, the regional GMV-age correlation was markedly different between the pipelines (Fig. 6).

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 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 (Fig. 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 (see Fig. 7).

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 (Fig. 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 (Fig. 4) and identification using the subject-specific multivariate

GMV pattern (Fig. 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 Fig. 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 (Fig. 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 (Fig. 1), similar to what has been previously reported for voxel-level analysis and using CAT and FSL-based pipelines (Zhou et al., 2022). Our age-prediction accuracy for CAT and fMRIPrep-FSL are comparable to previous reports, considering our dataset size and the wide age range (Eickhoff et al., 2021; Cole et al., 2017a). 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 (Callaert et al., 2014; Popescu et al., 2016; Rajagopalan and Pioro, 2015). 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 (Pomponio et al., 2020; Radua et al., 2020), 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 (Fig. 2). This result has important implications for data sharing and privacy issues

(White et al., 2022). 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 $r = 0.76$) (Fig. 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 (Fig. 5 (b)).

Differences between registration algorithms have been reported (Ou et al., 2014). 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 (Fig. 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 (Haynes et al., 2020). Effectively, using a data-template imposes higher similarity between the subjects' images, which we also observed for some pipelines (Fig. 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 (Fig. 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 (Fig. 5(b)).

The identification results (Fig. 2) were very similar to the pairwise similarity estimated using Pearson's correlation (Fig. 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 (Fig. 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 (Rao et al., 2022; Kruggel et al., 2010; Valverde et al., 2015).

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 (Peng et al., 2021). 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 (Jonsson et al., 2019). 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 (Fig. 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 (Farokhian et al., 2017b; Gennatas et al., 2017; Koops et al., 2020). 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 (Franke et al., 2010; Peng et al., 2021), 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study is supported by Deutsche Forschungsgemeinschaft, Germany (DFG, PA 3634/1-1 and EI 816/21-1), the National Institute of Mental Health, United States (R01-MH074457), the Helmholtz Portfolio Theme “Supercomputing and Modelling for the Human Brain”, the European Union’s Horizon 2020 Research and Innovation Program grant agreement 945539 (HBP SGA3) and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project-ID 431549029 – SFB 1451 project B05.

Data availability

All data used are from open datasets available online (maybe upon request/registration).

References

Amico Enrico, Goñi Joaquín, 2018. The quest for identifiability in human functional connectomes. Sci. Rep 8 (1), 8254. [PubMed: 29844466]

Ashburner John, 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95–113. [PubMed: 17761438]

Ashburner John, Friston Karl J., 2000. Voxel-based morphometry—The methods. NeuroImage 11 (6), 805–821. [PubMed: 10860804]

Ashburner John, Friston Karl J., 2011. Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. NeuroImage 55 (3), 954–967. [PubMed: 21216294]

Avants Brian B., Epstein CL, Grossman M, Gee JC, 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal 12 (1), 26–41. [PubMed: 17659998]

Avants Brian B., Tustison Nicholas J., Song Gang, Cook Philip A., Klein Arno, Gee James C., 2011a. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54 (3), 2033–2044. [PubMed: 20851191]

Avants Brian B., Tustison Nicholas J., Wu Jue, Cook Philip A., Gee James C., 2011b. An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9 (4), 381–400. [PubMed: 21373993]

Avants Brian B., Yushkevich Paul, Pluta John, Minkoff David, Korczykowski Marc, Detre John, Gee James C., 2010. The optimal template effect in hippocampus studies of diseased populations. NeuroImage 49 (3).

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

Battaglini Marco, Giorgio Antonio, Stromillo Maria L., Bartolozzi Maria L., Guidi Leonello, Federico Antonio, De Stefano Nicola, 2009. Voxel-wise assessment of progression of regional brain atrophy in relapsing-remitting multiple sclerosis. J. Neurol. Sci 282 (1–2), 55–60. [PubMed: 19286193]

Bourisly Ali K, El-Beltagi Ahmed, Cherian Jigi, Gejo Grace, Al-Jazzaf Abrar, Ismail Mohammad, 2015. 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. Neuroradiol. J 28 (5), 450–459, Publisher: SAGE Publications Ltd. [PubMed: 26306927]

Brewer James B., 2009. Fully-automated volumetric MRI with normative ranges: Translation to clinical practice. Behav. Neurol 21 (1–2), 21–28. [PubMed: 19847042]

Buckner Randy L., Krienen Fenna M., Castellanos Angela, Diaz Julio C., Yeo B.T. Thomas, 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. J. Neurophysiol 106 (5), 2322–2345. [PubMed: 21795627]

Callaert Dorothee V., Ribbens Annemie, Maes Frederik, Swinnen Stephan P., Wenderoth Nicole, 2014. Assessing age-related gray matter decline with voxel-based morphometry depends significantly on segmentation and normalization procedures. Front. Aging Neurosci 6.

Ceccarelli Antonia, Rocca Maria A., Pagani Elisabetta, Colombo Bruno, Martinelli Vittorio, Comi Giancarlo, Filippi Massimo, 2008. A voxel-based morphometry study of grey matter loss in MS patients with different clinical phenotypes. NeuroImage 42 (1), 315–322. [PubMed: 18501636]

Cole James H., Annus Tiina, Wilson Liam R., Remtulla Ridhaa, Hong Young T., Fryer Tim D., Acosta-Cabronero Julio, Cardenas-Blanco Arturo, Smith Robert, Menon David K., Zaman Shahid H., Nestor Peter J., Holland Anthony J., 2017a. Brain-predicted age in down syndrome is associated with beta amyloid deposition and cognitive decline. Neurobiol. Aging 56, 41–49. [PubMed: 28482213]

Cole James H., Franke Katja, 2017. Predicting age using neuroimaging: Innovative brain ageing biomarkers. Trends Neurosci. 40 (12), 681–690. [PubMed: 29074032]

Cole James H., Poudel Rudra P.K., Tsagkrasoulis Dimosthenis, Caan Matthan W.A., Steves Claire, Spector Tim D., Montana Giovanni, 2017b. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124. [PubMed: 28765056]

Cole James H., Ritchie SJ, Bastin ME, Valdés Hernández MC, Muñoz Maniega S, Royle N, Corley J, Pattie A, Harris SE, Zhang Q, Wray NR, Redmond P, Marioni RE, Starr JM, Cox SR, Wardlaw JM, Sharp DJ, Deary IJ, 2018. Brain age predicts mortality. Mol. Psychiatry 23 (5), 1385–1392. [PubMed: 28439103]

Colloby Sean. J., O'Brien John. T., Taylor John-Paul, 2014. Patterns of cerebellar volume loss in dementia with lewy bodies and alzheimer's disease: A VBM-DARTEL study. Psychiatry Res.: Neuroimaging 223 (3), 187–191.

Dadar Mahsa, Duchesne Simon, 2020. Reliability assessment of tissue classification algorithms for multi-center and multi-scanner data. NeuroImage 217, 116928. [PubMed: 32413463]

Dale Anders M., Fischl Bruce, Sereno Martin I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage 9 (2), 179–194. [PubMed: 9931268]

Dinsdale Nicola K., Bluemke Emma, Smith Stephen M., Arya Zobair, Vidaurre Diego, Jenkinson Mark, Namburete Ana I.L., 2021. Learning patterns of the ageing brain in MRI using deep convolutional networks. NeuroImage 224, 117401. [PubMed: 32979523]

Douaud Gwenaëlle, Smith Stephen, Jenkinson Mark, Behrens Timothy, Johansen-Berg Heidi, Vickers John, James Susan, Voets Natalie, Watkins Kate, Matthews Paul M., James Anthony, 2007. Anatomically related grey and white matter abnormalities in adolescent-onset Schizophrenia. Brain: J. Neurol 130 (Pt 9), 2375–2386.

Eickhoff Claudia R, Hoffstaedter Felix, Caspers Julian, Reetz Kathrin, Mathys Christian, Dogan Imis, Amunts Katrin, Schnitzler Alfons, Eickhoff Simon B, 2021. Advanced brain ageing in Parkinson's disease is related to disease duration and individual impairment. Brain Commun. 3 (3), fcab191. [PubMed: 34541531]

Esteban Oscar, Markiewicz Christopher J., Blair Ross W., Moodie Craig A., Isik A. Ilkay, Erramuzpe Asier, Kent James D., Goncalves Mathias, DuPre Elizabeth, Snyder Madeleine, Oya Hiroyuki, Ghosh Satrajit S., Wright Jessey, Durnez Joke, Poldrack Russell A., Gorgolewski Krzysztof J., 2019. Fmriprep: a robust preprocessing pipeline for functional MRI. Nature Methods 16 (1), 111–116. [PubMed: 30532080]

Fan Lingzhong, Li Hai, Zhuo Junjie, Zhang Yu, Wang Jiaojian, Chen Liangfu, Yang Zhengyi, Chu Congying, Xie Sangma, Laird Angela R., Fox Peter T., Eickhoff Simon B., Yu Chunshui, Jiang Tianzi, 2016. The human brainnetome atlas: A new brain Atlas based on connectional architecture. Cerebral Cortex 26 (8), 3508–3526. [PubMed: 27230218]

Farokhian Farnaz, Beheshti Iman, Sone Daichi, Matsuda Hiroshi, 2017a. Comparing CAT12 and VBM8 for detecting brain morphological abnormalities in temporal lobe epilepsy. Front. Neurol 8, 428. [PubMed: 28883807]

Farokhian Farnaz, Yang Chunlan, Beheshti Iman, Matsuda Hiroshi, Wu Shuicai, 2017b. Age-related gray and white matter changes in normal adult brains. Aging Dis. 8 (6), 899. [PubMed: 29344423]

Finn Emily S., Shen Xilin, Scheinost Dustin, Rosenberg Monica D., Huang Jessica, Chun Marvin M., Papademetris Xenophon, Constable R. Todd, 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neurosci. 18 (11), 1664–1671. [PubMed: 26457551]

Fonov Vladimir, Evans Alan C., Botteron Kelly, Almli C. Robert, McKinstry Robert C., Collins D. Louis, 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54 (1), 313–327. [PubMed: 20656036]

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

Franke Katja, Gaser Christian, 2019. Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained? Front. Neurol 10, 789. [PubMed: 31474922]

Franke Katja, Ziegler Gabriel, Klöppel Stefan, Gaser Christian, 2010. 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. [PubMed: 20070949]

Friston Karl J, Ashburner John T, Kiebel Stefan J, Nichols Thomas E, Penny William D, 2007. Statistical Parametric Mapping: The Analysis of Functional Brain Images, first ed. Academic Press.

Gaser Christian, Dahnke R, 2016. CAT-A computational anatomy toolbox for the analysis of structural MRI data.

Gennatas Efstathios D., Avants Brian B., Wolf Daniel H., Satterthwaite Theodore D., Ruparel Kosha, Ciric Rastko, Hakonarson Hakon, Gur Raquel E., Gur Ruben C., 2017. Age-related effects and sex differences in gray matter density, volume, mass, and cortical thickness from childhood to Young adulthood. J. Neurosci 37 (20), 5065–5073. [PubMed: 28432144]

Good Catriona D., Johnsrude Ingrid S., Ashburner John, Henson Richard N.A., Friston Karl J., Frackowiak Richard S.J., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. NeuroImage 14 (1), 21–36. [PubMed: 11525331]

Habes M, Janowitz D, Erus G, Toledo JB, Resnick SM, Doshi J, Van der Auwera S, Wittfeld K, Hegenscheid K, Hosten N, Biffar R, Homuth G, Völzke H, Grabe HJ, Hoffmann W, Davatzikos C, 2016. Advanced brain aging: relationship with epidemiologic and genetic risk factors, and overlap with alzheimer disease atrophy patterns. Transl. Psychiatry 6 (4), e775. [PubMed: 27045845]

Haynes Logan, Ip Amanda, Cho Ivy Y.K., Dimond Dennis, Rohr Christiane S., Bagshawe Mercedes, Dewey Deborah, Lebel Catherine, Bray Signe, 2020. Grey and white matter volumes in early childhood: A comparison of voxel-based morphometry pipelines. Develop. Cogn. Neurosci 46.

Holm Sture, 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat 6 (2), 65–70.

Johnson Eileanoir B., Gregory Sarah, Johnson Hans J., Durr Alexandra, Leavitt Blair R., Roos Raymund A., Rees Geraint, Tabrizi Sarah J., Scahill Rachael I., 2017. Recommendations for the use of automated gray matter segmentation tools: Evidence from Huntington's disease. Front. Neurol 8, 519. [PubMed: 29066997]

Jonsson BA, Bjornsdottir G, Thorgeirsson TE, Ellingsen LM, Walters G. Bragi, Gudbjartsson DF, Stefansson H, Stefansson K, Ulfarsson MO, 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nature Commun. 10 (1), 5409. [PubMed: 31776335]

Jovicich Jorge, Czanner Silvester, Han Xiao, Salat David, van der Kouwe Andre, Quinn Brian, Pacheco Jenni, Albert Marilyn, Killiany Ronald, Blacker Deborah, Maguire Paul, Rosas Diana, Makris Nikos, Gollub Randy, Dale Anders, Dickerson Bradford C., Fischl Bruce, 2009. 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. [PubMed: 19233293]

Katuwal Gajendra J., Baum Stefi A., Cahill Nathan D., Dougherty Chase C., Evans Eli, Evans David W., Moore Gregory J., Michael Andrew M., 2016. Inter-method discrepancies in brain volume estimation may drive inconsistent findings in Autism. Front. Neurosci 10, 439. [PubMed: 27746713]

Khagi Bijen, Lee Kun Ho, Choi Kyu Yeong, Lee Jang Jae, Kwon Goo-Rak, Yang Hee-Deok, 2021. VBM-based Alzheimer's disease detection from the region of interest of T1 MRI with supportive Gaussian smoothing and a Bayesian regularized neural network. Appl. Sci 11 (13), 6175.

Klauschen Frederick, Goldman Aaron, Barra Vincent, Meyer-Lindenberg Andreas, Lundervold Arvid, 2008. Evaluation of automated brain MR image segmentation and volumetry methods. Hum. Brain Map 30 (4), 1310–1327.

Klein Arno, Andersson Jesper, Ardekani Babak A., Ashburner John, Avants Brian B., Chiang Ming-Chang, Christensen Gary E., Collins D. Louis, Gee James, Hellier Pierre, Song Joo Hyun, Jenkinson Mark, Lepage Claude, Rueckert Daniel, Thompson Paul, Vercauteren Tom, Woods Roger P., Mann J. John, Parsey Ramin V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46 (3), 786–802. [PubMed: 19195496]

Koops Elouise A., de Kleine Emile, van Dijk Pim, 2020. Gray matter declines with age and hearing loss, but is partially maintained in tinnitus. Sci. Rep 10 (1), 21801, Number: 1 Publisher: Nature Publishing Group. [PubMed: 33311548]

Koutsouleris Nikolaos, Davatzikos Christos, Borgwardt Stefan, Gaser Christian, Bottlender Ronald, Frodl Thomas, Falkai Peter, Riecher-Rössler Anita, Möller Hans-Jürgen, Reiser Maximilian, Pantelis Christos, Meisenzahl Eva, 2014-09-01. Accelerated brain aging in Schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. Schizophrenia Bull. 40 (5), 1140–1153.

Kruggel Frithjof, Turner Jessica, Muftuler L. Tugan, 2010. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. NeuroImage 49 (3), 2123–2133. [PubMed: 19913626]

Li Meng, Yan Jianhao, Li Shumei, Wang Tianyue, Wen Hua, Yin Yi, Fu Shishun, Zeng Luxian, Tian Junzhang, Jiang Guihua, 2018. Altered gray matter volume in primary insomnia patients: a DARTEL-VBM study. Brain Imaging Behav. 12 (6), 1759–1767. [PubMed: 29411240]

Lin Ching-Hung, Chen Chun-Ming, Lu Ming-Kuei, Tsai Chon-Haw, Chiou Jin-Chern, Liao Jan-Ray, Duann Jeng-Ren, 2013. VBM reveals brain volume differences between Parkinson's disease and essential tremor patients. Front. Hum. Neurosci 7.

Malone Ian B., Leung Kelvin K., Clegg Shona, Barnes Josephine, Whitwell Jennifer L., Ashburner John, Fox Nick C., Ridgway Gerard R., 2015. Accurate automatic estimation of total intracranial volume: A nuisance variable with less nuisance. Neuroimage 104, 366–372. [PubMed: 25255942]

Matsuda Hiroshi, 2013. Voxel-based morphometry of brain MRI in normal aging and Alzheimer's Disease. Aging Dis. 4 (1), 29. [PubMed: 23423504]

Matsuda H, Mizumura S, Nemoto K, Yamashita F, Imabayashi E, Sato N, Asada T, 2012. 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: Am. J. Neuroradiol 33 (6), 1109–1114. [PubMed: 22300935]

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

More Shammi, Eickhoff Simon B., Caspers Julian, Patil Kaustubh R., 2021. Confound removal and normalization in practice: A neuroimaging based sex prediction case study. In: Dong Yuxiao, Ifrim Georgiana, Mladenić Dunja, Saunders Craig, Van Hoecke Sofie (Eds.), Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 3–18.

Nooner Kate Brody, Colcombe Stanley J., Tobe Russell H., Mennes Maarten, Benedict Melissa M., Moreno Alexis L., Panek Laura J., Brown Shaquanna, Zavitz Stephen T., Li Qingyang, Sikka Sharad, Gutman David, Bangaru Saroja, Schlachter Rochelle Tziona, Kamiel Stephanie M., Anwar Ayesha R., Hinz Caitlin M., Kaplan Michelle S., Rachlin Anna B., Adelsberg Samantha, Cheung Brian, Khanuja Ranjit, Yan Chaogan, Craddock Cameron C., Calhoun Vincent, Courtney William, King Margaret, Wood Dylan, Cox Christine L., Kelly A.M. Clare, Di Martino Adriana, Petkova Eva, Reiss Philip T., Duan Nancy, Thomsen Dawn, Biswal Bharat, Coffey Barbara, Hoptman Matthew J., Javitt Daniel C., Pomara Nunzio, Sidtis John J., Koplewicz Harold S., Castellanos Francisco Xavier, Leventhal Bennett L., Milham Michael P., 2012. The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. Front. Neurosci 6, 152. [PubMed: 23087608]

Ou Yangming, Akbari Hamed, Bilello Michel, Da Xiao, Davatzikos Christos, 2014. Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. IEEE Trans. Med. Imaging 33 (10), 2039–2065. [PubMed: 24951685]

Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, Blondel Mathieu, Prettenhofer Peter, Weiss Ron, Dubourg Vincent, Vanderplas Jake, Passos Alexandre, Cournapeau David, Brucher Matthieu, Perrot Matthieu, Duchesnay Édouard, 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res 12 (85), 2825–2830.

Peng Han, Gong Weikang, Beckmann Christian F., Vedaldi Andrea, Smith Stephen M., 2021. Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal 68, 101871. [PubMed: 33197716]

Poldrack Russell A., Huckins Grace, Varoquaux Gael, 2020. Establishment of best practices for evidence for prediction: A review. JAMA Psychiatry 77 (5), 534–540. [PubMed: 31774490]

Pomponio Raymond, Erus Guray, Habes Mohamad, Doshi Jimit, Srinivasan Dhivya, Mamourian Elizabeth, Bashyam Vishnu, Nasrallah Ilya M., Satterthwaite Theodore D., Fan Yong, Launer Lenore J., Masters Colin L., Maruff Paul, Zhuo Chuanjun, Völzke Henry, Johnson Sterling C., Fripp Jurgen, Koutsouleris Nikolaos, Wolf Daniel H., Gur Raquel, Gur Ruben, Morris John, Albert Marilyn S., Grabe Hans J., Resnick Susan M., Bryan R. Nick, Wolk David A., Shinohara Russell T., Shou Haochang, Davatzikos Christos, 2020. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage 208, 116450. [PubMed: 31821869]

Popescu Veronica, Schoonheim Menno M., Versteeg Adriaan, Chaturvedi Nimisha, Jonker Marianne, Menezes Renee Xavier de, Garre Francisca Gallindo, Uitdehaag Bernard M.J., Barkhof Frederik, Vrenken Hugo, 2016. Grey matter atrophy in multiple sclerosis: Clinical interpretation depends on choice of analysis method. PLoS One 11 (1), e0143942. [PubMed: 26745873]

Radua Joaquim, Vieta Eduard, Shinohara Russell, Kochunov Peter, Quidé Yann, Green Melissa J., Weickert Cynthia S., Weickert Thomas, Bruggemann Jason, Kircher Tilo, Nenadić Igor, Cairns Murray J., Seal Marc, Schall Ulrich, Henskens Frans, Fullerton Janice M., Mowry Bryan, Pantelis Christos, Lenroot Rhoshel, Cropley Vanessa, Loughland Carmel, Scott Rodney, Wolf Daniel, Satterthwaite Theodore D., Tan Yunlong, Sim Kang, Piras Fabrizio, Spalletta Gianfranco, Banaj Nerisa, Pomarol-Clotet Edith, Solanes Aleix, Albajes-Eizagirre Anton, Canales-Rodríguez Erick J., Sarro Salvador, Di Giorgio Annabella, Bertolino Alessandro, Stäblein Michael, Oertel Viola, Knöchel Christian, Borgwardt Stefan, du Plessis Stefan, Yun Je-Yeon, Kwon Jun Soo, Dannlowski Udo, Hahn Tim, Grotegerd Dominik, Alloza Clara, Arango Celso, Janssen Joost, Dí az Caneja Covadonga, Jiang Wenhao, Calhoun Vince, Ehrlich Stefan, Yang Kun, Cascella Nicola G., Takayanagi Yoichiro, Sawa Akira, Tomyshev Alexander, Lebedeva Irina, Kaleda Vasily, Kirschner Matthias, Hoschl Cyril, Tomecek David, Skoch Antonin, van Amelsvoort Therese, Bakker Geor, James Anthony, Preda Adrian, Weideman Andrea, Stein Dan J., Howells Fleur, Uhlmann Anne, Temmingh Henk, López-Jaramillo Carlos, Díaz-Zuluaga Ana, Fortea Lydia, Martinez-Heras Eloy, Solana Elisabeth, Llufriu Sara, Jahanshad Neda, Thompson Paul, Turner Jessica, van Erp Theo, Glahn David, Pearlson Godfrey, Hong Elliot, Krug Axel, Carr Vaughan, Tooney Paul, Cooper Gavin, Rasser Paul, Michie Patricia, Catts Stanley, Gur Raquel, Gur Ruben, Yang Fude, Fan Fengmei, Chen Jingxu, Guo Hua, Tan Shuping, Wang Zhiren, Xiang Hong, Piras Federica, Assogna Francesca, Salvador Raymond, McKenna Peter, Bonvino Aurora, King Margaret, Kaiser Stefan, Nguyen Dana, Pineda-Zapata Julian, 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. NeuroImage 218, 116956. [PubMed: 32470572]

Rajagopalan Venkateswaran, Pioro Erik P., 2015. Disparate Voxel Based Morphometry (VBM) results between SPM and FSL softwares in ALS patients with frontotemporal dementia: which VBM results to consider? BMC Neurol. 15, 32. [PubMed: 25879588]

Rajapakse JC, Giedd JN, Rapoport JL, 1997. Statistical approach to segmentation of single-channel cerebral MR images. IEEE Trans. Med. Imaging 16 (2), 176–186. [PubMed: 9101327]

Rao Vishwanatha M., Wan Zihan, Ma David J., Lee Pin-Yu, Tian Ye, Laine Andrew F., Guo Jia, 2022. Improving across-dataset brain tissue segmentation using transformer. arXiv:2201.08741 [cs, eess].

Rasmussen Carl Edward, Williams Christopher K.I., 2005. In: Bach Francis (Ed.), Gaussian Processes for Machine Learning. In: Adaptive Computation and Machine Learning series, MIT Press, Cambridge, MA, USA.

Santosa Fadil, Symes William W., 1986. Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comput 7 (4), 1307–1330.

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

Sepulcre Jorge, Sastre-Garriga Jaume, Cercignani Mara, Ingle Gordon T., Miller David H., Thompson Alan J., 2006. Regional gray matter atrophy in early primary progressive multiple sclerosis: A voxel-based morphometry study. Arch. Neurol 63 (8), 1175–1180. [PubMed: 16908748]

Shafto Meredith A., Tyler Lorraine K., Dixon Marie, Taylor Jason R., Rowe James B., Cusack Rhodri, Calder Andrew J., Marslen-Wilson William D., Duncan John, Dalgleish Tim, Henson Richard N., Brayne Carol, Matthews Fiona E., 2014. The Cambridge centre for ageing and neuroscience (cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC Neurol. 14, 204. [PubMed: 25412575]

Smith Stephen M., Jenkinson Mark, Woolrich Mark W., Beckmann Christian F., Behrens Timothy E.J., Johansen-Berg Heidi, Bannister Peter R., Luca Marilena De, Drobnjak Ivana, Flitney David E., Niazy Rami K., Saunders James, Vickers John, Zhang Yongyue, Stefano Nicola De, Brady J. Michael, Matthews Paul M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1, S208–S219. [PubMed: 15501092]

Sowell Elizabeth R., Peterson Bradley S., Thompson Paul M., Welcome Suzanne E., Henkenius Amy L., Toga Arthur W., 2003. Mapping cortical change across the human life span. Nature Neurosci. 6 (3), 309–315. [PubMed: 12548289]

Su Longfei, Wang Lubin, Shen Hui, Hu Dewen, 2011. Age-related classification and prediction based on MRI: A sparse representation method. Procedia Environ. Sci 8, 645–652.

Su Ting, Zhu Pei-Wen, Li Biao, Shi Wen-Qing, Lin Qi, Yuan Qing, Jiang Nan, Pei Chong-Gang, Shao Yi, 2022. Gray matter volume alterations in patients with strabismus and amblyopia: voxel-based morphometry study. Sci. Rep 12 (1), 458. [PubMed: 35013442]

Taylor Jason R., Williams Nitin, Cusack Rhodri, Auer Tibor, Shafto Meredith A., Dixon Marie, Tyler Lorraine K., Cam-Can, null, Henson Richard N., 2017. 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. [PubMed: 26375206]

Tibshirani Robert, 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Stat. Methodol 58 (1), 267–288.

Tipping Michael E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res 1 (Jun), 211–244.

Tisserand Danielle J., van Boxtel Martin P.J., Pruessner Jens C., Hofman Paul, Evans Alan C., Jolles Jelle, 2004. 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. [PubMed: 15115735]

Tohka Jussi, Zijdenbos Alex, Evans Alan, 2004. Fast and robust parameter estimation for statistical partial volume models in brain MRI. NeuroImage 23 (1), 84–97. [PubMed: 15325355]

Tustison Nicholas James, Avants Brian B., 2013. Explicit B-spline regularization in diffeomorphic image registration. Front. Neuroinform 7.

Tustison Nicholas J., Avants Brian B., Cook Philip A., Zheng Yuanjie, Egan Alexander, Yushkevich Paul A., Gee James C., 2010. N4ITK: Improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. [PubMed: 20378467]

Tustison Nicholas J., Johnson Hans J., Rohlfing Torsten, Klein Arno, Ghosh Satrajit S., Ibanez Luis, Avants Brian B., 2013. Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences. Front. Neurosci 7.

Valverde Sergi, Oliver Arnau, Cabezas Mariano, Roura Eloy, Lladó Xavier, 2015. Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. J. Magn. Reson. Imaging 41 (1), 93–101. [PubMed: 24459099]

Varikuti Deepthi P., Genon Sarah, Sotiras Aristeidis, Schwender Holger, Hoffstaedter Felix, Patil Kaustubh R., Jockwitz Christiane, Caspers Svenja, Moebus Susanne, Amunts Katrin, Davatzikos Christos, Eickhoff Simon B., 2018. Evaluation of non-negative matrix factorization of grey matter in age prediction. NeuroImage 173, 394–410. [PubMed: 29518572]

Vovk Vladimir, 2013. Kernel ridge regression. In: Empirical Inference. Springer, pp. 105–116.

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

Won Ji Hye, Kim Mansu, Youn Jinyoung, Park Hyunjin, 2020. 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.

Wright IC, McGuire PK, Poline J-B, Travere JM, Murray RM, Frith CD, Frackowiak RSJ, Friston KJ, 1995. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. NeuroImage 2 (4), 244–252. [PubMed: 9343609]

Yousef Hosam Abozaid, ElSerogy Yasser Mohamed Bader-Eldein, Abdelal Sherif Mohamed, Abdel-Rahman Shaza Ragab, 2020. Voxel-based morphometry in patients with mood disorder bipolar I mania in comparison to normal controls. Egypt. J. Radiol. Nucl. Med 51 (1), 9.

Zhang Y, Brady M, Smith S, 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57. [PubMed: 11293691]

Zhang Yin-Nan, Li Hui, Shen Zhi-Wei, Xu Chang, Huang Yue-Jun, Wu Ren-Hua, 2021 Healthy individuals vs patients with bipolar or unipolar depression in gray matter volume. World J. Clin. Cases 9 (6), 1304–1317. [PubMed: 33644197]

Zhou Xinqi, Wu Renjing, Zeng Yixu, Qi Ziyu, Ferraro Stefania, Xu Lei, Zheng Xiaoxiao, Li Jialin, Fu Meina, Yao Shuxia, Kendrick Keith M., Becker Benjamin, 2022 Choice of voxel-based morphometry processing pipeline drives variability in the location of neuroanatomical brain markers. Commun. Biol 5 (1), 1–12. [PubMed: 34987157]

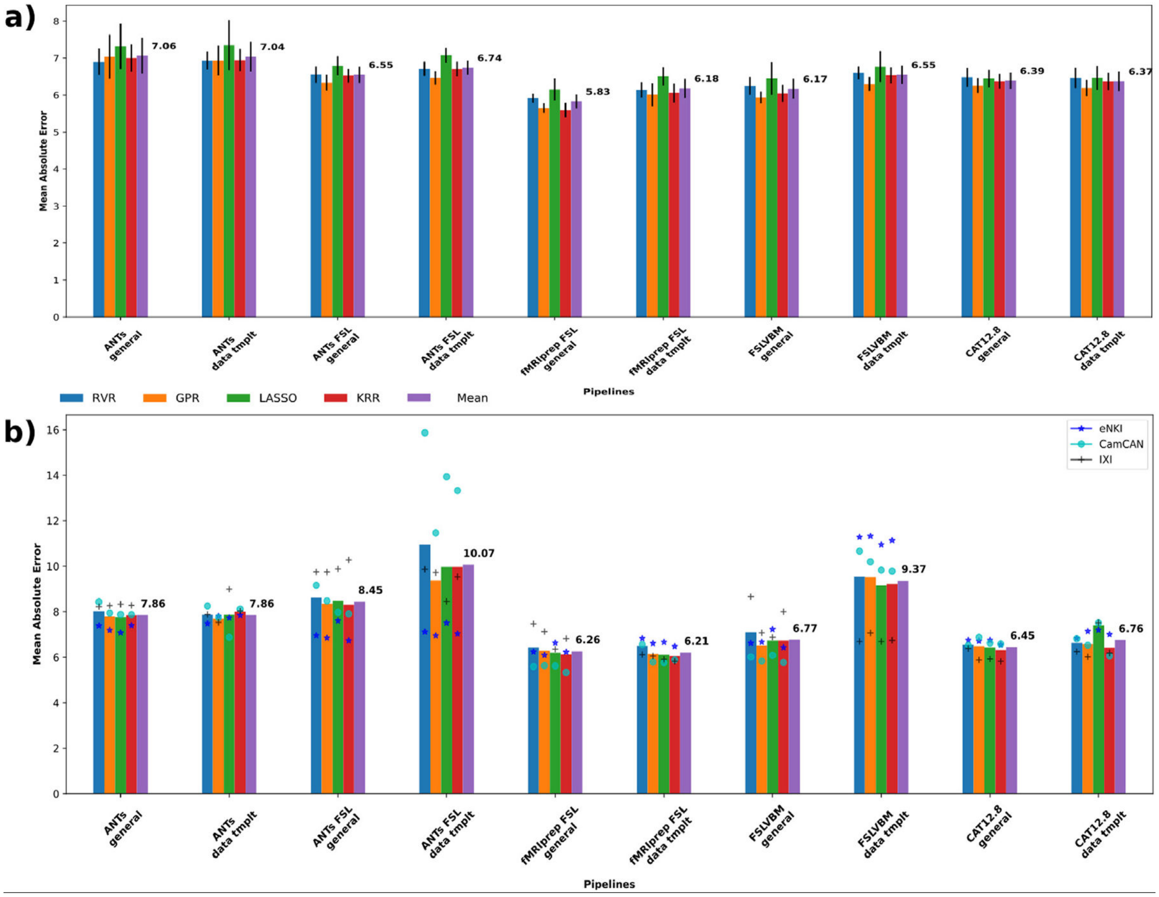

Fig. 1.

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.

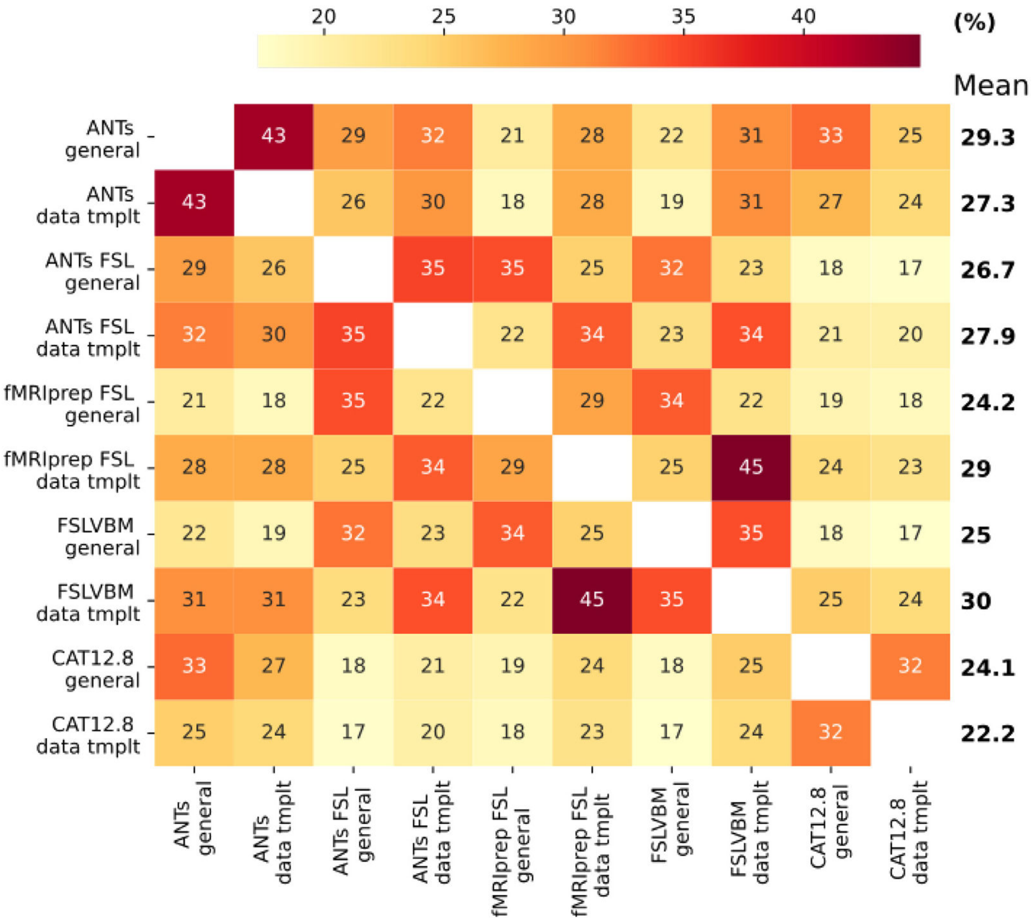

Fig. 2.

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.

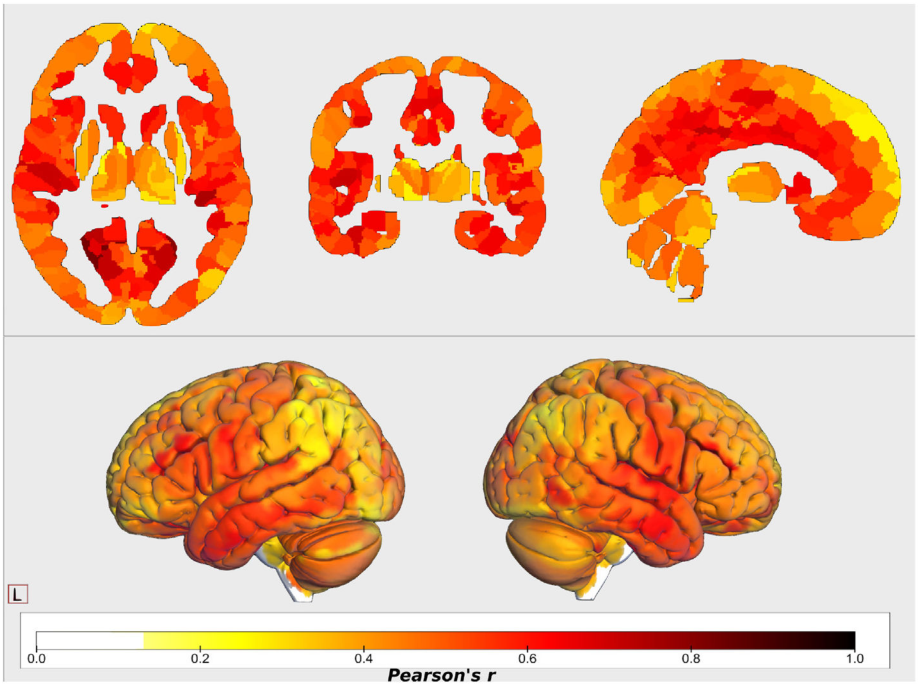

Fig. 3.

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.

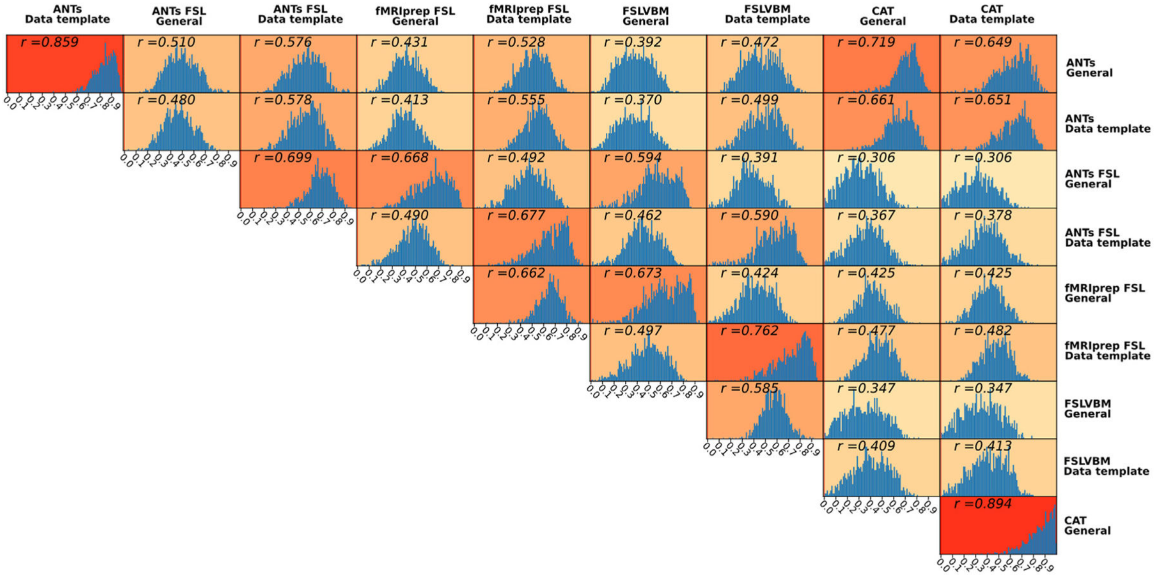

Fig. 4.

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 ($p < 0.05$).

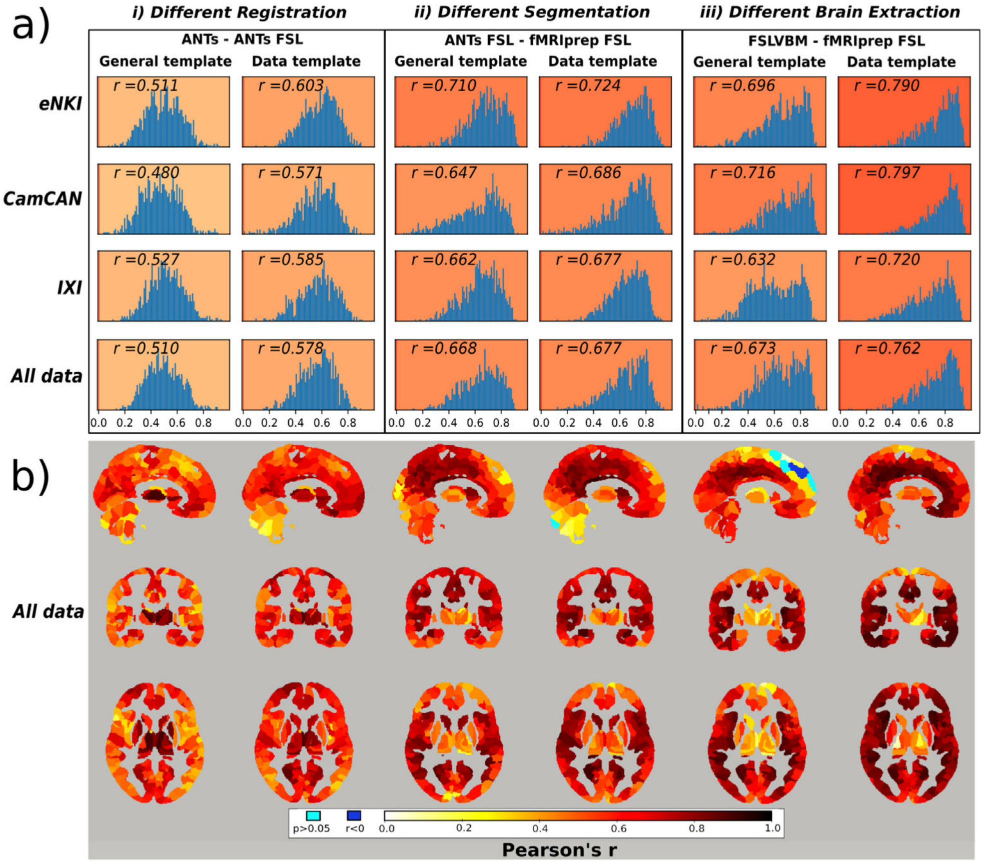

Fig. 5.

(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.

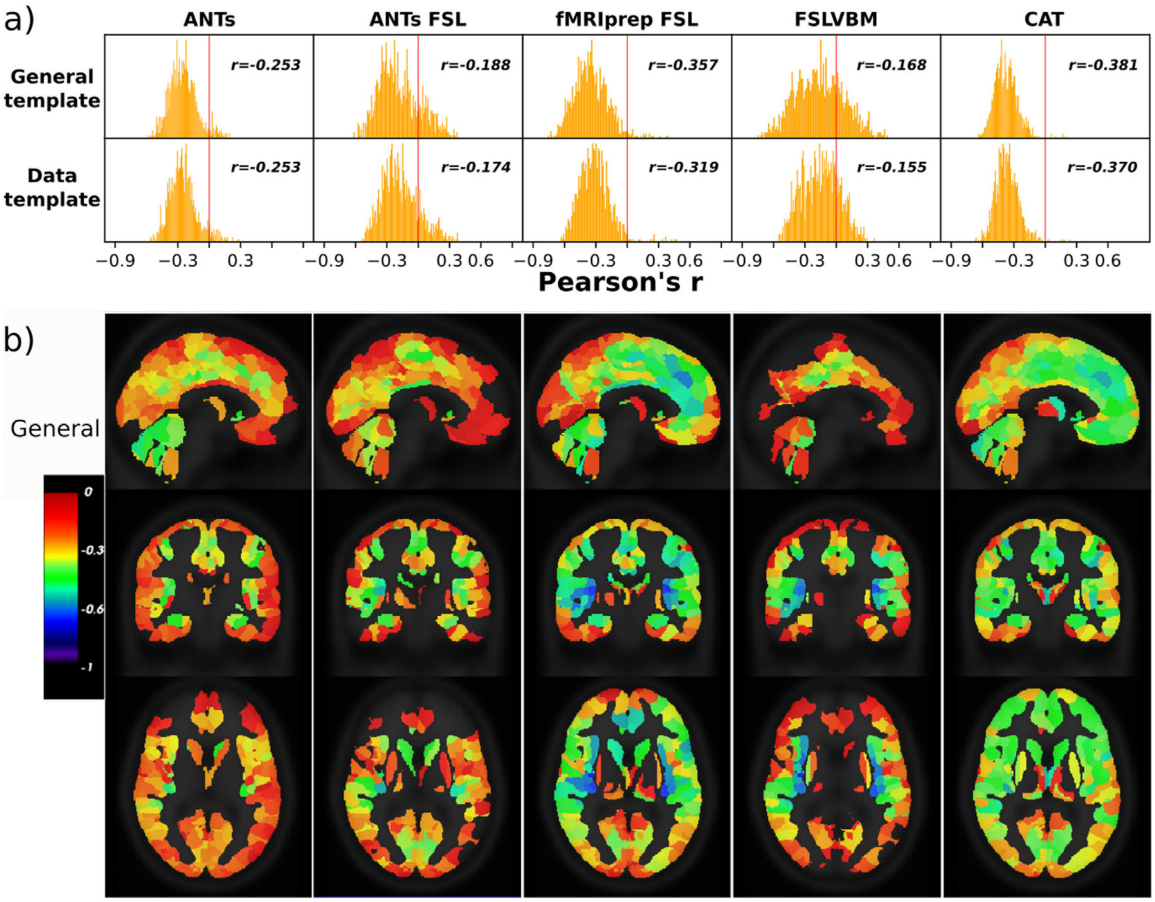

Fig. 6.

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 nonsignificant 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.

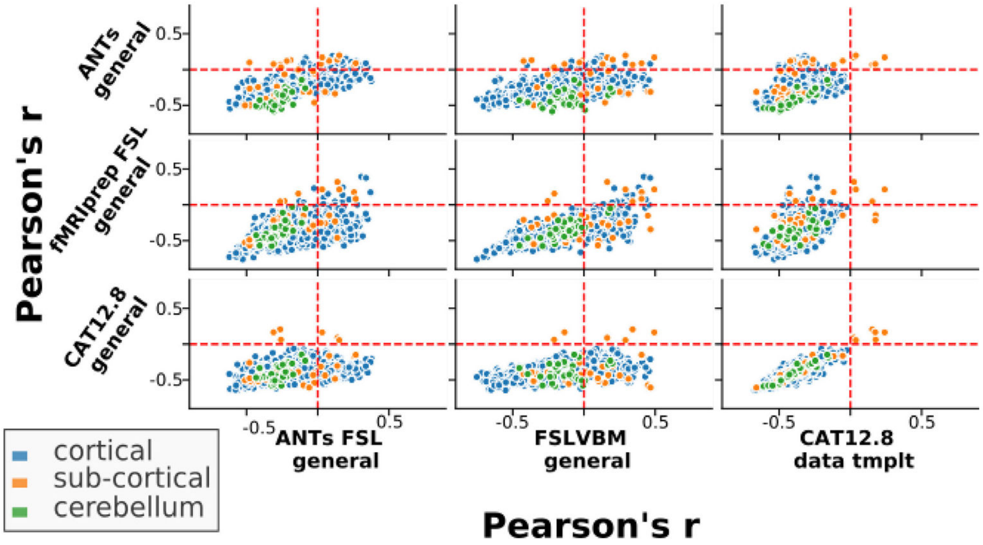

Fig. 7.

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.

Table 1

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

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

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.

	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 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.

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