

From the Institute for Systems Neuroscience  
at Heinrich Heine University Düsseldorf

**Building confound-free and generalizable machine learning  
workflows with neuroimaging data**

Dissertation

to obtain the academic title of Doctor of Philosophy (Ph.D.) in Medical Sciences  
from the Faculty of Medicine at Heinrich Heine University Düsseldorf

submitted by  
Shammi More  
(2024)

As an inaugural dissertation printed by permission of the  
Faculty of Medicine at Heinrich Heine University Düsseldorf

signed:

Dean: Prof. Dr. med. Nikolaj Klöcker

Examiners: Prof. Simon Eickhoff, Prof. Julian Caspers, Prof. Tim Hahn

Research is to see what everybody else has seen  
and to think what nobody else has thought.

- Albert Szent-Györgyi

Parts of this work have been published:

More, S., Eickhoff, S.B., Caspers, J., and Patil, K.R. (2020). “Confound removal and normalization in practice: A neuroimaging based sex prediction case study”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 3–18

More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S.B., Patil, K.R., Initiative, A.D.N., et al. 2023. Brain-age prediction: a systematic comparison of machine learning workflows. *NeuroImage*, 119947

Antonopoulos, G., More, S., Raimondo, F., Eickhoff, S.B., Hoffstaedter, F., and Patil, K.R. 2023. A systematic comparison of VBM pipelines and their application to age prediction. *Neuroimage*, 120292

Yeung, A.W.K., More, S., Wu, J., and Eickhoff, S.B. 2022. Reporting details of neuroimaging studies on individual traits prediction: a literature survey. *Neuroimage*, 119275



## Zusammenfassung

Die Magnetresonanztomographie ist ein leistungsfähiges bildgebendes Verfahren zur Untersuchung der Gehirnstruktur und -funktion, das unser Verständnis der normalen Gehirnfunktion sowie der zugrunde liegenden Mechanismen neurologischer und psychiatrischer Störungen verbessert. Techniken des maschinellen Lernens (ML) werden zunehmend mit Neuroimaging-Daten für die klinische Versorgung und die Forschung eingesetzt. ML-Arbeitsabläufe sind jedoch anfällig für Fehler, wie z. B. Überanpassung und verzerrte Ergebnisse, die zu falschen Interpretationen und Entscheidungen führen können. Daher müssen ML-Arbeitsabläufe sorgfältig konzipiert werden. In der vorliegenden Arbeit wurden zwei Schlüsselkomponenten des ML-Arbeitsablaufsdesign systematisch bewertet, die für die Entwicklung unvoreingenommener und verallgemeinerbarer ML-Modelle unerlässlich sind. Der erste Aspekt ist die effektive Beseitigung von Störsignalen, die für die Erstellung von unverfälschten Modellen ohne Störfaktoren wichtig ist. Der zweite Aspekt ist die Verwendung verschiedener Merkmalsräume und ML-Algorithmen für eine gegebene Aufgabe, um ein verallgemeinerbares Modell zu finden, sowie die Auswirkungen verschiedener Vorverarbeitungsentscheidungen auf die extrahierten Merkmale und die Modellleistung. In Studie 1 untersuchten wir zwei Confound-Regressionstechniken zur Abschwächung von Störsignalen in einem ML-Arbeitsablauf für die Aufgabe der Geschlechtsvorhersage unter Verwendung von Daten aus der funktionellen Magnetresonanztomographie im Ruhezustand. Wir fanden heraus, dass die Durchführung einer Confound-Regression im Rahmen einer Kreuzvalidierung bei der Confound-Regression wirksam war und eine bessere Schätzung der Generalisierungsleistung ergab als die Confound-Regression für die gesamten Daten. In Studie 2 untersuchten wir den Einfluss verschiedener Merkmalsräume, die aus strukturellen Magnetresonanztomographie-Daten (Volumen der grauen Substanz) und ML-Algorithmen abgeleitet wurden, auf die Leistung und Generalisierbarkeit der Altersvorhersage. Wir stellten fest, dass die Merkmalsräume und ML-Algorithmen einen erheblichen Einfluss auf die Vorhersageleistung haben, ebenso wie die Vorverarbeitungsalternativen und Merkmale aus verschiedenen Gewebetypen. Das Gehirn-Alter-Delta war bei neurodegenerativen Erkrankungen erhöht. Im Anschluss an Studie 2 wurde in Studie 3 die Auswirkung verschiedener Vorverarbeitungsalternativen auf die Schätzung des Volumens der grauen Substanz

bewertet, wobei die verschiedenen Pipelines unterschiedliche Altersvorhersageleistungen erbrachten. Studie 4 schließlich umfasste eine systematische Überprüfung bestehender psychometrischer Vorhersagestudien, wobei Trends in diesem Bereich aufgezeigt und große Kohorten sowie eine externe Validierung empfohlen wurden. Insgesamt unterstreichen unsere Ergebnisse die Bedeutung einer sorgfältigen Implementierung in jedem Schritt des ML-Arbeitsabläufe und empfehlen die Anwendung von Confound-Regression und eines Vorverarbeitungsschritts innerhalb der Kreuzvalidierung, die Erforschung verschiedener Merkmalsräume und ML-Algorithmen, die Verwendung großer Trainingskohorten zur Entwicklung optimaler und verallgemeinerbarer Arbeitsabläufe und die Durchführung einer externen Validierung.

## Summary

Magnetic resonance imaging (MRI) is a powerful neuroimaging technique to study brain structure and function, advancing our understanding of normal brain function as well as the underlying mechanisms of neurological and psychiatric disorders. Machine learning (ML) techniques have been increasingly used with neuroimaging data for clinical care and research. However, ML workflows are prone to errors, such as overfitting and biased outcomes, which can lead to wrong interpretations and conclusions. Hence, there is a need for careful designing of ML workflows. The current work systematically evaluated several key components of ML workflow design, essential for developing unbiased and generalizable ML models. The first aspect is the effective removal of confounding signals, which is important for creating confound-free unbiased models. The second aspect is the usage of different feature spaces and ML algorithms for a given task to find a generalizable model—additionally, the impact of various preprocessing choices on extracted features and model performance. In study 1, we investigated two confound regression techniques to mitigate confounding signals in an ML workflow for the sex prediction task using resting-state functional MRI data. We found that performing confound regression within cross-validation (CV) was effective in confound removal and gave a better generalization performance estimate than whole-data confound regression. In study 2, we assessed the impact of different feature spaces derived from structural MRI data (gray matter volume; GMV) and ML algorithms on age prediction performance and generalizability. We found a substantial impact of feature spaces and ML algorithms on prediction performance, along with an impact of preprocessing alternatives and features from different tissue types. Brain-age delta was elevated in neurodegenerative disease. Following study 2, in study 3, the impact of several preprocessing alternatives on GMV estimates was assessed, revealing varying age prediction performance from different pipelines. Lastly, study 4 involved a systematic review of existing psychometric prediction studies, highlighting trends in the field and advocating for large cohorts and external validation. Overall, our findings emphasize the importance of careful implementation at each step of ML workflow, recommending applying confound removal and any preprocessing step within CV, exploring various feature spaces and ML algorithms, utilizing large training cohorts for developing optimal and generalizable workflows, and performing external validation.

## List of abbreviations

<b>AD</b>	Alzheimer’s disease
<b>ANTs</b>	Advanced Normalization Tools
<b>BOLD</b>	blood-oxygen-level-dependent
<b>CAT</b>	Computational Anatomy Toolbox
<b>CSF</b>	cerebrospinal fluid
<b>CV</b>	cross-validation
<b>CVCR</b>	cross-validated confound regression
<b>FC</b>	functional connectivity
<b>fMRI</b>	functional magnetic resonance imaging
<b>FSL</b>	FMRIB Software Library
<b>GMV</b>	gray matter volume
<b>GPR</b>	Gaussian process regression
<b>HC</b>	healthy control
<b>KRR</b>	kernel ridge regression
<b>MAE</b>	mean absolute error
<b>MCI</b>	mild cognitive impairment
<b>ML</b>	machine learning
<b>MRI</b>	magnetic resonance imaging
<b>PCA</b>	principal component analysis
<b>ReHo</b>	regional homogeneity
<b>rs-fMRI</b>	resting-state functional magnetic resonance imaging
<b>RVR</b>	relevance vector regression

<b>sMRI</b>	structural magnetic resonance imaging
<b>SPM</b>	Statistical Parametric Mapping
<b>SVR</b>	support vector regression
<b>VBM</b>	voxel-based morphometry
<b>WDCR</b>	whole-data confound regression
<b>WMV</b>	white matter volume

# Contents

<b>1 Introduction</b>	<b>1</b>
1.1 Neuroimaging-based prediction	3
1.2 Machine learning workflows	6
1.3 Challenges	9
1.3.1 Confound removal	9
1.3.2 Designing of robust and generalizable workflows	11
1.3.3 Other general consideration in designing ML workflows	13
1.4 Ethics Protocols	14
1.5 Aims of Thesis	14
<b>2 Confound Removal and Normalization in Practice: A Neuroimaging Based Sex Prediction Case Study.</b> More, S., Eickhoff, S.B., Caspers, J., Patil, K.R., Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track, 12461:3–18 (2021)	<b>16</b>
<b>3 Brain-age prediction: a systematic comparison of machine learning workflows.</b> More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S.B. and Patil, K.R., NeuroImage, 119947 (2023)	<b>33</b>
<b>4 A systematic comparison of VBM pipelines and their application to age prediction.</b> Antonopoulos, G., More, S., Raimondo, F., Eickhoff, S.B., Hoffstaedter, F., and Patil, K.R., NeuroImage, 120292 (2023)	<b>51</b>
<b>5 Reporting Details of Neuroimaging Studies on Individual Traits Predictions: A Literature Survey.</b> Yeung, A.W.K., More, S., Wu, J., Eickhoff, S.B., NeuroImage, 119275 (2022)	<b>66</b>
<b>6 Discussion</b>	<b>74</b>
6.1 Machine learning workflow design	75
6.1.1 Try different feature spaces and ML algorithms	76

6.1.2	Control for bias . . . . .	81
6.1.2.1	Removal of confounding signal . . . . .	81
6.1.2.2	Mitigation of age bias . . . . .	83
6.1.3	Other general considerations . . . . .	84
6.1.3.1	Feature preprocessing and engineering . . . . .	85
6.1.3.2	Large training sample size and external validation . . . . .	86
6.1.3.3	Presence of data shift . . . . .	86
6.2	Interpretability and clinical relevance . . . . .	87
6.2.1	Interpretability of confound-free sex prediction model . . . . .	88
6.2.2	Clinical relevance of brain-age delta . . . . .	88
6.2.2.1	Higher brain-age delta in disease . . . . .	89
6.2.2.2	Delta-behavior correlations in healthy populations . . . . .	90
6.3	Conclusion . . . . .	92
<b>Bibliography . . . . .</b>		<b>93</b>

# 1 Introduction

A World Health Organization report highlights that approximately one billion people globally are impacted by a spectrum of neurological disorders, encompassing conditions such as epilepsy, Alzheimer’s disease (AD), stroke, and brain injuries (Bertolote, 2007). These disorders affect people worldwide, regardless of age, gender, education, or income. In the past 30 years, the absolute number of deaths has increased by 39%, and disability-adjusted life-years have increased by 15%, causing a huge economic burden (Feigin et al., 2020). This necessitates the advancement of methods and techniques to understand the human brain and methods for early detection of disease and treatment.

Neuroscience is a multidisciplinary field of study focused on unraveling the complexities of the nervous system, aiming to understand the intricate workings of the brain and its role in behavior, cognition, and various physiological functions. Neuroimaging is a powerful tool in this endeavor, providing techniques such as magnetic resonance imaging (MRI) and Computed Tomography to study brain structure and function. MRI is widely used in clinical practice to support clinicians in making diagnoses and planning treatments (Hashemi et al., 2012). Unlike Computed Tomography and Positron Emission Tomography, MRI does not use dangerous radiation or require an injection of radioactive substances, so it is considered safe and non-invasive. MRI allows us to study the brain in both healthy and diseased states, advancing our understanding of normal brain function as well as the underlying mechanisms of neurological and psychiatric disorders. Different MRI modalities can capture anatomical, diffusion, and functional characteristics of the brain, making it a versatile tool for neuroimaging research and clinical diagnosis. Anatomical or structural MRI (sMRI) provides detailed images of brain structures, while diffusion MRI measures the movement of water molecules, offering insights into white matter connectivity. Functional MRI (fMRI) detects changes in blood flow, enabling the observation of brain activity patterns. Together, these modalities help unravel the complex workings of the human brain and are invaluable in understanding neurological disorders and cognitive



processes.

Structural magnetic resonance imaging (sMRI): It is a non-invasive imaging technique used to examine the static anatomy of the brain by differentiating between tissue types (Ombao, 2016). This technique takes advantage of tissue-dependent differences in the proton's rate of relaxation in the presence of a radio-frequency pulse after placing the tissue in a powerful, uniform external magnetic field (Hashemi et al., 2012). Images measured this way are useful for their high spatial resolution and provide a good distinction between different tissue types that contain different proportions of water and fats. Different images can be generated to emphasize contrast related to different tissue characteristics. For example, T1-weighted MRI provides good contrast between gray matter and white matter tissues, with gray matter appearing as dark gray, white matter as lighter gray, and cerebrospinal fluid (CSF) appearing as the dark region. T2-weighted images show CSF as bright and gray matter lighter than white matter.

Functional magnetic resonance imaging (fMRI): It provides a proxy measure for brain activity by detecting changes associated with blood flow. This technique relies on the fact that cerebral blood flow and neuronal activation are coupled, i.e., when an area of the brain is activated, the blood flow to that region also increases (Soares et al., 2016). The most common approach towards fMRI uses the blood-oxygen-level-dependent (BOLD) contrast, which allows the measurement of the ratio of oxygenated to deoxygenated hemoglobin in the blood. The increase in blood flow leads to an increase in the ratio of oxygenated blood to deoxygenated blood in the region. Oxygenated hemoglobin takes longer to lose magnetization and hence causes stronger BOLD signals, while deoxygenated hemoglobin results in weaker BOLD signals. Therefore, a stronger BOLD signal reflects an increase in blood flow, which reflects an increase in neuronal activity in the brain region. Two common types of fMRI approaches are task-based fMRI and resting-state fMRI (rs-fMRI), each offering distinct insights into brain function (Biswal et al., 1995). In task-based fMRI, participants perform a behavioral or cognitive task in the scanner. The neuronal responses represented by the BOLD signals during the task are compared with the baseline task to establish a mapping between brain regions involved in the particular task execution. Conversely, in rs-fMRI, participants are instructed to relax in the scanner. It captures the spontaneous brain activity in the absence of tasks, shedding light on the brain's intrinsic organization (Fox and Raichle, 2007).

## 1.1 Neuroimaging-based prediction

Machine learning (ML) involves algorithms and statistical models that enable computers to learn from data, identify patterns, and make predictions. In the context of neuroimaging, ML utilizes these techniques to analyze vast amounts of brain imaging data, such as sMRI or fMRI, extracting intricate patterns useful for predicting brain-related conditions and disease progression. For example, ML models can be trained to learn the relationship between MRI-derived features and targets (for example, disease vs. healthy) and then used to make predictions on new unseen data (Du et al., 2012, Wang et al., 2015, Du et al., 2018, Nenning and Langs, 2022). This technology holds immense promise in assisting neuroscientists and clinicians by providing efficient tools for diagnosing neurological disorders, identifying neurological biomarkers, understanding brain function, predicting treatment outcomes, and ultimately advancing personalized medicine tailored to an individual’s brain characteristics (Caspers, 2021, Nenning and Langs, 2022).

Diverse features can be derived from different MRI modalities, which can be used to make these predictions. For example, cortical and subcortical measurements of volume, surface, and thickness values, or gray matter volume (GMV), white matter volume (WMV), CSF obtained through voxel-based morphometry (VBM) analysis from sMRI, can serve as essential inputs for training ML models (Fischl and Dale, 2000, Ashburner and Friston, 2000). The rs-fMRI data can provide measures for spontaneous brain activity at rest, such as local synchronization of rs-fMRI signals or regional homogeneity (ReHo), which measures the similarity of the time series of a set of voxels and thus reflects the temporal synchrony of the regional BOLD signal (Zang et al., 2004). Other features measure the intrinsic connectivity of the brain by measuring the temporal correlation in BOLD signal changes between different brain regions using functional connectivity (FC) matrices (Biswal et al., 1995, Fox and Raichle, 2007). Additionally, graph-theory representation of FC has been used to infer topological characteristics of brain networks, such as modularity, centrality, and small-worldedness, which can provide valuable insights (Wang et al., 2010, Kazeminejad and Sotero, 2019, Khosla et al., 2019). More recently, several studies have also begun to explore the predictive capacity of dynamic FC (Fong et al., 2019, Zhu et al., 2021). Similarly, FC can be derived from task-based fMRI data (Ooi et al., 2022). Since different MRI modalities offer complementary information, it is sometimes useful to use them together

to get better predictive performance (Pisharady et al., 2023, Cole, 2020, De Lange et al., 2020).

Using these features extracted from structural and functional MRI, ML models have correctly differentiated healthy control (HC) individuals from patients with neurodegenerative disorders such as AD (Klöppel et al., 2008, Guo et al., 2017), mild cognitive impairment (MCI) (Westman et al., 2011, Yu et al., 2017), multiple sclerosis (Weygandt et al., 2011; Weygandt et al., 2015), Parkinson’s disease (Marquand et al., 2013), neurodevelopment disorders such as autism spectrum disorder (Ecker et al., 2010, Abraham et al., 2017), neuropsychiatric disorders such as schizophrenia (Zarogianni et al., 2013, Venkataraman et al., 2012), and depression (Foland-Ross et al., 2015). This suggests that ML models trained with MRI data could be a valuable tool for the automatic diagnosing of diseases (Mateos-Pérez et al., 2018). It also allows studying which regions are associated with these diseases, revealing their imaging signatures. ML can also help in disease prognosis, predicting the likely course of the disease (Storelli et al., 2022; Moazami et al., 2021). For instance, studies have used ML to predict the progression of stable MCI to progressive MCI patients (Moradi et al., 2015), and conversion of MCI to AD (Westman et al., 2011, Davatzikos et al., 2011).

The applications described above use supervised methods in the sense that they involve training ML models using labeled data, where a target variable (e.g., disease status) is provided to guide the learning process. Unsupervised methods, which do not require a target variable but look for structure in the data, have also been successfully employed. Unsupervised ML algorithms have been used to find subgroups within diseases, for example, finding subtypes of multiple sclerosis that exhibited distinct treatment responses (Eshaghi et al., 2021). Consensus clustering has been used to find sub-groups of tumor patients (Choi et al., 2020) and patients with epilepsy (Lee et al., 2020). Identification of subtypes can help develop individualized precision treatment.

Another fundamental aim of neuroscience is understanding how brain characteristics are linked to cognitive and behavioral measures. There is evidence stating that inter-individual variation in functional and structural patterns co-vary with cognitive, behavioral, and demographic traits (Llera et al., 2019). Consequently, these patterns have been used to predict various individual traits and can help identify biomarkers for health and disease. For instance, FC has been used to predict cognitive abilities such as fluid intelligence (Finn et al., 2015), sustained attention (Rosenberg

et al., 2016), memory performance (Sasse et al., 2023, Meskaldji et al., 2016, Siegel et al., 2016) in healthy and clinical populations. It has also been used to predict personality traits such as neuroticism, extraversion, agreeableness, and openness (Nostro et al., 2018, Hsu et al., 2018). Additionally, numerous studies have used ML to predict demographic variables such as sex (Zhang et al., 2018, Weis et al., 2020) and age (Franke et al., 2010, Cole et al., 2017) and achieved good performance.

Studies have highlighted differences in cognition and psychopathology between the sexes (Seeman, 1997). For instance, variations in spatial perception, memory, and verbal skills (Miller and Halpern, 2014), a higher susceptibility of females to depression (Picco et al., 2017), and a greater incidence of autism among males (Werling and Geschwind, 2013) have been reported, indicating underlying differences in structural and functional brain organization between the sexes (Kaczkurkin et al., 2019). Therefore, sex prediction studies can help with the understanding of the neurobiology of sex differences, provide insights into risks and protective factors, and eventually help to develop sex-specific treatments (Zhang et al., 2018, Weis et al., 2020).

Since aging is a major risk factor for most neurodegenerative diseases, individual-level quantification of atypical aging can be helpful for early detection of disorders. Consequently, many studies have used ML methods to capture multivariate patterns of age-related changes in the brain associated with healthy aging (Ashburner, 2007, Franke et al., 2010, Cole et al., 2018, Varikuti et al., 2018, Franke and Gaser, 2019, Baecker et al., 2021b). ML models can be trained using neuroimaging data from healthy subjects to predict age. A higher positive difference between predicted age (brain-age) and chronological (true) age, i.e., brain-age delta or delta, indicates “older-appearing” brains. Therefore, brain-age prediction studies can help inform about abnormal brain aging by measuring the deviation of predicted age from chronological age. Higher delta has been reported in several common brain disorders (Kaufmann et al., 2019, Wrigglesworth et al., 2021, Sone et al., 2021). Higher delta has also been known to relate to several age-related risk factors such as weaker grip strength, poorer lung function, increased mortality risk, and poorer cognitive functions such as fluid intelligence, processing speed, semantic verbal fluency, visual attention, and cognitive flexibility (Cole et al., 2018, Boyle et al., 2021, Wrigglesworth et al., 2021). Thus, delta can potentially serve as a biomarker of brain integrity.

All these applications rely on a robust and reliable ML workflow design to give

correct predictions and interpretations. ML workflows involve several crucial steps, including selecting a suitable ML algorithm to learn the relationship between features and targets, getting enough training data, employing data transformation methods, feature selection techniques, and hyperparameter tuning (Scheinost et al., 2019, Lones, 2021). Collectively, these elements form an integrated ML workflow. Despite numerous successful demonstrations, ML workflows are susceptible to pitfalls such as overfitting and biased model outcomes due to various factors such as model complexity and non-representative training data, among others (Domingos, 2012, Lones, 2021, Mehrabi et al., 2021). Such models might not generalize well and reflect existing biases in the data, leading to erroneous interpretations and problematic conclusions. Therefore, careful and correct implementation of an ML workflow is crucial for its application in real-world scenarios. By recognizing the potential pitfalls and actively addressing them in the implementation process, we can harness the power of ML while minimizing its inherent risks. The following section outlines the steps involved in designing an ML workflow and addresses some of the challenges encountered in ML applications.

## 1.2 Machine learning workflows

An ML workflow comprises various steps, including 1) Problem definition, 2) Data collection and preparation, 3) Workflow definition, and 4) Model training and evaluation (Figure 1). Several choices are available for each step, making designing a robust ML workflow challenging.

**1. Problem definition:** The first step includes defining the target to predict (e.g., demographic variable, behavioral scores, or disease status) and the features to be used (e.g., neuroimaging-derived FC or GMV). One can also define confounds, i.e., variables related to both features and target, which one may choose not to model or consider these relationships in their analysis (Weber et al., 2022). For example, brain size can be a confound when predicting sex using GMV as brain size correlates with the target, i.e., sex (males have bigger brains than females, Ritchie et al., 2018), and brain size information is encoded in GMV features (Wiersch et al., 2023). Thus, if the study aims to find structural brain organization differences between sexes, it is essential to control for confounds to ensure that the model learns the true signal of interest, i.e., the feature-target relationship, and not the confound-target relationship.

**2. Data collection and preparation:** One needs to collect (or, in some cases, use

existing databases) and prepare the data for training and testing the ML model. The most fundamental assumption for the data is that it is composed of independent and identically distributed samples, i.e., each data point is assumed to be independent of the others and is drawn from the same underlying distribution (Bishop and Nasrabadi, 2006). Meeting this assumption lays a strong foundation for learning, enhancing the model’s ability to perform well on unseen data that share a similar distribution. One could control for some confounds at the data collection stage, e.g., controlling for sex by equally sampling males and females or controlling for age by balancing the age range in healthy and diseased groups. When that is not feasible, post-hoc methods may be employed for confound control (Tripepi et al., 2010, Snoek et al., 2019, Chyzyk et al., 2022). Data cleaning is an important part of data preparation, including imputing missing values, removing features with too many missing values, removing duplicate values, avoiding typos, and converting data types (Brownlee, 2020).

**3. Workflow definition:** It involves several key decisions. One must choose the model(s) for the task. Choosing an appropriate model depends on the type of problem, such as classification for predicting disease status or regression for predicting cognitive/behavioral scores, with several choices available for both. One can decide which model to use depending on the prior knowledge from literature, the assumed relationship between features and target (e.g., linear vs. non-linear), the nature of the data (number of samples and number of features), and the available computational resources.

One can choose to apply several optional data transformations or preprocessing steps to the features, such as confound removal, feature normalization (e.g., z-score, robust scaler), dimensionality reduction via feature selection (e.g., variance thresholding, information gain, high correlation filter, etc.), or feature engineering (e.g., principal component analysis (PCA), independent component analysis, etc.), which might help the training process (Bishop and Nasrabadi, 2006). For example, feature normalization brings all the features on the same scale, ensuring they contribute equally to the learning process, improving the stability of optimization algorithms. Dimensionality reduction can help remove irrelevant or redundant features, thus providing better-performing models. Deciding on these steps is not trivial, as each choice can substantially impact the outcome.

Since ML aims to create models that accurately predict outcomes on new unseen

data by learning generalizable information, testing the model on new unseen out-of-sample test data (also called external validation) is essential. However, when a dedicated test dataset is unavailable, a portion of the available data can serve as a proxy for test data, allowing the assessment of the model’s generalization performance, i.e., its ability to perform accurately on new, unseen data from the same distribution. Cross-validation (CV) is frequently employed as a model evaluation scheme for this purpose. In K-fold CV, the initial dataset is divided into K equally sized non-overlapping parts, where all subsets but one are used for training the model and the remaining subset for testing. The assignment of training and testing subsets is repeated K times, so all folds are used for test once. The average performance across all test folds is computed as an estimate of generalization performance (also called CV performance). If the model performs much better on the training set than the test set, then it is overfitting. An optimization strategy, such as random search or grid search, can be employed for optimizing hyperparameters (parameters that are not learned by data but rather tuned for a given predictive task) or feature preprocessing (e.g., feature selection). This is done in a nested CV (also known as double CV), which involves doing hyperparameter optimization and feature selection as an extra loop inside the main CV loop (Poldrack et al., 2020, Varoquaux et al., 2017, Cawley and Talbot, 2010).

**4. Model training and evaluation:** Model training involves using the training data to adjust the parameters and tune the model’s hyperparameters (from user-defined search space) to minimize the prediction error. The training procedure yields models with fixed parameters and hyperparameters, which can then be used to make predictions on the test data. It is crucial to treat hyperparameters and feature optimization (e.g., feature selection) as part of model training to avoid data leakage. Moreover, it is essential to check if the hyperparameters are hitting the boundaries in the defined search space and adjust them accordingly when necessary.

After the model has been trained, it must be evaluated to determine its performance. This is done by comparing the model’s predictions with the actual values in the test data using appropriate evaluation metrics, such as classification accuracy (or balanced accuracy), F1 score, and area under the receiver operating characteristic curve for classification, or mean absolute error (MAE) and  $R^2$  for regression. It is a good practice to report multiple metrics since different metrics can present different perspectives on the results and increase transparency.

## 1.3 Challenges

Designing a generalizable and unbiased ML workflow encompasses many challenges that demand careful consideration. Overfitting, a common problem, involves models fitting training data too well and performing poorly on new unseen data (Yarkoni and Westfall, 2017). This can happen because of a small sample size or high model complexity. Another common challenge is data leakage, a phenomenon where information from outside the training set is unintentionally included in the model, leading to an overestimated and unrealistic performance in practice (Kapoor and Narayanan, 2022). There can be several reasons for data leakage, such as using test data as part of training data and performing any preprocessing or tuning hyperparameters outside CV, among others. Another challenge is interpretability, i.e., the degree to which a human can understand the cause of a decision (Miller, 2019). Highly accurate models may be more complex and difficult to understand; simpler, more interpretable models may sacrifice some accuracy. Hence, a trade-off exists between the accuracy and interpretability of ML models (Dziugaite et al., 2020). The interpretability of a model can suffer from incorrect methods, for example, not controlling for confounds when investigating brain-behavior relationships, which can lead to biased predictions driven by confound-target relationships instead of feature-target relationships and thus misleading conclusions. Furthermore, establishing a robust and generalizable workflow is challenging as it involves intricate decisions about data preprocessing, feature selection, model design, hyperparameter tuning, and additional optimization criteria depending on the task. Addressing these challenges necessitates a holistic approach that blends domain knowledge and sound methodologies. The current work addressed some key challenges, including confound removal and designing a robust and generalizable workflow.

### 1.3.1 Confound removal

One of the significant challenges in ML is accounting for confounding effects. A confound is a variable that influences both the independent and the dependent variables (Pourhoseingholi et al., 2012). Features derived from neuroimaging data can contain information uniquely associated with the target (true signal-of-interest) but also contain information from nuisance sources, confounding the relationship between the



neuroimaging signal and the target. Common confounding variables in neuroimaging studies include age, sex, handedness, brain size, and in-scanner movement (Alfaro-Almagro et al., 2021). Failure to remove confounds can lead to biased predictions and interpretations. For example, in a sex prediction task using FC, brain size is a confound as it is associated with sex (males having bigger brain size than females) and is encoded in FC (Ritchie et al., 2018, Zhang et al., 2016). In such an instance, predictions can be biased as a successful outcome may be driven by the confounding signal (brain size differences) rather than the true signal of interest (FC differences). If a study aims to maximize model performance, then the confounding variables containing neurobiological effects of interest can be used as input features; however, if a study aims to identify true brain-behavior relationships, then it is important to control for confounding signals.

Several approaches exist to mitigate confounding variables. One could control for some confounds at the data collection stage by balancing the acquisition for confounds or using randomized controlled trials (Pourhoseingholi et al., 2012). However, in observational/epidemiological studies where data has already been collected, it is necessary to control for confounds in a post-hoc approach. These approaches include post-hoc counterbalancing, anti-mutual information sampling, and stratification using pooling analysis (Tripepi et al., 2010, Snoek et al., 2019, Chyzhyk et al., 2022). However, these methods often result in data loss and are not feasible with a small sample. A prevalent strategy is confound regression, which involves fitting a linear regression model on each feature separately with the confound as the predictor, and the corresponding residuals are used as new “confound-removed” features (Todd et al., 2013, Snoek et al., 2019).

Confound regression can be implemented through whole-data confound regression (WDCR) or cross-validated confound regression (CVCR). WDCR, although aggressive, suffers from data leakage as it constructs confound-removed features on the whole data before CV. CVCR, on the other hand, addresses this by performing CV-consistent confound regression within each CV fold. Though both methods are used in neuroimaging research, the impact of these approaches on generalization estimates and interpretability is unknown, along with their interaction with normalization methods (Snoek et al., 2019, Pervaiz et al., 2020). Employing rank-based inverse normal transformation for normalization after confound regression may reintroduce confounding

effects (Pain et al., 2018). This lack of knowledge of how to correctly perform confound removal and the interaction between confound regression and normalization (Figure 1.2) makes it difficult to design ML workflows. Lastly, the influence of covariate and confounding shifts on model building requires exploration.

In study 1, we addressed these gaps by empirically evaluating WDCR and CVCR for confound removal efficacy and generalization performance, investigating normalization interactions, and examining model deployment under covariate and confounding shifts. We apply these investigations to predict sex from rs-fMRI data, considering brain size and age as confounds, aiming to discern differences in functional organization between sexes while accounting for brain size differences.

### 1.3.2 Designing of robust and generalizable workflows

Designing an ML workflow for a specific task involves decisions about various choices at each step; not all can be predetermined without considering the data. In other words, data-driven decisions are essential to develop a robust and generalizable workflow.

Many factors can influence model performance, with the feature space being a primary consideration. Different feature spaces (Figure 1.1) can have different information content, leading to differential outcomes. Furthermore, different ML algorithms (Figure 1.3), each with its own inductive biases, contribute to disparate performance results. Every algorithm must embody some knowledge or assumptions to generalize beyond the provided data (Domingos, 2012). Formalized by Wolpert as the “no free lunch” theorem, according to which no algorithm can beat random guessing over all possible functions to be learned (Wolpert, 1996), highlighting that there is no single ML algorithm universally the best for all problems. So, it is recommended to try different algorithms to evaluate what works best for the task at hand (Domingos, 2012). Moreover, different combinations of feature spaces and ML algorithms can yield diverse outcomes.

For instance, to design a workflow for brain-age estimation, voxel-wise GMV data can be used directly, or additional pre-processing such as smoothing and/or resampling can be applied, or parcel-wise averages within a brain atlas can be used as features (Franke et al., 2010, Boyle et al., 2021, Varikuti et al., 2018, Eickhoff et al., 2021). Further dimensionality reduction methods, such as PCA, can improve the observations-

to-features and signal-to-noise ratios (Franke et al., 2010, Franke et al., 2013, Gaser et al., 2013). Choosing from a pool of ML algorithms like relevance vector regression (RVR), support vector regression (SVR), Gaussian process regression (GPR), and kernel ridge regression (KRR) is crucial as these choices can impact performance (Lee et al., 2021, Baecker et al., 2021a, Lange et al., 2022). Many studies predicting age from VBM-derived GMV have shown  $\sim 5\text{--}8$  years of prediction errors in healthy individuals. Despite the extensive work in this field, there remains a gap in understanding which feature spaces and ML algorithms can effectively capture the aging process and perform optimally for age prediction. Challenges arise due to the diversity in study setups and methodology, such as variations in training data, sample size, feature spaces, and ML algorithms, making it difficult to compare the results and draw valid conclusions.

There can be additional criteria to optimize for when predicting behavioral, demographic, or cognitive variables from neuroimaging data. For example, for brain-age estimation, the workflow should perform well on new samples from the same dataset (high within-dataset performance) and generalize well on data from a new site (high cross-dataset performance). The ability to make predictions that generalize across sites is crucial. It allows for the development of diagnostic tools, biomarkers, or predictive models that can be applied in diverse healthcare settings or research studies. It should have high test-retest reliability, i.e., estimated age must be reliable on repeated measurements, and exhibit longitudinal consistency, i.e., the predicted age should be proportionally higher for later scans assuming no significant health-related interventions between the measurements (Franke and Gaser, 2019, Cole and Franke, 2017, Sone and Beheshti, 2022). These objectives can make designing robust and generalizable workflows even more challenging. Overall, designing a generalizable workflow is intricate because of the many choices available at each step, especially when a workflow is expected to perform well in multiple criteria.

Consequently, in study 2, we studied the task of age prediction using GMV data to develop a robust and generalizable workflow through evaluation under different criteria important for real-world application. We examined 128 workflows encompassing 16 feature spaces derived from gray matter images (voxel-wise or parcel-wise) and eight ML algorithms leveraging extensive neuroimaging databases containing a broad age spectrum. We evaluated these workflows for their within-dataset and cross-dataset performances. Following this, we delved into the test-retest reliability and the

longitudinal consistency of predictions over time for some well-performing workflows. All these criteria are important to ensure real-world application of delta. Additionally, we measured the effectiveness of our top-performing workflow in a clinical setting. We examined the correlations between delta and behavioral/cognitive measures in healthy and clinical cohorts and various factors affecting these correlations. Further analyses were carried out to study the effects of preprocessing choices and the inclusion of features from various tissue types on predictive performance.

There are many preprocessing tools available for feature extraction from neuroimaging data, such as Statistical Parametric Mapping (SPM) (Friston, 2003), Computational Anatomy Toolbox (CAT) (Gaser et al., 2022), and FMRIB Software Library (FSL) (Smith et al., 2004). Prior studies have highlighted the variability in extracted features, such as cortical thickness estimates, introduced by the choice of a preprocessing pipeline for sMRI data (Tustison et al., 2014, Dickie et al., 2017). These inconsistencies in the results arise from several algorithmic and parametric differences that exist in the preprocessing tasks, such as image normalization, registration, and segmentation within pipelines (Bhagwat et al., 2021). Differences in feature spaces extracted by various preprocessing tools can impact their correlation with behavioral, cognitive, or demographic variables. Consequently, there has been a difference in the performance of the individual-centric prediction tasks using different preprocessing pipelines (Bhagwat et al., 2021, Tavares et al., 2020, Zhou et al., 2022). Therefore, in study 3, we studied the impact of 10 different VBM preprocessing tools on GMV estimation by comparing their performance for age prediction. By systematically examining the effects of various preprocessing tools on the derived features and subsequent predictive models, study 3 contributes valuable insights into the importance of methodological choices in neuroimaging analyses and highlights the necessity of considering preprocessing variations when interpreting results or building predictive models based on neuroimaging data.

### 1.3.3 Other general consideration in designing ML workflows

Our previous studies delved into investigating various factors impacting ML model performance in neuroimaging analysis, including preprocessing tools choices, feature spaces, feature preprocessing, and ML algorithms. There are numerous other factors, such as the training sample size and the CV strategy used (leave-one-out vs. K-fold

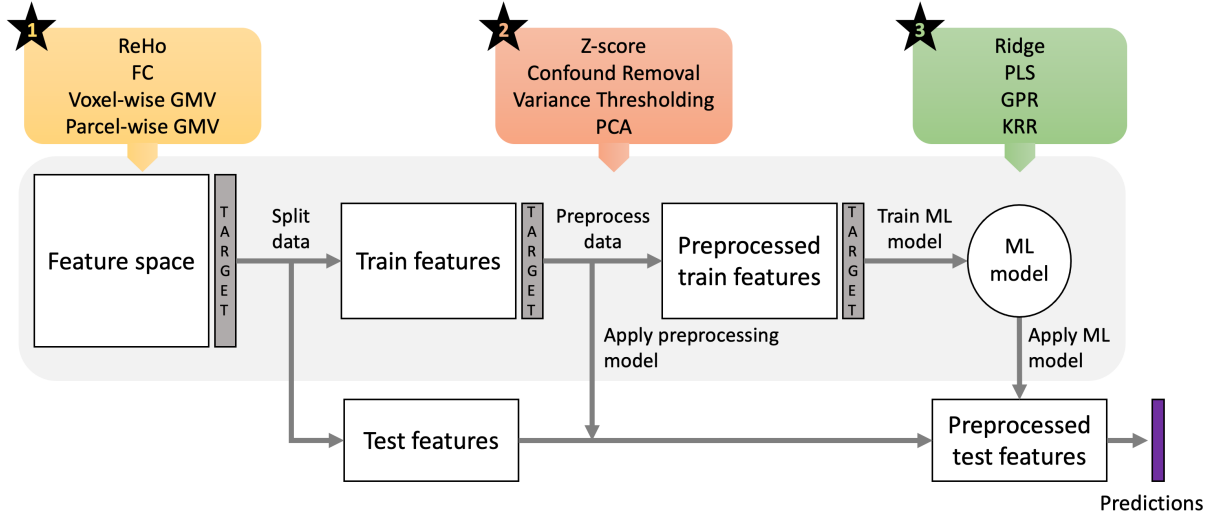


Figure 1: **Various steps in machine learning (ML) workflow design with some examples of (1) feature spaces, (2) preprocessing steps, and (3) ML algorithms.** First, the input data is split into training and test sets. Next, preprocessing steps are applied exclusively to the training features, and ML models are trained using these preprocessed training features and the target. Next, the preprocessing models from training data are applied to the test features. Finally, the trained ML model is applied to the preprocessed test features to get the test predictions.

CV), which can affect generalization estimates (Varoquaux, 2018, Scheinost et al., 2014, Poldrack et al., 2020). Additionally, the validation of models using external data holds pivotal importance in ensuring they are not overfitted and aids in evaluating their applicability in real-world scenarios. A comprehensive understanding of these factors is crucial to devising an improved study design. To achieve this goal, in study 4, we conducted a literature survey focusing on psychometric prediction, such as memory, fluid intelligence, and attention in healthy subjects. Our aim was to outline the current status and ongoing advancements concerning data, analysis methods, and reporting. This excluded papers related to sex and age prediction and clinical applications.

## 1.4 Ethics Protocols

The ethics protocols were approved by the Ethics Committee of Heinrich Heine University Düsseldorf (5193 and 2018-317-RetroDEuA).

## 1.5 Aims of Thesis

This work aims to assess several key components of ML workflows by predicting demographic traits, sex, and age using neuroimaging data. While the ultimate goal for ML in clinical application is to develop fair and trustworthy models to understand the

disease and deliver correct treatment, starting with reliable and clinically relevant targets such as sex and age can provide crucial understanding regarding key components of ML workflows.

In study 1, we evaluated the methods for confound removal to understand the effect of confounds in predictive modeling and the procedures to deal with them. This was studied using a sex prediction task (male vs. female) using ReHo and FC as features from rs-fMRI data, with brain size and age as confounds. The additional aim was the interpretability of the ML confound-free model to gain insights about brain regions involved in sex prediction. We aimed to answer an important biological question: “Are there differences in the functional organization of brains between males and females after controlling for the apparent difference in brain size?”.

In study 2, the aim was to establish a robust and reliable ML workflow for age prediction by evaluating several combinations for feature spaces derived from GMV (voxel-wise and parcel-wise) and ML algorithms and assessing them under different scenarios crucial for real-world applications. The additional aim was to explore the potential clinical value of the brain-age delta as a biomarker for brain health and factors affecting the estimation.

In study 3, we studied several preprocessing alternatives for VBM analysis commonly used for localized quantification of GMV and compared their utility for age estimation.

In study 4, we performed a comprehensive literature survey that examined previous studies investigating psychometric prediction based on neuroimaging data. By analyzing the patterns and findings from these studies, we aimed to identify established and novel concerns that can be effectively acknowledged and tackled in future studies.





**2 Confound Removal and Normalization in Practice:  
A Neuroimaging Based Sex Prediction Case  
Study.** More, S., Eickhoff, S.B., Caspers, J., Patil, K.R.,  
Machine Learning and Knowledge Discovery in Databases.  
Applied Data Science and Demo Track, 12461:3–18 (2021)

**Authorship contribution statement**

**Shammi More (Doctoral researcher, first author):** Formal analysis, Software, Validation, Visualization, Writing – original draft. **Simon B. Eickhoff:** Writing – review & editing, Supervision, Funding acquisition. **Julian Caspers:** Supervision, Writing – review & editing. **Kaustubh R. Patil (Corresponding author):** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.



# Confound Removal and Normalization in Practice: A Neuroimaging Based Sex Prediction Case Study

Shammi More<sup>1,2</sup> , Simon B. Eickhoff<sup>1,2</sup> , Julian Caspers<sup>3</sup>,  
and Kaustubh R. Patil<sup>1,2</sup> (  ) 

<sup>1</sup> Institute of Neuroscience and Medicine (INM-7),  
Forschungszentrum Jülich, Jülich, Germany  
{s.more,s.eickhoff,k.patil}@fz-juelich.de

<sup>2</sup> Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University  
Düsseldorf, Düsseldorf, Germany

<sup>3</sup> Department of Diagnostic and Interventional Radiology, University Hospital  
Düsseldorf, Düsseldorf, Germany  
julian.caspers@med.uni-duesseldorf.de

**Abstract.** Machine learning (ML) methods are increasingly being used to predict pathologies and biological traits using neuroimaging data. Here controlling for confounds is essential to get unbiased estimates of generalization performance and to identify the features driving predictions. However, a systematic evaluation of the advantages and disadvantages of available alternatives is lacking. This makes it difficult to compare results across studies and to build deployment quality models. Here, we evaluated two commonly used confound removal schemes—whole data confound regression (WDCR) and cross-validated confound regression (CVCR)—to understand their effectiveness and biases induced in generalization performance estimation. Additionally, we study the interaction of the confound removal schemes with Z-score normalization, a common practice in ML modelling. We applied eight combinations of confound removal schemes and normalization (pipelines) to decode sex from resting-state functional MRI (rfMRI) data while controlling for two confounds, brain size and age. We show that both schemes effectively remove linear univariate and multivariate confounding effects resulting in reduced model performance with CVCR providing better generalization estimates, i.e., closer to out-of-sample performance than WDCR. We found no effect of normalizing before or after confound removal. In the presence of dataset and confound shift, four tested confound removal procedures yielded mixed results, raising new questions. We conclude that CVCR is a better method to control for confounding effects in neuroimaging studies. We believe that our in-depth analyses shed light on choices associated with confound removal and hope that it generates more interest in this problem instrumental to numerous applications.

**Keywords:** Confound removal · Generalization · Interpretability · Sex classification · Neuroimaging application



# 1 Introduction

A critical challenge in applied machine learning is controlling for confounding effects as not removing them can lead to biased predictions and interpretations. This is especially true for biological data as common underlying processes introduce shared variance between the measurements, giving rise to confounding effects and blurring the boundaries between signals and confounds. Nevertheless, when confounds can be identified, removing their effects can lead to unbiased models and better understanding of the underlying biological processes.

In the field of neuroimaging, predictive analysis using machine learning has gained popularity for decoding phenotypes with a clear application to understand brain organization and its relationship to behavior and disease [9, 14, 41] with a twofold aim, (1) to establish brain-phenotype relationship by estimating the generalization performance, and (2) to identify brain regions explaining the variance of the phenotype. Cross-validation (CV) is employed for the first goal while the second goal is usually achieved by identifying predictive features, e.g., features with a high weight assigned by a linear model. Specifically, in addition to information uniquely associated with the target (true signal) neuroimaging features may also contain information from nuisance sources, e.g., brain size, confounding the relationship between the neuroimaging signal and the target. In this case, both goals can yield biased results as a successful prediction might be driven by the confounding signal rather than the true signal (Fig. 1a). Thus, the confounding effects need to be removed to estimate generalizability and to gain interpretability in an unbiased way. Various alternatives exist for confound removal and are integrated within ML pipelines. However, the pros and cons of these possibilities are not well understood.

Confounding can be controlled in the experiment design phase before data collection by randomization, restriction and matching [27]. However, this is not always feasible, e.g. when all the confounds are not known. Confounds can be controlled for after data acquisition. One way is to add them as additional predictors to capture the corresponding variance. However, this approach is not suitable for predictive modelling because it is designed to control in-sample rather than out-of-sample (OOS) properties. Another method is post-hoc counterbalancing i.e., taking a subset in which there is no empirical relationship between the confound and the target [35]. Advanced techniques such as the anti-mutual information sampling [10] and stratification using pooling analysis by the Mantel-Haenszel formula [38] have been proposed. However, these methods lead to data loss and are not feasible with a small sample and a large number of confounds. Specifically, when matching sexes according to brain size, these methods will represent extremes of the population and not the whole population. Of note, confound removal can be seen as supplementary to debiasing and fair learning [2, 16, 18] but here we do not investigate this angle.

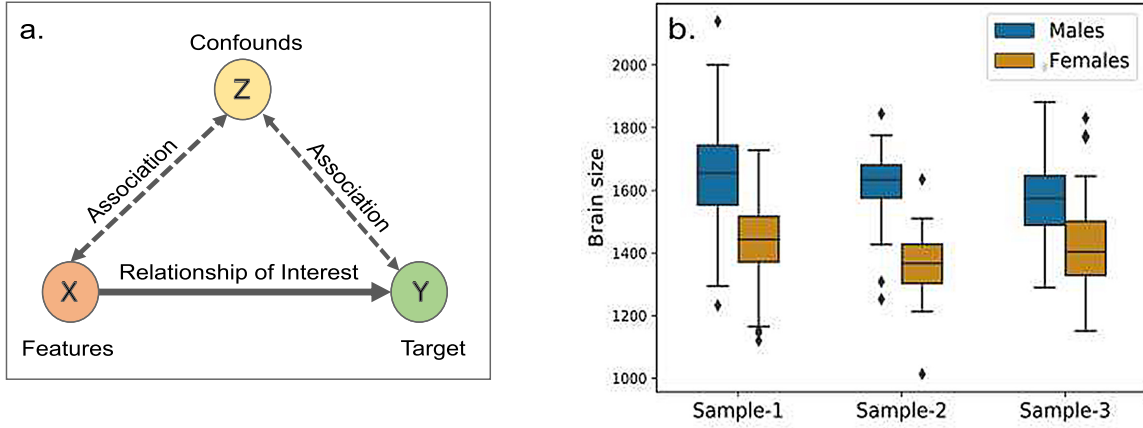
One of the most common confound control approaches while using all the data is “regressing out” their variance from the features before learning, referred to as confound regression [35] or image correction [28]. In this method, a linear regression model is fitted on each feature separately with the confounds as predictors, and the

corresponding residuals are used as new “confound-removed” features. This approach can be implemented in two possible ways. The first scheme is whole data confound regression (WDCR), regresses out confounds from the entire dataset at once [28, 35, 37] followed by CV to estimate the generalization performance. WDCR has yielded inconsistent results, from a substantial drop in performance [17, 37] to a similar or slightly lower performance compared to the models without confound control [28]. This discrepancy is possibly due to differences in the strength of the relationship between the confounds, the features, and the target and implementation differences. WDCR leads to “data-leakage” as the information from the whole data is used to create the confound-removed features before CV. However, the “aggressive” confound removal by WDCR has been proposed to be desirable [25].

To alleviate issues with WDCR, a CV-consistent scheme, cross-validated confound regression (CVCR) has been proposed in which the linear confound regression models are estimated within CV using only the training subset, and applied to both the training and the validation subsets. This avoids information leaking from training into validation sets. Although both WDCR and CVCR schemes have been used in neuroimaging studies [20, 35, 45], there is a lack of information regarding how they affect the generalization estimates and interpretability with one study recommending WDCR [25] while another recommending CVCR [35].

Moreover, whether to apply a feature normalization and standardization procedures, like Z-scoring (Zero mean and unit-variance features), before confound removal or after has not been investigated. It is known that in the specific case of normalization using rank-based inverse normal transformation (INT) after confound regression may reintroduce confounding effects [24]. Such reintroduction of confounding effects can be counterproductive for model generalizability and interpretability. Furthermore, the ability of an algorithm to learn from the data might differ depending upon when normalization is applied. This lack of understanding about the interaction between confound regression and normalization makes it difficult to design ML pipelines. Lastly, building models when one suspects a shift in the covariates and/or in the relationship between the confounds, the features and the target has not been studied. Several design possibilities can be imagined and need to be evaluated.

In this work we empirically investigate three facets of the confound removal issue, (1) evaluation of the two confound removal schemes, WDCR and CVCR, for their effectiveness in removing the confounding signal and estimation of generalization performance, (2) interaction of confound removal schemes with normalization, and (3) model deployment when covariate and confounding effect shift is suspected. We consider prediction of sex from resting-state functional magnetic resonance imaging (rfMRI) data while controlling for two confounds, brain size and age. We aim to answer an important biological question “are male and female brains functionally different after controlling for the apparent difference in brain size?”. With systematic evaluation of a real-world problem reporting positive as well as negative results, we hope to attract the attention of the machine learning community to the critical problem of confound removal.



**Fig. 1.** (a) Confounding effect: Confound Z influences both the features X and the target Y. In the presence of Z, the actual relationship between X and Y is masked. For sex classification, brain size is a confound (Z) as it is associated with both rfMRI features (X) and sex (Y). (b) Significant sex difference in brain size in the three data samples used in this study.

## 2 Sex Classification and Brain Size

There are reports on differences in cognition and psychopathology between sexes [33], such as differences in spatial tasks [22], females being more vulnerable to depression [26] and autism being more prevalent in males [42]. These differences may influence diagnostic practices and help developing sex-specific treatments, making understanding neurobiology of sex differences essential. Accordingly there has been an increasing interest in finding sex differences in structural and functional properties of the brain [29, 30, 41].

Functional magnetic resonance imaging (fMRI) is a non-invasive technique used to study functional—i.e. time dependent—changes in brain activity by taking 3D MRI images in succession. Even unregulated processes in the resting brain, i.e., resting-state fMRI (rfMRI), show stable and individualized synchronies [12]. Such functional activities have been related to cognition and several phenotypes, especially using the functional connectivity (FC) (see Sect. 4.2). Based on whole-brain FC, the sex prediction accuracy of 75–80% was achieved with discriminative features mainly located in the Default mode network (DMN) [41, 45]. Another study with a lower prediction accuracy of 62% found discriminative FC in motor, sensory, and association areas [6]. Smith and colleagues [34] reported a higher prediction accuracy of 87%. A recent study reported sex prediction accuracy of 98% using multi-label learning, i.e., sex in conjunction with nine other cognitive, behavioural and demographic variables [8].

Brain size is highly correlated with sex, with larger total brain volume in males compared to females [4, 29]; and is encoded in MRI data. Figure 1b shows the difference in brain size between sexes for the data samples used in the current study. In such a scenario, even if a model decodes sex from MRI data significantly above chance, there is no clear understanding of the unique contribution of the

functional features independent of brain size. It is likely that the prediction is driven partly by brain size in addition to the functional differences. Zhang and colleagues [45] have shown that the sex prediction accuracy drops from 80% to 70% after regressing out brain size from FC, indicating an apparent effect of brain size in sex prediction. Hence, there is clearly a need to study sex prediction using rfMRI while controlling for brain size.

### 3 Experimental Setup

#### 3.1 Study Design

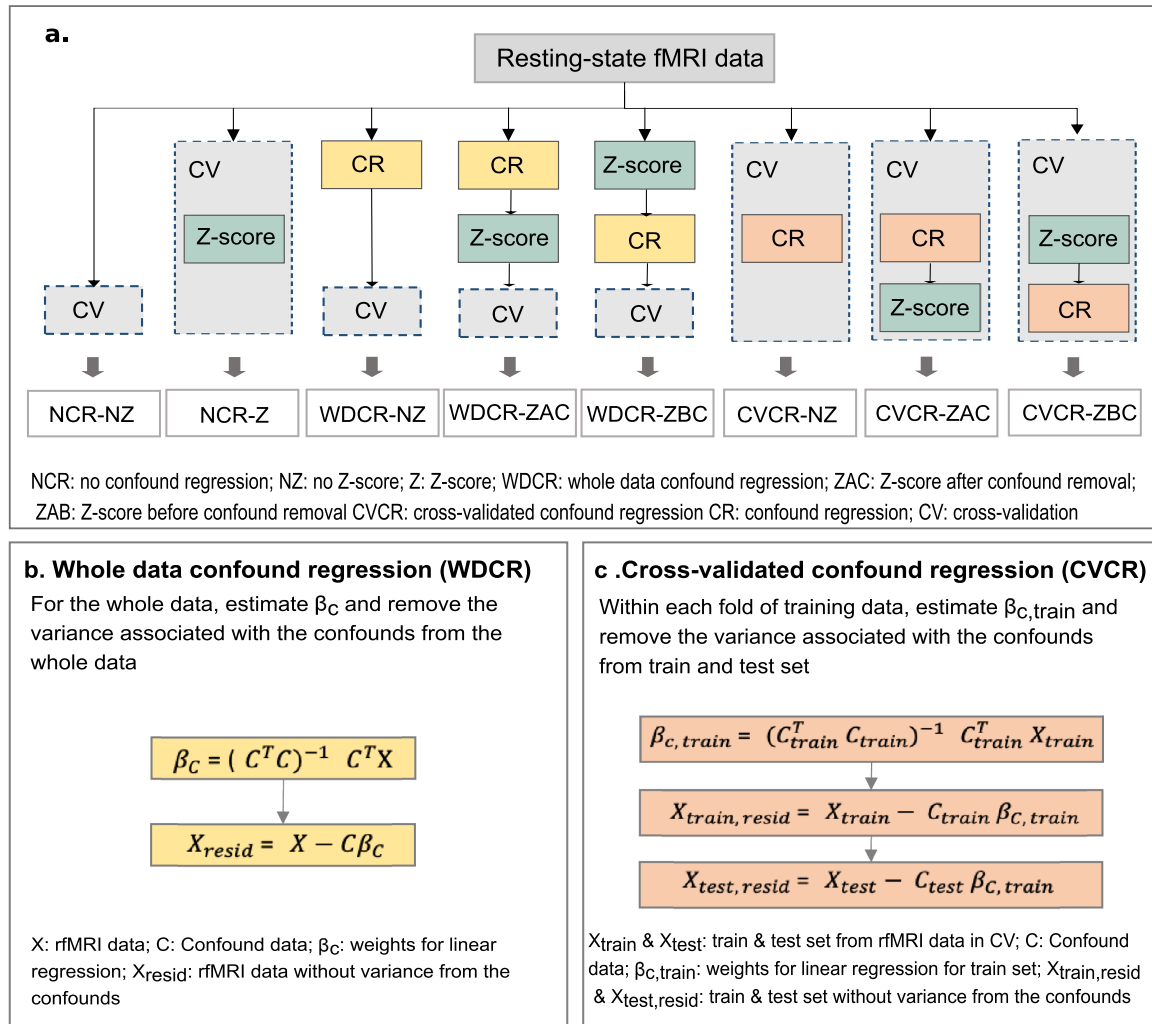
With a limited and contrasting literature, there is a lack of knowledge of how to perform confound removal. Here we aimed to evaluate two confound removal schemes (WDCR and CVCR) and their interaction with the commonly used Z-score feature normalization. We evaluated eight pipelines in total (Fig. 2a);

1. No confound removal, no Z-scoring (NCR-NZ)
2. No confound removal, with Z-scoring (NCR-Z)
3. WDCR, no Z-scoring (WDCR-NZ)
4. WDCR, Z-scoring after confound removal (WDCR-ZAC)
5. WDCR, Z-scoring before confound removal (WDCR-ZBC)
6. CVCR, no Z-scoring (CVCR-NZ)
7. CVCR, Z-scoring after confound removal (CVCR-ZAC)
8. CVCR, Z-scoring before confound removal (CVCR-ZBC)

We applied these pipelines for predicting an individual’s sex using features derived from rfMRI data while controlling for two confounds brain size and age. We performed two evaluations; (1) CV to estimate the generalization performance and compared it with prediction on an OOS dataset, and (2) OOS prediction with covariate and confound shift as a model deployment scenario. The prediction performance was evaluated using AUC, F1-score and balanced accuracy.

For evaluation-1, we used a publicly available database (HCP, see Sect. 4.1) and carefully derived sample-1 ( $N = 377$ ) and sample-2 ( $N = 54$ ). After standard preprocessing two types of features were extracted from rfMRI data, Regional Homogeneity (ReHo) and FC (see Sect. 4.2). Each feature set was analyzed separately using Ridge Regression and Partial Least Square Regression with all eight pipelines. The generalization performance was estimated on sample-1 using 10 times repeated 10-fold CV. The OOS performance was evaluated on sample-2. By comparing the CV and OOS results, we can comment on whether the CV procedure can reliably estimate the generalization performance.

As the confounds were linearly removed from the features in a univariate way (see Sect. 3.2) multivariate confounding effects might still remain. We, therefore, assessed the effectiveness of confound removal pipelines in removing univariate and multivariate confounding effects. The Pearson correlation between each



**Fig. 2.** a. The schematic diagram of various combinations of confound removal schemes and Z-score for confound removal evaluated in the study. b. Whole data confound regression (WDCR). c. Cross-validated confound regression (CVCR).

residual feature and the brain size was calculated to check for remaining univariate confounding effects. The adjusted  $r^2$  of the multiple linear regression model predicting the brain size using residual features was used to check for remaining multivariate confounding effects.

In neuroimaging studies it is common that the data is acquired on different scanners [40] and there may exist demographic differences between samples. Such differences can lead to covariate shift [19] and by extension confound shift. An ideal model should generalize well despite such differences. To evaluate this (evaluation-2), we employed an additional sample (sample-3;  $N = 484$ ) from a public dataset (eNKI, see Sect. 4.1) where demographics, scanner parameters and preprocessing are different than sample-1 and 2. We tested four ways to remove confounds from OOS data.

1. **Train-to-test:** The confound removal models from the train data were applied to the OOS data. This is the standard method.
2. **Test WDCR:** WDCR was performed on the OOS data.

3. **Test CVCR:** CVCR was performed on the OOS data, i.e. confound regression was performed within CV for OOS data and the residuals were retained.
4. **Train and test combined:** WDCR was performed on the combined train and OOS data. The data was then re-split into train and test.

Methods 2, 3 aimed to obtain confound-free OOS data, with the assumptions that confound-removed models can perform well on confound-removed OOS data as confounds are handled within a sample. Method 4 assumes that the confound removal linear models can capture variance from both train and OOS data. Note that 2, 3 and 4 can only be used with sufficiently large OOS data. WDCR models trained on sample-1 were used to predict the confound-removed OOS data. The sample-2 and sample-3 with similar and different properties to sample-1 respectively were the OOS datasets. Note that for method 1, 2 and 3 trained models (on sample-1) come from the above-mentioned pipelines used for evaluation-1.

### 3.2 Confound Regression

We tested two different versions of confound regression, WDCR and CVCR (Fig. 2b and c). In WDCR, using multiple linear regression we regressed out the confounds from each of the predictors from the entire dataset before the cross-validated procedure. Note that, this procedure uses information from the whole dataset leading to data-leakage. In CVCR, we regressed the confounds in a similar way to WDCR but the confound removal models were estimated on the training data and subsequently applied to both train and validation sets. In this way, there is no leakage from train to test.

### 3.3 Predictive Modelling

We used two prediction models, Ridge Regression and Partial Least Square regression. Ridge Regression (RR) uses a sum of the square penalty on the model parameters to reduce model complexity and prevent overfitting [15]. The balance between the fit and the penalty is defined using a hyper-parameter  $\lambda$  which we tuned in an inner CV loop. PLS Regression (PLS) performs dimensionality reduction and learning simultaneously, making it a popular choice when there are more features than observations, and/or when there is multicollinearity among the features. It has performed well in MRI-based estimations for cognitive, behavioural and demographic variables [8, 45]. PLS searches for a set of latent vectors that performs a simultaneous decomposition of predictors and the target such that these components explain the maximum covariance between them [1]. These latent vectors are then used for prediction. The hyperparameter for the PLS is the number of latent variables which was tuned in an inner CV loop.

## 4 Data Samples and Features

### 4.1 Data Samples

This study included three samples. Sample 1 and 2 are two independent subsets of the data provided by the Human Connectome Project (HCP) S1200 release



[39]. Sample-1 contained 377 subjects (age range: 22–37, mean age: 28.6 years; 182 females), sample-2 comprised 54 subjects (age range: 22–36, mean age: 28.9 years; 28 females). As the HCP data contains siblings and twins, the samples were constructed such that there were no siblings within or across the two samples, to avoid biases due to any similarity in the FC of the siblings. Within each of the two samples, males and females were matched for age, and education. Resting-state blood oxygen level-dependent (BOLD) data comprised 1200 functional volumes per subject, acquired on a Siemens Skyra 3T scanner with the following parameters: TR = 720 ms, TE = 33.1 ms, flip angle =  $52^\circ$ , voxel size =  $2 \times 2 \times 2 \text{ mm}^3$ , FoV =  $208 \times 180 \text{ mm}^2$ , matrix =  $104 \times 90$ , slices = 72. Sample-3 was obtained from the Enhanced Nathan Kline Institute–Rockland Sample (eNKI-RS) [23] with 484 subjects (age range: 6–85, mean age: 41.9 years; 311 females). Images were acquired on a Siemens TimTrio 3T scanner using BOLD contrast with the following parameters: TR = 1400 ms, TE = 30 ms, flip angle =  $65^\circ$ , voxel size =  $2 \times 2 \times 2 \text{ mm}^3$ , slices = 64. Subjects were asked to lie with eyes open, with “relaxed” fixation on a white cross (on a dark background), think of nothing in particular, and not to fall asleep. The CAT-12 toolbox (<http://www.neuro.uni-jena.de/cat/>) was used to calculate the brain size of each subject based on T1-weighted images. Note the stark differences between sample-1, 2 and sample-3 in terms of demographics as well as scanner parameters. This selection was made to elucidate the common scenario of data heterogeneity.

Two-sample t-test revealed significant sex differences in the brain size across all the samples ( $p < 0.001$ ; Fig. 1b). This clearly demonstrates that brain size is a confound in sex prediction. There was no difference in age between sexes in sample-1 but significant differences was observed in sample-2 and 3 ( $p < 0.001$ ). Age is not expected to be related to sex but was included as a control confound.

## 4.2 Pre-processing and Feature Extraction

After standard rfMRI pre-processing we extracted two types of features based on the voxel-wise time-series.

**Preprocessing.** The rfMRI data needs to be pre-processed so that the effects of motion in the scanner are removed as well as the brain of each subject is normalized to a standard brain template (e.g., MNI-152) so that they can be compared across subjects. For samples 1 and 2, the pre-processed, FIX-denoised and spatially normalized to the MNI-152 template data provided by the HCP S1200 release was used. There was no difference in the movement parameters (measured as mean framewise displacement) between males and females in both the samples. No further motion correction was performed. For sample-3, physical noise and effects of motion in the scanner were removed by using FIX (FMRIB’s ICA-based Xnoiseifier, version 1.061 as implemented in FSL 5.0.9; [13,31]). Unique variance related to the identified artefactual independent components and 24 movement parameters [32] were then regressed from the data. The FIX-denoised data were further preprocessed using SPM8 (Wellcome Trust Centre for Neuroimaging, London) and in-house Matlab scripts for movement correction and spatial normalization to the MNI-152 template [3].

**Regions of Interest (ROI).** The Dosenbach atlas was used to extract 160 ROIs from the whole-brain data. These ROIs are spheres of 10 mm diameter, identified from a series of meta-analyses of task-related fMRI studies and broadly cover much of the cerebral cortex and cerebellum [11]. This atlas has been utilized in several brain network analyses including for sex prediction [5, 45].

**Feature Space 1: Regional Homogeneity (ReHo)** measures the similarity of the time-series of a set of voxels and thus reflects the temporal synchrony of the regional BOLD signal [44]. ReHo for each subject and each of the 160 ROIs was calculated as the Kendall’s coefficient of concordance between all the time-series of the voxels within a given ROI resulting in 160 features per subject.

**Feature Space 2: Functional Connectivity (FC)** is the correlation between the time-series of different brain regions [36]. For each subject, the time series of all the voxels within a ROI were averaged and FC was calculated as the Pearson’s correlation coefficients between them for all pairs of ROI. These were then transformed using Fisher’s Z-score. Each subject had a feature vector of length 12,720 after vectorization of the lower triangle of the  $160 \times 160$  FC matrix.

## 5 Results

We compiled the results from two viewpoints. We first asked which of the pipelines incorporating confound removal provides more realistic generalization performance estimates. Then we assessed the efficacy of the confound removal schemes in a model deployment scenario with data heterogeneity.

### 5.1 Generalization Performance Estimates

CV is commonly used to estimate generalization performance. However, it is not without caveats [7]. Therefore, we compared CV performance of the pipelines with “true” OOS performance. In this case, the CV was performed on sample-1 and sample-2 was used as the OOS data. PLS generally performed better than RR, so in the following we focus on the PLS results.

As expected, the CV performance was highest without controlling for confounds (Table 1). AUC and F1-scores for sex prediction with ReHo were 0.838 and 0.754 and with FC were 0.874 and 0.787, respectively. Both the schemes WDCR and CVCR showed reduced performance in line with previous studies [25, 35]. As brain size is highly correlated to sex, regressing it out from every feature can remove sex-specific information, resulting in a lower performance.

WDCR provided lower generalization estimates than CVCR, with the balanced accuracy dropping close to chance level with WDCR. One might expect higher generalization performance with WDCR as it causes data leakage from the train to the validation set violating the crucial assumption of independence in cross-validated analysis. However, in this case, it leads to worse performance. This might be because WDCR is performed on the whole dataset and hence is more aggressive in removing the confounding signal than CVCR leading to poorer performance. When the trained models were applied to OOS data, we



**Table 1.** Comparison of the pipelines using RR and PLS. Models were trained on sample-1 and out-of-sample/test performance was tested on sample-2.

CR	Z-score	Feat.	Ridge regression						Partial least squares					
			CV: Sample-1			Test: Sample-2			CV: Sample-1			Test: Sample-2		
			AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.
NCR	NZ	ReHo	0.750	0.667	0.662	0.751	0.690	0.688	0.776	0.714	0.712	0.808	0.759	0.760
		FC	0.857	0.763	0.757	0.823	0.728	0.725	0.874	0.787	0.785	0.835	0.762	0.761
NCR	Z	ReHo	0.829	0.749	0.746	0.832	0.759	0.758	0.838	0.754	0.751	0.860	0.778	0.776
		FC	0.860	0.772	0.768	0.841	0.765	0.762	0.860	0.781	0.779	0.813	0.765	0.762
WDCR	NZ	ReHo	0.477	0.490	0.490	0.511	0.500	0.500	0.476	0.494	0.494	0.685	0.647	0.647
		FC	0.466	0.488	0.496	0.607	0.500	0.500	0.417	0.454	0.455	0.685	0.661	0.654
	ZAC	ReHo	0.528	0.523	0.522	0.501	0.500	0.500	0.553	0.548	0.546	0.735	0.685	0.683
		FC	0.467	0.482	0.483	0.611	0.500	0.500	0.409	0.444	0.446	0.677	0.578	0.577
	ZBC	ReHo	0.528	0.528	0.526	0.501	0.500	0.500	0.553	0.546	0.545	0.735	0.685	0.683
		FC	0.456	0.476	0.478	0.611	0.500	0.500	0.407	0.444	0.445	0.677	0.578	0.577
CVCR	NZ	ReHo	0.552	0.522	0.519	0.511	0.500	0.500	0.569	0.553	0.553	0.685	0.647	0.647
		FC	0.516	0.500	0.500	0.607	0.500	0.500	0.595	0.576	0.575	0.685	0.661	0.654
	ZAC	ReHo	0.632	0.589	0.585	0.577	0.611	0.518	0.668	0.637	0.634	0.694	0.666	0.665
		FC	0.543	0.532	0.529	0.661	0.592	0.582	0.588	0.565	0.563	0.705	0.595	0.595
	ZBC	ReHo	0.634	0.591	0.587	0.577	0.611	0.518	0.669	0.635	0.633	0.703	0.666	0.665
		FC	0.547	0.532	0.529	0.662	0.592	0.582	0.586	0.564	0.563	0.705	0.595	0.595

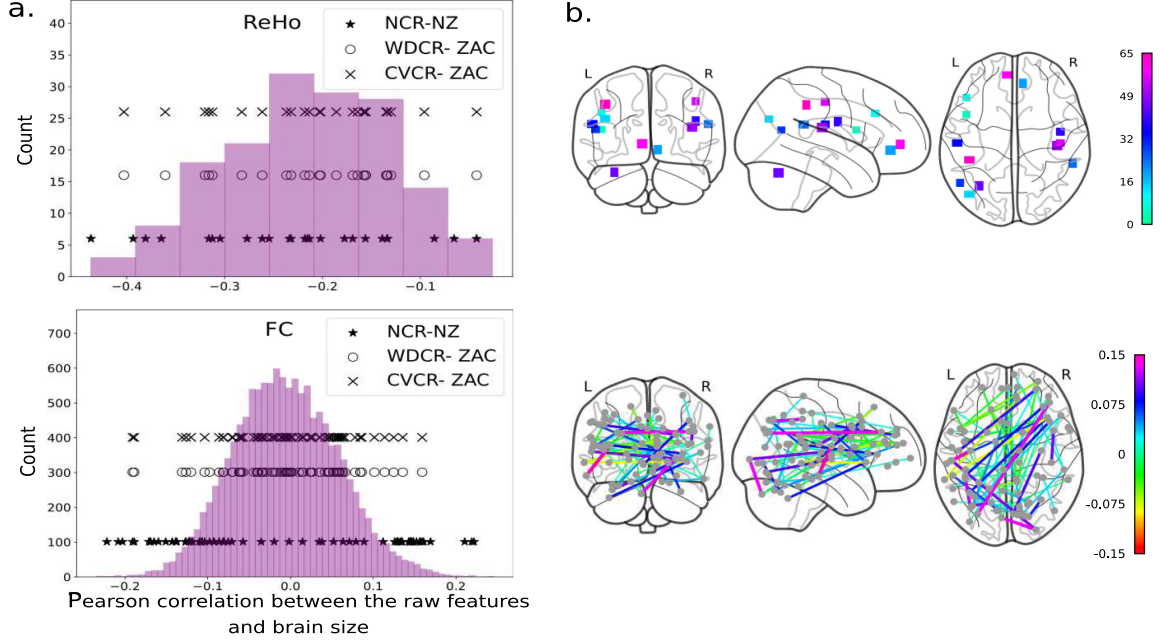
found that OOS performance was higher than the CV estimates for most of the pipelines. This might happen if the OOS data is easier to classify. The OOS performance was closer to the generalization performance estimated with CVCR. This result suggests that CVCR is a better way to do confound removal in predictive analyses with neuroimaging data.

We then checked whether the confound removal was happening as expected. First, in a univariate way we correlated the residuals (new features) with the confounding variables. We found no significant correlation with both confound removal schemes indicating effective univariate removal of the confounding signal from the features. However, as multivariate effects might still be remaining, we used multiple linear regression to predict brain size from the residual features. With CVCR and WDCR, these models on the training sets revealed negative adjusted  $r^2$ . This indicates that there were no remaining linear multivariate confounding effects with both WDCR and CVCR. Thus the models trained with the residual features contained no information from the confounds.

These trends were similar for both ReHo and FC. Z-scoring improved the model performance with ReHo but not with FC. There was no effect of Z-scoring the features before (raw features) or after (residuals) confound removal.

## 5.2 Predictive Features

One of the main objectives of a decoding analysis is to identify predictive features (brain regions) explaining the variance in phenotype. As the confounding effect can impact predictive features selection, it is important to compare them with and without confound removal. The Z-scored feature weights (the absolute value) averaged across CV runs were used to select predictive features. We found that predictive features with and without confound removal were different (Fig. 3).



**Fig. 3.** a. Pearson correlation between the raw features and the brain size as histograms. The dots show the correlations of the selected features (jittered); 25 for ReHo (top) and 70 for FC (bottom) for NCR-NZ, WDCR-ZAC and CVCR-ZAC pipelines. b. Brain regions associated with the selected features; ReHo (top, relative weights), and FC (bottom), both with the CVCR-ZAC pipeline.

We compared 25 ReHo and 70 FC features with highest absolute weights from 3 pipelines, NCR-NZ, WDCR-ZAC and CVCR-ZAC (Fig. 3a). The features selected without confound removal had relatively higher positive or negative correlation with brain size. However, after confound removal (WDCR and CVCR), for FC the features with lower correlation were selected. This suggests that the features selected after confound removal represent the functional signal predictive of sex. We then identified features selected after confound removal (CVCR-ZAC) but not selected without confound removal (NCR-NZ) (Fig. 3b). With ReHo, selected regions were in dorsolateral prefrontal cortex, inferior parietal lobule, occipital, ventromedial prefrontal cortex, precentral gyrus, post insula, parietal, temporoparietal junction and inferior cerebellum, in line with a study identifying regions in the inferior parietal lobule and precentral gyrus [43]. In contrast, another study found sex differences in right hippocampus and amygdala [21]. We found important FC features widespread across the entire brain with strong inter-hemispheric connections. In contrast to the study by Zhang and colleagues [45] we did not find many intra-network FC in the DMN. Z-score feature normalization before or after confound removal did not affect selected features.

### 5.3 Out-of-Sample Performance

To study how a model deployment would work, especially in the presence of data heterogeneity common in neuroimaging studies, we tested four different ways to remove confounds from the OOS data including, applying confound models

from train to OOS data using CVCR-ZAC pipeline, self-confound removal on the OOS data using WDCR and CVCR, and WDCR on the combined train and OOS data. The Z-score normalization was performed after the confound removal (ZAC) and PLS was used for prediction.

For sample-2, train-to-test confound removal showed best performance compared to other three methods (Table 2). This is expected as the properties of these two samples are expected to be similar (i.e., no data shift). Even though, residual correlations were observed in the OOS data after applying confound models from train data (Fig. 4a), the training models were confound-free so this performance cannot be driven by confounding effects.

For sample-3 (data shift expected), we observed mixed results. For ReHo, the combined WDCR model (learned on the train data) gave highest performance (Table 2b). However, significant correlation was present between the residual features and brain size in both train and OOS data (Fig. 4b). This might indicate that the performance is driven by confounding effects. A similar model using FC was lowest performing. With combined WDCR, it seems like the dataset with higher variance dominates leaving the other part correlated, indicating it might be suboptimal. Predictions on self-confound removed OOS data (sample-3) (Test WDCR and Test CVCR) were similar to when the confound models from sample-1 were applied (Table 2a). However, the OOS performance using ReHo dropped compared to CV while that of FC improved.

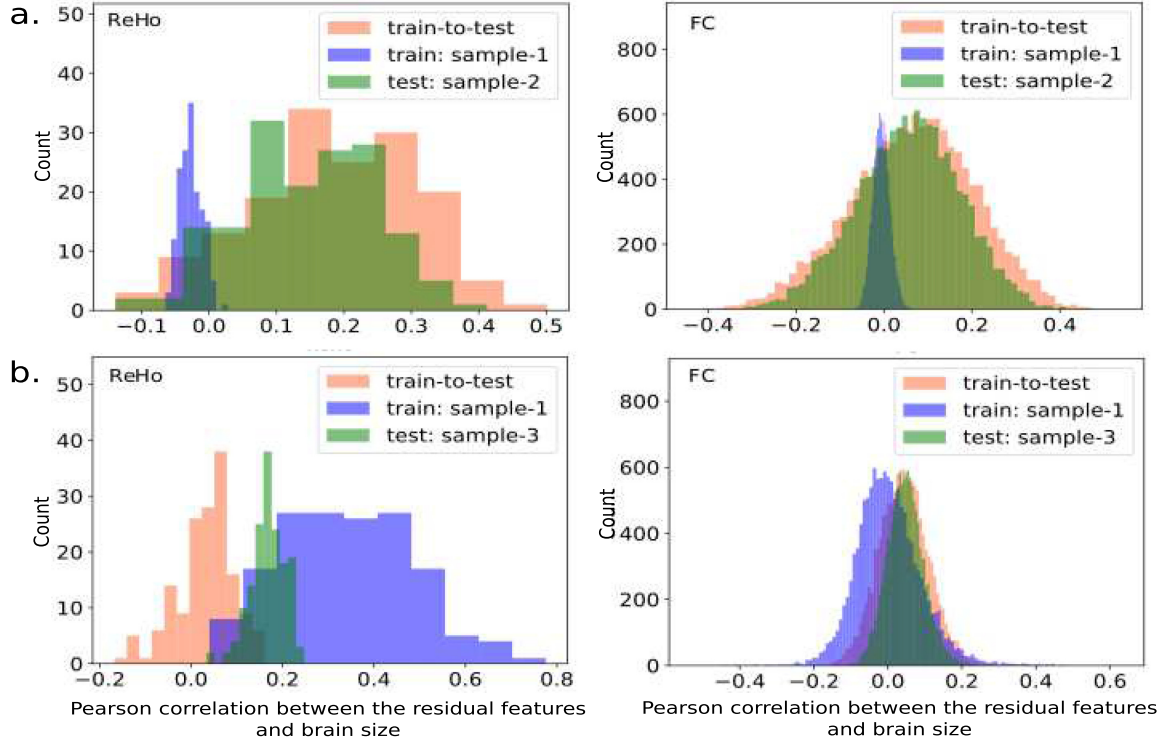
**Table 2.** Comparison of confound removal schemes on out-of-sample/test data. a. Confound models learned from the train data (sample-1) applied to test data (sample-2 and 3), WDCR and CVCR performed only on test data. b. WDCR on the combined train and test data.

a. Method		CV: Sample-1			Test: Sample-2			Test: Sample-3		
	Feat.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.
<b>Train-to test:</b>	ReHo	0.668	0.637	0.634	0.694	0.666	0.665	0.549	0.528	0.527
	CVCR-ZAC	0.588	0.565	0.563	0.705	0.595	0.595	0.637	0.628	0.619
<b>Test WDCR:</b>	ReHo	0.553	0.548	0.546	0.562	0.573	0.573	0.524	0.530	0.531
	WDCR-ZAC	0.409	0.444	0.446	0.632	0.576	0.576	0.635	0.592	0.595
<b>Test CVCR:</b>	ReHo	0.668	0.637	0.634	0.582	0.591	0.591	0.505	0.508	0.509
	CVCR-ZAC	0.588	0.565	0.563	0.603	0.578	0.577	0.634	0.597	0.601

b. Feat.		CV: Sample-1			Test: Sample-2			CV: Sample-1			Test: Sample-3		
		AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.
ReHo		0.533	0.538	0.538	0.580	0.558	0.560	0.870	0.788	0.786	0.614	0.577	0.502
FC		0.450	0.459	0.461	0.387	0.409	0.412	0.871	0.779	0.777	0.541	0.502	0.501

Taken together, we found that train-to-test application of confound removal models and self-confound removal to be better strategies but inconsistent across feature spaces. This raises questions regarding optimal confound removal strategies when data heterogeneity is present. Based on the results, we also speculate that covariate and confound shift is more pronounced in ReHo compared to FC.



**Fig. 4.** Correlation between the residual features and brain size: for out-of-sample/test data when training confound removal models were applied (orange), and for train (purple) and test (green) data when combined train and test WDCR was performed. (Color figure online)

## 6 Conclusion

In this study, several confound removal pipelines were tested on the task of rfMRI data based sex classification. As expected, the two confound removal schemes (WDCR and CVCR) could effectively remove the signal corresponding to confounds leading to a substantial drop in prediction performance compared to without confound removal. Analyses on the residual features after WDCR and CVCR revealed that there were no remaining univariate and multivariate confounding effects. Thus, both these confound removed models should not have confound-related information encoded. We found CVCR to be a better method compared to WDCR as CVCR estimated generalization performance was closer to OOS performance. As WDCR leads to data leakage, one might expect it to be over-optimistic. However, our results point to the opposite. This is likely due to the aggressive confound removal. Our findings provide further corroboration to the idea of applying data analysis operations within the CV loop. In this work we focused on the sex prediction problem and whether our results apply to other problems remains to be seen.

The Z-score normalization of the features before or after confound removal did not affect model performance. We recommend to normalize after confound removal, as some learning algorithms might benefit from well-scaled features.

We also found that the OOS performance was best when the confound models from the train data were used, provided that the sample properties between train and test are similar but results were inconsistent with data shift. Although we used multiple regression to test for remaining multivariate confounding effects, we are not aware of a method that can directly remove multivariate effects. This calls for further investigations and development of new methods.

**Acknowledgments.** This study was supported by the Deutsche Forschungsgemeinschaft (DFG, PA 3634/1-1 and EI 816/21-1), the Helmholtz Portfolio Theme “Supercomputing and Modelling for the Human Brain” and the European Union [Horizon 2020 Research and Innovation Programme under Grant Agreement No. 945539 (HBP SGA3)].

## References

1. Abdi, H.: Partial least squares regression and projection on latent structure regression (pls regression). *Wiley Interdiscip. Rev. Comput. Stat.* **2**(1), 97–106 (2010)
2. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Niebles, J.C., et al.: Representation learning with statistical independence to mitigate bias. [arXiv:1910.03676](https://arxiv.org/abs/1910.03676) (2019)
3. Ashburner, J., Friston, K.J.: Unified segmentation. *Neuroimage* **26**(3), 839–851 (2005)
4. Barnes, J., Ridgway, G.R., Bartlett, J., Henley, S.M., Lehmann, M., Hobbs, N., et al.: Head size, age and gender adjustment in mri studies: a necessary nuisance? *Neuroimage* **53**(4), 1244–1255 (2010)
5. Cao, M., Wang, J.H., Dai, Z.J., Cao, X.Y., Jiang, L.L., Fan, F.M., et al.: Topological organization of the human brain functional connectome across the lifespan. *Developmental Cognitive Neurosci.* **7**, 76–93 (2014)
6. Casanova, R., Whitlow, C., Wagner, B., Espeland, M., Maldjian, J.: Combining graph and machine learning methods to analyze differences in functional connectivity across sex. *The Open Neuroimaging Journal* **6**, 1 (2012)
7. Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Machine Learn. Res.* **11**, 2079–2107 (2010)
8. Chen, C., Cao, X., Tian, L.: Partial least squares regression performs well in mri-based individualized estimations. *Front. Neurosci.* **13**, 1282 (2019)
9. Chen, J., Patil, K.R., Weis, S., Sim, K., Nickl-Jockschat, T., Zhou, J., et al.: Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: An international machine learning study. *Biol. Psychiatry* **87**(3), 282–293 (2020)
10. Chyzhyk, D., Varoquaux, G., Thirion, B., Milham, M.: Controlling a confound in predictive models with a test set minimizing its effect. In: 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI), pp. 1–4. IEEE (2018)
11. Dosenbach, N.U., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., et al.: Prediction of individual brain maturity using FMRI. *Science* **329**(5997), 1358–1361 (2010)
12. Fox, M.D., Raichle, M.E.: Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* **8**(9), 700–711 (2007)
13. Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., et al.: Ica-based artefact removal and accelerated FMRI acquisition for improved resting state network imaging. *Neuroimage* **95**, 232–247 (2014)



14. Hahn, T., Nierenberg, A., Whitfield-Gabrieli, S.: Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Molecular Psychiatry* **22**(1), 37–43 (2017)
15. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **42**(1), 80–86 (2000)
16. Kilbertus, N., Ball, P.J., Kusner, M.J., Weller, A., Silva, R.: The sensitivity of counterfactual fairness to unmeasured confounding. [arXiv:1907.01040](https://arxiv.org/abs/1907.01040) (2019)
17. Kostro, D., Abdulkadir, A., Durr, A., Roos, R., Leavitt, B.R., Johnson, H., et al.: Correction of inter-scanner and within-subject variance in structural mri based automated diagnosing. *NeuroImage* **98**, 405–415 (2014)
18. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: *Advances in Neural Information Processing Systems*, pp. 4066–4076 (2017)
19. Landeiro, V., Culotta, A.: Robust text classification under confounding shift. *J. Artif. Intell. Res.* **63**, 391–419 (2018)
20. Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., et al.: Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage* **148**, 179–188 (2017)
21. Lopez-Larson, M.P., Anderson, J.S., Ferguson, M.A., Yurgelun-Todd, D.: Local brain connectivity and associations with gender and age. *Dev. Cogn. Neurosci.* **1**(2), 187–197 (2011)
22. Miller, D.I., Halpern, D.F.: The new science of cognitive sex differences. *Trends in Cognitive Sciences* **18**(1), 37–45 (2014)
23. Nooner, K.B., Colcombe, S., Tobe, R., Mennes, M., Benedict, M., Moreno, A., et al.: The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Front. Neurosci.* **6**, 152 (2012)
24. Pain, O., Dudbridge, F., Ronald, A.: Are your covariates under control? how normalization can re-introduce covariate effects. *Euro. J. Hum. Genet.* **26**(8), 1194–1201 (2018)
25. Pervaiz, U., Vidaurre, D., Woolrich, M.W., Smith, S.M.: Optimising network modelling methods for FMRI. *NeuroImage* **211**, 116604 (2020)
26. Picco, L., Subramaniam, M., Abdin, E., Vaingankar, J.A., Chong, S.A.: Gender differences in major depressive disorder: findings from the singapore mental health study. *Singapore Med. J.* **58**(11), 649 (2017)
27. Pourhoseingholi, M.A., Baghestani, A.R., Vahedi, M.: How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench* **5**(2), 79 (2012)
28. Rao, A., Monteiro, J.M., Mourao-Miranda, J., Initiative, A.D., et al.: Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage* **150**, 23–49 (2017)
29. Ritchie, S.J., Cox, S.R., Shen, X., Lombardo, M.V., Reus, L.M., Alloza, C., et al.: Sex differences in the adult human brain: evidence from 5216 uk biobank participants. *Cerebral Cortex* **28**(8), 2959–2975 (2018)
30. Ruigrok, A.N., et al.: A meta-analysis of sex differences in human brain structure. *Neurosci. Biobehav. Rev.* **39**, 34–50 (2014)
31. Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., et al.: Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* **90**, 449–468 (2014)
32. Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughhead, J., Calkins, M.E., et al.: An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage* **64**, 240–256 (2013)

33. Seeman, M.V.: Psychopathology in women and men: focus on female hormones. *Am. J. Psychiatry* **154**(12), 1641–1647 (1997)
34. Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., et al.: Resting-state fMRI in the human connectome project. *Neuroimage* **80**, 144–168 (2013)
35. Snoek, L., Miletić, S., Scholte, H.S.: How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage* **184**, 741–760 (2019)
36. Stephan, K., Friston, K., Squire, L.: Functional connectivity. *Encyclopedia of Neuroscience*, pp. 391–397 (2009)
37. Todd, M.T., Nystrom, L.E., Cohen, J.D.: Confounds in multivariate pattern analysis: theory and rule representation case study. *Neuroimage* **77**, 157–165 (2013)
38. Tripepi, G., Jager, K.J., Dekker, F.W., Zoccali, C.: Stratification for confounding-part 1: the mantel-haenszel formula. *Nephron Clin. Pract.* **116**(4), c317–c321 (2010)
39. Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., et al.: The human connectome project: a data acquisition perspective. *Neuroimage* **62**(4), 2222–2231 (2012)
40. Wachinger, C., Becker, B.G., Rieckmann, A., Pölsterl, S.: Quantifying confounding bias in neuroimaging datasets with causal inference. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) *MICCAI 2019*. LNCS, vol. 11767, pp. 484–492. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32251-9\\_53](https://doi.org/10.1007/978-3-030-32251-9_53)
41. Weis, S., Patil, K.R., Hoffstaedter, F., Nostro, A., Yeo, B.T., Eickhoff, S.B.: Sex classification by resting state brain connectivity. *Cerebral Cortex* **30**(2), 824–835 (2020)
42. Werling, D.M., Geschwind, D.H.: Sex differences in autism spectrum disorders. *Current Opinion Neurol.* **26**(2), 146 (2013)
43. Xu, C., Li, C., Wu, H., Wu, Y., Hu, S., Zhu, Y., et al.: Gender differences in cerebral regional homogeneity of adult healthy volunteers: a resting-state fMRI study. *BioMed research international* **2015** (2015)
44. Zang, Y., Jiang, T., Lu, Y., He, Y., Tian, L.: Regional homogeneity approach to fMRI data analysis. *Neuroimage* **22**(1), 394–400 (2004)
45. Zhang, C., Dougherty, C.C., Baum, S.A., White, T., Michael, A.M.: Functional connectivity predicts gender: evidence for gender differences in resting brain connectivity. *Human Brain Mapp.* **39**(4), 1765–1776 (2018)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



### **3 Brain-age prediction: a systematic comparison of machine learning workflows.** More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S.B. and Patil, K.R., NeuroImage, 119947 (2023)

#### **Authorship contribution statement**

**Shammi More (Doctoral researcher, first author):** Formal analysis, Software, Validation, Visualization, Writing – original draft. **Georgios Antonopoulos:** Data curation, Writing – review & editing. **Felix Hoffstaedter:** Data curation, Writing – review & editing. **Julian Caspers:** Supervision, Writing – review & editing. **Simon B. Eickhoff:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **Kaustubh R. Patil (Corresponding author):** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.





# Brain-age prediction: A systematic comparison of machine learning workflows<sup>☆</sup>

Shammi More<sup>a,b</sup>, Georgios Antonopoulos<sup>a,b</sup>, Felix Hoffstaedter<sup>a,b</sup>, Julian Caspers<sup>c</sup>,  
Simon B. Eickhoff<sup>a,b</sup>, Kaustubh R. Patil<sup>a,b,\*</sup>, for the Alzheimer's Disease Neuroimaging Initiative

<sup>a</sup> Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany

<sup>b</sup> Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

<sup>c</sup> Department of Diagnostic and Interventional Radiology, University Hospital Düsseldorf, Düsseldorf, Germany

## ARTICLE INFO

### Keywords:

Brain-age estimation  
Healthy aging  
Machine learning  
Regression analysis

## ABSTRACT

The difference between age predicted using anatomical brain scans and chronological age, i.e., the brain-age delta, provides a proxy for atypical aging. Various data representations and machine learning (ML) algorithms have been used for brain-age estimation. However, how these choices compare on performance criteria important for real-world applications, such as; (1) within-dataset accuracy, (2) cross-dataset generalization, (3) test-retest reliability, and (4) longitudinal consistency, remains uncharacterized. We evaluated 128 workflows consisting of 16 feature representations derived from gray matter (GM) images and eight ML algorithms with diverse inductive biases. Using four large neuroimaging databases covering the adult lifespan (total  $N = 2953$ , 18–88 years), we followed a systematic model selection procedure by sequentially applying stringent criteria. The 128 workflows showed a within-dataset mean absolute error (MAE) between 4.73–8.38 years, from which 32 broadly sampled workflows showed a cross-dataset MAE between 5.23–8.98 years. The test-retest reliability and longitudinal consistency of the top 10 workflows were comparable. The choice of feature representation and the ML algorithm both affected the performance. Specifically, voxel-wise feature spaces (smoothed and resampled), with and without principal components analysis, with non-linear and kernel-based ML algorithms performed well. Strikingly, the correlation of brain-age delta with behavioral measures disagreed between within-dataset and cross-dataset predictions. Application of the best-performing workflow on the ADNI sample showed a significantly higher brain-age delta in Alzheimer's and mild cognitive impairment patients compared to healthy controls. However, in the presence of age bias, the delta estimates in the patients varied depending on the sample used for bias correction. Taken together, brain-age shows promise, but further evaluation and improvements are needed for its real-world application.

## 1. Introduction

Precision and preventive medicine, e.g., early detection of Alzheimer's disease (AD), can benefit from individual-level quantification of atypical aging. Machine learning (ML) approaches, together with large neuroimaging datasets can provide such individualized predictions. Indeed, ML algorithms can capture the multivariate pattern of age-related changes in the brain associated with healthy or typical aging (Franke et al., 2010; Varikuti et al., 2018; Cole 2020; Beheshti et al., 2022; Hahn et al., 2022). Such a model can then be used to predict age, i.e., brain-age, from an unseen subject's image. Being a normative model, a large deviation between the chronological and the predicted

age is indicative of atypical aging. A higher positive difference between the brain-age and chronological age, i.e., brain-age delta (which we refer to simply as delta), indicates “older-appearing” brains. As an indicator of future risk of experiencing age-associated health issues, delta quantitatively relates to several age-related risk factors and general physical health, such as weaker grip strength, poorer lung function, history of stroke, greater frequency of alcohol intake, increased mortality risk (Cole et al., 2018; Cole, 2020), and poorer cognitive functions such as fluid intelligence, processing speed, semantic verbal fluency, visual attention, and cognitive flexibility (Cole et al., 2018; Boyle et al., 2021; Richard et al., 2018; Gaser et al., 2013; Cole et al., 2017). Overall, the delta can potentially serve as an omnibus biomarker of brain integrity

<sup>☆</sup> Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: [https://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf).

\* Corresponding author at: Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany.

E-mail address: [k.patil@fz-juelich.de](mailto:k.patil@fz-juelich.de) (K.R. Patil).

<https://doi.org/10.1016/j.neuroimage.2023.119947>.

Received 20 November 2022; Received in revised form 8 February 2023; Accepted 15 February 2023

Available online 16 February 2023.

1053-8119/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

and health if its reliability, given different ML workflow designs and other analyses, can be established.

Studies have shown global and local gray matter (GM) volume (GMV) loss (Good et al., 2001; Galluzzi et al., 2008; Giorgio et al., 2010) with aging and accelerated loss in neurodegenerative disorders (Good et al., 2001; Karas et al., 2004; Fjell et al., 2014). This makes GMV a clinically relevant candidate for the investigation of atypical aging via brain-age estimation (Franke et al., 2010; Cole et al., 2015). Brain-age prediction models tend to perform better using GMV than white matter volume (WMV) (Cole et al., 2017; Monté-Rubio et al., 2018), making GMV a promising candidate for further investigation. Furthermore, by reducing the methodological and data-related variance in a model's prediction error, the delta can better reflect a biological signal related to atypical aging.

A brain-age estimation workflow consists of a feature space and an ML algorithm, and several choices exist for each. For instance, voxel-wise data with additional smoothing and/or resampling or parcel-wise averages within a brain atlas can be used as features (Varikuti et al., 2018; Eickhoff et al., 2021). Further dimensionality reduction methods such as principal components analysis (PCA) can improve the observations-to-features ratio and signal-to-noise ratio (Franke et al., 2010; Franke et al., 2013; Gaser et al., 2013). One also needs to choose from a large pool of ML algorithms, such as relevance vector regression (RVR), and Gaussian process regression (GPR), many of which have shown success in brain-age estimation. These choices are known to affect performance (Gutierrez Becker et al., 2018; Baecker et al., 2021; de Lange et al., 2022).

Studies using voxel-based morphometry (VBM)-derived GMV to predict brain-age have claimed prediction errors of ~5–8 years in healthy individuals (Table S1). However, it is difficult to compare these studies as they differ in experimental setup and methodology, such as feature space used, ML algorithms, age range, and evaluation criteria. For a brain-age estimation model to be used in real-world applications, it must perform well on several evaluation criteria; (1) a model should generalize well on new data from the training site as well as on data from novel sites, (2) estimated age must be reliable on repeated measurements, and (3) it should also exhibit longitudinal consistency, i.e., the predicted age should be proportionally higher for later scans after a longer duration, assuming no significant change in lifestyle or health-related interventions between the measurements.

A critical aspect, especially for clinical application, is the commonly reported negative correlation between delta and true age (Beheshti et al., 2019; Smith et al., 2019; de Lange and Cole, 2020). This may result in spurious correlations between the delta and non-imaging measures when chronological age is not accounted for (Franke et al., 2013; Löwe et al., 2016). This age bias complicates or may even mislead downstream individualized decision-making. It can be mitigated using bias correction models; usually, linear regression predicting brain-age or delta using chronological age (Le et al., 2018; Liang et al., 2019; Smith et al., 2019; de Lange et al., 2022). The data source (within or cross-data) and size used to obtain bias correction models has substantial impact on quality of the model. Taken together, there is a gap in understanding the impact of the choices in designing brain-age workflows, and how they affect estimation and utility of individual-level delta.

To fill this gap, we systematically assessed 128 workflows consisting of 16 feature spaces derived from GM images and eight ML algorithms with diverse inductive biases. Using several large neuroimaging databases with a wide age range, we first evaluated these workflows for their within-dataset and cross-dataset performances. Next, we evaluated the test-retest reliability and longitudinal consistency of some top-performing workflows. Then, we assessed the performance of our best-performing workflow in a clinical sample. We investigated the correlations between delta and behavioral/cognitive measures in healthy and clinical cohorts and various factors affecting these correlations. We also compared our best-performing workflow with a publicly available

model, brainageR. Several follow-up analyses were performed to investigate the effect of preprocessing (CAT vs. SPM) and tissue type (GM vs. GM+WM+CSF) choices on prediction performance. Finally, given recent evidence that lower accuracy models may capture atypical aging better (Bashyam et al., 2020), we investigated relationship of model performance with delta and delta-behavior correlations.

## 2. Material and methods

### 2.1. Datasets

#### 2.1.1. MRI data

We used T1-weighted (T1w) magnetic resonance imaging (MRI) data from healthy subjects covering a wide age range (18–88 years, training data) from several large neuroimaging datasets (Table 1), including the Cambridge center for Ageing and Neuroscience (CamCAN,  $N = 651$ ) (Taylor et al., 2017), Information eXtraction from Images (IXI,  $N = 562$ ) (<https://brain-development.org/ixi-dataset/>), the enhanced Nathan Kline Institute-Rockland Sample (eNKI,  $N = 597$ ) (Nooner et al., 2012), the 1000 brains study (1000BRAINS;  $N = 1143$ ) (Caspers et al., 2014), Consortium for Reliability and Reproducibility (CoRR) (Zuo et al., 2014), the Open Access Series of Imaging Studies (OASIS-3) (LaMontagne et al., 2019), and the MyConnectome dataset (Poldrack et al., 2015). The inclusion criteria were age between 18 and 90 years, gender data available, and no current or past known diagnosis of neurological, psychiatric, or major medical conditions. The IXI dataset was acquired from multiple sites; however, we treat it as a single dataset representing typical data acquired in a noisy clinical setting. From the OASIS-3 dataset, we selected scans from healthy control subjects acquired on 3T scanners. Some other datasets used by brainageR were also used for a fair comparison with our best workflow. The corresponding details are provided in the Supplementary Table S8.

We used the Alzheimer's Disease Neuroimaging Initiative (ADNI; <https://adni.loni.usc.edu/>) database to evaluate the utility of brain-age in neurodegenerative disorders (Jack et al., 2008; Petersen et al., 2010). We included 3T T1w images from healthy control (HC,  $N = 209$ ), early and late mild cognitively impaired (EMCI,  $N = 237$ ; LMCI,  $N = 128$ ), and Alzheimer's disease (AD,  $N = 125$ ) subjects. For some of these subjects, data were available for the second timepoint 1–2 years apart (HC,  $N = 153$ ; EMCI,  $N = 197$ ; LMCI,  $N = 104$ ; AD,  $N = 61$ ) (Table 1d).

#### 2.1.2. Non-imaging data

We used various behavioral/cognitive measures to compute their correlations with delta. Fluid intelligence (FI;  $N = 631$ ) assessed by the Cattell Culture Fair test and reaction time for the motor learning task ( $N = 302$ ) from the CamCAN dataset (Taylor et al., 2017). From the eNKI dataset, we used the Color-Word Interference Test (CWIT) inhibition trial completion time ( $N = 340$ ), the Trail Making Test (TMT) number-letter switching condition completion time ( $N = 344$ ), Wechsler Abbreviated Scale of Intelligence (WASI-II) matrix reasoning scores ( $N = 347$ ), and WASI-II similarities scores ( $N = 347$ ) (Nooner et al., 2012).

Three cognitive tests from ADNI measuring disease severity were used; Mini-Mental State Examination (MMSE), Global Clinical Dementia Rating Scale (CDR), and Functional Assessment Questionnaire (FAQ).

All the datasets except the 1000BRAINS data are available publicly. 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.

### 2.2. Data preparation

For the main analysis all T1w images were preprocessed using the Computational Anatomy Toolbox (CAT) version 12.8

**Table 1**

Sample characteristics of the datasets used in the current study. Datasets used a. for training within-dataset models. b. for training cross-dataset models. c. to evaluate test-retest reliability and longitudinal consistency of brain-age delta and comparison with brainageR (note: for CoRR full sample, the demographics are reported for the last iteration). d. to evaluate performance in clinical samples. Abbreviations: CamCAN: the Cambridge center for ageing and Neuroscience, IXI: Information eXtraction from Images (includes 1.5 and 3T scans), eNKI: the enhanced Nathan Kline Institute-Rockland Sample, CoRR: Consortium for Reliability and Reproducibility, OASIS-3: the Open Access Series of Imaging Studies, ADNI: the Alzheimer's Disease Neuroimaging Initiative, HC: healthy control, EMCI and LMCI: early and late mild cognitively impaired, AD: Alzheimer's disease.

a.						
Train dataset	No. of subjects (N)	Males/Females	Age range	Mean $\pm$ S.D.	Median	
CamCAN	651	321/330	18 - 88	54.27 $\pm$ 18.58	54.50	
IXI	562	249/313	20 - 86	48.70 $\pm$ 16.44	48.85	
eNKI	597	188/409	18 - 85	48.25 $\pm$ 18.51	50.00	
1000BRAINS	1143	660/513	22 - 85	61.85 $\pm$ 12.39	63.60	
b.						
Train dataset	Train N	Test dataset	Test N			
IXI + eNKI + 1000BRAINS	2302	CamCAN	651			
CamCAN + eNKI + 1000BRAINS	2391	IXI	562			
IXI + CamCAN + 1000BRAINS	2356	eNKI	597			
IXI + CamCAN + eNKI	1810	1000BRAINS	1143			
IXI + CamCAN + eNKI + 1000BRAINS	2953	CoRR, OASIS-3, MyConnectome, ADNI	See below (c & d)			
c.						
Dataset	Data Filtering	N (sessions)	Males/Females	Age Range	Mean $\pm$ S.D.	Median
CoRR	Retest < 3 months	86 (2)	39/47	20.0 - 84.0	48.82 $\pm$ 18.28	49.00
	Retest 1 – 2 years	95 (2)	52/43	18.0 - 88.0	34.43 $\pm$ 22.51	20.00
	Retest 2 – 3.25 years	26 (2)	18/8	18.0 - 57.0	28.09 $\pm$ 11.89	24.50
	Full sample	107	51/56	18.0 – 88.0	49.99 $\pm$ 18.87	50.00
OASIS-3	Retest < 3 months	36 (2)	21/15	42.66 - 80.90	63.46 $\pm$ 8.80	62.93
	Retest 3- 4 years	127 (2)	52/75	46.04 - 86.21	65.59 $\pm$ 8.39	65.90
	Full sample	806	338/468	43.00 - 89.00	69.07 $\pm$ 9.06	69.00
MyConnectome	Retest < 3 years	1 (20)	1/0	45.39 - 48.02	45.73 $\pm$ 0.58	45.56
d.						
Dataset	Disease	N	Males/Females	Age Range	Mean $\pm$ S.D.	Median
ADNI (Timepoint-1)	HC	209	99/110	56.3 - 94.7	75.67 $\pm$ 6.94	75.50
	EMCI	237	128/109	55.7 - 88.7	70.88 $\pm$ 7.12	70.40
	LMCI	128	62/65	55.1 - 91.5	72.02 $\pm$ 7.89	72.55
	AD	125	65/60	56.0 - 91.0	74.68 $\pm$ 7.99	75.40
ADNI (Timepoint-2)	HC	153	70/83	57.3 - 95.8	75.89 $\pm$ 6.63	75.50
	EMCI	197	108/89	56.7 - 90.4	71.81 $\pm$ 7.04	71.10
	LMCI	104	51/53	56.1 - 92.5	73.36 $\pm$ 7.92	73.95
	AD	61	32/29	57.0 - 93.0	75.79 $\pm$ 7.83	76.80

(Gaser et al., 2022). To ensure accurate normalization and segmentation, initial affine registration of T1w images was done with higher than default accuracy (accstr = 0.8). After bias field correction and tissue class segmentation, accurate optimized Geodesic shooting (Ashburner and Friston, 2011) was used for normalization (regstr = 1). We used 1 mm Geodesic Shooting templates and outputted 1 mm isotropic images. The normalized GM segments were then modulated for linear and non-linear transformations.

For comparison with the brainageR model, we used the seven datasets used by brainageR (Table S8) and preprocessed them using CAT 12.8 (Section 2.9). To evaluate the effect of preprocessing and tissue types, we used the SPM12 based preprocessing as implemented by brainageR, which outputs three tissue segmentations (GM, WM, and CSF; see <https://github.com/james-cole/brainageR/>).

## 2.3. Workflows

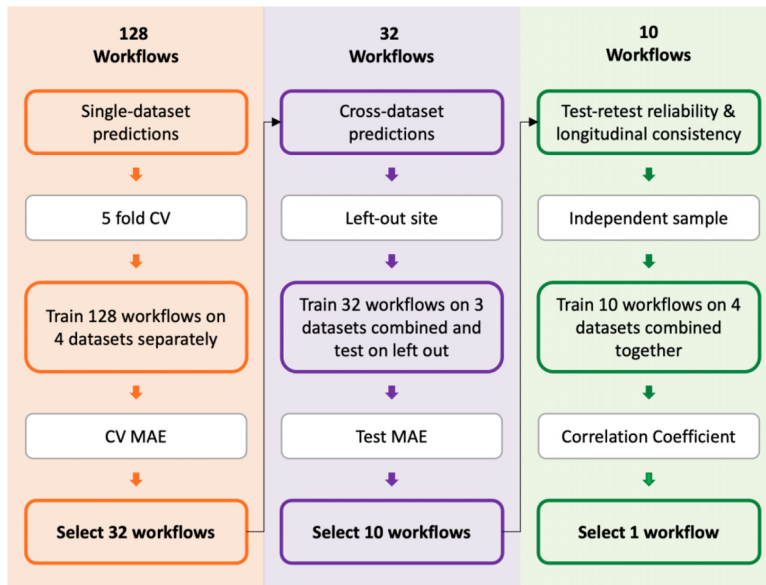
Each workflow consists of a feature representation and an ML algorithm. We evaluated 128 workflows constituting 16 feature representations and eight ML algorithms.

### 2.3.1. Feature representations

The 16 feature representations were derived from the CAT-preprocessed voxel-wise GM images. Using voxel-wise data can lead to overfitting due to the curse of dimensionality owing to a large number of features compared to the number of samples. Hence, we implemented two dimensionality reduction approaches previously used for brain-age prediction.

In the first strategy, we used voxel-wise GMV after smoothing and resampling (Franke et al., 2010), which may also improve the signal-to-noise ratio. In the second strategy, we used an atlas to summarize data from distinct brain regions (called parcels). This resulted in 16 feature representations.

1. SX\_RY: A whole-brain mask was used to select 238,955 voxels. Then, smoothing (S) with an X mm FWHM Gaussian kernel and resampling (R) using linear interpolation to Y mm spatial resolution were applied with  $X = \{0, 4, 8\}$  and  $Y = \{4, 8\}$ , resulting in six feature spaces (S0\_R4, S0\_R8, S4\_R4, S4\_R8, S8\_R4, S8\_R8; SX\_R4: 29,852 voxels and SX\_R8: 3747 voxels).
2. SX\_RY + PCA: Additionally, PCA (Jolliffe, 2002) was applied to each SX\_RY feature space while retaining 100% variance, creating an-



**Fig. 1.** The framework to select the best-performing workflow for brain-age prediction. A total of 128 workflows were first evaluated for their within-dataset prediction performance using five-fold cross-validation (CV). Next, 32 workflows were selected based on the CV mean absolute error (MAE) and assessed for cross-dataset prediction performance. Within-dataset and cross-dataset evaluations were performed using four datasets (CamCAN, IXI, eNKI and 1000BRAINS). Then, 10 workflows out of 32 were selected based on their test MAE and assessed for test-retest reliability and longitudinal consistency using OASIS-3 and CoRR datasets. The best-performing workflow was selected after considering all the evaluation criteria.

other six representations (S0\_R4 + PCA, S0\_R8 + PCA, S4\_R4 + PCA, S4\_R8 + PCA, S8\_R4 + PCA, S8\_R8 + PCA).

3. Parcel-wise: Four parcel-wise feature spaces were created by combining cortical {100, 400, 800, 1200} parcels (Schaefer et al., 2018) with 36 subcortical (Fan et al., 2016) and 37 cerebellum (Buckner et al., 2011) parcels. We calculated the mean GMV of all the voxels within each parcel (173, 473, 873, and 1273 features).

### 2.3.2. Machine learning algorithms

We included eight ML algorithms covering diverse inductive biases: ridge regression (RR), least absolute shrinkage and selection operator (LASSO) regression (LR), elastic net regression (ENR), kernel ridge regression (KRR), random forest regression (RFR), GPR, RVR with the linear kernel (RVRLin), and polynomial kernel of degree 1 (RVRpoly). These algorithms have been previously used in the prediction of age and other behavior variables from neuroimaging data (Franke et al., 2010; Gaser et al., 2013; Su et al., 2013; Cole et al., 2015; Varikuti et al., 2018; Jonsson et al., 2019; Liang et al., 2019; Zhao et al., 2019; He et al., 2020; Baecker et al., 2021; Boyle et al., 2021; Lee et al., 2021; Peng et al., 2021; Treder et al., 2021; Vidal-Pineiro et al., 2021; Beheshti et al., 2022; Cole, 2020) (Table S1). Details of these algorithms are provided in the Supplementary Methods.

Recently, deep-learning (DL) models have been applied for brain-age estimation with success (Jiang et al., 2019; Jonsson et al., 2019; Peng et al., 2021). However, in this work, we focus on conventional ML models for the following reasons: (1) ML models have shown competitive performance to DL models (Cole et al., 2017; He et al., 2020; Schulz et al., 2020; Grinsztajn et al., 2022), and (2) the resources required for ML are more readily available and thus still enjoy wider applicability with a lower computational footprint (Thompson et al., 2020; van Wylsberghe, 2021).

### 2.3.3. Learning setup and software

The ML algorithm's hyperparameters were estimated in a nested fashion using an inner cross-validation (CV) (Varoquaux et al., 2017). Before training, features with low variance were removed (threshold < 1e-5), and the remaining features were Z-scored to have zero mean and unit variance. Any preprocessing steps, including PCA, were applied in a CV-consistent fashion to avoid data leakage, i.e., the parameters were estimated on the training set and applied to both the training and the test set (More et al., 2021).

All the workflows were implemented in Python version 3.9.1 using the Julearn machine-learning library (<https://juaml.github.io/julearn/>), which in turn uses the scikit-learn library for the learning algorithms KRR, GPR, and RFR (<http://scikit-learn.org/>) (Pedregosa et al., 2011). LR, RR, and ENR were implemented using the Python wrapper for glmnet (<https://pypi.org/project/glmnet/>) (Friedman et al., 2010). RVRLin and RVRpoly were implemented using the scikit-rvm package (<https://github.com/JamesRitchie/scikit-rvm/>). The codes used for pre-processing, feature extraction, model training and prediction are available at [https://github.com/juaml/brainage\\_estimation](https://github.com/juaml/brainage_estimation).

### 2.4. Analysis setup

Given data acquisition and site-related biases, it is important to identify a workflow that shows high accuracy in different evaluation scenarios. For instance, a workflow that works well on a dataset might not work well on another dataset. To accommodate such real-world scenarios, we followed a systematic procedure where the workflows were subjected to increasingly stringent evaluations (Fig. 1). In brief, we first evaluated the within-dataset CV performance of the 128 workflows. Next, 32 workflows characterizing the overall pattern of performance were selected for cross-dataset evaluation. This selection was performed by uniformly sampling over the within-dataset CV performance. This allows for the possibility that workflows with low within-dataset performance might perform well in cross-dataset evaluation. Finally, the top 10 workflows out of the 32 were evaluated for their test-retest reliability and longitudinal consistency. After considering all the evaluation criteria, the best-performing workflow was chosen and used for application on ADNI data and comparison with brainageR. Specific analysis steps are described below.

#### 2.4.1. Within-dataset and cross-dataset evaluations

We evaluated the 128 workflows (see Section 2.3) separately on four datasets, CamCAN, IXI, eNKI, and 1000BRAINS. This scenario assumes that enough within-dataset training data are available and is widely used in brain-age estimation work (Ashburner, 2007; Su et al., 2013; Gutierrez Becker et al., 2018). To estimate a single out-of-sample brain-age for each subject, we used a 5-fold CV. For each hold-out (test) fold, the remaining 80% of the data were used for training and to obtain a generalization estimate using 5 times repeated 5-fold (5 × 5-fold) nested CV. All CV analysis was stratified by age to preserve the age distribution. It is important to obtain a single prediction per subject (as opposed to



multiple predictions per subject if the outer CV were repeated) for further meaningful analyses, such as correlation with non-imaging measures. Consequently, we computed two measures, test performance, and CV performance. The test performance was obtained by averaging over the outer 5 folds. The CV performance was obtained by first averaging over the inner 5  $\times$  5-fold CV and then over the outer 5-fold CV. Finally, the CV and test performance were averaged over the four datasets. The performance was evaluated using mean absolute error (MAE), Pearson's correlation between predicted and true (chronological) age, and the coefficient of determination  $R^2$ .

We followed a systematic procedure to select a subset of workflows while maintaining diversity in terms of CV performance. Specifically, the workflows were arranged in the increasing order of their average CV MAE and divided into 16 groups. Next, the top two workflows (with the lowest CV MAE) from each group were selected.

We tested these 32 selected workflows on cross-dataset to obtain sample-unbiased performance. This emulates the real-world scenario where data from the application site are not available, and the training and test data come from different sources with confounding effects, such as scanner hardware or operator inconsistencies (Jovicich et al., 2006; Chen et al., 2014). Three out of four datasets (CamCAN, IXI, eNKI and 1000BRAINS) were pooled to form the training data, and the hold-out dataset was used as the test data. A 5  $\times$  5-fold CV was performed on the training data to estimate the generalization performance with an internal CV for hyperparameter tuning. The CV performance was averaged over 5  $\times$  5-fold CV and then over the four hold-out datasets. The test performance was averaged over the four datasets. The performance was again evaluated using MAE, Pearson's correlation between predicted and true age, and the coefficient of determination  $R^2$ .

The 32 workflows were arranged in increasing order of their average test MAE, i.e., their average performance on the hold-out datasets, from which the top 10 workflows were selected.

#### 2.4.2. Test-retest reliability and longitudinal consistency

We then trained models using the 10 selected workflows with the four datasets combined as training data (IXI + eNKI + CamCAN + 1000BRAINS,  $N = 2953$ ; Supplementary Fig. S1). The test-retest reliability and longitudinal consistency of the delta were evaluated for the 10 models using the OASIS-3 and CoRR datasets.

To evaluate test-retest reliability, we used: two scans from the same subjects acquired within a delay of (1) less than three months (CoRR:  $N = 86$ , age range = 20–84 years, OASIS-3:  $N = 36$ , age range = 43–81), and (2) between 1 and 2 years (CoRR:  $N = 95$ , age range = 18–88). The concordance correlation coefficient (CCC) (Lin, 1989) between the delta (predicted age minus age at the scan time) from the two scans was calculated.

To evaluate longitudinal consistency, two scans from the same subjects acquired with a retest duration (1) between 2 and 3.25 years (CoRR:  $N = 26$ , age range = 18–57), and (2) between 3 and 4 years (OASIS-3:  $N = 127$ , age range = 46–86) were used. We computed Pearson's correlation between the difference in the predicted age and the difference in chronological age from the two scans. A higher positive correlation here would indicate higher longitudinal consistency.

By considering the results from the within- and cross-dataset analysis, test-retest reliability, and longitudinal consistency, we chose one best-performing workflow for further analysis.

#### 2.5. Bias correction

Many studies have reported age-dependency of the delta with over-prediction in young subjects and under-prediction in older subjects (Le et al., 2018; Liang et al., 2019), which renders the usage of delta as an individualized biomarker problematic. A common practice is to apply a statistical bias correction to remove the effect of age from either the predicted age or the delta (Le et al., 2018; Liang et al., 2019; Smith et al., 2019; Cole, 2020; de Lange and Cole, 2020). Note that

when calculating correlations of delta with non-imaging measures, bias correction is expected to be similar to partial correlation analysis when age is used as a covariate.

Several alternatives are available for bias correction (de Lange et al., 2019; Cole, 2020; de Lange and Cole, 2020; Smith et al., 2019; Beheshti et al., 2019)). We chose the method used by Cole and colleagues (Cole, 2020) as it does not use the chronological age of the test data, and thus avoids information leakage which can bias comparison between workflows by making low-performing workflows appear good (de Lange et al., 2022). Furthermore, this method is relevant for possible future applications like forensic investigations where test age is not available. A linear regression model was fitted with the out-of-sample (from the CV) predicted age as the dependent variable and chronological age as the independent variable using the training data. The predicted age in the test set was corrected by subtracting the resulting intercept and dividing by the slope.

#### 2.6. Correlation with cognitive measures

To understand the effect of bias correction and the impact of covariates on delta-behavior correlations, we performed correlations of behavior/cognitive measures from CamCAN and eNKI datasets (see Section 2.1.2) with (1) uncorrected delta, (2) uncorrected delta with age as a covariate, (3) corrected delta, and (4) corrected delta with age as a covariate. If the bias correction eliminates the antagonistic relation between delta and age, we expect (2), (3), and (4) to give similar correlations. Furthermore, to assess the impact of data used for learning bias correction models, we performed these analyses using delta obtained from within-dataset and cross-dataset predictions.

#### 2.7. Brain-age in clinical samples

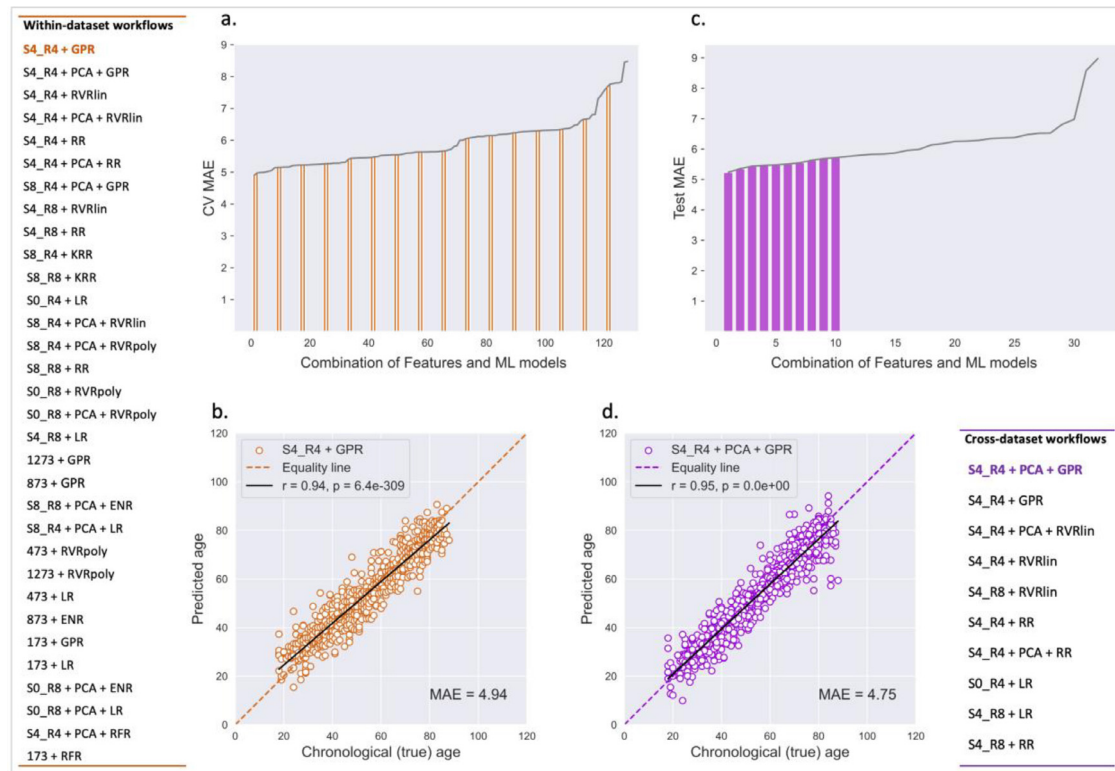
Next, we used the ADNI dataset (Jack et al., 2008; Petersen et al., 2010) to validate our best-performing workflow on clinical samples. We estimated and compared the delta between HC, EMCI, LMCI, and AD subjects (Table 1d).

Our best-performing workflow trained on the four datasets was used to obtain the predictions, followed by application of bias correction model (see Section 2.5). We compared two bias correction models, one derived using the CV predictions from the four training datasets and another using HC samples in ADNI data (Franke and Gaser, 2012). The group-wise corrected delta was compared using analysis of variance (ANOVA) followed by Bonferroni correction to counteract multiple comparisons. Emulating the scenario that application sites might have different numbers of HC samples, we learned bias correction models using HC sub-samples (0.1 to 0.9 fraction in steps of 0.1) drawn without replacement and applied them on the full HC and AD samples. This process was repeated 100 times to estimate the variance of mean corrected delta in HC and AD subjects.

Finally, we investigated associations between the corrected delta and three clinical test scores, MMSE, CDR, and FAQ. The correlations were computed using the whole sample and different diagnostic groups separately using Pearson's correlation with age as a covariate for both sessions separately.

#### 2.8. Relationship of MAE with delta and delta-behavior correlations

Here, we sought to select a workflow that provides accurate and reliable predictions. We reason that a workflow that accurately predicts the age of healthy individuals captures the typical brain aging process, and thus, a large delta in new data can be considered indicative of atypical aging. However, recent evidence shows that an overfitted brain-age model (high training accuracy) is not the most sensitive in identifying pathologies (Bashyam et al., 2020). This study showed that a relatively moderately fit model yielded brain-age deltas with more sig-



**Fig. 2.** Within-dataset and cross-dataset results. a. The line plot showing CV MAE (averaged across four datasets) for 128 workflows arranged in increasing order (names of all workflows are given in Table S2). The orange bars represent the MAEs of 32 selected workflows with their names in the table on left. b. The scatter plot between the chronological age and within-dataset predicted age for the CamCAN data using S4\_R4 + GPR workflow (MAE = 4.94 years and  $r = 0.94$ ,  $p = 6.4e-309$ ). c. The line plot showing test MAE (averaged across four runs) for the 32 workflows arranged in increasing order (names of all workflows are given in Table S3). The purple bars represent the MAEs of 10 selected workflows with their names in the table on the bottom right. d. The scatter plot between the chronological age and cross-dataset predicted age for the CamCAN data using S4\_R4 + PCA + GPR workflow (MAE = 4.75 years and  $r = 0.95$ ,  $p = 0.0e+00$ ).

nificant group differences and the larger effect sizes between control and disease groups in various brain pathologies.

To investigate this possibility, we trained the 32 workflows selected from the cross-dataset analysis with four datasets pooled together for training and applied to timepoint 2 ADNI data. To understand how the model performance varies with its utility, we compared the models' MAEs with the corrected mean delta in AD sample and examined whether it was related to the delta-behavior correlations. We then performed a similar analysis in two HC samples (CamCAN and eNKI) using corresponding within-dataset hold-out predictions.

## 2.9. Comparison with brainageR and effect of preprocessing and tissue types

We compared the performance of our best-performing workflow with an already available brain-age estimation model, brainageR. The brainageR model was trained on 3377 healthy individuals (age range = 18–92 years, mean  $\pm$  SD age =  $40.6 \pm 21.4$  years) from seven publicly available datasets using the GPR algorithm. It uses SPM12 to segment and normalize T1w images, from which GM, WM, and CSF vectors were extracted (using 0.3 probability masked brainageR-specific templates). PCA was used to reduce data dimensionality, and 435 components explaining 80% of the variance were retained. Note that brainageR uses three tissue types, while our focus is on GM.

To avoid bias due to different training data, for this comparison we used data from the same subjects used by brainageR (2 subjects could not be processed; Table S8). Next, using this training data, we trained our best-performing workflow using GMV extracted from CAT 12.8 and compared the performance with already trained brainageR model on three datasets, (1) CoRR ( $N = 107$ , sub-sampled to keep uniform dis-

tribution in age-range = 18–88 years, repeated 100 times; see Supplementary Methods for more details), (2) the OASIS-3 ( $N = 806$ ; first scan per subject, age-range = 43–89 years), and (3) the MyConnectome study (one subject scanned 20 times in a period of 3 years; age range = 45–48 years). Additionally, we used sub-samples from OASIS-3 with test-retest durations of (1) less than 3 months ( $N = 36$ , 43–81 years) and (2) between 3 and 4 years ( $N = 127$ , 46–86 years) to evaluate test-retest reliability and longitudinal consistency, respectively (see Section 2.4.2).

Next, we compared how the preprocessing and tissue types affect model performance. Following our focus on GMV, we compared; (1) CAT-preprocessed GMV, (2) SPM-preprocessed GMV, and (3) SPM-preprocessed GM, WM, and CSF images following brainageR. The latter investigates whether WM and CSF features provide complementary information leading to better predictions. For this, we performed within-dataset evaluation on IXI and CamCAN datasets (see Section 2.4.1).

## 3. Results

### 3.1. Within-dataset and cross-dataset predictions

For within-dataset analysis, the CV performance (average over 125 estimates—inner  $5 \times 5$ -fold CV, repeated 5 times, see Section 2.4.1) and test performance based on single prediction per subject from the outer CV, were calculated. These were then averaged separately over four datasets.

The average CV MAE (4.90–8.48 years) and the average test MAE (4.73–8.38 years) (Fig. 2a, Table S2) were similar, indicating that the nested CV generalization estimates are indeed indicative of their test performance. The correlation between the true and predicted age on the test data ranged from 0.81 to 0.93, while the age bias (correlation

**Table 2**

The performance metric for the best workflow on different datasets. A. Within-dataset prediction (using S4\_R4 + GPR) b. Cross-dataset prediction (using S4\_R4 + PCA + GPR). Abbreviations: MAE: mean absolute error between true and predicted age, MSE: mean squared error between true and predicted age,  $R^2$ : the proportion of variance of predicted age explained by the independent variables in the model, Corr (true, pred): Pearson's correlation between true and predicted age, Age bias: Pearson's correlation between true age and brain-age delta.

Datasets	N	a. Within-dataset results					b. Cross-dataset results				
		MAE	MSE	$R^2$	Corr (true, pred)	Age bias	MAE	MSE	$R^2$	Corr (true, pred)	Age bias
CamCAN	651	4.94	39.54	0.89	$r = 0.94, p = 6.4e-309$	$r = -0.42, p = 6.8e-29$	4.75	38.35	0.89	$r = 0.95, p = 0.0e+00$	$r = -0.23, p = 3.1e-09$
IXI	562	4.76	35.20	0.87	$r = 0.93, p = 2.9e-252$	$r = -0.48, p = 3.5e-33$	6.08	57.35	0.79	$r = 0.94, p = 1.2e-267$	$r = -0.18, p = 2.2e-05$
eNKI	597	5.20	44.85	0.87	$r = 0.93, p = 8.1e-267$	$r = -0.47, p = 1.4e-33$	4.97	39.65	0.88	$r = 0.94, p = 9.7e-288$	$r = -0.49, p = 3.6e-38$
1000- BRAINS	1143	4.04	26.65	0.83	$r = 0.91, p = 0.0e+00$	$r = -0.50, p = 2.0e-73$	5.13	41.03	0.73	$r = 0.90, p = 0.0e+00$	$r = -0.15, p = 2.0e-07$

**Table 3**

Concordance correlation coefficient (CCC) between brain-age delta from two sessions at different test-retest durations and their respective mean absolute error (MAE) between true and predicted age for CoRR and OASIS-3 datasets for the top 10 workflows.

Retest duration Age range (years)	CoRR dataset						OASIS-3 dataset		
	< 3 months ( $N = 86$ ; 20.0 - 84.0)			1 - 2 years ( $N = 95$ ; 18.0 - 88.0)			< 3 months ( $N = 36$ ; 42.66 - 80.90)		
	MAE (ses-1)	MAE (ses-2)	CCC	MAE (ses-1)	MAE (ses-2)	CCC	MAE (ses-1)	MAE (ses-2)	CCC
Workflows									
S4_R4 + PCA + GPR	4.808	5.008	0.97	4.374	4.204	0.95	4.2	3.801	0.80
S4_R4 + GPR	4.928	5.112	0.97	4.738	4.49	0.96	4.24	3.935	0.82
S4_R4 + PCA + RVRlin	5.811	5.757	0.97	5.156	5.072	0.96	5.288	5.223	0.83
S4_R4 + RVRlin	5.815	5.76	0.97	5.141	5.065	0.96	5.234	5.177	0.83
S4_R8 + RVRlin	6.375	6.265	0.95	5.444	5.33	0.96	5.109	5.2	0.77
S4_R4 + RR	5.64	5.653	0.98	5.174	5.277	0.97	4.918	4.71	0.85
S4_R4 + PCA + RR	5.742	5.732	0.98	5.288	5.404	0.97	4.988	4.744	0.85
S0_R4 + LR	6.281	6.359	0.96	6.251	6.293	0.94	4.949	5.161	0.86
S4_R8 + LR	6.763	6.676	0.97	6.497	6.434	0.97	5.811	5.896	0.79
S4_R8 + RR	6.232	6.185	0.97	5.975	6.016	0.97	5.332	5.328	0.81

between true age and delta) ranged from  $-0.22$  to  $-0.83$  (Table S2). Overall, all workflows showed a high similarity in their predictions (correlations  $0.83$ – $0.99$  averaged across the four datasets; Fig. S2). The top 20 workflows showed comparable CV and test MAE with a difference of less than 0.4 years.

Well-performing workflows primarily consisted of voxel-wise smoothed and resampled feature spaces with and without PCA, with S4\_R4 (smoothed with a 4 mm FWHM kernel and resampled to 4 mm spatial resolution) generally performing better. Some workflows with PCA performed similarly to their respective non-PCA version but not all (see Supplementary Table S2). GPR, KRR, RR, and both RVR algorithms generally ranked high. Most algorithms performed worse with parcel-wise features, while RFR generally exhibited the worst performance.

The workflow S4\_R4 + GPR performed the best (see Table 2a for its performance on each of the four datasets). This workflow showed the lowest average CV MAE with a high  $R^2$  and a high correlation between true and predicted age (Fig. 2b) but a relatively high age bias (Fig. S3). The second-best workflow, S4\_R4 + PCA + GPR, performed similarly to the best workflow. Other workflows with the S4\_R4 feature space, with or without PCA, together with the KRR, RVRpoly, and RVRlin algorithms, performed comparably. From the 128 workflows, we selected 32 workflows while preserving diversity in terms of CV MAE.

The 32 workflows selected for cross-dataset analysis showed the average CV ( $5 \times 5$ -fold on training data) MAE (4.28–7.39 years) lower than the test (hold-out dataset) MAE (5.23–8.98 years) (Fig. 2c). The test-set correlation between true and predicted age ranged from  $0.82$  to  $0.93$ , while the age bias ranged from  $-0.27$  to  $-0.75$  (Table S3). All workflows showed a high similarity in their predictions (correlations  $0.83$ – $0.99$  averaged across the four runs). Due to this high similarity, the averaged predictions, i.e., ensemble, from 32 workflows were not better than the top-performing workflow (Fig. S2). The workflows that performed well within-dataset also performed well in cross-dataset predictions (Fig. S6). These results indicate that the corresponding models could generalize well to data from a new unseen site.

We selected 10 workflows with the lowest test MAE for further analysis. These workflows consisted of only voxel-wise feature spaces

(S4\_R4, S4\_R8, and S0\_R4) with and without PCA. The ML algorithms included GPR, RVRlin, RR, and LR. The best-performing workflow was the S4\_R4 + PCA + GPR with the lowest average test MAE, a high  $R^2$ , a high correlation between true and predicted age (Fig. 2d), and moderate age bias (Fig. S3), see Table 2b for its performance on all four datasets), followed by the S4\_R4 + GPR workflow.

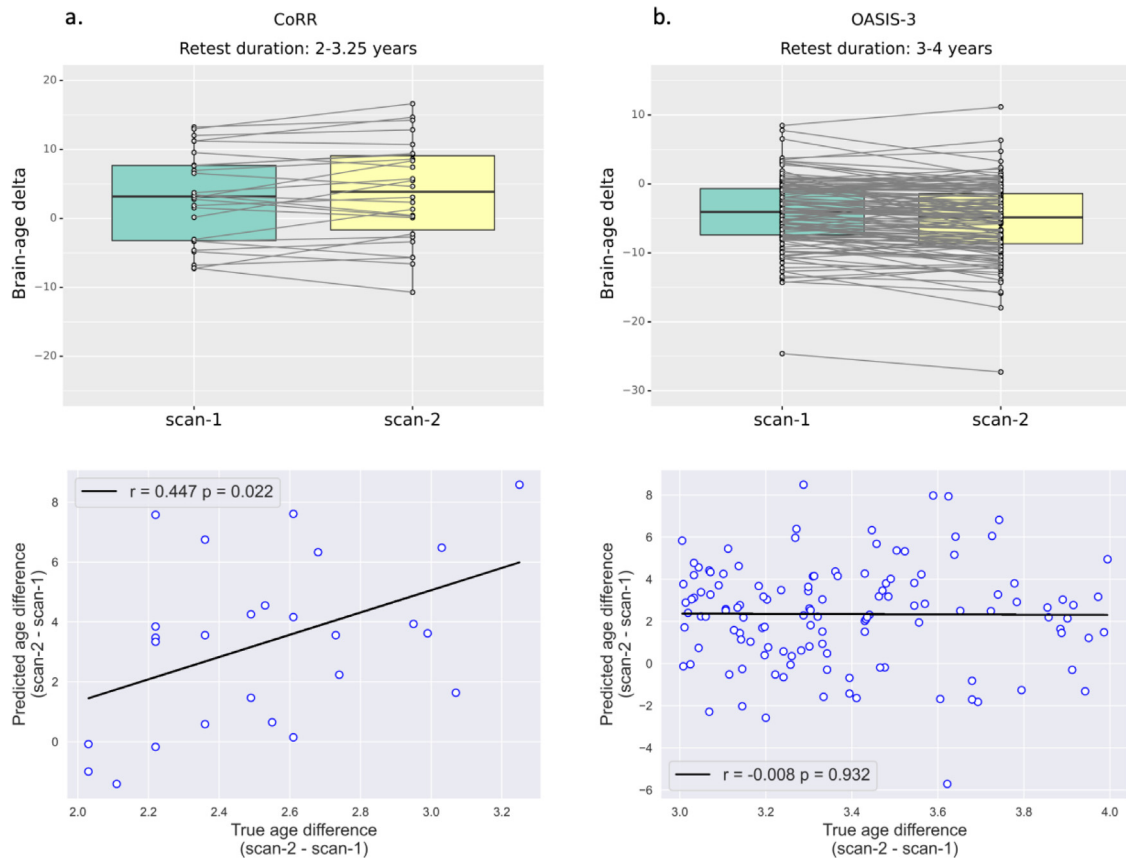
### 3.3. Test-retest reliability and longitudinal consistency

The test-retest reliability and longitudinal consistency of the top 10 workflows selected from the cross-dataset evaluation were evaluated using the CoRR and OASIS-3 datasets.

For the short retest duration of less than three months, all 10 workflows showed high test-retest reliability (CoRR: CCC =  $0.95$ – $0.98$ , age range =  $20$ – $84$  years; OASIS-3: CCC =  $0.77$ – $0.86$ , age range =  $43$ – $81$  years). For the longer retest duration of 1–2 years in the CoRR dataset, CCC ranged between  $0.94$ – $0.97$  (age range =  $18$ – $88$  years) (Table 3). These results show that the age was reliably estimated by the selected workflows.

Next, we evaluated the longitudinal consistency as the correlation between the difference in the predicted age and the difference in the chronological age (Fig. 3, Table S4). Six workflows out of 10 showed a significant positive linear relationship at the retest duration of 2–3.25 years ( $r$  between  $0.451$ – $0.437$ ,  $p < 0.05$ ) in the CoRR dataset. These workflows included the S4\_R4 feature space with and without PCA with the GPR, RVRlin, and RR algorithms. In contrast, none of the workflows showed a linear relationship in the OASIS-3 dataset (retest duration 3–4 years).

Although the workflows showed similar test-retest reliability and longitudinal consistency, the workflow S4\_R4 + PCA + GPR showed the lowest MAE on these sub-samples (Tables 3, S4). Therefore, considering all the analysis scenarios, within-dataset, cross-dataset, test-retest reliability, and longitudinal consistency, although other workflows were also competitive, we deemed the S4\_R4 + PCA + GPR workflow as well-performing and chose it for further analysis.



**Fig. 3.** Longitudinal consistency. (top) The brain-age delta from two scans of the same subjects and (bottom) the scatter plot between the difference in chronological age and the difference in predicted age between two scans acquired within a retest duration of a. 2–3.25 years (CoRR dataset) b. 3–4 years (OASIS-3 dataset).

### 3.4. Bias correction and correlation with behavioral/cognitive measures

In the CamCAN data, FI was negatively correlated with age ( $r = -0.661$ ,  $p = 1.92e-80$ ), while motor learning reaction time was positively correlated with age ( $r = 0.544$ ,  $p = 1.11e-24$ ). In the eNKI data, CWIT inhibition trial completion time ( $r = 0.361$ ,  $p = 6.50e-12$ ) and TMT number-letter switching trial completion time ( $r = 0.279$ ,  $p = 1.45e-07$ ) were positively correlated with age. On the other hand, WASI matrix reasoning scores were negatively correlated ( $r = -0.240$ ,  $p = 6.03e-06$ ), and WASI similarities scores were not correlated ( $r = 0.052$ ,  $p = 0.332$ ) with age (Table 4).

As several ways have been proposed to obtain the correlation between delta and behavior, e.g., using bias-corrected delta or using age as a covariate, we evaluated several alternatives (see Section 2.6).

#### 3.4.1. Within-dataset predictions

Within-dataset hold-out predictions, i.e., single prediction per subject, were derived using the chosen workflow (S4\_R4 + PCA + GPR). The bias correction model was estimated using the CV predictions on the same dataset. In both datasets, there was no residual age bias after bias correction: CamCAN,  $r = -0.17$ ,  $p = 1.13e-05$  and  $r = 0.00$ ,  $p = 0.999$ ; and eNKI,  $r = -0.20$ ,  $p = 4.53e-07$  and  $r = 0.001$ ,  $p = 0.986$ , before and after correction, respectively (Fig. S3).

We first calculated the correlation between the uncorrected delta and behavioral measures using age as a covariate (Table 4a). In the CamCAN data, a higher delta was associated with lower FI ( $r = -0.154$ ,  $p = 0.0001$ ) and higher motor learning reaction time ( $r = 0.181$ ,  $p = 0.002$ ). In the eNKI data, a higher delta was associated with lower response inhibition and selective attention, as indicated by a higher CWIT inhibition trial completion time ( $r = 0.109$ ,  $p = 0.045$ ). There were no correlations between delta and intelligence scores (WASI matrix reason-

ing and similarities). The results with age, age<sup>2</sup>, and gender as covariates showed a similar trend (Table S5a).

Next, we repeated this analysis with the corrected delta (Table 4a) and expected results similar to using uncorrected delta with age as a covariate. We indeed found similar correlations with FI ( $r = -0.157$ ,  $p = 7.24e-05$ ) and motor learning reaction time ( $r = 0.186$ ,  $p = 0.001$ ) in the CamCAN data, but no significant correlation with CWIT inhibition trial completion time ( $r = 0.094$ ,  $p = 0.084$ ) in the eNKI data. The correlations using corrected delta with covariate were highly similar to uncorrected delta with covariate (Table 4a).

#### 3.4.2. Cross-dataset predictions

Cross-dataset predictions were derived for the CamCAN and eNKI datasets using the S4\_R4 + PCA + GPR workflow trained on the IXI + eNKI + 1000BRAINS ( $N = 2302$ ) and IXI + CamCAN + 1000BRAINS ( $N = 2356$ ) datasets, respectively.

In the CamCAN data, the bias correction model was successful with age bias before and after correction  $r = -0.23$ ,  $p = 3.06e-09$  and  $r = -0.04$ ,  $p = 0.263$ , respectively. However, the correction was not successful in the eNKI data; the age bias was  $r = -0.49$ ,  $p = 3.62e-38$  and  $r = -0.35$ ,  $p = 8.39e-19$  before and after correction, respectively (Fig. S3). This result indicates that the bias correction might not always work well when applied to cross-dataset.

Using age as a covariate on the uncorrected delta, we did not find a significant delta-behavior correlation in the CamCAN data. In the eNKI data, a higher delta was associated with lower response inhibition and selective attention, as indicated by a higher CWIT inhibition trial completion time ( $r = 0.208$ ,  $p = 0.0001$ ) and lower cognitive flexibility indicated by a higher TMT completion time ( $r = 0.147$ ,  $p = 0.006$ ) (Table 4b). There were no correlations between delta and intelligence



**Table 4**  
Correlation of brain-age delta with various behavioral measures with and without bias correction. a. From within-dataset predictions. b. From cross-dataset predictions. Age was used as a covariate. Abbreviations: CWIT: Color-Word Interference Test, TMT: Trail Making Test, WASI-II: Wechsler Abbreviated Scale of Intelligence.

a. From within-dataset predictions				
Dataset	Behavioral measure	N	Correlation with age	
No bias correction				
(a) No covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = -0.661, p = 1.9e-80$	
	Motor Learning (Reaction time)	302	$r = 0.544, p = 1.1e-24$	
With covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = -0.043, p = 0.282$	$r = -0.154, p = 0.0001$
	Motor Learning (Reaction time)	302	$r = 0.089, p = 0.122$	$r = 0.181, p = 0.002$
After bias correction				
(c) No covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = -0.157, p = 7.2e-05$	$r = -0.154, p = 0.0001$
	Motor Learning (Reaction time)	302	$r = 0.186, p = 0.001$	$r = 0.180, p = 0.002$
With covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = 0.094, p = 0.084$	$r = 0.110, p = 0.043$
	Motor Learning (Reaction time)	302	$r = 0.022, p = 0.690$	$r = 0.033, p = 0.545$
(d) With covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = -0.019, p = 0.728$	$r = -0.029, p = 0.590$
	Motor Learning (Reaction time)	302	$r = -0.023, p = 0.667$	$r = -0.021, p = 0.698$
b. From cross-dataset predictions				
Dataset	Behavioral measure	N	Correlation with age	
No bias correction				
(a) No covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = 0.071, p = 0.074$	$r = -0.073, p = 0.066$
	Motor Learning (Reaction time)	302	$r = -0.023, p = 0.689$	$r = 0.083, p = 0.151$
With covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = 0.005, p = 0.931$	$r = 0.065, p = 0.230$
	Motor Learning (Reaction time)	302	$r = -0.007, p = 0.898$	$r = 0.039, p = 0.469$
After bias correction				
(c) No covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = 0.005, p = 0.931$	$r = 0.065, p = 0.230$
	Motor Learning (Reaction time)	302	$r = -0.007, p = 0.898$	$r = 0.039, p = 0.469$
With covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = 0.005, p = 0.931$	$r = 0.065, p = 0.230$
	Motor Learning (Reaction time)	302	$r = -0.007, p = 0.898$	$r = 0.039, p = 0.469$
(d) With covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = 0.005, p = 0.931$	$r = 0.065, p = 0.230$
	Motor Learning (Reaction time)	302	$r = -0.007, p = 0.898$	$r = 0.039, p = 0.469$
(e) With covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = 0.005, p = 0.931$	$r = 0.065, p = 0.230$
	Motor Learning (Reaction time)	302	$r = -0.007, p = 0.898$	$r = 0.039, p = 0.469$
(f) With covariate				
CamCAN	Fluid Intelligence (Cattel test)	631	$r = 0.005, p = 0.931$	$r = 0.065, p = 0.230$
	Motor Learning (Reaction time)	302	$r = -0.007, p = 0.898$	$r = 0.039, p = 0.469$

scores (WASI matrix reasoning and similarities). The results with age, age<sup>2</sup>, and gender as covariates showed a similar trend (Table S5b).

Since there was a residual correlation between corrected delta and age, the correlations with behavior without age as a covariate can be unreliable. We, therefore, do not discuss correlations of the corrected delta without age as a covariate, but they are reported in Table 4 for completeness. Additionally, as expected, the correlations using corrected delta with age as a covariate were similar to uncorrected delta with covariate (Table 4b).

### 3.5. Predictions in the ADNI sample

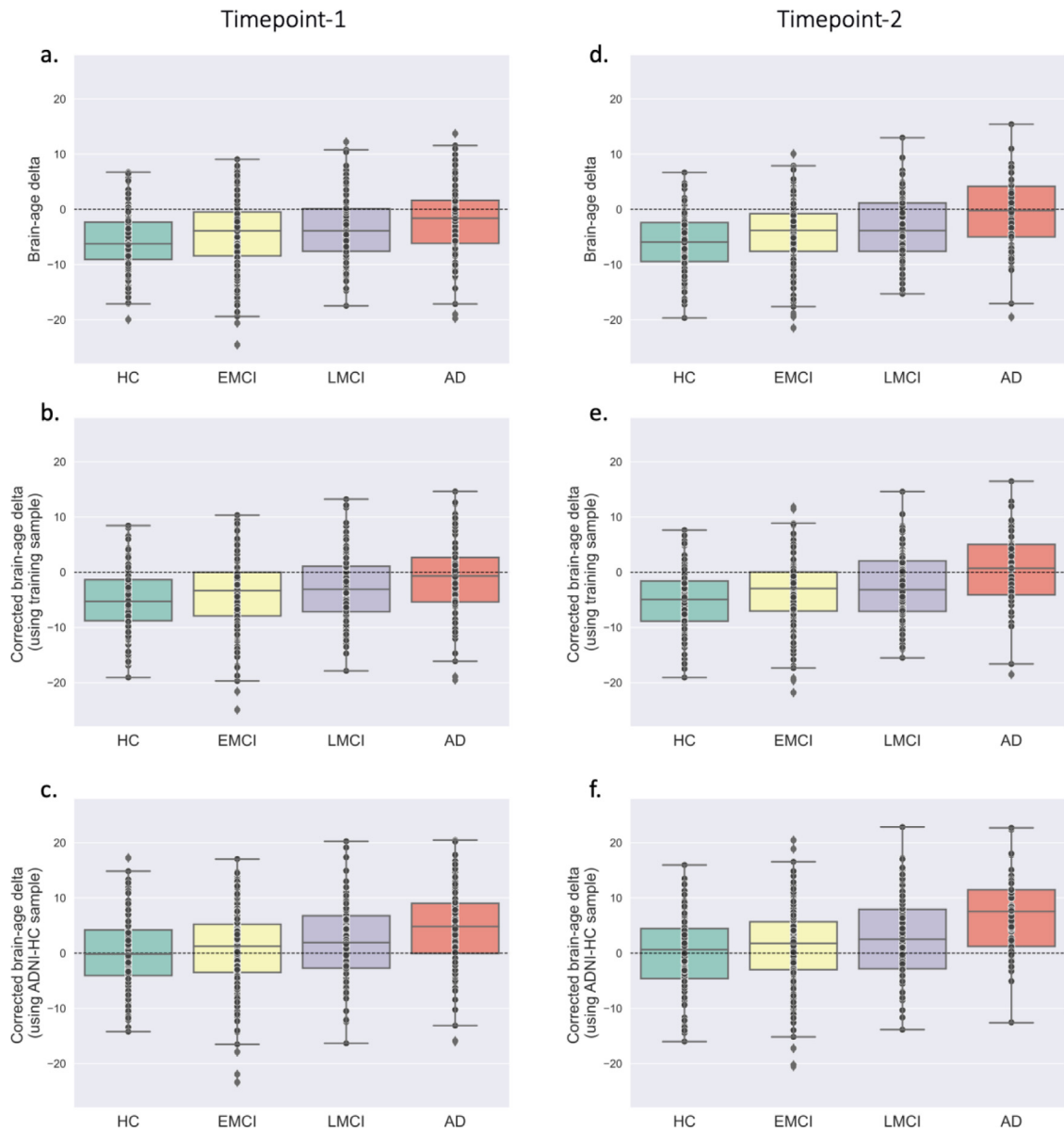
At timepoint 1, the mean uncorrected delta was -5.97 years in HC, -4.39 in EMCI, -3.57 in LMCI, and -2.13 in AD (Fig. 4a). In other words, the model underestimated age. The slope and intercept derived from the bias correction model using the training data (CV predictions) could not entirely correct for the under-estimation and age bias (Fig. 4b). Bias correction using the whole ADNI HC sample removed the bias (average delta, HC = 0, EMCI = 0.85, LMCI = 2.09, AD = 4.47 years) (Fig. 4c). ANOVA revealed that the corrected delta differed significantly across the groups ( $F = 12.94, p = 3.10e-08$ ), and post-hoc t-tests revealed significant differences between AD and HC ( $p = 1.16e-08$ ), EMCI ( $p = 1.87e-05$ ), LMCI ( $p = 0.043$ ), and HC and LMCI ( $p = 0.022$ ) after Bonferroni correction. At timepoint 2, the pattern was similar to timepoint 1 but with higher corrected delta values (EMCI = 1.15 years, LMCI = 2.88, AD = 6.59 years) (Fig. 4e-f, Table 5). These results demonstrate that our model could capture the range of normal structural variation related to age in healthy subjects and deviance in both MCI and AD patients.

The correlations between HC sample-corrected delta and various clinical test scores were calculated with age as a covariate (Table 6). At timepoint 1, the delta was negatively correlated with MMSE ( $r = -0.255, p = 0.016$ ) and positively correlated with FAQ ( $r = 0.275, p = 0.005$ ) in the entire sample. No correlations were found in individual diagnostic groups or could not be calculated due to insufficient score data. At timepoint 2, the delta was negatively correlated with MMSE ( $r = -0.303, p = 2.40e-12$ ) and positively correlated with CDR ( $r = 0.270, p = 7.35e-10$ ) and FAQ ( $r = 0.331, p = 2.31e-14$ ) in the whole sample. In the AD group, the delta was positively correlated with FAQ ( $r = 0.298, p = 0.021$ ) but not with MMSE or CDR. In the LMCI group, the delta was positively correlated with FAQ ( $r = 0.309, p = 0.002$ ), negatively correlated with MMSE ( $r = -0.227, p = 0.022$ ), and not correlated with CDR. In the EMCI group, the delta positively correlated with CDR ( $r = 0.153, p = 0.034$ ) but not MMSE and FAQ scores. No correlations were found in the HC group. The correlations with age, age<sup>2</sup>, and gender as covariates were similar (Table S6).

We also found that the size of HC sample used for bias correction considerably impacts the mean corrected delta in AD subjects (Fig. S7). Specifically, with fewer HC subjects, the variance of the corrected delta in AD was much higher in both sessions, e.g., at the timepoint 1 when using 21 HC samples, the mean AD delta ranged between ~1–12 years and converged to 4.47 years as the sub-samples approached the complete sample.

### 3.6. Relationship of MAE with delta and delta-behavior correlations

Using 32 workflows selected from the cross-dataset evaluation, we analyzed whether model performance (MAE) was associated with their brain-behavior correlations. The corrected mean delta in AD ranged from 5.43 to 10.01 years, with some relatively poor performing models yielding a higher delta in AD (Table S7). Lower accuracy (higher MAE) was associated with stronger delta-MMSE correlation (Fig. 5c). In contrast, lower MAE was associated with a stronger brain-behavior correlations in the two healthy samples, delta-motor learning reaction time in CamCAN, and delta-CWIT inhibition trial completion time in eNKI datasets (Fig. 5a & b).



**Fig. 4.** Brain-age delta in the clinical population. The box plot compares the delta between healthy control (HC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and Alzheimer's disease (AD) from the ADNI sample at (left) timepoint-1 and (right) timepoint-2. Box plot with a & d. uncorrected delta. b & e. corrected delta using the CV predictions from the training set. c & f. corrected delta using the predictions from HC-ADNI subjects.

**Table 5**

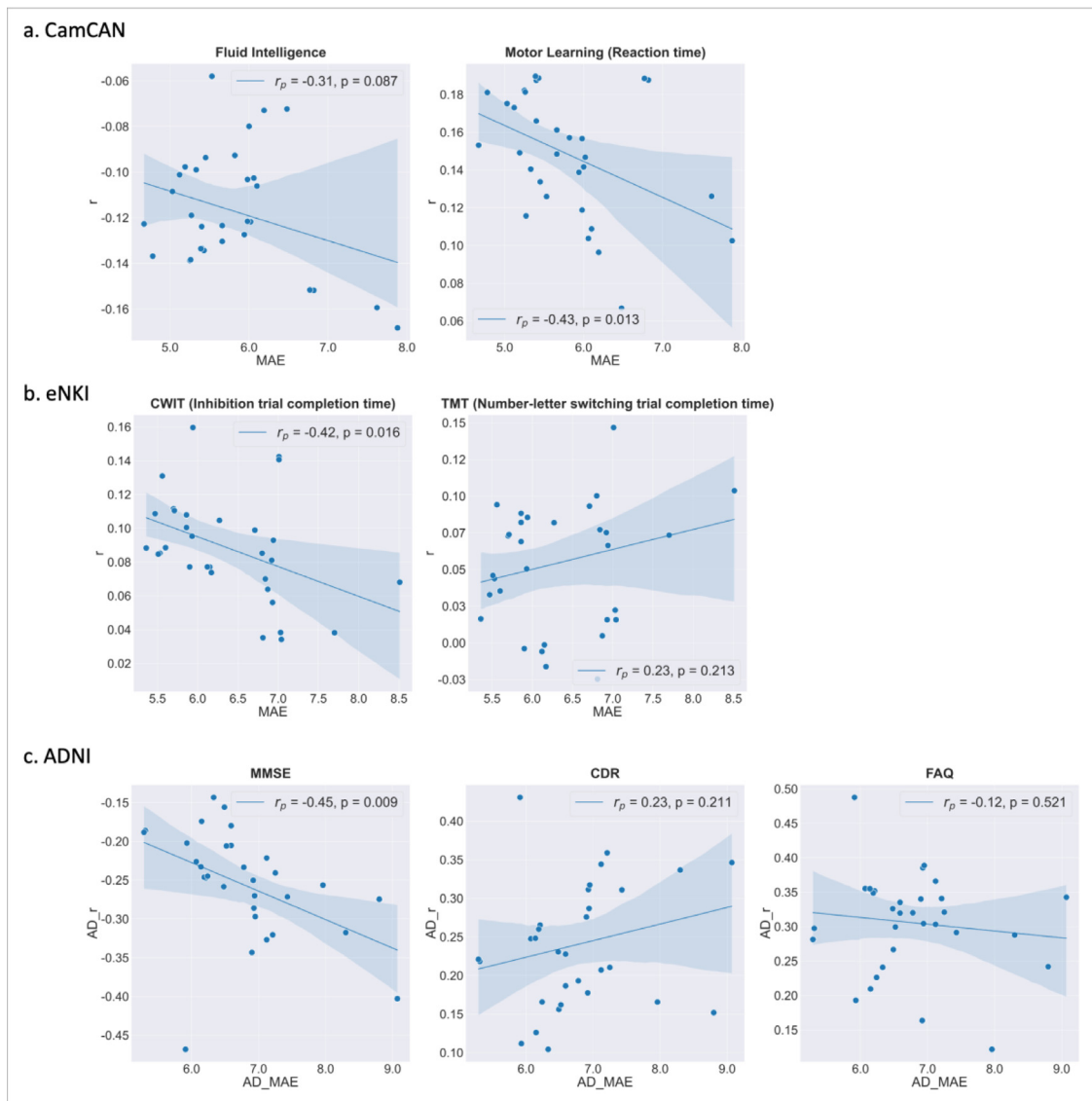
Prediction performance on the ADNI data from two timepoints using the best-performing (S4\_R4 + PCA + GPR) workflow. Abbreviations: HC: healthy control, EMCI and LMCI: early and late mild cognitive impairment, AD: Alzheimer's disease.

Time-point	ADNI sample	N	MAE	MSE	Corr (true, pred)	Mean delta	Mean corrected delta (train samples)	Mean corrected delta (ADNI-HC samples)
1	HC	209	6.56	61.19	$r = 0.76, p = 4.67\text{e-}40$	-5.97	-5.18	0.00
	EMCI	237	5.76	52.30	$r = 0.72, p = 1.07\text{e-}38$	-4.39	-3.78	0.85
	LMCI	127	5.56	46.52	$r = 0.75, p = 4.30\text{e-}24$	-3.57	-2.86	2.09
	AD	125	5.18	44.29	$r = 0.66, p = 5.00\text{e-}17$	-2.13	-1.20	4.47
2	HC	153	6.56	62.73	$r = 0.73, p = 5.46\text{e-}27$	-6.05	-5.27	0.00
	EMCI	197	5.57	50.82	$r = 0.73, p = 1.23\text{e-}34$	-4.32	-3.66	1.15
	LMCI	104	5.68	47.75	$r = 0.72, p = 6.54\text{e-}18$	-3.25	-2.44	2.88
	AD	61	5.31	44.12	$r = 0.59, p = 6.09\text{e-}07$	-0.76	0.31	6.59

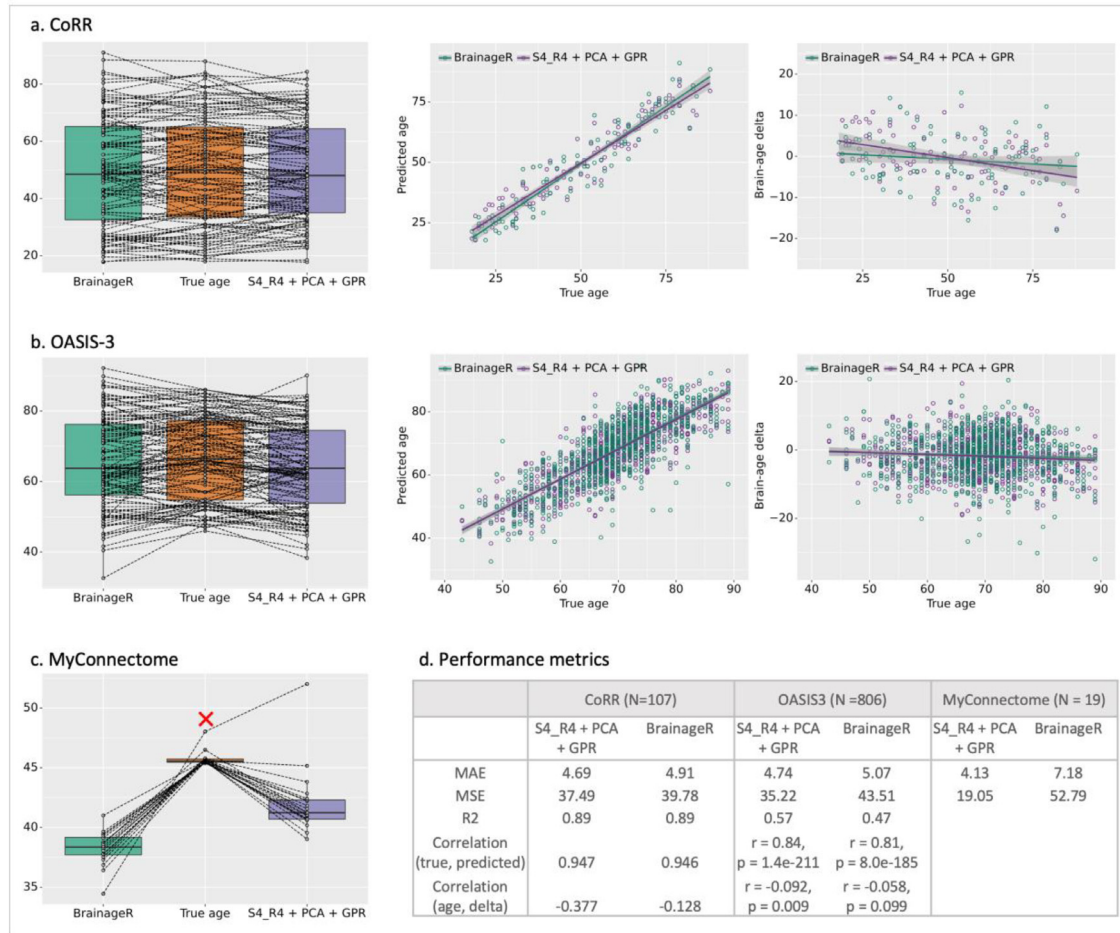
**Table 6**

Pearson's correlation coefficients between corrected brain-age delta using S4\_R4 + PCA + GPR workflow and cognitive measures (MMSE, CDR, and FAQ) using age as a covariate from the ADNI sample. The correlations were computed for the whole sample and each diagnostic group (HC, EMCI, LMCI and AD) separately from two timepoints. Abbreviations: MMSE: Mini-Mental State Examination, CDR: Global Clinical Dementia Rating Scale, FAQ: Functional Assessment Questionnaire; HC: healthy control, EMCI and LMCI: early and late mild cognitive impairment, AD: Alzheimer's disease.

	Timepoint-1			Timepoint-2		
	MMSE	CDR	FAQ	MMSE	CDR	FAQ
<b>HC</b>	$N = 68$ $r = -0.202, p = 0.101$	$N = 67$ $r = 0.025, p = 0.841$	$N = 74$ $r = 0.153, p = 0.196$	$N = 153$ $r = -0.065, p = 0.427$	$N = 147$ $r = -0.019, p = 0.819$	$N = 149$ $r = 0.070, p = 0.399$
<b>EMCI</b>	$N = 3$ n.a.	$N = 3$ n.a.	$N = 3$ n.a.	$N = 196$ $r = -0.079, p = 0.272$	$N = 194$ <b><math>r = 0.153, p = 0.034</math></b>	$N = 193$ $r = 0.091, p = 0.211$
<b>LMCI</b>	$N = 2$ n.a.	$N = 2$ n.a.	$N = 2$ n.a.	$N = 103$ <b><math>r = -0.227, p = 0.022</math></b>	$N = 102$ $r = 0.115, p = 0.253$	$N = 103$ <b><math>r = 0.309, p = 0.002</math></b>
<b>AD</b>	$N = 17$ $r = -0.435, p = 0.092$	$N = 17$ $r = 0.221, p = 0.412$	$N = 26$ $r = 0.244, p = 0.240$	$N = 61$ $r = -0.186, p = 0.155$	$N = 61$ $r = 0.218, p = 0.094$	$N = 61$ <b><math>r = 0.298, p = 0.021</math></b>
<b>Whole sample</b>	$N = 90$ <b><math>r = -0.255, p = 0.016</math></b>	$N = 89$ $r = 0.114, p = 0.290$	$N = 105$ <b><math>r = 0.275, p = 0.005</math></b>	$N = 513$ <b><math>r = -0.303, p = 2.40e-12</math></b>	$N = 504$ <b><math>r = 0.270, p = 7.35e-10</math></b>	$N = 506$ <b><math>r = 0.331, p = 2.31e-14</math></b>



**Fig. 5.** Correlation between MAE and delta-behavioral correlations obtained using 32 workflows a. CamCAN ( $N = 302$ ) b. eNKI ( $N = 340$ ) c. ADNI ( $N = 61$ ). For CamCAN and eNKI data, the within-dataset delta-behavior correlations with age as a covariate were used. For ADNI data, we used the delta-behavior correlations using corrected delta (corrected using the HC sample) with age as a covariate.



**Fig. 6.** Comparison of our best workflow (S4\_R4 + PCA + GPR) with the brainageR model on a. CoRR dataset (left) the box plot comparing predicted age from two models with true age using a sub-sample of 107 subjects, (center) the scatter plot between the chronological (true) age and predicted age, (right) the scatter plot between the chronological (true) age and brain-age delta. b. OASIS-3 dataset (for visual clarity, the box plot is created using a random sub-sample;  $N = 120$ ) c. MyConnectome dataset (the red cross indicates the outlier scan that was removed from the analysis; final  $N = 19$ ). d. Performance metrics for all datasets. For the CoRR dataset, the table shows average values from 100 iterations of sub-sampled data, but the plots are from one iteration.

### 3.7. Comparison with brainageR and effect of preprocessing and tissue types

Next, we compared the S4\_R4 + PCA + GPR workflow and the brainageR model both trained on the same data using the CoRR, OASIS-3, and MyConnectome datasets (Fig. 6).

In CoRR dataset, S4\_R4 + PCA + GPR (mean MAE = 4.69,  $r = 0.947$ , bias  $r = -0.377$ ) performed better than brainageR (mean MAE = 4.91,  $r = 0.946$ , bias  $r = -0.128$ ) in MAE (paired  $t$ -test:  $t = -8.04$ ,  $p = 1.97e-12$ ) but brainageR showed a lower mean age bias (Steiger's Z test (Steiger, 1980)  $z = -3.31$ ,  $p = 0$ ; Figs. 6a & S8). There was no significant difference between the mean true and predicted age correlations from two models ( $z = 0.133$ ,  $p = 0.447$ ).

S4\_R4 + PCA + GPR (MAE = 4.74,  $r = 0.836$ , bias  $r = -0.092$ ) also showed lower MAE than brainageR (MAE = 5.07,  $r = 0.805$ , bias  $r = -0.058$ ) on the OASIS-3 dataset (Fig. 6b). The predicted ages (paired  $t$ -test:  $t = -1.37$ ,  $p = 0.17$ ) and the bias ( $z = -1.031$ ,  $p = 0.151$ ) of the two models were similar but the  $r$  value for our model was significantly higher ( $z = 3.101$ ,  $p = 0.001$ ). Test-retest reliability on a sub-sample of the OASIS-3 dataset (retest duration < 3 months) was higher for brainageR (CCC = 0.94 vs. 0.82 for S4\_R4 + PCA + GPR). Both models did not show longitudinal consistency at a retest duration of 3–4 years.

Additionally, S4\_R4 + PCA + GPR workflow (MAE = 4.13) performed significantly better than brainageR (MAE = 7.18) on the MyConnectome

dataset (paired  $t$ -test:  $t = 9.60$ ,  $p = 1.66e-08$ ; Fig. 6c). Note that one outlier scan (true age = 48) was excluded from this analysis (final  $N = 19$ ).

To gain insight into the impact of preprocessing, we compared within-dataset performance of our workflow using SPM preprocessing on IXI and CamCAN datasets. On both datasets, CAT-derived GM features performed better (IXI: MAE = 4.85 years; CamCAN: MAE = 5.01) than SPM-derived GM features (IXI: MAE = 6.25; CamCAN: MAE = 5.82) (Table 7). SPM-derived features from three tissue types performed better (IXI: MAE = 5.08; CamCAN: MAE = 4.88) than using only SPM-derived GM features, indicating that different tissue types carry complementary information (Table 7).

## 4. Discussion

### 4.1. Effect of feature space and ML algorithm

The wide range of options available for designing brain-age estimation workflows makes it challenging to disentangle the effect of feature space and ML algorithms. To this end, we investigated 128 workflows constituting combinations of 16 feature representations (voxel-wise and parcel-wise) extracted from GMV images and eight ML algorithms.

Previous studies have shown that the age prediction MAE ranges between ~5–8 years for broad age range data (18–90 years) when using GMV features (Table S1). Our workflows showed performance in a similar range, with some of the workflows generalizing well to data from a



**Table 7**

Comparison of within-dataset performance between models trained with CAT-preprocessed GM features ( $S4\_R4 + PCA + GPR$ ; our framework), SPM-preprocessed GM features ( $S4\_R4_{SPM} + PCA + GPR$ ) and SPM-preprocessed GM+WM+CSF features ( $S4\_R4_{SPM}^{WM+CSF} + PCA + GPR$ ) on IXI and CamCAN data. Abbreviations: MAE: mean absolute error, MSE: mean squared error, Corr (true, pred): Pearson's correlation between true age and predicted age, Age bias: Pearson's correlation between true age and brain-age delta.

	Workflow	MAE	MSE	Corr (true, pred)	Age bias
<b>IXI (N = 562)</b>	$S4\_R4 + PCA + GPR$	4.85	36.89	$r = 0.93, p = 1.03e-247$	$r = -0.21, p = 7.39e-07$
	$S4\_R4_{SPM} + PCA + GPR$	6.25	62.34	$r = 0.88, p = 1.15e-181$	$r = -0.40, p = 1.61e-22$
	$S4\_R4_{SPM}^{WM+CSF} + PCA + GPR$	5.08	40.80	$r = 0.92, p = 3.98e-234$	$r = -0.27, p = 1.64e-10$
<b>CamCAN (N = 650)</b>	$S4\_R4 + PCA + GPR$	5.01	40.89	$r = 0.94, p = 6.45e-307$	$r = -0.17, p = 1.14e-05$
	$S4\_R4_{SPM} + PCA + GPR$	5.82	56.83	$r = 0.92, p = 3.87e-258$	$r = -0.30, p = 2.66e-15$
	$S4\_R4_{SPM}^{WM+CSF} + PCA + GPR$	4.88	39.77	$r = 0.94, p = 8.29e-308$	$r = -0.25, p = 1.53e-10$

new site. Specifically, the MAE ranged between 4.90–8.48 years in CV and 4.73–8.38 years in test data for within-dataset analysis and for cross-dataset analysis between 4.28–7.39 years and 5.23–8.98 years in CV and test data, respectively. The test MAE and  $R^2$  were highly correlated for both within-dataset and cross-dataset analysis (Tables S2 & S3, Fig. S5). The workflows showed high positive correlations between chronological age and predicted age for within-dataset ( $r$  between 0.81–0.93) and cross-dataset ( $r$  between 0.82–0.93) analyses. The workflows that performed well in within-dataset analysis also performed well in cross-dataset analysis. The lower cross-dataset CV MAE (4.28–7.39 years) compared to within-dataset CV MAE (4.90–8.48 years) might be because of the larger sample sizes in the cross-dataset analysis or possible overfitting in smaller samples. This corroborates previous studies showing lower errors with larger training sets (Baecker et al., 2021; de Lange et al., 2022), contrary to others that have shown a negative correlation between sample size and CV performance estimates (Wolters et al., 2015; Varoquaux, 2018). The age range of the training and test data affects the performance estimates. Specifically, when using a narrow age range, performance metrics such as MAE and RMSE are usually better than broad age range evaluations (Cole, 2020; Peng et al., 2021; de Lange et al., 2022). However, the lower errors and hence smaller brain-age delta values in those cases are not necessarily due to better model performance but rather because the predictions are closer to the mean age of the group. Here, our focus was on broad age range models, and the errors we obtained are within the range of what has been previously shown.

Our results showed that the choice of feature space and the ML algorithm both affect the prediction error. In general, feature spaces derived from voxel-wise GMV such as  $S4\_R4$ ,  $S4\_R8$ , and  $S0\_R4$  in combination with GPR, KRR, RVRpoly, and RVRlin algorithms performed well in the within-dataset analysis. The results were similar with PCA retaining 100% variance for some workflows but not all, especially the regularized models (LR and ENR) showed lower performance after PCA (see Supplementary Table S2). This might be because of the different biases of ML algorithms, e.g., due to regularization. It is possible that the sparsity-inducing penalization in addition to PCA leads to lower accuracy models. Some of these selected workflows also performed well on cross-dataset analysis. Specifically, the voxel-wise GMV features smoothed with a 4 mm FWHM kernel and resampled to a spatial resolution of 4 mm, without and with PCA ( $S4\_R4$  and  $S4\_R4 + PCA$ ) together with the GPR algorithm performed best in both the within-dataset and cross-dataset analyses. A previous study has reported a voxel size of 3.73 mm<sup>3</sup> and a smoothing kernel of 3.68 mm as the optimal parameters for processing GM images for brain-age prediction with a performance similar to our workflows (Lancaster et al., 2018). In general, parcel-wise features performed worse than voxel-wise features irrespective of the ML algorithm used, suggesting that the GMV summarized from parcels leads to a loss of age-related information. Our results align with a recent study comparing several ML models (GPR-dot product kernel, RVR-linear kernel, and SVR-linear kernel) trained on region-based and voxel-based features with or without PCA on a narrower age range (47–73 years) (Baecker et al., 2021). They found minimal differences in performance

due to the ML algorithms with voxel-based features performing better than region-based features.

Our results also indicate that the non-linear algorithm (GPR with RBF kernel) and the kernel-based algorithms (KRR and RVR) outperformed linear algorithms such as RR and LR. Surprisingly, the non-linear RFR algorithm performed the worst irrespective of the feature space used (Fig. S4). This suggests that capturing distributional information using the RBF kernel, as we did using GPR, and use of kernels that capture the similarity between the GMV features in an invariant manner (e.g., Pearson correlation) is beneficial. These results corroborate a recent study that comprehensively evaluated 22 regression algorithms (test MAE between 4.63–7.14 years) in broad age range data (18–94 years) using GMV features and found SVR, KRR, and GPR with a diverse set of kernels to perform well (Beheshti et al., 2022).

In sum, the smoothed and resampled voxel-wise data (such as  $S4\_R4$ ,  $S4\_R8$ ) with either a non-linear or a kernel-based algorithm (GPR with RBF kernel, KRR with polynomial kernel degree (1 or 2), and RVR with linear and polynomial degree 1 kernels) are well suited for brain-age estimation. Sometimes, especially with a large number of features, PCA might help improve performance (Franke et al., 2010; Baecker et al., 2021). However, we found the performance of these workflows with and without PCA to be similar. Therefore, one could use the features directly for immediate interpretability of the models; on the other hand, if computation is a constraint, then the PCA retaining 100% variance could be used without affecting the performance.

Future studies can investigate options to improve model generalizability, such as data harmonization to remove site effects and considerations for population structure (e.g., over-representative of the Caucasian population in the datasets used).

#### 4.2. Test-retest reliability and longitudinal consistency

The brain-age estimates must be reliable within a subject. We found the delta to be reliable over a short scan delay (CoRR: CCC = 0.95–0.98, age range = 20–84; OASIS-3: CCC = 0.76–0.85, age range = 43–80). The reliability of delta within a short scan duration has been reported in previous studies. For example, one study showed an intraclass correlation coefficient (ICC) of 0.96 between deltas from subjects scanned an average of  $28.35 \pm 1.09$  days apart ( $N = 20$ , mean age at first scan =  $34.05 \pm 8.71$ ) (Cole et al., 2017). Another study showed an ICC of 0.93 in young adults from the OASIS-3 dataset ( $N = 20$ , age range = 19–34) scanned within a short delay of less than 90 days (Franke and Gaser 2012). Another study found an ICC of 0.81 with a mean interval of 79 days between scans ( $N = 20$ , chronological age = 45 years) (Elliott et al., 2021).

Longitudinal consistency, i.e., chronologically proportionate increase in predicted age, is crucial for real-world application. Previous studies have shown lifestyle interventions, such as meditation and exercise (Luders et al., 2016; Steffener et al., 2016), can have positive effects on brain-age, while factors such as smoking and alcohol intake can have adverse effects (Bittner et al., 2021). For instance, 18 months of lifestyle intervention, including diet change and physical activity,

showed attenuated brain-age in a longitudinal sample which correlated with improvement in several physiological measures (Levakov et al., 2022). Thus, lifestyle can lead to different longitudinal brain-age trajectories. However, in our analyses, we assumed that there were no such interventions over the retest duration as the datasets did not provide such information. With this assumption, we expected brain-age to increase proportionally with chronological age.

In support of this assumption, we found a positive linear relationship between the difference in predicted age and the difference in chronological age at a retest duration of 2–3.25 years ( $N = 26$ ;  $r = 0.447$ ,  $p = 0.022$ ) in the CoRR dataset. However, there was no correlation in the OASIS-3 dataset with a retest duration of 3–4 years ( $N = 127$ ;  $r = -0.008$ ,  $p = 0.932$ ). Thus, the evidence of longitudinal consistency was weak. This can be speculatively explained by the maximum test-retest duration of 3–4 years which lies within the range of the MAE for the OASIS-3 dataset (MAE session-1: 5.08 and session-2: 5.86 years, Table S4). Taken together, the high reliability supports the use of brain-age in clinical settings; however, further evaluations are needed to establish longitudinal consistency.

#### 4.3. Effect of bias correction

Most brain-age estimation workflows produce biased results, i.e., overestimation at younger ages and underestimation at older ages (Liang et al., 2019). Therefore, correcting this age bias is important to facilitate individual-level decisions. Here, we adopted a bias correction model that does not use the chronological age of test samples for correction (Cole, 2020), as using chronological age can hamper fair comparison between workflows (de Lange et al., 2022).

The tested workflows generally showed negative associations between chronological age and delta for both within-dataset ( $r$  between  $-0.22$  to  $-0.83$ ) and cross-dataset ( $r$  between  $-0.27$  to  $-0.75$ ) predictions. However, this age bias was less pronounced in more accurate models (Fig. S5). This result is in line with the previous work (de Lange et al., 2022) that showed that if input features are not informative enough to predict age, predictions will be closer to the median or mean age, leading to this bias. Additionally, we found that the data used to estimate the bias correction models can significantly impact the corrected delta. Specifically, within-dataset-derived models corrected the age bias more adequately than cross-dataset models (Fig. S3). This discrepancy might be due to the difference in data properties, e.g., scanner-specific idiosyncrasy, between the training and the out-of-site test data. Our results suggest that a bias correction model might not always work well when applied to a new site, even when the training data itself consists of multiple sites. Consequently, using part of the test data to correct the age bias in the remaining test data works well (as seen in the ADNI data analysis, Section 3.5). However, this might not be feasible when the test sample is small or in the extreme case, a single test subject is available.

How much data is needed for learning a bias correction model is an important but unexplored question. We investigated this by learning bias correction models from sub-samples of the HC subjects from ADNI data. Smaller samples led to higher variance in the efficacy of bias correction models when applied to AD patients (Varoquaux, 2018). For instance, at the smallest sample size ( $N = 21$ ), the average corrected delta of the AD patients varied from 1 to 12 years (Fig. S7, ADNI timepoint 1). It is likely that different studies use different samples for bias correction, so the results should be interpreted and compared with caution. This result shows the importance of using large samples for bias correction and emphasizes careful analysis and reporting of the results.

#### 4.4. Correlation with behavior

Using the selected workflow we observed that the correlation of delta with behavioral measures is sensitive to whether the delta was adjusted for age, either via bias correction or using it as a covariate. For instance,

the uncorrected delta was not correlated with FI and motor learning reaction time (in CamCAN data) or CWIT inhibition trial completion time (in eNKI data); however, significant correlations were obtained using age-adjusted delta (Table 4). Thus, it is important to control for age when analyzing correlations between delta and behavioral measures.

Using out-of-sample predictions from within-dataset analysis, we found that a higher uncorrected delta (with age as a covariate) was associated with lower FI, higher motor learning reaction time (from CamCAN data), and lower response inhibition and selective attention, indicated by higher CWIT inhibition trial completion time (from eNKI data). We expected these correlations to be similar to correlations calculated using corrected delta (de Lange and Cole, 2020), as there was no significant age bias. In the CamCAN data, the behavioral correlations using uncorrected delta with age as a covariate and corrected delta were quite similar (FI:  $r = -0.154$ ,  $p = 0.0001$  vs.  $r = -0.157$ ,  $p = 7.24e-05$ ; motor learning reaction time:  $r = 0.181$ ,  $p = 0.002$  vs.  $r = 0.186$ ,  $p = 0.001$ ). However, the correlation of CWIT inhibition trial completion time with uncorrected delta with age as a covariate was significant but not when using the corrected delta ( $r = 0.109$ ,  $p = 0.045$  vs.  $r = 0.094$ ,  $p = 0.084$ ). This slight difference could potentially be explained by the small effect size and differences inherent in the two methods used for correction.

We also found that there was disagreement between delta-behavior correlations from within-dataset and cross-dataset predictions with age as a covariate. For instance, CamCAN showed significant correlations with FI and motor learning reaction time with within-dataset delta but not with cross-dataset delta. On the other hand, eNKI showed significant correlations only with CWIT inhibition trial completion time using within-dataset delta, but a significant correlation with TMT completion time was found using cross-dataset delta. These results indicate that the subtle differences in predictions can impact behavioral correlations, even though the two predictions were highly correlated (CamCAN:  $r = 0.961$ , eNKI:  $r = 0.962$ ; Fig. S6). Thus, the delta-behavior correlations, whether using within-dataset or cross-dataset delta, should be interpreted with caution.

Taken together, within-dataset data yields better bias correction models, as we observed in two scenarios, behavioral correlations and delta estimation. However, when enough data are not available, the resulting models may fail to correct the age bias, leading to high variability in the mean delta (Fig. S7). We therefore caution the practitioners and recommend carefully assessing bias correction models, e.g., using bootstrap analysis, before application. We observed that subtle differences in predicted age (within-dataset vs. cross-dataset) lead to different behavioral correlations, which can question the impact of the workflow used for prediction, the analysis method used for computing behavioral correlation (corrected delta versus covariates) and their interaction. Future studies should focus on disentangling such intricacies before applying the brain-age paradigm in practice.

#### 4.5. Higher brain-age delta in neurodegenerative disorders

Neurodegenerative disorders such as AD, MCI, and Parkinson's disease (PD) are accompanied by brain atrophy. Many studies have shown a decrease in global and local GMV in MCI and AD (Good et al., 2001; Karas et al., 2004; Fjell et al., 2014) and also in a broad range of neuropsychiatric disorders (Kaufmann et al., 2019). Consequently, an increased delta, i.e., older appearing brains, has been reported in patients with MCI (3–8 years) and AD (~10 years) (Franke and Gaser 2012; Gaser et al., 2013; Varikuti et al., 2018). We assessed the delta in HC, EMCI, LMCI, and AD patients by applying our best-performing workflow followed by a bias correction model estimated on HC. We found that brain aging is advanced by ~4.5–7 years in AD, ~2–3 years in LMCI, and ~1 year in EMCI (timepoint 1-timepoint 2; Table 5). Furthermore, the delta was correlated with measures associated with disease severity and cognitive impairment in MCI and AD patients. Thus, in line with previous studies, brain-age delta confirmed its potential to indicate accelerated brain aging in neurodegenerative diseases based on structural

MRI data (Franke and Gaser, 2012; Varikuti et al., 2018; Cole et al., 2020; Eickhoff et al., 2021; Lee et al., 2021).

We also show that different workflows can lead to different delta estimates in AD and, consequently, different correlations with cognitive measures (Table S7). In addition, the mean corrected delta in the patient group depends on the type (within-dataset or cross-dataset) and size of sample used for bias correction (Figure S7). Thus, the results should be interpreted with caution when comparing different studies.

#### 4.6. Relationship of MAE with delta and delta-behavior correlations

The utility of age prediction models lies in their application to capture atypical aging. However, to achieve this, it is imperative to minimize the methodological variance, due to decisions in feature space and ML algorithms, by building accurate models so that the resulting brain-age delta captures biological variance. A recent study has shown that delta from overfitted models (i.e., with higher training accuracy) results in smaller differences in AD vs. CN, while delta from a model with comparatively lower (training) accuracy captures biological variance (Bashyam et al., 2020). However, our analyses and model selection was based on nested cross-validation. Therefore, our accurate models cannot be considered overfitted.

In healthy samples, higher accuracy (lower MAE) was associated with higher delta-motor learning reaction time (CamCAN) and delta-CWIT inhibition trial completion time (eNKI) associations. In contrast, in AD patients, models with lower accuracy (higher MAE) showed a stronger delta-MMSE correlation. This observation that some less accurate models can better capture the delta-behavioral correlation better in AD is in line with a previous study (Bashyam et al., 2020) (Fig. 5 and Table S7). These contrasting observations in healthy and patient cohorts make it difficult to develop a model selection strategy based on delta-behavioral correlations.

The corrected mean delta in AD (corrected using the CN sample, indicative of separation between CN and AD), for the 32 workflows ranged from 5.43 to 10.01 years. Some moderately accurate models, e.g., S0\_R4 + LR (delta = 7.27, MAE = 5.91 years), showed a high delta for AD and a strong correlation with AD scales (Table S7). However, the model with the highest delta (173 + RFR: delta = 10.01, MAE: 9.07 years) showed a comparatively weaker correlation with behavior. Moreover, similarly performing models (S0\_R4 + LR: delta = 7.27, MAE = 5.91 years vs. S8\_R4 + KRR: delta = 7.17, MAE = 6.59 years) showed quite different correlation with behavior. This indicates a non-linear relationship between the models' MAEs, deltas, and behavioral correlations.

Based on these results, we speculate that perhaps using adequately regularized models in the patient population can be beneficial even if they show a lower accuracy. It might be possible that regularization pushes the models to focus on fewer specific features containing typical aging-related signal. This in turn could lead to lower accuracy models (as it downweights some features) but also leads to delta estimates that are more informative of atypical aging.

Taken together, comparing models based on their performance on patient data and delta-behavior correlations is a promising but open topic. In particular, it is unclear which delta-behavioral correlation to use, and generalizability of models across behavioral scores, samples, and disorders remains unknown. Further studies are needed to define appropriate procedures for model selection based on such criteria.

#### 4.7. Comparison with brainageR and effect of preprocessing and tissue types

Using the same training data as brainageR, our workflow outperformed brainageR in terms of MAE in three datasets; CoRR ( $N = 107$ ; mean MAE = 4.69 vs. 4.91), OASIS-3 ( $N = 806$ ; MAE = 4.74 vs. 5.07), and MyConnectome ( $N = 19$ ; MAE = 4.13 vs. 7.18). However, the bias of our model was similar or higher than that of brainageR and its test-retest

reliability was lower (OASIS-3,  $N = 36$ ; CCC = 0.82 vs. CCC = 0.94). Overall, our workflow showed lower MAE, higher correlation between true and predicted age but also higher age bias compared to brainageR. These differences are likely driven by differences in preprocessing, and the use of three tissue types by brainageR as opposed to us using only GM. To investigate this further, we performed two additional analyses.

Different VBM tools can provide different GMV estimates, influencing the estimated association with age (Tavares et al., 2019; Antonopoulos et al., 2023). The CAT-derived GMV features performed better than SPM preprocessing (both with S4\_R4 + PCA for feature extraction together with the GPR algorithm for learning) in terms of MAE (e.g., IXI: MAE = 4.85 vs. 6.25), the correlation between true and predicted age ( $r = 0.93$  vs.  $0.88$ ,  $p < 1e-6$  both) and age bias ( $r = -0.21$  vs.  $r = -0.40$ ,  $p < 1e-6$  both) (Table 7). We further found that the predictions when using three tissue types from SPM (GM, WM, and CSF) were better (IXI: MAE = 5.08,  $r = 0.92$ ,  $p < 1e-6$ , bias:  $r = -0.27$ ,  $p < 1e-6$ ). This is in line with a previous study that showed a slight performance improvement when using both GM and WM compared to only GM (Cole et al., 2017). Features from different tissue types may carry complementary information regarding age, providing better predictions and lower age bias. Many previous studies have used GM and WM together as features (Franke and Gaser, 2012; Cole et al., 2017; Cole et al., 2018, 2020), and others have used all three tissue types (Monté-Rubio et al., 2018; Xifra-Porxas et al., 2021; Hobday et al., 2022). CAT-derived GMV performed similarly to SPM-derived three tissue types with slightly lower age bias for the former (Table 7), showing the suitability of GM for this task following its clinical relevance in neurodegenerative disorders (Karas et al., 2004; Wu et al., 2021). Further studies are needed to cleanly disentangle the effect of tissue types on different performance criteria investigated here.

## 5. Conclusion

Numerous choices exist for designing a workflow for age prediction. The systematic evaluation of different workflows on the same data in different scenarios (within-dataset, cross-dataset, test-retest reliability, and longitudinal consistency) revealed a substantial impact of feature representation and ML algorithm choices. Notably, voxel-wise GM features, especially smoothed with a 4 mm FWHM kernel and resampled to a spatial resolution of 4 mm (S4\_R4), were better than parcel-wise features. Additionally, performing PCA did not affect the prediction performance, but it can help reduce computational resources. ML algorithms, including Gaussian process regression with the radial basis kernel, kernel ridge regression with polynomial kernel degree 1 or 2, and relevance vector machine with linear and polynomial degree 1 kernels, performed well. Overall, some workflows performed well on out-of-site data and showed high test-retest reliability but only moderate longitudinal reliability. Consistent with the literature, we found a higher delta in Alzheimer's and mild cognitive impairment patients after correcting the delta with a large sample of controls. Our results provide evidence for the potential future application of delta as a biomarker but also caution regarding analysis setup and data used for behavioral correlations and bias correction. Findings from the current study can serve as guidelines for future brain-age prediction studies.

## 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.

## Declaration of Competing Interest

The authors report no competing interests



## Credit authorship contribution statement

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

## Data availability

The codes used for preprocessing, feature extraction, model training and prediction are available at [https://github.com/juaml/brainage\\_estimation](https://github.com/juaml/brainage_estimation). All the datasets used except the 1000BRAINS data are available publicly (might require registration and approval).

## Acknowledgments

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

Data for the MyConnectome project were obtained from the OpenNeuro database (ds000031).

The clinical data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). The ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of the ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). For up-to-date information, see [www.adni-info.org](http://www.adni-info.org). Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health ([www.fnih.org](http://www.fnih.org)). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.neuroimage.2023.119947](https://doi.org/10.1016/j.neuroimage.2023.119947).

## References

- Antonopoulos, G., More, S., Raimondo, F., Eickhoff, S.B., Hoffstaedter, F. and Patil, K.R. 2023. A systematic comparison of VBM pipelines and their application to age prediction. *BioRxiv*.
- Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. *Neuroimage* 38 (1), 95–113.
- Ashburner, J., Friston, K.J., 2011. Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. *Neuroimage* 55 (3), 954–967.
- Baecker, L., Dafflon, J., da Costa, P.F., et al., 2021. Brain age prediction: a comparison between machine learning models using region- and voxel-based morphometric data. *Hum. Brain Mapp.* 42 (8), 2332–2346.
- Bashyam, V.M., Erus, G., Doshi, J., et al., 2020. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain J. Neurol.* 143 (7), 2312–2324.
- Beheshti, I., Ganaie, M.A., Paliwal, V., Rastogi, A., Razzak, I., Tanveer, M., 2022. Predicting brain age using machine learning algorithms: a comprehensive evaluation. *IEEE J. Biomed. Health Inform.* 26 (4), 1432–1440.
- Beheshti, I., Nugent, S., Potvin, O., Duchesne, S., 2019. Bias-adjustment in neuroimaging-based brain age frameworks: a robust scheme. *Neuroimage Clin.* 24, 102063.
- Bittner, N., Jockwitz, C., Franke, K., et al., 2021. When your brain looks older than expected: combined lifestyle risk and BrainAGE. *Brain Struct. Funct.* 226 (3), 621–645.
- Boyle, R., Jollans, L., Rueda-Delgado, L.M., et al., 2021. Brain-predicted age difference score is related to specific cognitive functions: a multi-site replication analysis. *Brain Imaging Behav.* 15 (1), 327–345.
- Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., Yeo, B.T.T., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106 (5), 2322–2345.
- Caspers, S., Moebus, S., Lux, S., et al., 2014. Studying variability in human brain aging in a population-based German cohort-rationale and design of 1000BRAINS. *Front. Aging Neurosci.* 6, 149.
- Chen, J., Liu, J., Calhoun, V.D., et al., 2014. Exploration of scanning effects in multi-site structural MRI studies. *J. Neurosci. Methods* 230, 37–50.
- Cole, J.H., 2020. Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. *Neurobiol. Aging* 92, 34–42.
- Cole, J.H., Leech, R., Sharp, D.J. Alzheimer’s Disease Neuroimaging Initiative, 2015. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Ann. Neurol.* 77 (4), 571–581.
- Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., et al., 2017a. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage* 163, 115–124.
- Cole, J.H., Raffel, J., Friede, T., et al., 2020. Longitudinal assessment of multiple sclerosis with the brain-age paradigm. *Ann. Neurol.* 88 (1), 93–105.
- Cole, J.H., Ritchie, S.J., Bastin, M.E., et al., 2018. Brain age predicts mortality. *Mol. Psychiatry* 23 (5), 1385–1392.
- Cole, J.H., Underwood, J., Caan, M.W.A., et al., 2017b. Increased brain-predicted aging in treated HIV disease. *Neurology* 88 (14), 1349–1357.
- Eickhoff, C.R., Hoffstaedter, F., Caspers, J., et al., 2021. Advanced brain ageing in Parkinson’s disease is related to disease duration and individual impairment. *Brain Commun.* 3 (3), fcab191.
- Elliott, M.L., Belsky, D.W., Knodt, A.R., et al., 2021. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. *Mol. Psychiatry* 26 (8), 3829–3838.
- Fan, L., Li, H., Zhuo, J., et al., 2016. The human brainnetome atlas: a new brain atlas based on connectome architecture. *Cereb. Cortex* 26 (8), 3508–3526.
- Fjell, A.M., McEvoy, L., Holland, D., Dale, A.M., Walhovd, K.B. Alzheimer’s Disease Neuroimaging Initiative, 2014. What is normal in normal aging? Effects of aging, amyloid and Alzheimer’s disease on the cerebral cortex and the hippocampus. *Prog. Neurobiol.* 117, 20–40.
- Franke, K., Gaser, C., 2012. Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer’s disease. *GeroPsych* 25 (4), 235–245 (Bern).
- Franke, K., Gaser, C., Manor, B., Novak, V., 2013. Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Front. Aging Neurosci.* 5, 90.
- Franke, K., Ziegler, G., Klöppel, S., Gaser, C. Alzheimer’s Disease Neuroimaging Initiative, 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.
- Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33 (1), 1–22.
- Galluzzi, S., Beltramello, A., Filippi, M., Frisoni, G.B., 2008. Aging. *Neurolog. Sci.* 29 (Suppl 3), 296–300.
- Gaser, C., Dahnke, R., Thompson, P.M., Kurth, F., Luders, E. and Alzheimer’s Disease Neuroimaging Initiative 2022. CAT – a computational anatomy toolbox for the analysis of structural MRI data. *Biorxiv*.
- Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H. Alzheimer’s Disease Neuroimaging Initiative, 2013. Brainage in mild cognitive impaired patients: predicting the conversion to alzheimer’s disease. *PLoS One* 8 (6), e67346.
- Giorgio, A., Santelli, L., Tomassini, V., et al., 2010. Age-related changes in grey and white matter structure throughout adulthood. *Neuroimage* 51 (3), 943–951.
- Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* 14 (1 Pt 1), 21–36.
- Grinsztajn, L., Oyallon, E. and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on tabular data? *arXiv*.



- Gutierrez Becker, B., Klein, T., Wachinger, C., Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing, 2018. Gaussian process uncertainty in age estimation as a measure of brain abnormality. *Neuroimage* 175, 246–258.
- Hahn, T., Ernsting, J., Winter, N.R., et al., 2022. An uncertainty-aware, shareable, and transparent neural network architecture for brain-age modeling. *Sci. Adv.* 8 (1), eabg9471.
- He, T., Kong, R., Holmes, A.J., et al., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage* 206, 116276.
- Hobday, H., Cole, J.H., Stanyard, R.A., et al., 2022. Tissue volume estimation and age prediction using rapid structural brain scans. *Sci. Rep.* 12 (1), 12005.
- Jack, C.R., Bernstein, M.A., Fox, N.C., et al., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* 27 (4), 685–691.
- Jiang, H., Lu, N., Chen, K., et al., 2019. Predicting brain age of healthy adults based on structural MRI parcellation using convolutional neural networks. *Front. Neurol.* 10, 1346.
- Jolliffe, I.T., 2002. *Principal Component Analysis*, 2nd ed. Springer-Verlag, New York.
- Jonsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., et al., 2019. Deep learning based brain age prediction uncovers associated sequence variants. *Biorxiv*.
- Jovicich, J., Czanner, S., Greve, D., et al., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage* 30 (2), 436–443.
- Karas, G.B., Scheltens, P., Rombouts, S.A.R.B., et al., 2004. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *Neuroimage* 23 (2), 708–716.
- Kaufmann, T., van der Meer, D., Doan, N.T., et al., 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat. Neurosci.* 22 (10), 1617–1623.
- LaMontagne, P.J., Benzinger, T.L.S., Morris, J.C., et al., 2019. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*.
- Lancaster, J., Lorenz, R., Leech, R., Cole, J.H., 2018. Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Front. Aging. Neurosci.* 10, 28.
- de Lange, A.-M.G., Anatórk, M., Rokicki, J., et al., 2022. Mind the gap: performance metric evaluation in brain-age prediction. *Hum. Brain Mapp.*
- de Lange, A.-M.G., Cole, J.H., 2020. Commentary: correction procedures in brain-age prediction. *Neuroimage Clin.* 26, 102229.
- de Lange, A.-M.G., Kaufmann, T., van der Meer, D., et al., 2019. Population-based neuroimaging reveals traces of childbirth in the maternal brain. *Proc. Natl. Acad. Sci. U.S.A.* 116 (44), 22341–22346.
- Le, T.T., Kuplicki, R.T., McKinney, B.A., et al., 2018. A nonlinear simulation framework supports adjusting for age when analyzing brainage. *Front. Aging. Neurosci.* 10, 317.
- Lee, W.H., Antoniadis, M., Schnack, H.G., Kahn, R.S., Frangou, S., 2021. Brain age prediction in schizophrenia: does the choice of machine learning algorithm matter? *Psychiatry research. Neuroimaging* 310, 111270.
- Levakov, G., Kaplan, A., Meir, A.Y., et al., 2022. The effect of 18 months lifestyle intervention on brain age assessed with resting-state functional connectivity. *medRxiv*.
- Liang, H., Zhang, F., Niu, X., 2019. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. *Hum. Brain Mapp.* 40 (11), 3143–3152.
- Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45 (1), 255–268.
- Löwe, L.C., Gaser, C., Franke, K., Alzheimer's Disease Neuroimaging Initiative, 2016. The effect of the APOE genotype on individual brainage in normal aging, mild cognitive impairment, and Alzheimer's disease. *PLoS One* 11 (7), e0157514.
- Luders, E., Cherbuin, N., Gaser, C., 2016. Estimating brain age using high-resolution pattern recognition: younger brains in long-term meditation practitioners. *Neuroimage* 134, 508–513.
- Monté-Rubio, G.C., Falcón, C., Pomarol-Clotet, E., Ashburner, J., 2018. A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods. *Neuroimage* 178, 753–768.
- More, S., Eickhoff, S.B., Caspers, J., Patil, K.R., 2021. Confound removal and normalization in practice: a neuroimaging based sex prediction case study. In: Dong, Y., Ifrim, G., Mladenici, D., Saunders, C., Van Hoecke, S. (Eds.), *ECML PKDD 2020: Demo Track*. Lecture Notes in Computer Science. Springer International Publishing, Ghent, Belgium, pp. 3–18.
- Nooner, K.B., Colcombe, S.J., Tobe, R.H., et al., 2012. The NKI-rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Front. Neurosci.* 6, 152.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*
- Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021. Accurate brain age prediction with lightweight deep neural networks. *Med. Image Anal.* 68, 101871.
- Petersen, R.C., Aisen, P.S., Beckett, L.A., et al., 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology* 74 (3), 201–209.
- Poldrack, R.A., Laumann, T.O., Koyejo, O., et al., 2015. Long-term neural and physiological phenotyping of a single human. *Nat. Commun.* 6, 8885.
- Richard, G., Kolskär, K., Sanders, A.-M., et al., 2018. Assessing distinct patterns of cognitive aging using tissue-specific brain age prediction based on diffusion tensor imaging and brain morphometry. *PeerJ* 6, e5908.
- Schaefer, A., Kong, R., Gordon, E.M., et al., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* 28 (9), 3095–3114.
- Schulz, M.-A., Yeo, B.T.T., Vogelstein, J.T., et al., 2020. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat. Commun.* 11 (1), 4238.
- Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E., Miller, K.L., 2019. Estimation of brain age delta from brain imaging. *Neuroimage* 200, 528–539.
- Steffener, J., Habeck, C., O'Shea, D., Razlighi, Q., Bherer, L., Stern, Y., 2016. Differences between chronological and brain age are related to education and self-reported physical activity. *Neurobiol. Aging* 40, 138–144.
- Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* 87 (2), 245–251.
- Su, L., Wang, L., Hu, D., 2013. Predicting the age of healthy adults from structural MRI by sparse representation. In: Yang, J., Fang, F., Sun, C. (Eds.), *Intelligent Science and Intelligent Data Engineering*. Lecture notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 271–279.
- Tavares, V., Prata, D., Ferreira, H.A., 2019. Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study. *J. Neurosci. Methods* 334, 108565.
- Taylor, J.R., Williams, N., Cusack, R., et al., 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.
- Thompson, N.C., Greenewald, K., Lee, K. and Manso, G.F. 2020. The computational limits of deep learning. *arXiv*.
- Treder, M.S., Shock, J.P., Stein, D.J., du Plessis, S., Seedat, S., Tsvetanov, K.A., 2021. Correlation constraints for regression models: controlling bias in brain age prediction. *Front. Psychiatry* 12, 615754.
- Varikuti, D.P., Genon, S., Sotiras, A., et al., 2018. Evaluation of non-negative matrix factorization of grey matter in age prediction. *Neuroimage* 173, 394–410.
- Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* 180 (Pt A), 68–77.
- Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *Neuroimage* 145 (Pt B), 166–179.
- Vidal-Pineiro, D., Wang, Y., Krogsrud, S.K., et al., 2021. Individual variations in “brain age” relate to early-life factors more than to longitudinal brain change. *Elife* 10.
- Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci. Biobehav. Rev.* 57, 328–349.
- Wu, Z., Peng, Y., Hong, M., Zhang, Y., 2021. Gray matter deterioration pattern during Alzheimer's disease progression: a regions-of-interest based surface morphometry study. *Front. Aging Neurosci.* 13, 593898.
- van Wynsberghe, A., 2021. Sustainable AI: AI for sustainability and the sustainability of AI. *AI Ethics*.
- Xifra-Porras, A., Ghosh, A., Mitsis, G.D., Boudrias, M.-H., 2021. Estimating brain age from structural MRI and MEG data: insights from dimensionality reduction techniques. *Neuroimage* 231, 117822.
- Zhao, Y., Klein, A., Castellanos, F.X., Milham, M.P., 2019. Brain age prediction: cortical and subcortical shape covariation in the developing human brain. *Neuroimage* 202, 116149.
- Zuo, X.-N., Anderson, J.S., Bellec, P., et al., 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci. Data* 1, 140049.

**4 A systematic comparison of VBM pipelines and their application to age prediction.** Antonopoulos, G., More, S., Raimondo, F., Eickhoff, S.B., Hoffstaedter, F., and Patil, K.R., NeuroImage, 120292 (2023)

**Authorship contribution statement**

**Georgios Antonopoulos:** Formal analysis, Software, Validation, Visualization, Writing – original draft. **Shammi More (Doctoral researcher):** 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 (Corresponding author):** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.



# 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 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.

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

(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 ( $48.1 \pm 17.2$  years, 630 female). CamCAN (Taylor et al., 2017; Shafto et al., 2014):  $n = 634$  aging individuals without serious psychiatric conditions or cognitive impairment ( $54.8 \pm 18.4$  years, 320 female). IXI (<https://brain-development.org/ixi-dataset/>): multisite sample of  $n = 582$  normal and healthy subjects ( $49.4 \pm 16.7$  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). CAT provides a complete VBM pipeline including denoising with spatial-adaptive nonlocal means, bias-correction, skull-stripping, and linear and nonlinear spatial registration. Images are segmented by an adaptive maximum a-posteriori approach (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

**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

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  $43 > \text{Idiff} > 29$ . fMRIPrep-FSL and FSLVBM, both with data-template, had the highest  $\text{Idiff} = 45$ , followed by the two ANTs pipelines ( $\text{Idiff} = 43$ ). The two CAT pipelines had the lowest mean  $\text{Idiff}$  values, with the data-template pipeline being the lowest. FSLVBM with data-template had the highest mean  $\text{Idiff}$ . Pipelines using FSL for registration and modulation, with a general template, had a mean  $\text{Idiff} = 33.7$ . The same pipelines with a data-template showed mean  $\text{Idiff} = 37.7$ .

ANTs-FSL and fMRIPrep-FSL, when both using a general template had  $\text{Idiff} = 35$  and when using a data-template  $\text{Idiff} = 34$ . Finally, ANTs and ANTs-FSL, which differ in registration (and modulation), had  $\text{Idiff} = 29$  when both used general templates and  $\text{Idiff} = 30$  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.

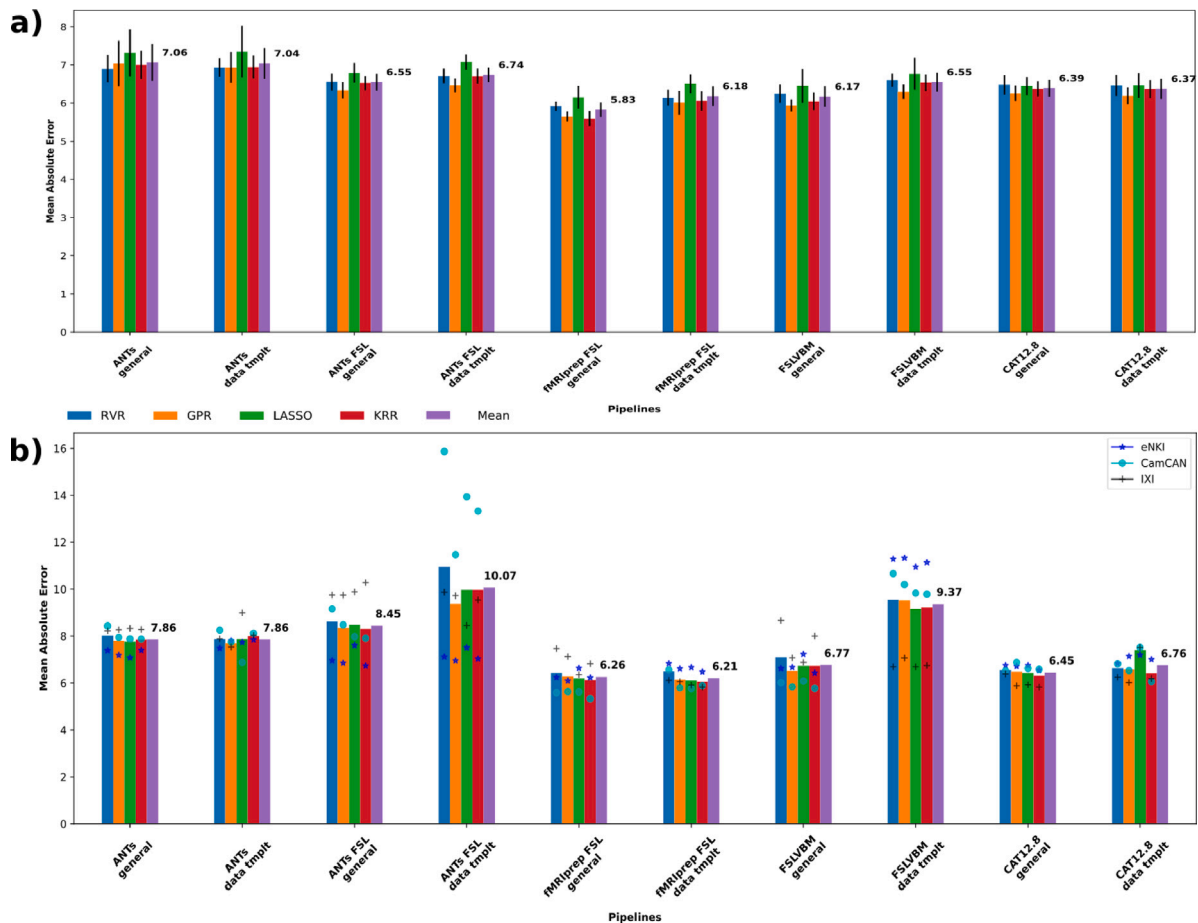
Regionwise similarity between ANTs and ANTs-FSL that differed only in **registration** (and therefore in modulation) using the general template was moderate to low, average  $r = 0.51$ . When using data-specific templates, the similarity was higher for all data ( $0.58$ ) but also for each of the three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



**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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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).

**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

### 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



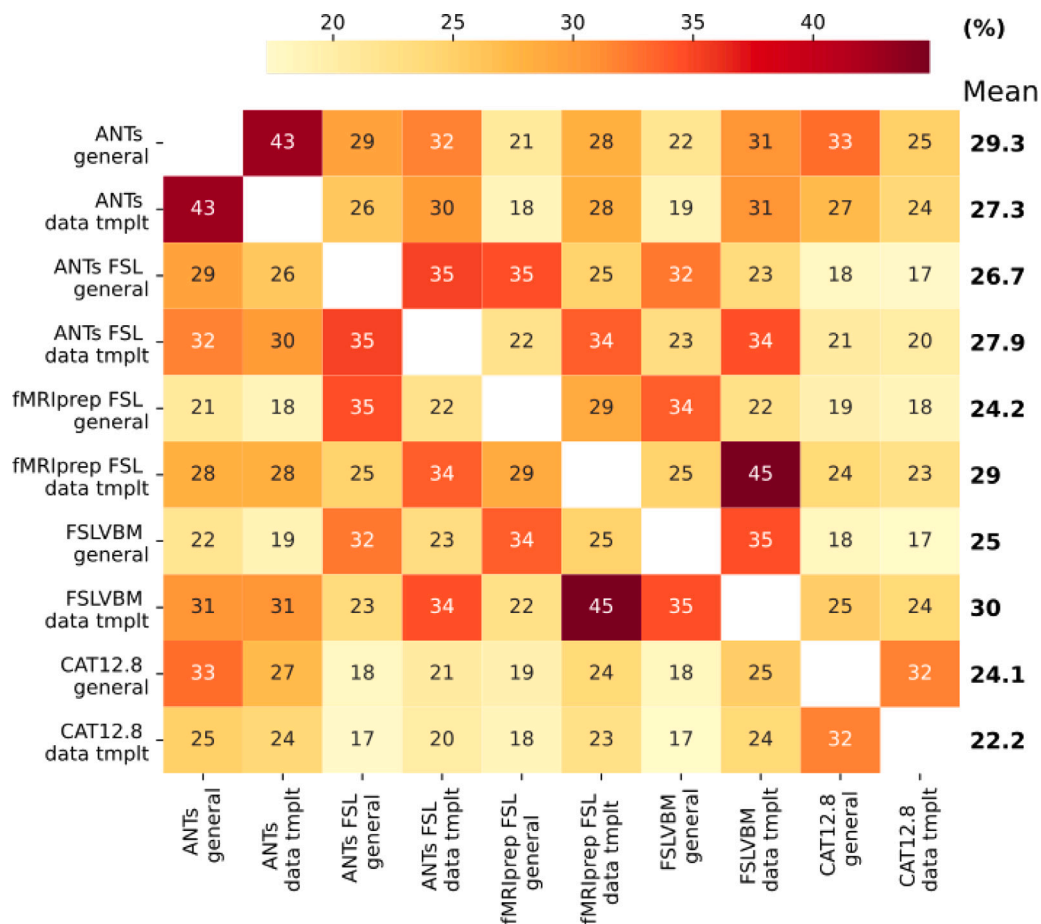


Fig. 2. Identification performance in terms of differential identifiability. We used Pearson's coefficient to calculate similarity between subjects. The highest mean Idiff was found for FSLVBM data-template followed by ANTs general template. The two CAT pipelines showed the lowest mean Idiff values.

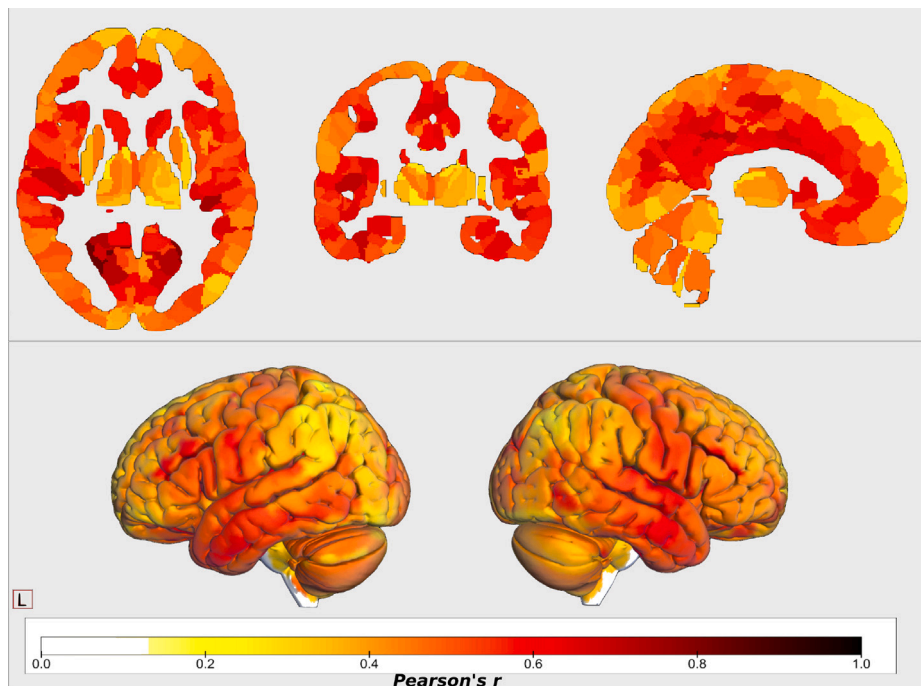
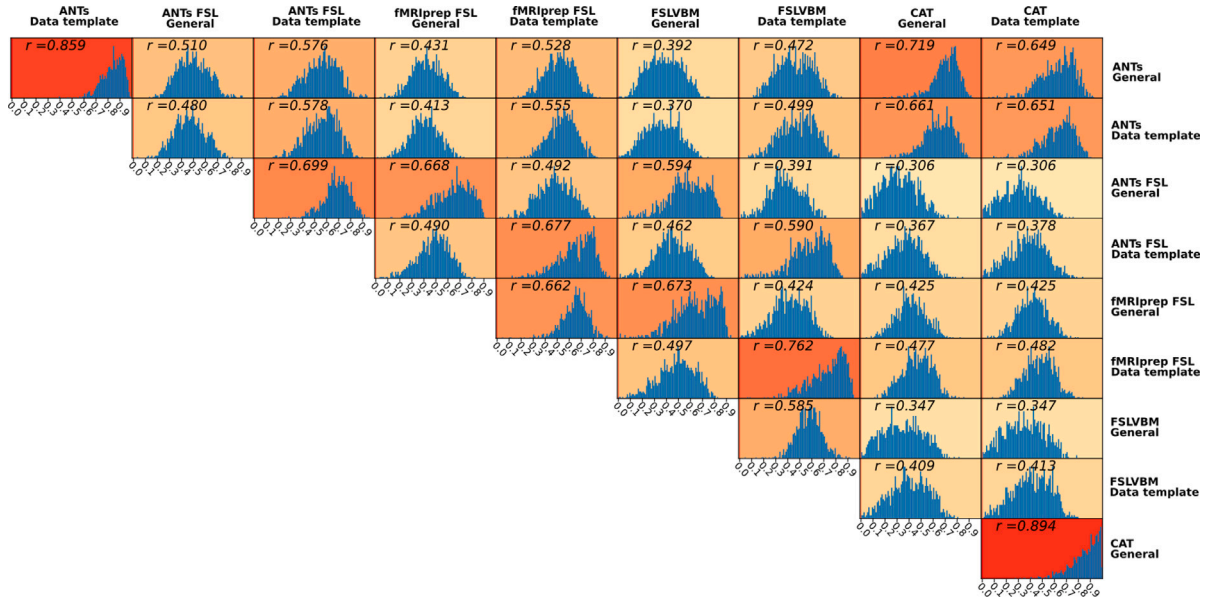


Fig. 3. Median values of regional correlations calculated across subjects of all pairwise combinations of pipelines. The frontal lobe, subcortical regions and cerebellum showed lower similarity. First, correlations between regional GMVs across subjects were calculated for each pipeline pair. The median 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 ( $p < 0.05$ ).

**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

(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).

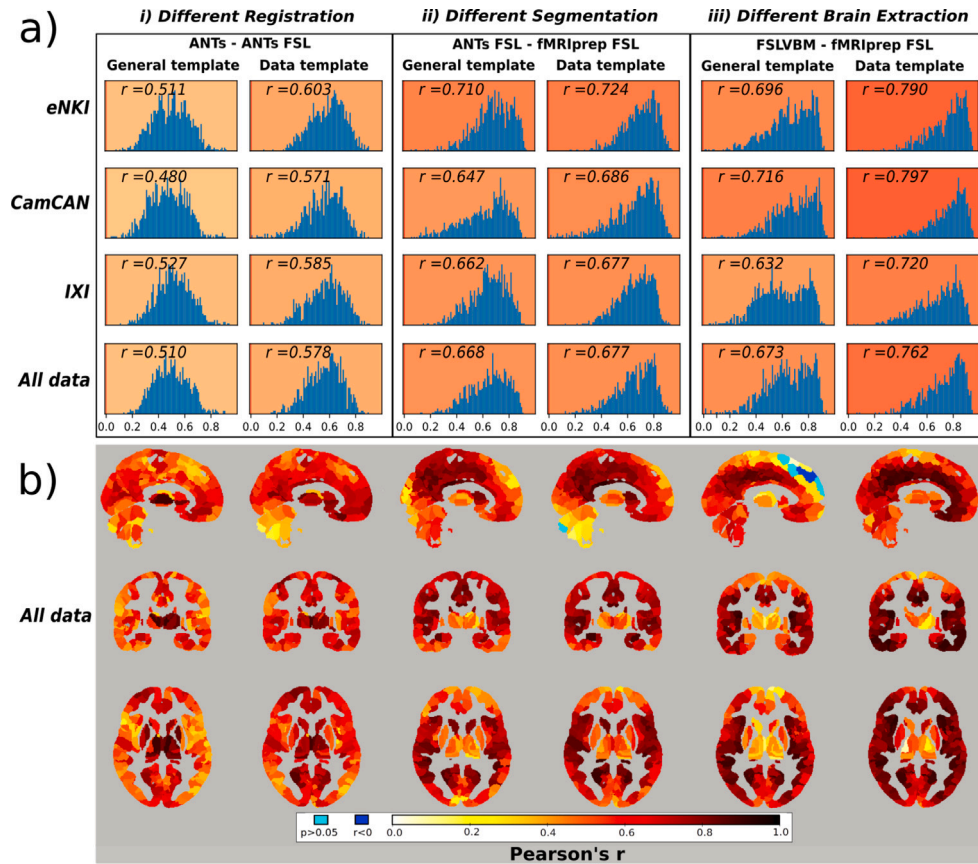
### 3.3.3. Effect of parcel size

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



**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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.

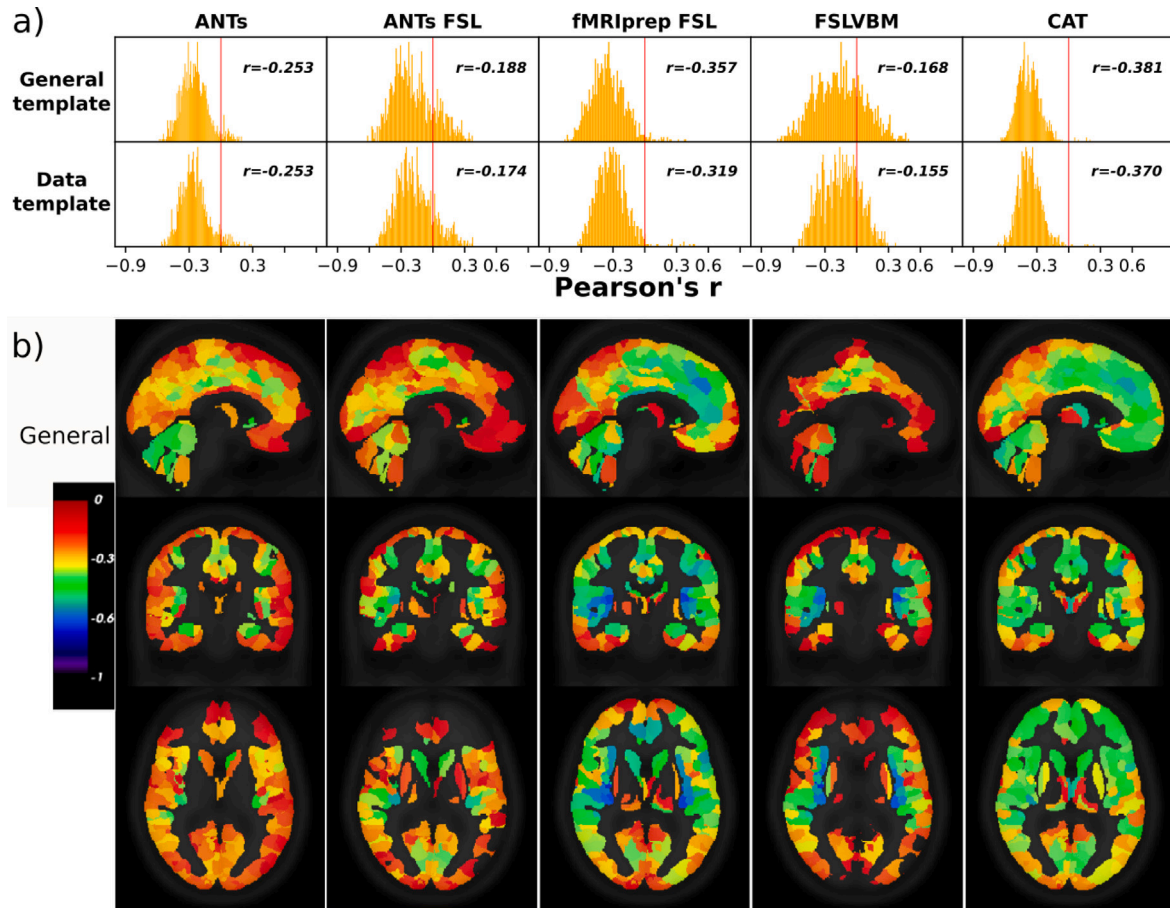
Overall, we made the following observations based on analysis of the GMV estimates from different perspectives. The differences in individuals' brain-age predictions confirmed that different VBM pipelines produce different GMVs (Fig. 1, Tables S.2 & S.3). The systematic differences between the pipelines 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

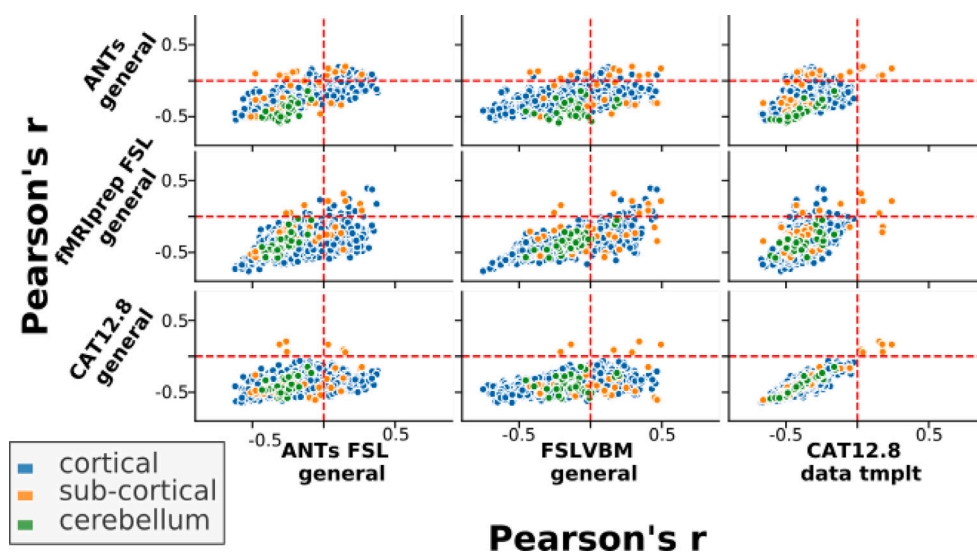
to what has been previously reported for voxel-level analysis and using CAT and FSL-based pipelines (Zhou et al., 2022). Our age-prediction accuracy for CAT and fMRIPrep-FSL are comparable to previous reports, considering our dataset size and the wide age range (Eickhoff et al., 2021; Cole et al., 2017a). 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





**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.

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; Kooops 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.

## 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/vb\\_m\\_comparison\\_codes](https://jugit.fz-juelich.de/g.antonopoulos/vb_m_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

## Data availability

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

## 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.

## Appendix A. Supplementary data

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

## References

- Amico, Enrico, Goñi, Joaquín, 2018. The quest for identifiability in human functional connectomes. *Sci. Rep.* 8 (1), 8254.
- Ashburner, John, 2007. A fast diffeomorphic image registration algorithm. *NeuroImage* 38 (1), 95–113.
- Ashburner, John, Friston, Karl J., 2000. Voxel-based morphometry—The methods. *NeuroImage* 11 (6), 805–821.
- Ashburner, John, Friston, Karl J., 2011. Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. *NeuroImage* 55 (3), 954–967.
- Avants, Brian B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12 (1), 26–41.
- 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.
- 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.
- Avants, Brian B., Yushkevich, Paul, Pluta, John, Minkoff, David, Korczynski, 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.
- 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.
- Brewer, James B., 2009. Fully-automated volumetric MRI with normative ranges: Translation to clinical practice. *Behav. Neurol.* 21 (1–2), 21–28.
- 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.
- 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.
- Cole, James H., Annus, Tiina, Wilson, Liam R., Remtulla, Ridhaa, Hong, Young T., Fryer, Tim D., Acosta-Cabrero, 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.
- Cole, James H., Franke, Katja, 2017. Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends Neurosci.* 40 (12), 681–690.
- 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.



- Cole, James H., Ritchie, S.J., Bastin, M.E., Valdés Hernández, M.C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., Wray, N.R., Redmond, P., Marioni, R.E., Starr, J.M., Cox, S.R., Wardlaw, J.M., Sharp, D.J., Deary, I.J., 2018. Brain age predicts mortality. *Mol. Psychiatry* 23 (5), 1385–1392.
- 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.
- Dale, Anders M., Fischl, Bruce, Sereno, Martin I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage* 9 (2), 179–194.
- Dinsdale, Nicola K., Bluemke, Emma, Smith, Stephen M., Arya, Zobair, Vidaurre, Diego, Jenkinson, Mark, Namburete, Ana L.L., 2021. Learning patterns of the ageing brain in MRI using deep convolutional networks. *NeuroImage* 224, 117401.
- 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.
- 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, Jesse, Durmez, Joke, Poldrack, Russell A., Gorgolewski, Krzysztof J., 2019. Fmrip: a robust preprocessing pipeline for functional MRI. *Nature Methods* 16 (1), 111–116.
- 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 connective architecture. *Cerebral Cortex* 26 (8), 3508–3526.
- Farokhan, 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.
- Farokhan, 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.
- 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.
- 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.
- Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., Collins, D.L., 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.
- 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.
- 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.
- 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.
- Habes, M., Janowitz, D., Erus, G., Toledo, J.B., Resnick, S.M., Doshi, J., Van der Auwera, S., Wittfeld, K., Hegenscheid, K., Hosten, N., Biffar, R., Homuth, G., Völzke, H., Grabe, H.J., 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.
- 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, Eileanor B., Gregory, Sarah, Johnson, Hans J., Durr, Alexandra, Leavitt, Blair R., Roos, Raymond 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.
- Jonsson, B.A., Björnsdóttir, G., Thorgeirsson, T.E., Ellingsen, L.M., Walters, G. Bragi, Gudbjartsson, D.F., Stefansson, H., Stefansson, K., Ulfarsson, M.O., 2019. Brain age prediction using deep learning uncovers associated sequence variants. *Nature Commun.* 10 (1), 5409.
- 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.
- 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.
- Khagi, Bijan, 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.
- 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.
- Koutsouleris, Nikolaos, Davatzikos, Christos, Borgwardt, Stefan, Gaser, Christian, Botlender, 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.
- 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.
- 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.
- Matsuda, Hiroshi, 2013. Voxel-based morphometry of brain MRI in normal aging and Alzheimer's Disease. *Aging Dis.* 4 (1), 29.
- 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.
- More, Shammie, 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, Shammie, 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, Mladenici, 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,

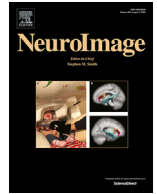
- Thomsen, Dawn, Biswal, Bharat, Coffey, Barbara, Hoptman, Matthew J., Javitt, Daniel C., Pomara, Nunzio, Sidtis, John J., Koplewicz, Harold S., Castellanos, Francisco Xavier, Leventhal, Bennett L., Milham, Michael P., 2012. The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Front. Neurosci.* 6, 152.
- 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.
- 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.
- Poldrack, Russell A., Huckins, Grace, Varoquaux, Gael, 2020. Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry* 77 (5), 534–540.
- 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., Frapp, 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.
- 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.
- 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, Croyley, 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-Eizaguirre, 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, Di az Caneja, Covadonga, Jiang, Wenhao, Calhoun, Vince, Ehrlich, Stefan, Yang, Kun, Casella, Nicola G., Takayanagi, Yoichiro, Sawa, Akira, Tomyshew, 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, Martínez-Heras, Eloy, Solana, Elisabeth, Llufríu, Sara, Jahanshad, Neda, Thompson, Paul, Turner, Jessica, van Erp, Theo, Glahn, David, Pearson, 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.
- 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.
- Rajapakse, J.C., Giedd, J.N., Rapoport, J.L., 1997. Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Trans. Med. Imaging* 16 (2), 176–186.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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 Bostel, 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.
- 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.
- 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.
- 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.
- Variakuti, Deepthi P., Genon, Sarah, Sotiras, Aristides, 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.
- 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, I.C., McGuire, P.K., Poline, J.-B., Travers, J.M., Murray, R.M., Frith, C.D., Frackowiak, R.S.J., Friston, K.J., 1995. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *NeuroImage* 2 (4), 244–252.
- 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.
- 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.
- 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.



**5 Reporting Details of Neuroimaging Studies on Individual Traits Predictions: A Literature Survey.** Yeung, A.W.K., More, S., Wu, J., Eickhoff, S.B., NeuroImage, 119275 (2022)

**Authorship contribution statement**

**Andy Wai Kan Yeung (Corresponding author):** Conceptualization, Methodology, Writing – original draft. **Shammi More (Doctoral researcher):** Data curation, Writing –review & editing. **Jianxiao Wu:** Data curation, Writing –review & editing. **Simon B. Eickhoff (Corresponding author):** Conceptualization, Methodology, Writing –review & editing.



# Reporting details of neuroimaging studies on individual traits prediction: A literature survey

Andy Wai Kan Yeung<sup>a,\*</sup>, Shammi More<sup>b,c</sup>, Jianxiao Wu<sup>b,c</sup>, Simon B. Eickhoff<sup>b,c,\*</sup>

<sup>a</sup> Oral and Maxillofacial Radiology, Applied Oral Sciences and Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong, China

<sup>b</sup> Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany

<sup>c</sup> Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

## ARTICLE INFO

### Keywords:

Individual trait  
Prediction  
Neuroimaging  
Predictive modeling  
Survey

## ABSTRACT

Using machine-learning tools to predict individual phenotypes from neuroimaging data is one of the most promising and hence dynamic fields in systems neuroscience. Here, we perform a literature survey of the rapidly work on phenotype prediction in healthy subjects or general population to sketch out the current state and ongoing developments in terms of data, analysis methods and reporting. Excluding papers on age-prediction and clinical applications, which form a distinct literature, we identified a total 108 papers published since 2007. In these, memory, fluid intelligence and attention were most common phenotypes to be predicted, which resonates with the observation that roughly a quarter of the papers used data from the Human Connectome Project, even though another half recruited their own cohort. Sample size (in terms of training and external test sets) and prediction accuracy (from internal and external validation respectively) did not show significant temporal trends. Prediction accuracy was negatively correlated with sample size of the training set, but not the external test set. While known to be optimistic, leave-one-out cross-validation (LOO CV) was the prevalent strategy for model validation ( $n = 48$ ). Meanwhile, 27 studies used external validation with external test set. Both numbers showed no significant temporal trends. The most popular learning algorithm was connectome-based predictive modeling introduced by the Yale team. Other common learning algorithms were linear regression, relevance vector regression (RVR), support vector regression (SVR), least absolute shrinkage and selection operator (LASSO), and elastic net. Meanwhile, the amount of data from self-recruiting studies (but not studies using open, shared dataset) was positively correlated with internal validation prediction accuracy. At the same time, self-recruiting studies also reported a significantly higher internal validation prediction accuracy than those using open, shared datasets. Data type and participant age did not significantly influence prediction accuracy. Confound control also did not influence prediction accuracy after adjusted for other factors. To conclude, most of the current literature is probably quite optimistic with internal validation using LOO CV. More efforts should be made to encourage the use of external validation with external test sets to further improve generalizability of the models.

## Introduction

Individual traits prediction (e.g. cognition abilities, personality traits, emotional feeling, and motor performance) using neuroimaging data is an upcoming hotspot in cognitive neuroscience (Shen et al., 2017; Sui et al., 2020). The term prediction refers to the ability to predict outcomes successfully in data sets other than the original one used to construct the model (Poldrack et al., 2020). It is better for translational or prediction purposes than the traditional univariate brain mapping analysis, as the latter focused on within-sample fit of correlational relationships that tends to be overfitting and not generalizable (Poldrack et al.,

2020). The overall scheme usually begins with collecting structural or functional (resting-state or task-induced) neuroimaging data and personal trait measures from a large sample. Then the neuroimaging data should be preprocessed and entered into a machine-learning model. The model will be trained to find out the link between the neuroimaging data (brain features) and the personal traits. Finally, the trained model can be generalized to predict the traits in a new sample. Its accuracy can be computed by comparing with the ground truth (reality) (Eickhoff and Langner, 2019). In short, there are four stages: model building, internal validation, external validation, and generalizability and transposability (Bzdok and Ioannidis, 2019). There are many ap-

\* Corresponding authors at: Oral and Maxillofacial Radiology, Applied Oral Sciences and Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong, China (Andy Wai Kan Yeung); Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany (Simon B. Eickhoff)

E-mail addresses: [ndyeung@hku.hk](mailto:ndyeung@hku.hk) (A.W.K. Yeung), [Simon.Eickhoff@uni-duesseldorf.de](mailto:Simon.Eickhoff@uni-duesseldorf.de) (S.B. Eickhoff).

<https://doi.org/10.1016/j.neuroimage.2022.119275>.

Received 28 February 2022; Received in revised form 27 April 2022; Accepted 29 April 2022

Available online 2 May 2022.

1053-8119/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

proaches to individual traits prediction. One famous approach is called the connectome-based predictive modeling (CPM) approach, developed by (Finn et al., 2015), the term CPM established by (Rosenberg et al., 2016), its protocol published by (Shen et al., 2017), codes deposited at <https://github.com/YaleMRRC/CPM>. Moreover, (Finn et al., 2015) introduced the now widely used Shen 268 atlas, which was produced based on the parcellation method and the 100-, 200-, and 300-node atlases introduced by (Shen et al., 2013). In that study by Rosenberg et al., it was reported that functional connectivity derived from task-induced functional magnetic resonance imaging (fMRI) could be used to train a model that predicted a previously unseen individual's performance in sustained attention and even symptoms of attention deficit hyperactivity disorder based on his/her resting-state fMRI signals (Rosenberg et al., 2016). The predictive modeling field in neuroscience has seen a rapid growth and accumulated many papers since then.

Of course, there are a number of factors that affect the validity or generalizability of predictive models in neuroimaging, spanning from sample size, processing, features, learning, to validation. To begin with, it was recommended that a dataset of over 100 individuals should be used for predictive modeling (Scheinost et al., 2019). Data from (He et al., 2020) even suggested that 500–1000 subjects should be the minimum. Small sample sizes would lead to underestimated errors and vibration effects, meaning that methodological choices could have a drastic impact on the analysis outcome based on few samples (Varoquaux, 2018). Subject recruitment and financial constraints could be potential issues, and might be circumvented by the use of large, open, shared datasets as training or test set. During data processing, potential confounding factors should be accounted for, such as physiological and head motion artifacts (Murphy et al., 2013). At the stage of features input, one needs to consider what data to be entered. For instance, for a model that predicts behavior based on brain connectivity data, connectomes from multiple sources could improve the prediction accuracy compared to a single connectome (Gao et al., 2019). Finally, external validation is the best practice, meaning testing the model with an independently collected (external) data set (Scheinost et al., 2019). Out-of-sample generalization and later cross-validation (CV) is less ideal, as the portion of the sample taken out from the same dataset will inevitably share similar subject and imaging features with the training set and create bias. Since generally it is relatively difficult to obtain a separate test set, doing a CV has been a popular approach, meaning that the whole dataset is divided into subsets that train and test the model respectively. CV is generally fine, but it should be noted that CV in small samples may render the models too optimistic (Whelan and Garavan, 2014).

In this work, we performed an updated general literature survey on the study design and analytic pipeline of the individual traits prediction among healthy individuals or general population (not purely clinical), and aimed to evaluate the published studies on individual traits prediction based on regression, to reveal if their generalizability could be undermined by the caveats mentioned above.

## Methods

### Literature search strategy

PubMed and Web of Science Core Collection online databases were queried on 16 December 2021 with the following search string: (((("machine learning") OR ("predict\* model\*") OR ("support vector machine\*") OR ("LASSO\*") OR ("elastic net\*") OR ("random forest\*") OR ("cross validat\*") OR ("artificial intelligen\*")) AND ((brain behavior\*) OR (brain behavior\*) OR (neuromarker\*) OR (brain biomarker\*) OR ("individual difference\*")))). The search covered "all fields" for PubMed and "title/abstract/keywords" for Web of Science. We also performed reference tracing from the yielded publications and previous review articles. A total of 7153 publications were identified after removing duplicates. The full text of these them were inspected and publications were excluded due to the following reasons: irrelevant (e.g. within-sample cor-

relation instead of predictive modeling;  $n = 6018$ ), classification instead of regression (e.g. sex classification;  $n = 692$ ), involved patients only ( $n = 154$ ), review/opinion paper ( $n = 121$ ), method papers ( $n = 17$ ), age prediction instead of individual traits ( $n = 43$ ), conference abstract without full text ( $n = 0$ ), and unspecific phenotype ( $n = 0$ ). In the end, 108 articles entered the survey (Supplementary Table 1). For completeness, a list of excluded papers could be found in Supplementary Table 2.

To assess the reporting details and identify patterns/trends among these papers, we examined the content of them carefully. The surveyed contents involved sampling, processing strategy, feature selection, learning algorithm, and validation. Sample size is a critical aspect of the papers, as smaller samples may be underpowered and overfit the models, and hence producing false positives (Varoquaux et al., 2017). Meanwhile, papers dealing with large open-source neuroimaging datasets should report the dataset details well enough, as each dataset has its unique demographic factor, imaging and behavioral measures (Horien et al., 2021). For processing, accounting for confounds such as head motion is an important step in modeling, as their presence may make the model less meaningful (Rao et al., 2017). Other details of processing such as dimensionality reduction, feature selection, learning algorithm, hyperparameter tuning, and validation strategy were also evaluated and recorded as these are important for fellow researchers to replicate their results. Finally, prediction accuracy was noted to evaluate the model performances. Because of these rationales, the parameters recorded for each study were listed in the following paragraph.

### Parameters recorded

The following parameters were recorded for each study: sample size (training set and test set), data source, type of subject (minor vs adult), amount of data for each subject (number of volumes), input data (e.g. what kind of connectome and matrix size), data type (task, rest, naturalistic, vs structural), target phenotype (e.g. intelligence), processing strategy, reference to the Yale approach (connectome-based predictive modeling, c.f. (Finn et al., 2015; Rosenberg et al., 2016; Shen et al., 2017)), brain atlas referred to, confounding variables accounted for (e.g. head motion), dimensionality reduction if relevant, feature selection, learning algorithm, hyperparameter tuning, validation strategy (e.g. external validation or CV), and prediction accuracy (from internal and external validation, respectively). The temporal trends of the statistics were tested across studies if they were continuous variables (e.g. prediction accuracy), and across years if they were categorical (e.g. ratio of studies using external test set). Additional analyses were performed for fMRI and structural MRI (sMRI) studies separately.

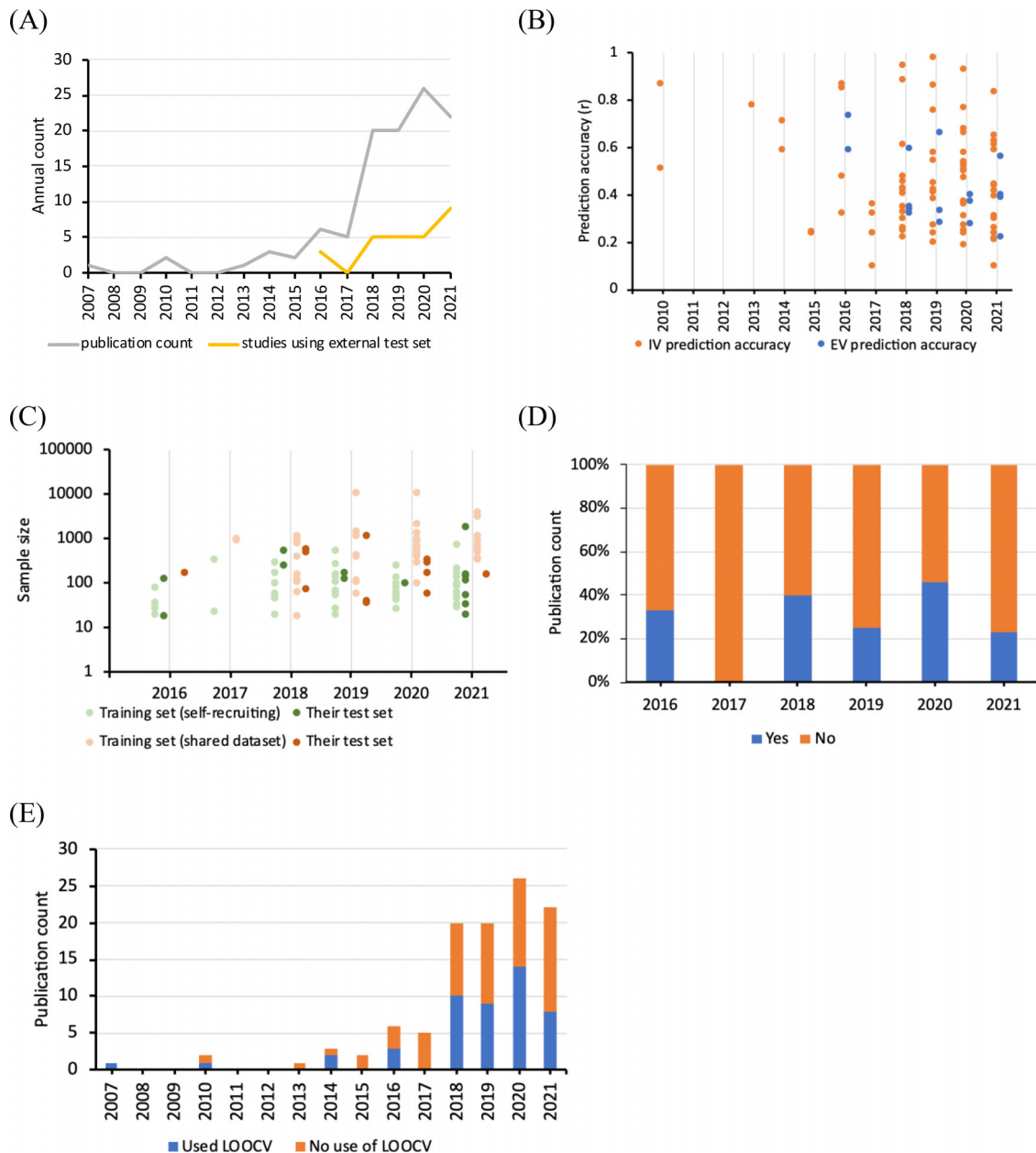
## Results

### General bibliographic information

The annual publication count showed a sharp increase in year 2018 (Fig. 1A). Prior to 2018, there were fewer than 6 papers published per year.

### Prediction accuracy

In brief, 81 studies reported Pearson  $r$  as the prediction accuracy value from internal validation, and 16 studies reported so from external validation. The accuracy from internal validation ranged from 0.098 to 0.978, whereas the accuracy from external validation ranged from 0.220 to 0.736. Though the prediction accuracy from either validation method seemed to show a slight decreasing trend by year (Fig. 1B), no significant linear correlation was observed (Pearson correlation test, internal validation:  $n = 81$ ,  $r = -0.201$ ,  $p = 0.071$ ; external validation:  $n = 16$ ,  $r = -0.482$ ,  $p = 0.059$ ). For the studies that did not report any Pearson  $r$  as the prediction accuracy, Spearman  $\rho$  was the most popular metric



**Fig. 1.** Graphical summary of the surveyed articles. (A) Annual publication count and studies using external test set. (B) Prediction accuracy of models using internal validation (IV) and external validation (EV) respectively. (C) Sample sizes of training set and external test set. Studies published in 2015 or before were not plotted as they did not recruit external test set. (D) Proportion of studies mentioning that they followed the Yale approach (“connectome-based predictive modeling” [CPM], e.g. from (Finn et al., 2015; Rosenberg et al., 2016; Shen et al., 2017)). (E) The use of leave-one-out cross-validation (LOOCV).

( $n = 11$ ). Other reported metrics included standardized mean squared error, mean absolute error (MAE), root mean square error (RMSE), prediction  $R^2$ , and adjusted  $R^2$ .

#### Sample size

Before year 2016, the few surveyed studies only recruited subjects for their training set, without external test set or involvement of open, shared dataset. The mean sample size of their training set was 64, 25.5, 40, 49, and 185.5 for year 2007, 2010, 2013, 2014, and 2015, respectively. Fig. 1C illustrated the sample size since year 2016. For training set, self-recruiting studies had a mean sample size of 108 during the period of 2016–2021, whereas studies using open, shared datasets had a much larger mean sample size of 1140. For test set, however, the former group and the latter group had a similar mean sample size (251 vs 278).

The sample size of self-recruiting studies did not show significant linear correlation with year (Pearson correlation test, training set:  $n = 52$ ,  $r = 0.078$ ,  $p = 0.581$ ; test set:  $n = 14$ ,  $r = 0.158$ ,  $p = 0.590$ ). The same held true for studies using open, shared datasets (training set:  $n = 45$ ,  $r = 0.106$ ,  $p = 0.489$ ; test set:  $n = 12$ ,  $r = -0.093$ ,  $p = 0.773$ ).

#### Selection of data source and data type

In terms of data source of the training set, 61 studies recruited their own subjects, whereas the Human Connectome Project (HCP) was used by 21 studies (Table 1). HCP may refer to various versions of the HCP dataset, such as “unrelated 100”, S500, S900, and S1200. Readers should be aware that the more recent datasets (e.g. S1200) not only had a larger sample size, but also contained updated data on family structures of the subjects (e.g., relationships as twins or non-twin siblings, but excluding

**Table 1**  
Data sources of the training set.

Data source	Number of studies
Recruited subjects	61
HCP S1200	6
HCP S900	6
HCP S500	4
ABCD	4
PNC	4
HCP (unclear version)	3
SLIM	3
Previously trained model	2
HCP “unrelated 100”	2
UESTC	2
UKB	2

ABCD, Adolescent Brain Cognitive Development Study. HCP, Human Connectome Project. PNC, Philadelphia Neurodevelopmental Cohort. SLIM, SLIM dataset from Liu et al. (2017). UESTC, University of Electronic Science and Technology of China. UKB, UK Biobank. The data sources below were each used once and hence not listed in the table: ABIDE-II (Autism Brain Imaging Data Exchange), ADNI-2 (Alzheimer’s Disease Neuroimaging Initiative), ADNI-GO, AHAB-2 (Adult Health and Behavior project—Phase 2), ATR dataset from Yamashita et al. (2015), BBP (Behavioral Brain Research Project of Chinese Personality), BCAS (Brain and Cognition Aging Study), CamCAN (Cambridge centre for Ageing and Neuroscience), CBDC (Cognition and Brain Development in Children), DIAMOND (Dimensions of Affect, Mood, and Neural Circuitry Underlying Distress Study), Duke Neurogenetics Study, GUSTO (Growing Up in Singapore Towards healthy Outcomes), IMAGEN dataset from Schumann et al. (2010), NKI-RS (Nathan Kline Institute Rockland Sample), OASIS-3 (Open Access Series of Imaging Studies), PING (Pediatric Imaging, Neurocognition, and Genetics), PIP (Pittsburgh Imaging Project), TTC (Tokyo TEEN Cohort Study), UNC Early Brain Development Study.

**Table 2**  
Frequency of common types of input data.

Input data	Frequency (n)
Resting state functional connectivity (RSFC) connectome	32
Both task-induced FC and RSFC connectome	15
Task-induced FC connectome	10

For other less common types of input data, please refer to Supplementary Table 1 for the details of each study.

birth order). Therefore, the umbrella term HCP did not necessarily imply an identical sample used across the studies. Among the 21 studies using HCP data, S1200 was the most popular dataset (Table 1). Meanwhile, different target phenotypes were investigated, with some of the recurring ones being fluid intelligence/intelligence quotient ( $n = 15$ ), attention ( $n = 12$ ), and memory ( $n = 11$ ).

Regarding input data, resting state functional connectivity (RSFC) connectome was much more common than task-induced FC connectome or the combination of both (Table 2). The brain atlas used for connectome data were also diverse, with the canonical Shen 268 atlas (Finn et al., 2015; Shen et al., 2013) being most prevalent (Table 3).

#### Dealing with confounding factors during data processing

Over half of the studies (61 out of 108) did not control for potential confounding factors such as age, sex, head motion. Studies controlled for them mainly entered them as regressors in the regression models.

#### Varied feature selection and learning

Thirty-two papers (32.3%) published since 2016 followed the Yale approach pioneered by Finn et al. (2015) mentioned in the Introduction, which achieved a brain-behavior prediction by means of an approach called connectome-based predictive modeling (CPM) (Fig. 1D). In year 2020, almost half of the studies followed this approach. It basically in-

volved a linear regression and the feature selection method was typically choosing FCs with significant correlation (e.g.  $p < 0.01$ ) with the predicted measure [and then the sums of selected positive or negative edges (the summary measure), is used as input features for linear regression]. Other papers had very diverse feature selection methods, with two recurring methods including feature selection from regions-of-interest (ROIs,  $n = 4$ ) and principal component analysis (PCA,  $n = 2$ ). Some common learning algorithm used by these non-CPM papers were multiple linear regression, relevance vector regression (RVR), support vector regression (SVR), partial least squares regression (PLSR), least absolute shrinkage and selection operator (LASSO), and elastic net. Most studies did not require hyperparameters tuning, and nested k-fold CV [in descending order of frequency: ( $n = 8$ ) 10-fold, ( $n = 6$ ) 3-fold, ( $n = 3$ ) 5-fold, and ( $n = 2$ ) 20-fold] was the predominant choice. See Supplementary Table 1 for details.

#### Diversity of validation

For validation strategy, 48 studies (44.4%) involved leave-one-out cross-validation (LOO CV). Twenty-four of these LOO CV papers mentioned they used the Yale approach, suggesting a dependency of this CV strategy on the CPM modeling approach. The annual ratio of studies using LOO CV fluctuated around 50% and showed no obvious trend against year (Pearson correlation test on period 2013–2021,  $n = 9$ ,  $r = 0.341$ ,  $p = 0.369$ ; Fig. 1E). Ten-fold CV and 4-fold CV were involved in 19 and 10 studies respectively. Meanwhile, 27 studies involved external test sets (26 were cross-dataset whereas one was cross-site) and they were published in 2016–2021 (Fig. 1A). The proportion of studies using an external test set has remained largely constant across years with no discernible trend (Pearson correlation test,  $n = 6$ ,  $r = 0.037$ ,  $p = 0.945$ ). Readers should be aware of the few data points used in these two tests.

#### Potential influencing factors on prediction accuracy

Table 4 shows that sample size could influence the prediction accuracy. Precisely, the smaller the sample size of the training set, the higher the internal validation prediction accuracy was found ( $n = 81$ ,  $r = -0.265$ ,  $p = 0.017$ ). On the contrary, the sample size of the external test set did not show significant correlation with external validation prediction accuracy. Meanwhile, the amount of data from self-recruiting studies (but not studies using open, shared dataset) was positively correlated with internal validation prediction accuracy ( $n = 30$ ,  $r = 0.651$ ,  $p < 0.001$ ). At the same time, self-recruiting studies also reported a significantly higher internal validation prediction accuracy than those using open, shared datasets (Mean  $\pm$  SD, self-recruiting:  $n = 46$ ,  $0.509 \pm 0.229$ , shared dataset:  $n = 35$ ,  $0.386 \pm 0.190$ ,  $p = 0.012$ ). Besides, internal validation reported higher accuracy than external validation within studies that used both types of validation. Meanwhile, studies using models that were uncontrolled for confounds reported a significantly higher internal validation prediction accuracy than those using models that were controlled for confounds (Mean  $\pm$  SD, controlled studies:  $n = 33$ ,  $0.395 \pm 0.197$ ; uncontrolled studies:  $n = 48$ ,  $0.498 \pm 0.228$ ,  $p = 0.038$ ). Data type and participant age did not significantly influence prediction accuracy. Table 4 shows the detailed results of the statistical tests. It should be noted that some studies may provide more than one data point whereas some studies may have missing data for the statistical tests, and hence the  $n$  reported may not correspond to the number of studies involved. Readers should refer to Supplementary Table 3 for the data used.

When the significant factors were considered together by partial correlation tests, it was found that training set sample size remained significant after adjusted for confound control, but became insignificant after considering participant source or amount of data. Meanwhile, participant source remained significant after adjusted for confound control, but became insignificant after considering training set sample size or amount of data. In turn, amount of data for studies recruiting subjects



**Table 3**

Brain atlases referred by studies using connectome data.

Brain atlas	No. of nodes	Coverage (whole brain, cortex only, cerebellum only, etc.)	Functionally defined vs anatomically defined	Number of studies
Shen 268 atlas, see (Finn et al., 2015; Shen et al., 2013)	268	Whole brain	Functionally defined	29
(Power et al., 2011)	264	Whole brain	Functionally defined	8
(Fan et al., 2016)	246	Whole brain	Anatomically defined	5
Independent component analysis (ICA) components	Variable	Variable	Variable	5
(Dosenbach et al., 2010)	160	Whole brain	Functionally defined	4
(Tzourio-Mazoyer et al., 2002)	116	Cortex only	Anatomically defined	3
(Glasser et al., 2016)	360	Cortex only	Functionally and anatomically defined	3
(Gordon et al., 2016)	333	Cortex only	Functionally defined	3
(Schaefer et al., 2018)	100–1000 (400 version used in 2 studies)	Cortex only	Functionally defined	2
(Desikan et al., 2006)	68	Cortex only	Anatomically defined	2
(Gilmore et al., 2012)	78	Cortex only	Anatomically defined	1
(Diedrichsen, 2006)	28	Cerebellum	Anatomically defined	1
(Fischl et al., 2002)	37 (14 used in 1 study)	Whole brain	Anatomically defined	1
(Buckner et al., 2011)	7 or 17	Cerebellum	Functionally defined	1
(Feng et al., 2019)	52	Whole brain (neonatal)	Anatomically defined	1
(Yeo et al., 2011)	114 (39 used in 1 study)	Cortex only	Functionally defined	1
(Destrieux et al., 2010)	148	Cortex only	Anatomically defined	1

Some studies referred to multiple atlas and they were counted within the table.

**Table 4**

Influencing factors of prediction accuracy.

Factor	Test	Stat	P value
Sample size	Pearson correlation		
a. of external test set (prediction accuracy from external validation)		$n = 17, r = -0.302$ (i.e. larger test set, lower prediction accuracy)	0.239
b. of training set (prediction accuracy from internal validation)		$n = 81, r = -0.265$ (i.e. larger test set, lower prediction accuracy)	0.017
Amount of data (total number of volumes per individual)	Pearson correlation		
a. for studies recruiting subjects		$n = 30, r = 0.651$ (i.e. more data, higher prediction accuracy)	< 0.001
b. for studies using open, shared dataset		$n = 28, r = -0.095$ (i.e. less data, higher prediction accuracy)	0.629
Data type (task, rest, naturalistic, structural vs mixed)	One-way ANOVA	Mean $\pm$ SD Task ( $n = 20$ ): $0.510 \pm 0.218$ , rest ( $n = 36$ ): $0.421 \pm 0.166$ , structural ( $n = 18$ ): $0.410 \pm 0.270$ , mixed ( $n = 10$ ): $0.567 \pm 0.243$ (No study used naturalistic data)	0.129
Participant source (self-recruiting vs open, shared dataset)	T-test	Mean $\pm$ SD Self-recruiting ( $n = 46$ ): $0.509 \pm 0.229$ , shared dataset ( $n = 35$ ): $0.386 \pm 0.190$	0.012
Participant age (involved minor vs adult only)	T-test	Mean $\pm$ SD Involved minor ( $n = 16$ ): $0.476 \pm 0.233$ , adult only ( $n = 65$ ): $0.451 \pm 0.219$	0.688
Control for confounds (yes vs no)	T-test	Mean $\pm$ SD Yes ( $n = 33$ ): $0.395 \pm 0.197$ , no ( $n = 48$ ): $0.498 \pm 0.228$	0.038
Validation type (internal vs external)			
a. for studies that involved both types	Paired <i>t</i> -test	Mean $\pm$ SD ( $n = 13$ ) Internal: $0.536 \pm 0.242$ , external: $0.427 \pm 0.158$	0.014
b. across all studies	T-test	Mean $\pm$ SD Internal ( $n = 81$ ): $0.456 \pm 0.220$ , external ( $n = 16$ ): $0.426 \pm 0.153$	0.606

Unless otherwise specified, prediction accuracy referred to Pearson's correlation *r* value resulted from internal validation. Studies without reporting *r* value were omitted. It should be noted that some studies may provide more than one data point whereas some studies may have missing data for the statistical tests, and hence the *n* reported may not correspond to the number of studies involved. Readers should refer to Supplementary Table 3 for the data used.

remained significant after adjusted for training set sample size and confound control. On the contrary, confound control and validation type (for studies that involved both internal and external validation) were not significant after adjusted for other factors. Readers should refer to Supplementary Table 4 for the full results of the partial correlation tests.

Additional analyses for only fMRI studies have shown that, the amount of data from self-recruiting studies (but not studies using open, shared dataset) was positively correlated with internal validation prediction accuracy ( $n = 28, r = 0.654, p < 0.001$ ). Self-recruiting studies

also reported a significantly higher internal validation prediction accuracy than those using open, shared datasets (Mean  $\pm$  SD, self-recruiting:  $n = 31, 0.534 \pm 0.211$ , shared dataset:  $n = 27, 0.375 \pm 0.150, p = 0.002$ ). Internal validation reported higher accuracy than external validation within studies that used both types of validation ( $n = 10$ , Mean  $\pm$  SD, internal:  $0.524 \pm 0.255$ , external:  $0.408 \pm 0.146, p = 0.045$ ). Sample size did not correlate with prediction accuracy (Supplementary Table 5). No partial correlation test was conducted for this subset, as there were very little or no overlap between studies involving these significant factors.

Meanwhile, additional analyses for only sMRI studies have shown that none of the factors correlated with prediction accuracy (Supplementary Table 6).

## Discussion

Based on 108 neuroimaging studies on individual traits prediction published mainly in the late 2010s, it was found that sample size of the training set was negatively correlated with prediction accuracy from studies using internal validation. Meanwhile, amount of data of recruited subjects was positively correlated with internal validation prediction accuracy. Recurring target phenotypes were memory, attention, and intelligence. Half of the studies recruited their own subjects whereas HCP was the dominant open, shared dataset to be used. The most typical method for working with connectome data was “FCs with significant correlation (e.g.  $p < 0.01$ ) with the predicted measure were selected as features”. The most popular learning algorithms were CPM, multiple linear regression, RVR, SVR, PLSR, LASSO, and elastic net. Most studies did not require hyperparameters tuning, and nested k-fold CV was the predominant choice for those required. LOO CV was the commonest validation strategy. Only a quarter of studies used external validation.

Our results showed a negative correlation between internal validation prediction accuracy and sample size. This negative correlation was similar to what was reported for mental disorders and health (Sui et al., 2020). Small sample size could lead to overfitting and hence the higher prediction accuracy particularly for CV cases, so that the trained model might explain little of the variance from an external test set (Varoquaux et al., 2017). This is particularly problematic for studying patients with uncommon diseases or medical conditions, or evaluating clinical outcomes of certain treatments (Gabrieli et al., 2015). However, the use of external validation could overcome this problem, as such negative correlation vanished when the external validation prediction accuracy and the sample size of the external test set were evaluated.

Meanwhile, the small sample size issue could be partially addressed by using large open neuroimaging datasets. Currently there are multiple open datasets available to researchers, covering structural MRI, diffusion MRI, resting-state MRI, task-based fMRI, behavioral data, genomics data, and occasionally physiological and angiographic data from a single subject up to 100,000 subjects (Horien et al., 2021; Madan, 2021). Examples included HCP, UK Biobank, and Adolescent Brain Cognitive Development (ABCD) study. However, the use of these large datasets tended to encourage some researchers to use CV or hold-out test sets (a priori split of the dataset) that could be optimistic. It will be more challenging, but the results will be more robust if researchers share data and evaluate model performance on new sites or unseen datasets. Also, researchers should know how the data have been pre-processed and manipulated, so that it could better match the characteristics of the neuroimaging data from their own recruited subjects. Otherwise, the trained model might not give good predictions on an external test set. The users of HCP data should report precisely which dataset was used, as different datasets went through different processing pipelines and contained different subjects.

Here, we reported a 25% of surveyed studies using external test sets. Consistent to a previous review reporting that only 25% of predictive modeling studies on treatment response to addictions (alcohol and substance use) included external validation (Yip et al., 2020). The small number of studies reporting external validation / unseen test sets could be due to generalization failure (of the models) or lack of additional independent data (Sui et al., 2020). It was not possible for us to know if generalization failure did occur for the models, but such failure, if existed, could be accounted by model selection-related issues and homogeneous sample (Boeke et al., 2020). As very few datasets actually collected behavioral measures by implementing the same psychometric tests, it remains to be investigated whether similar behavioral measure (e.g. fluid intelligence/intelligence quotient) from different datasets can

be predicted with the same predictive model. External validation is recommended, as it will avoid confusion from reporting in-sample model fit indices as predictive accuracy and avoid inappropriate CV procedure such as post hoc CV (Poldrack et al., 2020). Therefore, open data sharing initiatives should be encouraged to make external validation more feasible beyond a single laboratory or study site, before the models would be ultimately tested in a large-scale, diverse population-level (Woo et al., 2017).

The connectome data and brain atlases used by the surveyed studies were heterogeneous. This created a variation in the methodology used by different studies, rendering it a potential confounding factor in comparing results across different predictive models. Together with the studies using recruited subjects instead of open datasets, the variability of the analysis pipeline would influence the results as it would for single dataset or group analysis (Botvinik-Nezer et al., 2020; Carp, 2012).

## Conclusion

Based on this work, it was found that the literature currently largely fails to adhere to the recommended best practices, for instance, as outlined by (Scheinost et al., 2019; Woo et al., 2017). Few studies employed external validation for their trained predictive model. Without external validation, internal validation requires very careful planning and considerations with regard to sample size and CV method, which might be subjects of debate to avoid predictive models being too optimistic. Therefore, the authors recommended that future predictive modeling studies should always consider incorporating external validation. When using training and test sets from the same datasets, it is crucial to make them completely independent.

## Funding

SBE acknowledges funding by the European Union's Horizon 2020 Research and Innovation Program (grant agreements 945539 (HBP SGA3) and 826421 (VBC)), the Deutsche Forschungsgemeinschaft (DFG, SFB 1451 & IRTG 2150) and the National Institute of Health (R01 MH074457).

## Data and code availability statement

Data used in this study is provided in the Supplementary Tables 1–3. The code for the partial correlations used in this study can be found at: [https://github.com/jadecci/partialcorr\\_factors](https://github.com/jadecci/partialcorr_factors).

## Conflict of interest

None to declare.

## Credit authorship contribution statement

**Andy Wai Kan Yeung:** Conceptualization, Methodology, Writing – original draft. **Shammi More:** Data curation, Writing – review & editing. **Jianxiao Wu:** Data curation, Writing – review & editing. **Simon B. Eickhoff:** Conceptualization, Methodology, Writing – review & editing.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119275.

## References

- Boeke, E.A., Holmes, A.J., Phelps, E.A., 2020. Toward robust anxiety biomarkers: a machine learning approach in a large-scale sample. *Biol. Psychiatry: Cognit. Neurosci. Neuroimaging* 5, 799–807.
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582, 84–88.



- Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., Yeo, B.T., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106, 2322–2345.
- Bzdok, D., Ioannidis, J.P., 2019. Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci.* 42, 251–262.
- Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Front. Neurosci.* 6, 149.
- Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980.
- Destrieux, C., Fischl, B., Dale, A., Hagren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53, 1–15.
- Diedrichsen, J., 2006. A spatially unbiased atlas template of the human cerebellum. *Neuroimage* 33, 127–138.
- Dosenbach, N.U., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., 2010. Prediction of individual brain maturity using fMRI. *Science* 329, 1358–1361.
- Eickhoff, S.B., Langner, R., 2019. Neuroimaging-based prediction of mental traits: road to utopia or Orwell? *PLoS Biol.* 17, e3000497.
- Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A.R., 2016. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cereb. Cortex* 26, 3508–3526.
- Feng, L., Li, H., Oishi, K., Mishra, V., Song, L., Peng, Q., Ouyang, M., Wang, J., Slinger, M., Jeon, T., 2019. Age-specific gray and white matter DTI atlas for human brain at 33, 36 and 39 postmenstrual weeks. *Neuroimage* 185, 685–698.
- Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* 18, 1664–1671.
- Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355.
- Gabrieli, J.D., Ghosh, S.S., Whitfield-Gabrieli, S., 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* 85, 11–26.
- Gao, S., Greene, A.S., Constable, R.T., Scheinost, D., 2019. Combining multiple connectomes improves predictive modeling of phenotypic measures. *Neuroimage* 201, 116038.
- Gilmore, J.H., Shi, F., Woolson, S.L., Knickmeyer, R.C., Short, S.J., Lin, W., Zhu, H., Hamer, R.M., Styner, M., Shen, D., 2012. Longitudinal development of cortical and subcortical gray matter from birth to 2 years. *Cereb. Cortex* 22, 2478–2485.
- Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., 2016. A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171–178.
- Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E., 2016. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* 26, 288–303.
- He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage* 206, 116276.
- Horien, C., Noble, S., Greene, A.S., Lee, K., Barron, D.S., Gao, S., O'Connor, D., Salehi, M., Dadashkarimi, J., Shen, X., 2021. A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nat. Hum. Behav.* 5, 185–193.
- Liu, W., Wei, D., Chen, Q., Yang, W., Meng, J., Wu, G., Bi, T., Zhang, Q., Zuo, X.-N., Qiu, J., 2017. Longitudinal test-retest neuroimaging data from healthy young adults in southwest China. *Sci. Data* 4, 170017.
- Madan, C.R., 2021. Scan once, analyse many: using large open-access neuroimaging datasets to understand the brain. *Neuroinformatics* doi:10.1007/s12021-12021-09519-12026, [Epub ahead of print].
- Murphy, K., Birn, R.M., Bandettini, P.A., 2013. Resting-state fMRI confounds and cleanup. *Neuroimage* 80, 349–359.
- Poldrack, R.A., Huckins, G., Varoquaux, G., 2020. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* 77, 534–540.
- Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., 2011. Functional network organization of the human brain. *Neuron* 72, 665–678.
- Rao, A., Monteiro, J.M., Mourao-Miranda, J., Initiative, A.S.D., 2017. Predictive modelling using neuroimaging data in the presence of confounds. *Neuroimage* 150, 23–49.
- Rosenberg, M.D., Finn, E.S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.T., Chun, M.M., 2016. A neuromarker of sustained attention from whole-brain functional connectivity. *Nat. Neurosci.* 19, 165–171.
- Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* 28, 3095–3114.
- Scheinost, D., Noble, S., Horien, C., Greene, A.S., Lake, E.M., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D.S., 2019. Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage* 193, 35–45.
- Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., Conrod, P., Dalley, J., Flor, H., Gallinat, J., 2010. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry* 15, 1128–1139.
- Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., Constable, R.T., 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* 12, 506–518.
- Shen, X., Tokoglu, F., Papademetris, X., Constable, R.T., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* 82, 403–415.
- Sui, J., Jiang, R., Bustillo, J., Calhoun, V., 2020. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biol. Psychiatry* 88, 818–828.
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliet, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289.
- Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* 180, 68–77.
- Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *Neuroimage* 145, 166–179.
- Whelan, R., Garavan, H., 2014. When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biol. Psychiatry* 75, 746–748.
- Woo, C.-W., Chang, L.J., Lindquist, M.A., Wager, T.D., 2017. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* 20, 365–377.
- Yamashita, M., Kawato, M., Imamura, H., 2015. Predicting learning plateau of working memory from whole-brain intrinsic network connectivity patterns. *Sci. Rep.* 5, 7622.
- Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106, 1125–1165.
- Yip, S.W., Kiluk, B., Scheinost, D., 2020. Toward addiction prediction: an overview of cross-validated predictive modeling findings and considerations for future neuroimaging research. *Biol. Psychiatry: Cognit. Neurosci. Neuroimaging* 5, 748–758.

## 6 Discussion

The development of usable ML models is a multifaceted endeavor influenced by several critical factors, including data quality, feature engineering, model selection, and interpretability. Overlooking these factors can introduce risks, such as biased and inaccurate predictions, compromised trust in the model’s decisions, and potential ethical concerns. Hence, careful consideration of these factors is essential to building reliable and trustworthy ML models (Scheinost et al., 2019).

In this work, we examined several key factors integral to the development of unbiased and generalizable ML models, ensuring their utility in real-world scenarios. The first is the effective removal of confounding signals so that models are unbiased. To study this, we tested several confound removal workflows on the task of sex classification using ReHo and FC features derived from rs-fMRI data with brain size and age as confounds in study 1. We addressed the biological question of whether there are differences in the functional organization of brains between males and females after controlling for brain size. The second is the usage of different feature spaces and ML algorithms for a given task to find a generalizable model. To study this, we investigated several ML workflows using various combinations of feature spaces from GMV data and ML algorithms to investigate their effect on age prediction performance in study 2. The aim was to find a generalizable and reliable workflow for age prediction by evaluating it under various criteria important for real-world application. We also investigated the potential of brain-age delta or delta, i.e., the difference in predicted and chronological age, as a biomarker and the factors influencing its estimation. As various VBM pipelines exist for GMV estimation, in study 3, we extended the investigation from study 2 to examine the effect of GMV estimates from several VBM alternatives on age prediction performance. Finally, in study 4, we conducted a literature survey of psychometric prediction studies using neuroimaging data. This offered a comprehensive summary of the field’s current state and advancements, highlighting additional factors to consider when designing ML workflows.

The discussion encompasses results from various studies and is structured as

follows. In the initial segment of our discussion, we delve into two critical facets of ML workflow design. First, we underscore the significance of exploring diverse feature spaces and ML algorithms to find a generalizable model. Additionally, we examine the influence of different preprocessing techniques on feature extraction and their consequent impact on predictive performance, drawing insights from studies 1, 2, and 3. Second, we address the concept of mitigating confounding bias and age bias to foster the development of unbiased models, citing findings from studies 1 and 2. Next, we discuss other general considerations integral to the design of ML workflows. These considerations encompass feature preprocessing and engineering (as observed in studies 1 and 2), training sample size (observed in studies 2 and 4), external validation (from study 4), and data shift (noted in studies 2 and 4), all of which exert an influence on the generalizability of ML workflows. The latter part of the discussion centers on interpretability and clinical relevance. We scrutinize the interpretability of the confound-free sex prediction model as demonstrated in study 1. Next, we delve into the clinical implications of the delta, touching upon its relevance in capturing deviance in neurodegenerative disorders and its correlation with behavioral/cognitive measures in healthy and diseased populations, drawing insights from study 2.

## **6.1 Machine learning workflow design**

The overarching goal of ML is to develop unbiased and generalizable models for the task at hand; however, the modeling process involves a series of pivotal decisions. When working with imaging data, the variety of features that can be extracted is extensive. For instance, in the field of computer vision, several kinds of features, such as color histograms, texture features, edge detection, corner detection, and shape descriptors, can be helpful for tasks such as image classification or object detection (Viola and Jones, 2001, Lienhart and Maydt, 2002). Given the variety of features, it is difficult to know which will be best for a given task. Similarly, in the neuroimaging domain, a plethora of features can be derived, and identifying the optimal set for a given task often necessitates a data-driven approach. Additionally, the choice of neuroimaging preprocessing tools can introduce variations in extracted features, potentially influencing model performance. Moreover, within ML workflows, the preprocessing steps undertaken on the features or targets, such as confound removal, Z-score normalization, feature selection, etc., also impact model performance. The choice of the ML algorithm can affect the learned relationship between

features and the target of interest, which can significantly affect the generalizability of the models. Furthermore, factors such as training set sample size and differences in data properties between the train and test set can affect how well the models perform on out-of-sample test data from a new site. Thus, constructing a robust and reliable ML model involves careful consideration of all these intricate decisions and their collective impact on model performance and generalizability.

### 6.1.1 Try different feature spaces and ML algorithms

The choice of feature space plays a vital role in predictive analysis. Various feature spaces can capture distinct types of information from neuroimaging data, leading to diverse outcomes in predictive tasks. Moreover, the selection of ML algorithms can significantly impact the ability to learn the true relationship between these features and target variables. Thus, it becomes imperative to systematically explore many feature spaces and ML algorithm combinations in neuroimaging studies to obtain optimal predictive models and get valuable insights.

A variety of features can be used for sex classification. Some studies have adopted a classification approach based on sMRI (Feis et al., 2013, Rosenblatt, 2016, Zhang et al., 2020, Ebel et al., 2023) or fMRI (Smith et al., 2013b, Ktena et al., 2018, Zhang et al., 2018, Weis et al., 2019) data. Studies with fMRI have generally employed whole-brain FC based on pre-defined regions of interest (ROI) or brain parcellations and achieved a sex prediction accuracy of roughly 75–83% (Satterthwaite et al., 2015, Weis et al., 2020, Zhang et al., 2018, Zhang et al., 2020). Using ReHo, a prediction accuracy of 91% has been shown (Zhang et al., 2020). The choice of the algorithm in previous studies includes SVM (Zhang et al., 2020, Weis et al., 2020), partial least squares regression (Zhang et al., 2018, Chen et al., 2019), random forests classifier (Chen et al., 2019), logistic regression (Al Zoubi et al., 2020). Our results from study-1 are consistent with the existing literature demonstrating CV accuracy of 75–78% and out-of-sample test accuracy of 76–78% without controlling for brain size. In contrast, one study observed a lower prediction accuracy of 62% (Casanova et al., 2012). This might be because of a smaller sample size of only 148 subjects and a high feature dimensionality of FC. A recent study reported a high sex prediction accuracy of 98% (Chen et al., 2019). This high accuracy might be because the study used the HCP1200 dataset (Van Essen et al., 2013), which includes sibling data. Since siblings exhibit similar FC patterns, high prediction accuracy can be achieved if

siblings are not grouped together either in the training or the test sets. Furthermore, the study employed group-independent component analysis (Smith et al., 2013a) to derive six ROI definitions as features before splitting the data into train and test sets. To be noted, in our study, we found slightly higher accuracy with ReHo features compared to FC and partial least squares outperforming ridge, indicating the effect of feature space and ML algorithms on prediction error.

The initial exploration of age prediction using GMV within a single cohort was documented in 2007 (Ashburner, 2007). Subsequently, there has been a surge in brain-age prediction studies aiming to assess the efficacy of delta as a potential biomarker for brain health (Cole et al., 2017, Beheshti et al., 2022). One of the crucial challenges with developing a brain-age estimation framework is selecting input feature space. Various imaging modalities offer distinct insights; for example, fluorodeoxyglucose-positron emission tomography scans reveal details about the brain’s glucose metabolism, while sMRI data provide information about the anatomy/structure of the brain. T1-weighted MRI images have been extensively used in brain-age estimation studies. The two commonly used feature extraction approaches from T1-weighted images include (i) voxel-wise methods which use gray matter, white matter, CSF signal intensities as brain features (Franke et al., 2010, Gaser et al., 2013, Cole et al., 2015, Becker et al., 2018, Varikuti et al., 2018, Sone et al., 2022); and (ii) region-wise methods, which use cortical and subcortical measurements of volume, surface, and thickness values as brain features (Aycheh et al., 2018, Zhao et al., 2019, Lee et al., 2021, Vidal-Pineiro et al., 2021, Elliott et al., 2021, Lange et al., 2022). Dimensionality reduction through unsupervised methods like PCA is commonly employed on voxel-based data, which removes redundant information and helps in reducing computational cost and increasing accuracy (Franke et al., 2010, Becker et al., 2018, Baecker et al., 2021a). Although both kinds of features are used widely, one study comparing ML models using voxel-and region-based morphometric data found voxel-based features to perform better than the region-based features (Baecker et al., 2021a). In our study 2, comparing 128 workflows constituting 16 feature spaces extracted from GMV images (voxel-wise and parcel-wise) and eight ML algorithms (linear and non-linear) for age prediction, we also found voxel-wise features generally performed better than parcel-wise features. This suggests that sometimes, summarizing information, like using average GMV from voxels in different parcels or regions, can

cause information loss, leading to lower prediction performance.

Another important step in developing a brain-age estimation framework is choosing an ML model. The most widely used regression algorithms include RVR (Franke et al., 2010, Gaser et al., 2013, Baecker et al., 2021a), GPR (Cole et al., 2018, Becker et al., 2018 Baecker et al., 2021a), SVR (Lancaster et al., 2018, Sone et al., 2021), and eXtreme Gradient Boosting (Lange et al., 2022, Butler et al., 2021). Overall, the available ML models for brain-age prediction differ with regard to complexity and computational resources and have been shown to influence prediction accuracy (Beheshti et al., 2022). Recent studies have compared the performance of commonly used models to guide on the most suitable model choices for brain-age prediction (in narrow age range: MAE = 2.6-2.7 and 3.7-4.7 years (Lee et al., 2021, Baecker et al., 2021a) and in broad age range: MAE = 7.2-7.7 and 4.6-7.1 years (Lee et al., 2021, Beheshti et al., 2022)). From our study 2, we found that either non-linear or kernel-based algorithms (GPR, KRR, and RVR) are well suited for brain-age estimation. These results align with a study that comprehensively evaluated 22 ML algorithms in broad age range data using GMV features and found SVR, KRR, and GPR with a diverse set of kernels to perform well (Beheshti et al., 2022).

We found voxel-wise GMV features smoothed with a 4 mm FWHM kernel and resampled to a spatial resolution of 4 mm, with PCA retaining 100% variance, and the GPR model (S4\_R4 + PCA + GPR) was the best-performing workflow on the evaluated criteria and was selected for the downstream analysis. This is in line with another study reporting a voxel size of 3.73 mm<sup>3</sup> and a smoothing kernel of 3.68 mm as the optimal parameters for processing GM images for brain-age prediction with a performance similar to our workflows (Lancaster et al., 2018).

To note, we evaluated these workflows on four criteria in contrast to other studies evaluating either one or two. Moreover, we used multiple large cohorts for training and testing the models. On the first criterion, within-dataset performance, the MAE ranged between 4.9–8.5 years and 4.7–8.4 years in CV and left-out-test data for 128 workflows. On the second criterion, cross-dataset performance, the MAE ranged between 4.3–7.4 years and 5.2–9.0 years in CV and out-of-sample test data. The third and fourth criteria, i.e., the test-retest reliability and longitudinal consistency, also varied for different combinations of feature space and ML algorithm. All these criteria are important facets of any biomarker (Cole and Franke, 2017). We found the delta reliable over a short scan delay of less than three months (concordance correlation coefficient = 0.76–0.98; Lawrence and Lin, 1989) in

two test datasets. This aligns with other studies which have shown intraclass correlation coefficient between 0.81-0.96 in different samples with different age groups (Cole et al., 2017, Franke and Gaser, 2012, Elliott et al., 2021). For the last criterion, longitudinal consistency, we found a significant positive linear relationship between the difference in predicted age and the difference in chronological age at a retest duration of 2–3.25 years ( $r = 0.45$ – $0.44$ ) in one dataset and no correlation with a retest duration of 3–4 years in another test dataset. Thus, the evidence for longitudinal consistency was weak. Previous research suggests that lifestyle interventions like meditation and exercise positively impact brain-age (Luders et al., 2016, Steffener et al., 2016, Levakov et al., 2023), while habits such as smoking and alcohol intake may have adverse effects (Bittner et al., 2021, Cole, 2020), influencing longitudinal brain-age trajectories. One study found no association between cross-sectional brain-age and the rate of brain change measured longitudinally, questioning the validity of brain age as a reliable marker for ongoing brain aging changes within an individual (Vidal-Pineiro et al., 2021). Thus, further studies on longitudinal brain age are therefore necessary.

In general, we observed MAE of  $\sim 4.7$  years in our healthy population, which compares favorably with existing literature (Franke et al., 2010, Cole et al., 2015, Lancaster et al., 2018, Boyle et al., 2021, Baecker et al., 2021a, Eickhoff et al., 2021). However, we would like to acknowledge here that this error (MAE) encompasses both the generalization error of the models and genuine biological deviation, and it is challenging to determine their respective contributions. So, there is still a need to develop more accurate models. Recent work suggests that by using large training datasets ( $\sim 10000$  subjects or more) and complex models such as deep learning, the prediction error can go down to  $\sim 3$  years (Levakov et al., 2020, He et al., 2021b, He et al., 2021a, Tanveer et al., 2023), likely reflecting biological variability.

We also conducted experiments to explore the potential performance improvement gained by incorporating additional features from various tissue types. Studies have shown different patterns in both the global and regional GMV, WMV, and CSF alterations in the young and older groups with aging (Good et al., 2001, Ge et al., 2002, Farokhian et al., 2017). Therefore, features from different tissue types may offer complementary information related to age, leading to better predictions. As anticipated, predictions using three tissues, GMV, WMV, and CSF combined as features, were better than GMV only in our study (for example, MAE = 5.08 vs. 6.23). However, one should be cautious



about large dimensions of features compared to the sample size, which might lead to overfitting (Hastie et al., 2009). To address this, we used PCA, keeping 100% variance on the features, thus reducing the number of features to 450 only. Our findings are consistent with a previous study that showed a slight performance improvement when using both GMV and WMV compared to only using GMV (Cole et al., 2017). Notably, combining features from different tissue types has been popular in brain-age estimation studies (Franke and Gaser, 2012, Cole et al., 2018, Hobday et al., 2022). Overall, our results from both study-1 and study-2 provide evidence for the impact of the choice of feature space and the ML algorithm on the prediction performance.

In study 2, we used the CAT toolbox (Gaser et al., 2022), one of the standard VBM analysis choices, to derive estimates of GMV, WMV, and CSF. However, there are several alternatives available, such as SPM (Ashburner and Friston, 2000) and FSL (Smith et al., 2004), exhibiting differential specificity in GMV estimation (Bhagwat et al., 2021). VBM analysis involves a series of essential preprocessing steps, encompassing brain extraction, segmentation, spatial registration or normalization, and modulation. VBM tools offer different algorithms with several configurable options for each preprocessing step. These differences can lead to differences in the GMV estimates, which can influence the estimated association with age (Tavares et al., 2020, Zhou et al., 2022). A study demonstrated that GMV and WMV estimates obtained through SPM12 and CAT12 differed, further impacting their relationship with age (Tavares et al., 2020). Another recent study performing a comprehensive comparison between CAT12, two FSL-based and one FSL-dependent hybrid pipelines has shown that the choice of preprocessing pipeline impacts sex and age prediction performances (Zhou et al., 2022). We found evidence supporting that different preprocessing tools can give differential age prediction outcomes. In study 2, we found that CAT-derived GMV performed better than SPM-derived GMV with lower MAE, higher correlation between true and predicted age, and lower age bias, i.e., the correlation between age and delta.

To delve deeper, in study 3, we evaluated 10 VBM pipelines, including two *off-the-shelf* pipelines, CAT (version 12.8, r1813) and FSLVBM (uses FSL tools, version 6.0), and three modularly constructed pipelines, including Advanced Normalization Tools (ANTs, version 2.2.0), ANTs-FSL (uses ANTs for brain extraction and segmentation, FSL for registration) and fMRIPrep-FSL (uses ANTs for brain extraction, FSL for segmentation and registration), each of these implemented using a general template (e.g., MNI-152) and

a study-/data-specific template. Using three large datasets covering the adult lifespan acquired in different scanners and protocols, the systematic differences between the VBM pipelines were confirmed by the high accuracy when predicting the pipelines using their respective GMV estimates. There was a substantial impact of GMV derived from different VBM pipelines on within-dataset and cross-dataset age prediction performance, with fMRIPrep-FSL and CAT-derived GMV estimates performing the best.

In summary, results from both studies reveal the significant impact of different preprocessing or feature extraction tools on GMV estimates, which influenced the prediction performance. The results highlighted the importance of testing different combinations of feature spaces and ML algorithms in a data-driven fashion and evaluating them on multiple criteria to find an accurate and generalizable workflow.

### **6.1.2 Control for bias**

Controlling for bias in ML workflows is critical to ensure fairness, equity, and accuracy in the predictions and decisions. Biases can arise from various sources, including non-representative training data, imbalances in class distribution, the presence of confounds, or incomplete information in the features (missing variable bias), among other potential sources (Mehrabi et al., 2021, Larrazabal et al., 2020, Li et al., 2022). Our studies addressed two specific biases and outlined strategies to deal with them effectively.

#### **6.1.2.1 Removal of confounding signal**

If one wants to establish a brain-phenotype relationship by estimating generalization performance and identifying brain regions explaining the variance in phenotype, it is important to control for confounding signals that can mask the true relationship between brain and phenotype. Brain size is highly correlated with sex, with a larger total brain volume in males compared to females, and is encoded in neuroimaging features such as ReHo and FC (Ruigrok et al., 2014, Ritchie et al., 2018). Hence, brain size is a confound in the sex classification task if one is interested in studying the difference in functional organization between sexes. Regressing out brain size signal from every feature can remove sex-specific information from the features, therefore forcing the prediction performance to be weaker. In (Zhang et al., 2018), authors have shown that the sex prediction accuracy drops from 80% to 70% after regressing out brain size from FC. In our study 1, all three

datasets showed significant brain size differences between sexes, and consequently, we saw the highest model performance for sex classification with workflow not controlling for confounds. The out-of-sample test accuracy dropped from 76-78% to 56-67% after confound removal; however, above-chance sex classification performance indicates that models can capture the difference in functional organization between sexes independent of variations in brain size.

The two confound removal approaches investigated, WDCR and CVCR, showed reduced performance in line with previous studies (Pervaiz et al., 2020, Snoek et al., 2019). We subsequently validated the effectiveness of these confound removal methods. There were no correlations between each residual feature and brain size in a univariate fashion with both schemes. We checked for any remaining multivariate confounding effects using multiple linear regression to predict brain size from the residual features and observed negative adjusted  $R^2$  with both schemes. Thus, there was no signal from brain size in the residual features after confound removal, and hence, the models should not encode any confound-related information.

We observed lower generalization estimates with WDCR compared to CVCR. In fact, with WDCR, the accuracy dropped to a chance level. This is contrary to expectations as WDCR uses the whole sample before CV to remove confounding signals, causing data leakage from the training sample to the testing/validation sample; therefore, we expected higher generalization performance. However, in this case, that actually made the model perform worse. This could be because WDCR aggressively removes confounding signals from the data, leading to chance-level performance. On the other hand, out-of-sample performance was closer to the generalization performance estimated with CVCR. Consistent with our findings, other studies demonstrated that WDCR led to pessimistic model performance estimates, notably below chance (Todd et al., 2013, Snoek et al., 2019). They demonstrated that this occurs when the “signal” in the data, operationalized as the width of the feature-target correlation distribution, is lower than would be expected by chance (Snoek et al., 2019), similar to findings by (Jamalabadi et al., 2016). WDCR reduces the width of the correlation distribution, leading to lower model performance, and this effect is exacerbated by higher confound-target correlations and by a larger number of features. They showed CVCR yielded significantly above-chance model performance and nearly unbiased model performance in the simulations and different datasets with different numbers of features

and the strength of the confound. CVCR removes all variance associated with the confound in the train set and may show reduced performance in some scenarios (Snoek et al., 2019).

Therefore, we concluded that CVCR is better for confound removal than WDCR. Moreover, since the sex classification performance after confound removal was still high, one could conclude that there are differences in the functional organization of brains between sexes, as captured from ReHo and FC after removing brain size differences. Another important observation was the disparity between important features from the model trained without confound removal and those trained after confound removal (using WDCR and CVCR), implying that interpretations derived from these models would be different (for more details, refer to section 6.2.1).

#### **6.1.2.2 Mitigation of age bias**

Numerous brain-age estimation studies have reported age bias, a phenomenon wherein brain-age or predicted age is over-predicted in young subjects, under-predicted in older subjects, and subjects closer to the mean of training data are predicted more accurately (Liang et al., 2019, Cole, 2020); thus causing a negative correlation between chronological age and delta. This age bias complicates the use of delta in clinical contexts, as it can lead to misleading correlations between delta and behavioral or cognitive measures and erroneous interpretations while comparing delta between different clinical groups. To mitigate this age bias, an additional bias correction step can be applied to the predicted age or delta to regress out the effect of age. Generally, a linear regression model is fitted with the predictions on CV-derived training data as the dependent variable and chronological age as the independent variable. The predicted age in the CV-derived test set is corrected by subtracting the resulting intercept and dividing by the slope (Cole, 2020). Training bias correction models in a CV-consistent fashion helps avoid information leakage from the test to training data. There are several alternatives available for statistical bias correction (Lange and Cole, 2020); the one we used does not use the chronological age of the test data for correction, while others use test labels in correction (Smith et al., 2019, Lange et al., 2019, Beheshti et al., 2019), causing data leakage and not suitable for real-world use.

Our workflows showed negative associations between chronological age and delta for both within-dataset and cross-dataset predictions (ranging between -0.2 to -0.8), with

more accurate models displaying lower age bias. Speculatively, this age bias may be attributed to missing or omitted variables bias, which occurs when a statistical model leaves out relevant independent variables that are a determinant of the dependent variable (Wilms et al., 2021). In other words, when the input features lack sufficient information to predict age, predictions tend to cluster around the median or mean age, thus introducing age bias, also demonstrated in another recent study (Lange et al., 2022). Consequently, we observed that adding features from additional tissue types reduced the age bias in our study.

Our results show that bias correction models work well in within-dataset analysis, i.e., when the train and test sets are derived from the same dataset or site, but residual bias remains in the predictions from cross-dataset analysis, i.e., when bias correction models are derived from the training set and applied to out-of-sample test data from a new site. This discrepancy may arise because of differences in data properties, e.g., scanner-specific idiosyncrasy (Jovicich et al., 2006, Chen et al., 2014), between the training and the test data. Additionally, we observed that the effectiveness of bias correction models was influenced by the sample size of the within-dataset used for correction. Specifically, we found that smaller samples used for bias correction led to high variance in mean corrected delta (see section 6.2.2.1). This aligns with previous studies demonstrating greater variability in model performance with small sample sizes (Varoquaux, 2018). Overall, the choice of data source (within-data or cross-data) and the sample size used for bias correction substantially impact the quality of the model, affecting the corrected prediction values. This eventually affects the observed delta-behavior correlations (see section 6.2.2.2).

With these findings, we emphasize the importance of selecting an appropriate bias mitigation strategy to ensure the predictions are bias-free, thereby ensuring accurate and equitable decision-making.

### 6.1.3 Other general considerations

There can be several other factors that can affect the generalizability of an ML model, for instance, employing feature preprocessing and engineering, such as Z-score normalization (Ali et al., 2014), PCA (Jolliffe, 2002), and other feature selection techniques (Chandrashekar and Sahin, 2014, Mwangi et al., 2014), can help improve model performance. Additionally, other factors such as training set sample size and

difference in data properties between the train and test set can affect how well the models perform on out-of-sample test data from a new site (Hastie et al., 2009). This section delves into specific observations derived from our studies in these contexts.

#### **6.1.3.1 Feature preprocessing and engineering**

Several preprocessing steps can be applied to features prior to model training, which can help improve data quality and improve model performance. One common technique is Z-score normalization, which transforms the features by subtracting the mean value of a feature from each data point and then dividing it by the standard deviation of that feature, thus centering the data around a mean of zero and scaling it to have a standard deviation of one (Ali et al., 2014). It helps mitigate the magnitude differences between features, ensuring that all features contribute equally to the learning process, aids algorithms that rely on distance or magnitude comparisons to work effectively, and makes the coefficients or feature importance scores comparable and easier to interpret. In study 1, we observed that Z-scoring improved the model performance for sex classification with ReHo but not with FC. Additionally, the Z-score normalization of the features before or after confound removal did not affect model performance. However, since some learning algorithms might benefit from well-scaled features (Anggoro and Supriyanti, 2019, Fei et al., 2021), we recommend normalizing features after confound removal.

For high-dimensional neuroimaging data, employing dimensionality reduction techniques can improve the observations-to-features ratio. One method is variance thresholding, which is a feature selection technique that filters out low-variance features that are less informative for predictive modeling. Some feature engineering methods, such as PCA, can transform high-dimensional data into a lower-dimensional space while retaining the variance in the original features (Jolliffe, 2002, Lever et al., 2017). Another commonly employed approach in neuroimaging involves resampling voxel-wise data to lower spatial resolution (Franke et al., 2010) or using a brain atlas to summarize data from distinct brain regions or parcels (Fan et al., 2016, Yeo et al., 2011, Buckner et al., 2011). In study 2, we observed that smoothed and resampled voxel-wise GMV outperformed parcel-wise GMV, suggesting that summarizing information can result in a loss of valuable information in certain cases. Interestingly, smoothed and resampled voxel-wise GMV with and without PCA yielded similar results, contrary to other studies that have shown performance improvement with PCA (Franke et al., 2010, Franke and

Gaser, 2012). This could be attributed to prior dimensionality reduction through resampling. These results highlight the importance of feature preprocessing and engineering in performance improvement in some cases.

#### **6.1.3.2 Large training sample size and external validation**

A large training sample is of paramount importance in ML. It can help improve generalization capabilities by providing a more representative and diverse set of data points, enabling the model to capture underlying patterns in the data and reduce the risk of overfitting (Hastie et al., 2009). As articulated by Domingos, a key rule is “more data beats a cleverer algorithm,” emphasizing the critical role of training sample size (Domingos, 2012). In study 2, we observed lower CV generalization errors with a higher sample size in the cross-dataset analysis as it had a larger sample pooled from multiple datasets compared to the single cohort within-dataset analysis. Additionally, bias correction models worked effectively with large sample sizes (see section 6.2.2.1). This highlights the impact of the training set sample size on the estimation of generalization performance and corroborates with previous studies showing lower errors with larger training datasets (Baecker et al., 2021a, Lange et al., 2022). On the contrary, in study 4, our literature review on existing psychometric prediction research showed an intriguing negative relationship between prediction accuracy and sample size, similar to some other studies (Sui et al., 2020, Varoquaux, 2018, Wolfers et al., 2015). This pattern was particularly noticeable in studies employing CV within single cohorts. Since only 25 percent of the surveyed studies used external test sets, it was not possible to assess whether highly accurate models were overfitted. The higher prediction accuracies observed in smaller samples may not necessarily imply superior models; rather, they could be attributed to publication bias. Nevertheless, this negative correlation did not reach statistical significance when comparing external test accuracy and the external test sample size, suggesting that employing external validation is a valuable approach to address this issue. .

#### **6.1.3.3 Presence of data shift**

Neuroimaging studies frequently involve data acquisition from various scanners, which might cause systematic differences related to different scanning platforms (Jovicich et al., 2006, Kruggel et al., 2010) between the training and the out-of-sample test



sample. Additionally, demographic differences between samples might exist, leading to dataset shift and confound shift (Landeiro and Culotta, 2018). An ideal model should generalize well despite such differences. From study 1, we found that in the absence of data shift, i.e., when sample properties between train and test are similar, the out-of-sample performance was best when confound models from the train data were applied to test data. On the other hand, the test performance was much lower in the presence of data shift. Even though residual correlations were observed between features and confound in the out-of-sample test data after applying confounding models, the training models were confounding-free, so this performance cannot be driven by confounding effects. Similarly, from study 2, we found the workflows gave a lower performance on out-of-sample test data from cross-dataset analysis compared to within-dataset analysis. Additionally, the bias correction models derived from the cross-dataset did not correct for the age bias adequately. These results indicate that ML workflows might show reduced performance on new test samples in the presence of data shift.

Overall, the findings from the four studies emphasize the significance of careful implementation at each step of ML workflow design. It highlights various factors impacting the predictive performance of ML workflows, including preprocessing tools, feature space and preprocessing steps applied to features, ML algorithm choices, and the presence of data shifts. They highlight the significance of conducting data preprocessing within the CV loop, utilizing large samples, and external validation if possible.

## 6.2 Interpretability and clinical relevance

Interpretability is the degree to which a human can understand the cause of a decision (Miller, 2019). The higher the interpretability of an ML model, the easier it is for someone to comprehend why certain decisions or predictions have been made. It aids trust in the decisions, which is especially important for critical tasks such as clinical diagnosis. Inherently interpretable models can provide valuable insights into brain-behavior relationships by investigating feature importance scores. The advancement in interpretable ML/explainable AI has led to local model-agnostic interpretability methods (Molnar, 2019, Carvalho et al., 2019). While exploring model interpretability was not our primary focus, we did investigate significant brain regions associated with sex prediction. Additionally, we sought to evaluate the clinical

significance of delta.

### **6.2.1 Interpretability of confound-free sex prediction model**

Removing confounding effects is crucial for obtaining unbiased results; otherwise, an ML model might mostly rely on confounds, rendering signals of interest redundant. We compared the predictive features from two models: a model trained without removing the confounding signal of brain size and another confound-free model for sex prediction. As anticipated, we observed differences in the predictive features between these two models. Specifically, we noticed that the features selected by the model without confound removal exhibited stronger positive or negative correlations with brain size. Conversely, in models incorporating confound removal techniques (WDCR and CVCR), the selected features displayed lower correlations with brain size. This suggests that the features selected after accounting for confounding signals can capture the functional patterns associated with sex differences. With ReHo, the performance was slightly better compared to FC, and selected regions were in the dorsolateral prefrontal cortex, inferior parietal lobule, occipital, ventromedial prefrontal cortex, precentral gyrus, post insula, parietal, temporoparietal junction, and inferior cerebellum, in line with a study identifying regions in the inferior parietal lobule and precentral gyrus (Xu et al., 2015). These regions are associated with a diverse array of cognitive and functional processes that have been shown to exhibit sex-related differences (Miller and Halpern, 2014). We found important FC features widespread across the entire brain with strong interhemispheric connections, suggesting sex-related variations in neural function and connectivity involve a global network and integration of information between the two brain hemispheres.

### **6.2.2 Clinical relevance of brain-age delta**

Brain-age estimations derived from sMRI features offer an intuitive measure of the brain’s intricate aging patterns. The disparity between predicted and chronological age (delta) can serve as a valuable metric for assessing deviations from typical brain aging trajectories. Various diseases, including neurological conditions such as AD, MCI (Franke et al., 2010, Franke and Gaser, 2012, Gaser et al., 2013), Parkinson’s disease (Eickhoff et al., 2021, Beheshti et al., 2020), traumatic brain injury (Cole et al., 2015,

Savjani et al., 2017), epilepsy (Sone et al., 2021, Pardoe et al., 2017), multiple sclerosis (Cole et al., 2020, Høgestøl et al., 2019), and stroke (Egorova et al., 2019, Richard et al., 2020), as well as psychiatric disorders such as schizophrenia (Lee et al., 2021, Koutsouleris et al., 2014), bipolar disorder (Hajek et al., 2019, Van Gestel et al., 2019), major depressive disorder (Han et al., 2021a, Han et al., 2021b), autism spectrum disorder (Becker et al., 2018, Lombardi et al., 2020), and attention deficit hyperactivity disorder (Kaufmann et al., 2019), have shown higher brain-age. Studies suggest that preclinical stages of some diseases, such as clinical high risk for psychosis (CHR) and early-stage first-episode psychosis (FEP) (preclinical stage of schizophrenia) and MCI (preclinical stage of AD), display neuroanatomical changes and already show increased delta. Moreover, higher brain-age has been shown to relate to cognitive aging, multiple aspects of physiological aging such as grip strength, lung function, lifestyle factors such as smoking and alcohol consumption, and mortality in older adults (Gaser et al., 2013, Liem et al., 2017, Anatürk et al., 2021, Boyle et al., 2021, Franke and Gaser, 2012, Cole et al., 2018, Cole, 2020). On the other hand, lower brain-age has been shown to relate to protective effects of medication (Luders et al., 2016), practicing music (Rogenmoser et al., 2018), or having higher levels of education or physical activity (Steffener et al., 2016). Thus, delta holds promise as a marker for general brain health, early detection of brain disorders, and evaluating the effects of lifestyle changes and medications (Franke and Gaser, 2019, Cole and Franke, 2017). We explored the clinical utility of delta by applying brain-age models to neurodegenerative disorder and by computing the relationship between delta and behavioral/cognitive measures in healthy and diseased populations.

#### **6.2.2.1 Higher brain-age delta in disease**

For age prediction, our selected workflow (S4\_R4 + PCA + GPR) showed high within-dataset performance, cross-dataset performance, test-retest reliability, and moderate longitudinal consistency in the healthy population. These findings illustrate that the brain-age model can effectively capture the typical structural changes associated with healthy aging. Neurodegenerative disorders, such as AD and MCI, are characterized by progressive structural and functional disruptions in the brain, causing a decline in global and local GMV (Good et al., 2001, Fjell et al., 2014). Consequently, patients with neurodegenerative disorders have older-appearing brains, which brain-age

prediction models should be able to capture. We tested this by comparing the delta between HC, early MCI, late MCI, and AD groups. We found advanced brain aging with neurodegenerative disorders, with the mean corrected delta significantly higher in the AD (6.6-4.5 years) and late MCI (2.9-2.1 years) groups compared to HC. Our results align with previous studies, which have reported an increased delta of 3–8 years in MCI and  $\sim 10$  years with AD patients (Franke and Gaser, 2012, Gaser et al., 2013, Varikuti et al., 2018, Beheshti et al., 2022). Furthermore, the corrected delta correlated with disease severity and cognitive impairment measures, such as the Mini-Mental State Examination, Global Clinical Dementia Rating Scale, and Functional Assessment Questionnaire in MCI and AD patients, in line with other studies (Franke and Gaser, 2012, Gaser et al., 2013, Löwe et al., 2016, Beheshti et al., 2018). Thus, the delta confirmed its potential to indicate accelerated brain aging in neurodegenerative diseases.

Furthermore, we demonstrated that the delta estimates in different groups were dependent on the workflow, i.e., the feature space and ML algorithm used, which consequently affected the observed relationship with cognitive measures. Moreover, the choice of data for bias correction, whether within-dataset or cross-dataset, impacted the delta estimates. Within-dataset correction worked more effectively, although it was also influenced by the size of the within-dataset. We tested the impact of within-dataset sample size on the effectiveness of bias correction by using different sub-samples of within-dataset HC subjects to correct the age bias in HC and AD groups. We found high variance in the mean corrected delta using small sample sizes. As a result, it is imperative to exercise caution when comparing findings across different research studies as they differ in experimental setup and methodology, such as feature spaces, ML algorithms, and different methods and sample sizes for bias correction, leading to differences in the outcomes.

#### **6.2.2.2 Delta-behavior correlations in healthy populations**

Previous studies have shown delta is predictive of mortality and correlates with age-sensitive physiological measures, including grip strength, lung function, walking speed, blood pressure, and allostatic load in the aging population (Cole et al., 2018, Cole, 2020). Delta is significantly increased in AD, MCI, and Parkinson’s disease (Franke and Gaser, 2012, Eickhoff et al., 2021). Most studies have shown an association of delta with cognitive variables in a clinical population. It is important to check if delta can capture cognitive

and behavior variability associated with healthy aging. Due to the presence of age bias, it is essential to control for age when analyzing correlations between delta and behavioral measures; otherwise, it will give spurious correlations. One could either use age as a covariate while using the uncorrected delta or apply the bias correction method to get corrected predictions and then use the corrected delta for further analysis (Le et al., 2018).

We identified a weak but statistically significant association between delta and several cognitive and motor performance measures using CV predictions from within-dataset analysis. Specifically, we observed that higher uncorrected delta values (while controlling for age as a covariate) were correlated to lower fluid intelligence, higher motor learning reaction time, and lower response inhibition and selective attention abilities. It is worth noting that these correlations exhibited slight variations when using corrected delta values. The reason could be the difference between the two methods to control for age bias; when using age as a covariate, the whole sample is used, while the linear regression for bias correction uses CV-derived training data, leading to correction using fewer data points. Moreover, our investigation also showed a disparity between delta-behavior correlations derived from within-dataset predictions and those obtained through cross-dataset predictions, even though they were highly correlated. In the cross-dataset analysis, delta values did not exhibit significant correlations with fluid intelligence and motor learning reaction time; however, the higher delta was correlated with lower response inhibition, selective attention abilities, and lower executive functioning. One previous multi-site study has shown that a higher delta is associated with lower general cognitive status, processing speed, visual attention, cognitive flexibility status, and semantic verbal fluency (Boyle et al., 2021). These findings collectively suggest that the delta can capture variability in cognitive and behavioral functioning in the healthy population. Nevertheless, the estimates of the delta are sensitive to the ML workflow used and data used for bias correction, leading to disparities in the observed delta-behavior associations.

Our results provide further evidence for the potential future application of delta as a biomarker while drawing attention to factors influencing delta estimates. It is important to note that there are remaining challenges in the field before brain-age estimation can be used as a general screening tool in clinics (Butler et al., 2021, Kumari and Sundarajan, 2023, Dempsey et al., 2023).

## 6.3 Conclusion

This work addressed challenges encountered in designing a robust, generalizable, and bias-free machine learning workflow. We emphasized the significance of confound removal and the impact of confound regression strategies on prediction performance and model interpretability, noting their limitations in the presence of data shifts. The study demonstrates the importance of performing confound regression within a cross-validation framework, akin to other preprocessing steps, to get generalizable performance estimates using a sex classification task. Furthermore, we demonstrated the importance of evaluating different feature spaces and machine learning algorithms in predictive analysis and evaluating them under multiple criteria to find a robust and generalizable workflow. Voxel-wise gray matter volume features and the Gaussian process regression model exhibited superior performance in age prediction across various criteria important for practical applicability. The studies highlight the effect of neuroimaging preprocessing tools for feature extraction, preprocessing steps on features, training sample size, and data shifts on model performance and downstream analyses. Lastly, by shedding light on the trends and issues in current psychometric prediction research, we advocate adopting large sample sizes and external validation. Collectively, these insights contribute to a more informed and effective approach to designing ML workflows and stress the need to exercise caution during the design process, meticulous result analysis, and reporting.

## Bibliography

- Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., and Varoquaux, G. 2017. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*. 147, 736–745.
- Al Zoubi, O., Misaki, M., Tsuchiyagaito, A., Zotev, V., White, E., Investigators, T.1., Paulus, M., and Bodurka, J. 2020. Predicting sex from resting-state fMRI across multiple independent acquired datasets. *BioRxiv*, 2020–08.
- Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J.L., Bastiani, M., Miller, K.L., Nichols, T.E., and Smith, S.M. 2021. Confound modelling in UK Biobank brain imaging. *NeuroImage*. 224, 117002.
- Ali, P.J.M., Faraj, R.H., Koya, E., Ali, P.J.M., and Faraj, R.H. 2014. Data normalization and standardization: a technical report. *Mach Learn Tech Rep*. 1(1), 1–6.
- Anatürk, M., Kaufmann, T., Cole, J.H., Suri, S., Griffanti, L., Zsoldos, E., Filippini, N., Singh-Manoux, A., Kivimäki, M., Westlye, L.T., et al. 2021. Prediction of brain age and cognitive age: Quantifying brain and cognitive maintenance in aging. *Human brain mapping*. 42(6), 1626–1640.
- Anggoro, D. and Supriyanti, W. 2019. Improving accuracy by applying Z-score normalization in linear regression and polynomial regression model for real estate data. *International Journal of Emerging Trends in Engineering Research*. 7(11), 549–555.
- Antonopoulos, G., More, S., Raimondo, F., Eickhoff, S.B., Hoffstaedter, F., and Patil, K.R. 2023. A systematic comparison of VBM pipelines and their application to age prediction. *Neuroimage*, 120292.
- Ashburner, J. 2007. A fast diffeomorphic image registration algorithm. *Neuroimage*. 38(1), 95–113.
- Ashburner, J. and Friston, K.J. 2000. Voxel-based morphometry—the methods. *Neuroimage*. 11(6), 805–821.



- Aycheh, H.M., Seong, J.-K., Shin, J.-H., Na, D.L., Kang, B., Seo, S.W., and Sohn, K.-A. 2018. Biological brain age prediction using cortical thickness data: a large scale cohort study. *Frontiers in aging neuroscience*. 10, 252.
- Baecker, L., Dafflon, J., Da Costa, P.F., Garcia-Dias, R., Vieira, S., Scarpazza, C., Calhoun, V.D., Sato, J.R., Mechelli, A., and Pinaya, W.H. 2021a. Brain age prediction: A comparison between machine learning models using region-and voxel-based morphometric data. *Human brain mapping*. 42(8), 2332–2346.
- Baecker, L., Garcia-Dias, R., Vieira, S., Scarpazza, C., and Mechelli, A. 2021b. Machine learning for brain age prediction: Introduction to methods and clinical applications. *EBioMedicine*. 72.
- Becker, B.G., Klein, T., Wachinger, C., Initiative, A.D.N., et al. 2018. Gaussian process uncertainty in age estimation as a measure of brain abnormality. *NeuroImage*. 175, 246–258.
- Beheshti, I., Ganaie, M., Paliwal, V., Rastogi, A., Razzak, I., and Tanveer, M. 2022. Predicting Brain Age Using Machine Learning Algorithms: A Comprehensive Evaluation. *IEEE Journal of Biomedical and Health Informatics*. 26(4), 1432–1440.
- Beheshti, I., Maikusa, N., and Matsuda, H. 2018. The association between “brain-age score” (BAS) and traditional neuropsychological screening tools in Alzheimer’s disease. *Brain and Behavior*. 8(8), e01020.
- Beheshti, I., Mishra, S., Sone, D., Khanna, P., and Matsuda, H. 2020. T1-weighted MRI-driven brain age estimation in Alzheimer’s disease and Parkinson’s disease. *Aging and disease*. 11(3), 618.
- Beheshti, I., Nugent, S., Potvin, O., and Duchesne, S. 2019. Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme. *NeuroImage: Clinical*. 24, 102063.
- Bertolote, J. 2007. Neurological disorders affect millions globally: WHO report. *World Neurology*. 22(1).
- Bhagwat, N., Barry, A., Dickie, E.W., Brown, S.T., Devenyi, G.A., Hatano, K., DuPre, E., Dagher, A., Chakravarty, M., Greenwood, C.M., et al. 2021. Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *GigaScience*. 10(1), giaa155.
- Bishop, C.M. and Nasrabadi, N.M. 2006. Pattern recognition and machine learning. Vol. 4. (4). Springer.

- Biswal, B., Zerrin Yetkin, F., Haughton, V.M., and Hyde, J.S. 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic resonance in medicine*. 34(4), 537–541.
- Bittner, N., Jockwitz, C., Franke, K., Gaser, C., Moebus, S., Bayen, U.J., Amunts, K., and Caspers, S. 2021. When your brain looks older than expected: combined lifestyle risk and BrainAGE. *Brain Structure and Function*. 226, 621–645.
- Boyle, R., Jollans, L., Rueda-Delgado, L.M., Rizzo, R., Yener, G.G., McMorro, J.P., Knight, S.P., Carey, D., Robertson, I.H., Emek-Savaş, D.D., et al. 2021. Brain-predicted age difference score is related to specific cognitive functions: a multi-site replication analysis. *Brain imaging and behavior*. 15, 327–345.
- Brownlee, J. 2020. Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.
- Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., and Yeo, B.T. 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of neurophysiology*. 106(5), 2322–2345.
- Butler, E.R., Chen, A., Ramadan, R., Le, T.T., Ruparel, K., Moore, T.M., Satterthwaite, T.D., Zhang, F., Shou, H., Gur, R.C., et al. (2021). *Pitfalls in brain age analyses*. Tech. rep. Wiley Online Library.
- Carvalho, D.V., Pereira, E.M., and Cardoso, J.S. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*. 8(8), 832.
- Casanova, R., Whitlow, C., Wagner, B., Espeland, M., and Maldjian, J. 2012. Combining graph and machine learning methods to analyze differences in functional connectivity across sex. *The open neuroimaging journal*. 6, 1.
- Caspers, J. 2021. Translation of predictive modeling and AI into clinics: a question of trust. *European Radiology*. 31(7), 4947–4948.
- Cawley, G.C. and Talbot, N.L. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*. 11, 2079–2107.
- Chandrashekar, G. and Sahin, F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering*. 40(1), 16–28.
- Chen, C., Cao, X., and Tian, L. 2019. Partial least squares regression performs well in MRI-based individualized estimations. *Frontiers in neuroscience*. 13, 1282.

- Chen, J., Liu, J., Calhoun, V.D., Arias-Vasquez, A., Zwiers, M.P., Gupta, C.N., Franke, B., and Turner, J.A. 2014. Exploration of scanning effects in multi-site structural MRI studies. *Journal of neuroscience methods*. 230, 37–50.
- Choi, S.W., Cho, H.-H., Koo, H., Cho, K.R., Nenning, K.-H., Langs, G., Furtner, J., Baumann, B., Woehrer, A., Cho, H.J., et al. 2020. Multi-habitat radiomics unravels distinct phenotypic subtypes of glioblastoma with clinical and genomic significance. *Cancers*. 12(7), 1707.
- Chyzyk, D., Varoquaux, G., Milham, M., and Thirion, B. 2022. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience*. 11.
- Cole, J.H. 2020. Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. *Neurobiology of aging*. 92, 34–42.
- Cole, J.H. and Franke, K. 2017. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*. 40(12), 681–690.
- Cole, J.H., Leech, R., Sharp, D.J., and Initiative, A.D.N. 2015. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of neurology*. 77(4), 571–581.
- Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., and Montana, G. 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*. 163, 115–124.
- Cole, J.H., Raffel, J., Friede, T., Eshaghi, A., Brownlee, W.J., Chard, D., De Stefano, N., Enzinger, C., Pirpamer, L., Filippi, M., et al. 2020. Longitudinal assessment of multiple sclerosis with the brain-age paradigm. *Annals of neurology*. 88(1), 93–105.
- Cole, J.H., Ritchie, S.J., Bastin, M.E., Hernández, V., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., et al. 2018. Brain age predicts mortality. *Molecular psychiatry*. 23(5), 1385–1392.
- Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., and Trojanowski, J.Q. 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of aging*. 32(12), 2322–e19.
- De Lange, A.-M.G., Anatórk, M., Suri, S., Kaufmann, T., Cole, J.H., Griffanti, L., Zsoldos, E., Jensen, D.E., Filippini, N., Singh-Manoux, A., et al. 2020. Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study. *NeuroImage*. 222, 117292.

- Dempsey, D.A., Deardorff, R., Wu, Y.-C., Yu, M., Apostolova, L.G., Brosch, J., Clark, D.G., Farlow, M.R., Gao, S., Wang, S., et al. 2023. BrainAGE Estimation: Influence of Field Strength, Voxel Size, Race, and Ethnicity. *medRxiv*, 2023–12.
- Dickie, E., Hodge, S., Craddock, R., Poline, J.-B., and Kennedy, D. 2017. Tools matter: comparison of two surface analysis tools applied to the ABIDE dataset. *Research Ideas and Outcomes*. 3, e13726.
- Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*. 55(10), 78–87.
- Du, W., Calhoun, V.D., Li, H., Ma, S., Eichele, T., Kiehl, K.A., Pearlson, G.D., and Adali, T. 2012. High classification accuracy for schizophrenia with rest and task fMRI data. *Frontiers in human neuroscience*. 6, 145.
- Du, Y., Fu, Z., and Calhoun, V.D. 2018. Classification and prediction of brain disorders using functional connectivity: promising but challenging. *Frontiers in neuroscience*. 12, 525.
- Dziugaite, G.K., Ben-David, S., and Roy, D.M. 2020. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*.
- Ebel, M., Domin, M., Neumann, N., Schmidt, C.O., Lotze, M., and Stanke, M. 2023. Classifying sex with volume-matched brain MRI. *Neuroimage: Reports*. 3(3), 100181.
- Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E.M., Brammer, M.J., Murphy, C., Murphy, D.G., Consortium, M.A., et al. 2010. Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. *Neuroimage*. 49(1), 44–56.
- Egorova, N., Liem, F., Hachinski, V., and Brodtmann, A. 2019. Predicted brain age after stroke. *Frontiers in aging neuroscience*. 11, 348.
- Eickhoff, C.R., Hoffstaedter, F., Caspers, J., Reetz, K., Mathys, C., Dogan, I., Amunts, K., Schnitzler, A., and Eickhoff, S.B. 2021. Advanced brain ageing in Parkinson’s disease is related to disease duration and individual impairment. *Brain communications*. 3(3), fcab191.
- Elliott, M.L., Belsky, D.W., Knodt, A.R., Ireland, D., Melzer, T.R., Poulton, R., Ramrakha, S., Caspi, A., Moffitt, T.E., and Hariri, A.R. 2021. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. *Molecular psychiatry*. 26(8), 3829–3838.

- Eshaghi, A., Young, A.L., Wijeratne, P.A., Prados, F., Arnold, D.L., Narayanan, S., Guttman, C.R., Barkhof, F., Alexander, D.C., Thompson, A.J., et al. 2021. Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nature communications*. 12(1), 2078.
- Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A.R., et al. 2016. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*. 26(8), 3508–3526.
- Farokhian, F., Yang, C., Beheshti, I., Matsuda, H., and Wu, S. 2017. Age-related gray and white matter changes in normal adult brains. *Aging and disease*. 8(6), 899.
- Fei, N., Gao, Y., Lu, Z., and Xiang, T. (2021). “Z-score normalization, hubness, and few-shot learning”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 142–151.
- Feigin, V.L., Vos, T., Nichols, E., Owolabi, M.O., Carroll, W.M., Dichgans, M., Deuschl, G., Parmar, P., Brainin, M., and Murray, C. 2020. The global burden of neurological disorders: translating evidence into policy. *The Lancet Neurology*. 19(3), 255–265.
- Feis, D.-L., Brodersen, K.H., Cramon, D.Y. von, Luders, E., and Tittgemeyer, M. 2013. Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. *Neuroimage*. 70, 250–257.
- Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., and Constable, R.T. 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*. 18(11), 1664–1671.
- Fischl, B. and Dale, A.M. 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*. 97(20), 11050–11055.
- Fjell, A.M., McEvoy, L., Holland, D., Dale, A.M., Walhovd, K.B., Initiative, A.D.N., et al. 2014. What is normal in normal aging? Effects of aging, amyloid and Alzheimer’s disease on the cerebral cortex and the hippocampus. *Progress in neurobiology*. 117, 20–40.
- Foland-Ross, L.C., Sacchet, M.D., Prasad, G., Gilbert, B., Thompson, P.M., and Gotlib, I.H. 2015. Cortical thickness predicts the first onset of major depression in adolescence. *International Journal of Developmental Neuroscience*. 46, 125–131.
- Fong, A.H.C., Yoo, K., Rosenberg, M.D., Zhang, S., Li, C.-S.R., Scheinost, D., Constable, R.T., and Chun, M.M. 2019. Dynamic functional connectivity during task performance

- and rest predicts individual differences in attention across studies. *NeuroImage*. 188, 14–25.
- Fox, M.D. and Raichle, M.E. 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience*. 8(9), 700–711.
- Franke, K. and Gaser, C. 2012. Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer’s disease. *GeroPsych*.
- Franke, K. and Gaser, C. 2019. Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained? *Frontiers in neurology*, 789.
- Franke, K., Gaser, C., Manor, B., and Novak, V. 2013. Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Frontiers in aging neuroscience*. 5, 90.
- Franke, K., Ziegler, G., Klöppel, S., Gaser, C., Initiative, A.D.N., et al. 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.
- Friston, K.J. 2003. Statistical parametric mapping. *Neuroscience databases: a practical guide*, 237–250.
- Gaser, C., Dahnke, R., Thompson, P.M., Kurth, F., Luders, E., and Initiative, A.D.N. 2022. CAT—A computational anatomy toolbox for the analysis of structural MRI data. *bioRxiv*, 2022–06.
- Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., and Initiative, A.D.N. 2013. BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer’s disease. *PloS one*. 8(6), e67346.
- Ge, Y., Grossman, R.I., Babb, J.S., Rabin, M.L., Mannon, L.J., and Kolson, D.L. 2002. Age-related total gray matter and white matter changes in normal adult brain. Part I: volumetric MR imaging analysis. *American journal of neuroradiology*. 23(8), 1327–1333.
- Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., and Frackowiak, R.S. 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage*. 14(1), 21–36.
- Guo, H., Zhang, F., Chen, J., Xu, Y., and Xiang, J. 2017. Machine learning classification combining multiple features of a hyper-network of fMRI data in Alzheimer’s disease. *Frontiers in neuroscience*. 11, 615.



- Hajek, T., Franke, K., Kolenic, M., Capkova, J., Matejka, M., Propper, L., Uher, R., Stopkova, P., Novak, T., Paus, T., et al. 2019. Brain age in early stages of bipolar disorders or schizophrenia. *Schizophrenia bulletin*. 45(1), 190–198.
- Han, L.K., Dinga, R., Hahn, T., Ching, C.R., Eyler, L.T., Aftanas, L., Aghajani, M., Aleman, A., Baune, B.T., Berger, K., et al. 2021a. Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group. *Molecular psychiatry*. 26(9), 5124–5139.
- Han, S., Chen, Y., Zheng, R., Li, S., Jiang, Y., Wang, C., Fang, K., Yang, Z., Liu, L., Zhou, B., et al. 2021b. The stage-specifically accelerated brain aging in never-treated first-episode patients with depression. *Human brain mapping*. 42(11), 3656–3666.
- Hashemi, R.H., Bradley, W.G., and Lisanti, C.J. 2012. MRI: the basics: The Basics. Lippincott Williams & Wilkins.
- Hastie, T., Tibshirani, R., Friedman, J.H., and Friedman, J.H. 2009. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. Springer.
- He, S., Grant, P.E., and Ou, Y. 2021a. Global-local transformer for brain age estimation. *IEEE Transactions on medical imaging*. 41(1), 213–224.
- He, S., Pereira, D., Perez, J.D., Gollub, R.L., Murphy, S.N., Prabhu, S., Pienaar, R., Robertson, R.L., Grant, P.E., and Ou, Y. 2021b. Multi-channel attention-fusion neural network for brain age estimation: Accuracy, generality, and interpretation with 16,705 healthy MRIs across lifespan. *Medical image analysis*. 72, 102091.
- Hobday, H., Cole, J.H., Stanyard, R.A., Daws, R.E., Giampietro, V., O’Daly, O., Leech, R., and Váša, F. 2022. Tissue volume estimation and age prediction using rapid structural brain scans. *Scientific Reports*. 12(1), 12005.
- Høgestøl, E.A., Kaufmann, T., Nygaard, G.O., Beyer, M.K., Sowa, P., Nordvik, J.E., Kolskår, K., Richard, G., Andreassen, O.A., Harbo, H.F., et al. 2019. Cross-sectional and longitudinal MRI brain scans reveal accelerated brain aging in multiple sclerosis. *Frontiers in neurology*. 10, 450.
- Hsu, W.-T., Rosenberg, M.D., Scheinost, D., Constable, R.T., and Chun, M.M. 2018. Resting-state functional connectivity predicts neuroticism and extraversion in novel individuals. *Social cognitive and affective neuroscience*. 13(2), 224–232.
- Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., and Gais, S. 2016. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Human brain mapping*. 37(5), 1842–1855.

- Jolliffe, I.T. 2002. Principal component analysis for special types of data. Springer.
- Jovicich, J., Czanner, S., Greve, D., Haley, E., Der Kouwe, A. van, Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., et al. 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage*. 30(2), 436–443.
- Kaczurkin, A.N., Raznahan, A., and Satterthwaite, T.D. 2019. Sex differences in the developing brain: insights from multimodal neuroimaging. *Neuropsychopharmacology*. 44(1), 71–85.
- Kapoor, S. and Narayanan, A. 2022. Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048*.
- Kaufmann, T., Meer, D. van der, Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A., et al. 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience*. 22(10), 1617–1623.
- Kazeminejad, A. and Sotero, R.C. 2019. Topological properties of resting-state fMRI functional networks improve machine learning-based autism classification. *Frontiers in neuroscience*. 12, 1018.
- Khosla, M., Jamison, K., Ngo, G.H., Kuceyeski, A., and Sabuncu, M.R. 2019. Machine learning in resting-state fMRI analysis. *Magnetic resonance imaging*. 64, 101–121.
- Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr, C.R., Ashburner, J., and Frackowiak, R.S. 2008. Automatic classification of MR scans in Alzheimer’s disease. *Brain*. 131(3), 681–689.
- Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M., et al. 2014. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophrenia bulletin*. 40(5), 1140–1153.
- Kruggel, F., Turner, J., Muftuler, L.T., Initiative, A.D.N., et al. 2010. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. *Neuroimage*. 49(3), 2123–2133.
- Ktena, S.I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., and Rueckert, D. 2018. Metric learning with spectral graph convolutions on brain connectivity networks. *NeuroImage*. 169, 431–442.
- Kumari, L.S. and Sundarajan, R. 2023. A review on brain age prediction models. *Brain Research*, 148668.

- Lancaster, J., Lorenz, R., Leech, R., and Cole, J.H. 2018. Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Frontiers in aging neuroscience*. 10, 28.
- Landeiro, V. and Culotta, A. 2018. Robust text classification under confounding shift. *Journal of Artificial Intelligence Research*. 63, 391–419.
- Lange, A.-M.G. de, Anatórk, M., Rokicki, J., Han, L.K., Franke, K., Alnæs, D., Ebmeier, K.P., Draganski, B., Kaufmann, T., Westlye, L.T., et al. 2022. Mind the gap: Performance metric evaluation in brain-age prediction. *Human Brain Mapping*. 43(10), 3113–3129.
- Lange, A.-M.G. de and Cole, J.H. 2020. Commentary: Correction procedures in brain-age prediction. *NeuroImage: Clinical*. 26.
- Lange, A.-M.G. de, Kaufmann, T., Meer, D. van der, Maglanoc, L.A., Alnæs, D., Moberget, T., Douaud, G., Andreassen, O.A., and Westlye, L.T. 2019. Population-based neuroimaging reveals traces of childbirth in the maternal brain. *Proceedings of the National Academy of Sciences*. 116(44), 22341–22346.
- Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., and Ferrante, E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*. 117(23), 12592–12594.
- Lawrence, I. and Lin, K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268.
- Le, T.T., Kuplicki, R.T., McKinney, B.A., Yeh, H.-W., Thompson, W.K., Paulus, M.P., and Investigators, T.1. 2018. A nonlinear simulation framework supports adjusting for age when analyzing BrainAGE. *Frontiers in aging neuroscience*. 10, 317.
- Lee, H.M., Gill, R.S., Fadaie, F., Cho, K.H., Guiot, M.C., Hong, S.-J., Bernasconi, N., and Bernasconi, A. 2020. Unsupervised machine learning reveals lesional variability in focal cortical dysplasia at mesoscopic scale. *NeuroImage: Clinical*. 28, 102438.
- Lee, W.H., Antoniadis, M., Schnack, H.G., Kahn, R.S., and Frangou, S. 2021. Brain age prediction in schizophrenia: Does the choice of machine learning algorithm matter? *Psychiatry Research: Neuroimaging*. 310, 111270.
- Levakov, G., Kaplan, A., Meir, A.Y., Rinott, E., Tsaban, G., Zelicha, H., Blüher, M., Ceglarek, U., Stumvoll, M., Shelef, I., et al. 2023. The effect of weight loss following 18 months of lifestyle intervention on brain age assessed with resting-state functional connectivity. *Elife*. 12, e83604.

- Levakov, G., Rosenthal, G., Shelef, I., Raviv, T.R., and Avidan, G. 2020. From a deep learning model back to the brain—Identifying regional predictors and their relation to aging. *Human brain mapping*. 41(12), 3235–3252.
- Lever, J., Krzywinski, M., and Altman, N. 2017. Points of significance: Principal component analysis. *Nature methods*. 14(7), 641–643.
- Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L.Q.R., Holmes, A.J., Ge, T., Patil, K.R., Jabbi, M., Eickhoff, S.B., et al. 2022. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*. 8(11), eabj1812. DOI: 10.1126/sciadv.abj1812.
- Liang, H., Zhang, F., and Niu, X. (2019). *Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders*. Tech. rep. Wiley Online Library.
- Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., et al. 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage*. 148, 179–188.
- Lienhart, R. and Maydt, J. (2002). “An extended set of haar-like features for rapid object detection”. In: *Proceedings. international conference on image processing*. Vol. 1. IEEE, pp. I–I.
- Llera, A., Wolfers, T., Mulders, P., and Beckmann, C.F. 2019. Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *Elife*. 8, e44443.
- Lombardi, A., Amoroso, N., Diacono, D., Monaco, A., Tangaro, S., and Bellotti, R. 2020. Extensive evaluation of morphological statistical harmonization for brain age prediction. *Brain sciences*. 10(6), 364.
- Lones, M.A. 2021. How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497*.
- Löwe, L.C., Gaser, C., Franke, K., and Initiative, A.D.N. 2016. The effect of the APOE genotype on individual BrainAGE in normal aging, mild cognitive impairment, and Alzheimer’s disease. *PloS one*. 11(7), e0157514.
- Luders, E., Cherbuin, N., and Gaser, C. 2016. Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners. *Neuroimage*. 134, 508–513.

- Marquand, A.F., Filippone, M., Ashburner, J., Girolami, M., Mourao-Miranda, J., Barker, G.J., Williams, S.C., Leigh, P.N., and Blain, C.R. 2013. Automated, high accuracy classification of parkinsonian disorders: a pattern recognition approach. *PloS one*. 8(7), e69237.
- Mateos-Pérez, J.M., Dadar, M., Lacalle-Aurioles, M., Iturria-Medina, Y., Zeighami, Y., and Evans, A.C. 2018. Structural neuroimaging as clinical predictor: A review of machine learning applications. *NeuroImage: Clinical*. 20, 506–522.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*. 54(6), 1–35.
- Meskaldji, D.-E., Preti, M.G., Bolton, T.A., Montandon, M.-L., Rodriguez, C., Morgenthaler, S., Giannakopoulos, P., Haller, S., and Van De Ville, D. 2016. Prediction of long-term memory scores in MCI based on resting-state fMRI. *NeuroImage: Clinical*. 12, 785–795.
- Miller, D.I. and Halpern, D.F. 2014. The new science of cognitive sex differences. *Trends in cognitive sciences*. 18(1), 37–45.
- Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*. 267, 1–38.
- Moazami, F., Lefevre-Utile, A., Papaloukas, C., and Soumelis, V. 2021. Machine learning approaches in study of multiple sclerosis disease through magnetic resonance images. *Frontiers in immunology*. 12, 700582.
- Molnar, C. 2019. Interpretable machine learning: a guide for making black box models explainable. 2019. URL <https://christophm.github.io/interpretable-ml-book>.
- Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A.D.N., et al. 2015. Machine learning framework for early MRI-based Alzheimer’s conversion prediction in MCI subjects. *Neuroimage*. 104, 398–412.
- More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S.B., Patil, K.R., Initiative, A.D.N., et al. 2023. Brain-age prediction: a systematic comparison of machine learning workflows. *NeuroImage*, 119947.
- More, S., Eickhoff, S.B., Caspers, J., and Patil, K.R. (2020). “Confound removal and normalization in practice: A neuroimaging based sex prediction case study”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 3–18.
- Mwangi, B., Tian, T.S., and Soares, J.C. 2014. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*. 12, 229–244.

- Nenning, K.-H. and Langs, G. 2022. Machine learning in neuroimaging: from research to clinical practice. *Die Radiologie*, 1–10.
- Nostro, A.D., Müller, V.I., Varikuti, D.P., Pläschke, R.N., Hoffstaedter, F., Langner, R., Patil, K.R., and Eickhoff, S.B. 2018. Predicting personality from network-based resting-state functional connectivity. *Brain Structure & Function*. 223(6), 2699–2719. DOI: 10.1007/s00429-018-1651-z.
- Ombao, H. 2016. Handbook of neuroimaging data analysis. Chapman and Hall/CRC. DOI: 10.1201/9781315373652.
- Ooi, L.Q.R., Chen, J., Zhang, S., Kong, R., Tam, A., Li, J., Dhamala, E., Zhou, J.H., Holmes, A.J., and Yeo, B.T. 2022. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *NeuroImage*. 263, 119636.
- Pain, O., Dudbridge, F., and Ronald, A. 2018. Are your covariates under control? How normalization can re-introduce covariate effects. *European Journal of Human Genetics*. 26(8), 1194–1201.
- Pardoe, H.R., Cole, J.H., Blackmon, K., Thesen, T., Kuzniecky, R., Investigators, H.E.P., et al. 2017. Structural brain changes in medically refractory focal epilepsy resemble premature brain aging. *Epilepsy research*. 133, 28–32.
- Pervaiz, U., Vidaurre, D., Woolrich, M.W., and Smith, S.M. 2020. Optimising network modelling methods for fMRI. *Neuroimage*. 211, 116604.
- Picco, L., Subramaniam, M., Abdin, E., Vaingankar, J.A., and Chong, S.A. 2017. Gender differences in major depressive disorder: findings from the Singapore Mental Health Study. *Singapore medical journal*. 58(11), 649.
- Pisharady, P.K., Eberly, L.E., Adanyeguh, I.M., Manousakis, G., Guliani, G., Walk, D., and Lenglet, C. 2023. Multimodal MRI improves diagnostic accuracy and sensitivity to longitudinal change in amyotrophic lateral sclerosis. *Communications Medicine*. 3(1), 84.
- Poldrack, R.A., Huckins, G., and Varoquaux, G. 2020. Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*. 77(5), 534–540.
- Pourhoseingholi, M.A., Baghestani, A.R., and Vahedi, M. 2012. How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench*. 5(2), 79.
- Richard, G., Kolskår, K., Ulrichsen, K.M., Kaufmann, T., Alnæs, D., Sanders, A.-M., Dørum, E.S., Sánchez, J.M., Petersen, A., Ihle-Hansen, H., et al. 2020. Brain age

- prediction in stroke patients: Highly reliable but limited sensitivity to cognitive performance and response to cognitive training. *NeuroImage: Clinical*. 25, 102159.
- Ritchie, S.J., Cox, S.R., Shen, X., Lombardo, M.V., Reus, L.M., Alloza, C., Harris, M.A., Alderson, H.L., Hunter, S., Neilson, E., et al. 2018. Sex differences in the adult human brain: evidence from 5216 UK biobank participants. *Cerebral cortex*. 28(8), 2959–2975.
- Rogenmoser, L., Kernbach, J., Schlaug, G., and Gaser, C. 2018. Keeping brains young with making music. *Brain Structure and Function*. 223, 297–305.
- Rosenberg, M.D., Finn, E.S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.T., and Chun, M.M. 2016. A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience*. 19(1), 165–171. DOI: 10.1038/nn.4179.
- Rosenblatt, J.D. 2016. Multivariate revisit to “sex beyond the genitalia”. *Proceedings of the National Academy of Sciences*. 113(14), E1966–E1967.
- Ruigrok, A.N., Salimi-Khorshidi, G., Lai, M.-C., Baron-Cohen, S., Lombardo, M.V., Tait, R.J., and Suckling, J. 2014. A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*. 39, 34–50.
- Sasse, L., Larabi, D.I., Omidvarnia, A., Jung, K., Hoffstaedter, F., Jocham, G., Eickhoff, S.B., and Patil, K.R. 2023. Intermediately synchronised brain states optimise trade-off between subject specificity and predictive capacity. *Communications biology*. 6(1), 705.
- Satterthwaite, T.D., Wolf, D.H., Roalf, D.R., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E.D., Elliott, M.A., Smith, A., Hakonarson, H., et al. 2015. Linked sex differences in cognition and functional connectivity in youth. *Cerebral cortex*. 25(9), 2383–2394.
- Savjani, R.R., Taylor, B.A., Acion, L., Wilde, E.A., and Jorge, R.E. 2017. Accelerated changes in cortical thickness measurements with age in military service members with traumatic brain injury. *Journal of neurotrauma*. 34(22), 3107–3116.
- Scheinost, D., Noble, S., Horien, C., Greene, A.S., Lake, E.M., Salehi, M., Gao, S., Shen, X., O’Connor, D., Barron, D.S., et al. 2019. Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage*. 193, 35–45.
- Scheinost, D., Stoica, T., Wasylink, S., Gruner, P., Saksa, J., Pittenger, C., and Hampson, M. 2014. Resting state functional connectivity predicts neurofeedback response. *Frontiers in Behavioral Neuroscience*. 8, 338. DOI: 10.3389/fnbeh.2014.00338.



- Seeman, M.V. 1997. Psychopathology in women and men: focus on female hormones. *American Journal of Psychiatry*. 154(12), 1641–1647.
- Siegel, J.S., Ramsey, L.E., Snyder, A.Z., Metcalf, N.V., Chacko, R.V., Weinberger, K., Baldassarre, A., Hacker, C.D., Shulman, G.L., and Corbetta, M. 2016. Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. *Proceedings of the National Academy of Sciences*. 113(30), E4367–E4376.
- Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., et al. 2013a. Resting-state fMRI in the human connectome project. *Neuroimage*. 80, 144–168.
- Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., et al. 2004. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*. 23, S208–S219.
- Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E., and Miller, K.L. 2019. Estimation of brain age delta from brain imaging. *Neuroimage*. 200, 528–539.
- Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., Nichols, T.E., Robinson, E.C., Salimi-Khorshidi, G., Woolrich, M.W., et al. 2013b. Functional connectomics from resting-state fMRI. *Trends in cognitive sciences*. 17(12), 666–682.
- Snoek, L., Miletić, S., and Scholte, H.S. 2019. How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage*. 184, 741–760.
- Soares, J.M., Magalhães, R., Moreira, P.S., Sousa, A., Ganz, E., Sampaio, A., Alves, V., Marques, P., and Sousa, N. 2016. A Hitchhiker’s guide to functional magnetic resonance imaging. *Frontiers in neuroscience*. 10, 515.
- Sone, D. and Beheshti, I. 2022. Neuroimaging-based brain age estimation: A promising personalized biomarker in neuropsychiatry. *Journal of Personalized Medicine*. 12(11), 1850.
- Sone, D., Beheshti, I., Maikusa, N., Ota, M., Kimura, Y., Sato, N., Koepp, M., and Matsuda, H. 2021. Neuroimaging-based brain-age prediction in diverse forms of epilepsy: a signature of psychosis and beyond. *Molecular psychiatry*. 26(3), 825–834.
- Sone, D., Beheshti, I., Shinagawa, S., Niimura, H., Kobayashi, N., Kida, H., Shikimoto, R., Noda, Y., Nakajima, S., Bun, S., et al. 2022. Neuroimaging-derived brain age is associated with life satisfaction in cognitively unimpaired elderly: A community-based study. *Translational psychiatry*. 12(1), 25.

- Steffener, J., Habeck, C., O'Shea, D., Razlighi, Q., Bherer, L., and Stern, Y. 2016. Differences between chronological and brain age are related to education and self-reported physical activity. *Neurobiology of aging*. 40, 138–144.
- Storelli, L., Azzimonti, M., Gueye, M., Vizzino, C., Preziosa, P., Tedeschi, G., De Stefano, N., Pantano, P., Filippi, M., and Rocca, M.A. 2022. A deep learning approach to predicting disease progression in multiple sclerosis using magnetic resonance imaging. *Investigative Radiology*. 57(7), 423–432.
- Sui, J., Jiang, R., Bustillo, J., and Calhoun, V. 2020. Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biological Psychiatry*. 88(11), 818–828. DOI: 10.1016/j.biopsych.2020.02.016.
- Tanveer, M., Ganaie, M., Beheshti, I., Goel, T., Ahmad, N., Lai, K.-T., Huang, K., Zhang, Y.-D., Del Ser, J., and Lin, C.-T. 2023. Deep learning for brain age estimation: A systematic review. *Information Fusion*.
- Tavares, V., Prata, D., and Ferreira, H.A. 2020. Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and Alzheimer's disease study. *Journal of Neuroscience Methods*. 334, 108565.
- Todd, M.T., Nystrom, L.E., and Cohen, J.D. 2013. Confounds in multivariate pattern analysis: theory and rule representation case study. *Neuroimage*. 77, 157–165.
- Tripepi, G., Jager, K.J., Dekker, F.W., and Zoccali, C. 2010. Stratification for confounding—part 1: The Mantel-Haenszel formula. *Nephron Clinical Practice*. 116(4), c317–c321.
- Tustison, N.J., Cook, P.A., Klein, A., Song, G., Das, S.R., Duda, J.T., Kandel, B.M., Strien, N. van, Stone, J.R., Gee, J.C., et al. 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage*. 99, 166–179.
- Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., et al. 2013. The WU-Minn human connectome project: an overview. *Neuroimage*. 80, 62–79.
- Van Gestel, H., Franke, K., Petite, J., Slaney, C., Garnham, J., Helmick, C., Johnson, K., Uher, R., Alda, M., and Hajek, T. 2019. Brain age in bipolar disorders: Effects of lithium treatment. *Australian & New Zealand Journal of Psychiatry*. 53(12), 1179–1188.
- Varikuti, D.P., Genon, S., Sotiras, A., Schwender, H., Hoffstaedter, F., Patil, K.R., Jockwitz, C., Caspers, S., Moebus, S., Amunts, K., et al. 2018. Evaluation of

- non-negative matrix factorization of grey matter in age prediction. *Neuroimage*. 173, 394–410.
- Varoquaux, G. 2018. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*. 180, 68–77.
- Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*. 145, 166–179.
- Venkataraman, A., Whitford, T.J., Westin, C.-F., Golland, P., and Kubicki, M. 2012. Whole brain resting state functional connectivity abnormalities in schizophrenia. *Schizophrenia research*. 139(1-3), 7–12.
- Vidal-Pineiro, D., Wang, Y., Krogsrud, S.K., Amlie, I.K., Baaré, W.F., Bartres-Faz, D., Bertram, L., Brandmaier, A.M., Dreven, C.A., Düzel, S., et al. 2021. Individual variations in ‘brain age’ relate to early-life factors more than to longitudinal brain change. *elife*. 10, e69995.
- Viola, P. and Jones, M. (2001). “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Ieee, pp. I–I.
- Wang, J., Zuo, X., and He, Y. 2010. Graph-based network analysis of resting-state functional MRI. *Frontiers in systems neuroscience*, 16.
- Wang, W.-Y., Yu, J.-T., Liu, Y., Yin, R.-H., Wang, H.-F., Wang, J., Tan, L., Radua, J., and Tan, L. 2015. Voxel-based meta-analysis of grey matter changes in Alzheimer’s disease. *Translational neurodegeneration*. 4(1), 1–9.
- Weber, K.A., Teplin, Z.M., Wager, T.D., Law, C.S., Prabhakar, N.K., Ashar, Y.K., Gilam, G., Banerjee, S., Delp, S.L., Glover, G.H., et al. 2022. Confounds in neuroimaging: A clear case of sex as a confound in brain-based prediction. *Frontiers in Neurology*. 13.
- Weis, S., Hodgetts, S., and Hausmann, M. 2019. Sex differences and menstrual cycle effects in cognitive and sensory resting state networks. *Brain and cognition*. 131, 66–73.
- Weis, S., Patil, K.R., Hoffstaedter, F., Nostro, A., Yeo, B.T., and Eickhoff, S.B. 2020. Sex classification by resting-state brain connectivity. *Cerebral cortex*. 30(2), 824–835.
- Werling, D.M. and Geschwind, D.H. 2013. Sex differences in autism spectrum disorders. *Current opinion in neurology*. 26(2), 146.
- Westman, E., Simmons, A., Zhang, Y., Muehlboeck, J.-S., Tunnard, C., Liu, Y., Collins, L., Evans, A., Mecocci, P., Vellas, B., et al. 2011. Multivariate analysis of MRI data

- for Alzheimer’s disease, mild cognitive impairment and healthy controls. *Neuroimage*. 54(2), 1178–1187.
- Weygandt, M., Hackmack, K., Pfüller, C., Bellmann–Strobl, J., Paul, F., Zipp, F., and Haynes, J.-D. 2011. MRI pattern recognition in multiple sclerosis normal-appearing brain areas. *PloS one*. 6(6), e21138.
- Weygandt, M., Hummel, H.-M., Schregel, K., Ritter, K., Allefeld, C., Dommes, E., Huppke, P., Haynes, J., Wuerfel, J., and Gärtner, J. 2015. MRI-based diagnostic biomarkers for early onset pediatric multiple sclerosis. *NeuroImage: Clinical*. 7, 400–408.
- Wiersch, L., Hamdan, S., Hoffstaedter, F., Votinov, M., Habel, U., Clemens, B., Derntl, B., Eickhoff, S.B., Patil, K.R., and Weis, S. 2023. Accurate sex prediction of cisgender and transgender individuals without brain size bias. *Scientific Reports*. 13(1), 13868.
- Wilms, R., Mäthner, E., Winnen, L., and Lanwehr, R. 2021. Omitted variable bias: a threat to estimating causal relationships. *Methods in Psychology*. 5, 100075.
- Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., and Marquand, A.F. 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience & Biobehavioral Reviews*. 57, 328–349.
- Wolpert, D.H. 1996. The lack of a priori distinctions between learning algorithms. *Neural computation*. 8(7), 1341–1390.
- Wrigglesworth, J., Ward, P., Harding, I.H., Nilaweera, D., Wu, Z., Woods, R.L., and Ryan, J. 2021. Factors associated with brain ageing-a systematic review. *BMC neurology*. 21(1), 312.
- Xu, C., Li, C., Wu, H., Wu, Y., Hu, S., Zhu, Y., Zhang, W., Wang, L., Zhu, S., Liu, J., et al. 2015. Gender differences in cerebral regional homogeneity of adult healthy volunteers: a resting-state fMRI study. *BioMed research international*. 2015.
- Yarkoni, T. and Westfall, J. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*. 12(6), 1100–1122.
- Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., et al. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*.

- Yeung, A.W.K., More, S., Wu, J., and Eickhoff, S.B. 2022. Reporting details of neuroimaging studies on individual traits prediction: a literature survey. *Neuroimage*, 119275.
- Yu, R., Zhang, H., An, L., Chen, X., Wei, Z., and Shen, D. 2017. Connectivity strength-weighted sparse group representation-based brain network construction for M CI classification. *Human brain mapping*. 38(5), 2370–2383.
- Zang, Y., Jiang, T., Lu, Y., He, Y., and Tian, L. 2004. Regional homogeneity approach to fMRI data analysis. *Neuroimage*. 22(1), 394–400.
- Zarogianni, E., Moorhead, T.W., and Lawrie, S.M. 2013. Towards the identification of imaging biomarkers in schizophrenia, using multivariate pattern classification at a single-subject level. *NeuroImage: Clinical*. 3, 279–289.
- Zhang, C., Cahill, N.D., Arbabshirani, M.R., White, T., Baum, S.A., and Michael, A.M. 2016. Sex and age effects of functional connectivity in early adulthood. *Brain connectivity*. 6(9), 700–713.
- Zhang, C., Dougherty, C.C., Baum, S.A., White, T., and Michael, A.M. 2018. Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. *Human brain mapping*. 39(4), 1765–1776.
- Zhang, X., Liang, M., Qin, W., Wan, B., Yu, C., and Ming, D. 2020. Gender differences are encoded differently in the structure and function of the human brain revealed by multimodal MRI. *Frontiers in Human Neuroscience*. 14, 244.
- Zhao, Y., Klein, A., Castellanos, F.X., and Milham, M.P. 2019. Brain age prediction: Cortical and subcortical shape covariation in the developing human brain. *Neuroimage*. 202, 116149.
- Zhou, X., Wu, R., Zeng, Y., Qi, Z., Ferraro, S., Xu, L., Zheng, X., Li, J., Fu, M., Yao, S., et al. 2022. Choice of voxel-based morphometry processing pipeline drives variability in the location of neuroanatomical brain markers. *Communications Biology*. 5(1), 913.
- Zhu, J., Li, Y., Fang, Q., Shen, Y., Qian, Y., Cai, H., and Yu, Y. 2021. Dynamic functional connectome predicts individual working memory performance across diagnostic categories. *NeuroImage: Clinical*. 30, 102593.

## Acknowledgements

This journey would not have been possible without the support and encouragement of many exceptional people who have played pivotal roles, both professionally and personally. First and foremost, I extend my deepest gratitude to Professor Simon B. Eickhoff for providing me with the opportunity to work in his esteemed research group. He has been an incredible mentor, and his dedication to excellence, scientific rigor, and methodological insights serve as an ongoing source of inspiration. Thank you for providing a safe and respectful work environment. I am thankful to Dr. Julian Caspers for his constructive feedback throughout my Ph.D. journey.

I give my most sincere gratitude to Dr. Kaustubh Patil, whose unwavering supervision, guidance, and support have been instrumental in my Ph.D. journey. His mentorship has fostered my growth as an independent researcher, encouraging me to question, think critically, and communicate ideas effectively. His commitment to perfectionism and a keen eye for detail have profoundly shaped my approach to research. Thank you for being patient with me and pushing me to be better.

I would also like to express my gratitude to my esteemed colleagues and collaborators, especially Felix Hoffstaedter, Georgios Antonopoulos, and Jianxio Wu, for their invaluable contributions that made my work possible. I owe a debt of gratitude to the entire Applied Machine Learning group for their engaging and insightful discussions, fun hackathon sessions, and social events. Special thanks to the administration group (Julia, Ute, and Anna) for always helping with tedious paperwork seamlessly and the data platform group for providing ample resources and support for the successful completion of projects.

The camaraderie among my fellow Ph.D. students has been a tremendous source of joy throughout this journey. I want to extend a heartfelt thank you to Lya Paas, Kyesam Jung, Julia Amunts, Marisa Heckner, Lisa Mochalski, Lisa Wiersch, Eliana Nicolaisen, and Mostafa Mahdipour. Your friendship brought an element of fun and adventure to this experience, from introducing me to the traditions of Karneval to enjoying the festive

Christmas celebrations and engaging in Ph.D. social events together. I have learned so much from each of you, and your support has been invaluable in helping me navigate the challenges of doctoral research. Thank you for making this journey so memorable and meaningful.

I'm also deeply grateful to my friends who, despite the distance, have been my strongest pillars of strength. A very special shoutout to Suvaranalata Xanthate, Monika Grewal, Kuldeep Gemini, Radhika Ranjan, Sumiti Saharan, Mithun James and Dennis Thomas. Thank you for constantly boosting my confidence. Your encouragement and belief in me have made a world of difference.

A special heartfelt thank you to Vaibhav Narang. Your presence has brought joy, appreciation, and motivation into my life, inspiring me to embark on this journey in the first place. Your unwavering belief in me has been a source of incredible strength and inspiration during my moments of doubt. Finally, my deepest appreciation and eternal gratitude go to my parents and my parents-in-law for their unflagging support and profound understanding. This journey has been enriched by the contributions of all these remarkable individuals, and I am deeply grateful for their presence in my life.