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

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

sMRI structural magnetic resonance imaging

SPM Statistical Parametric Mapping

SVR support vector regression

VBM voxel-based morphometry

WDCR whole-data confound regression

WMV white matter volume

Contents

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, Chyzhyk 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 ∼5–8 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

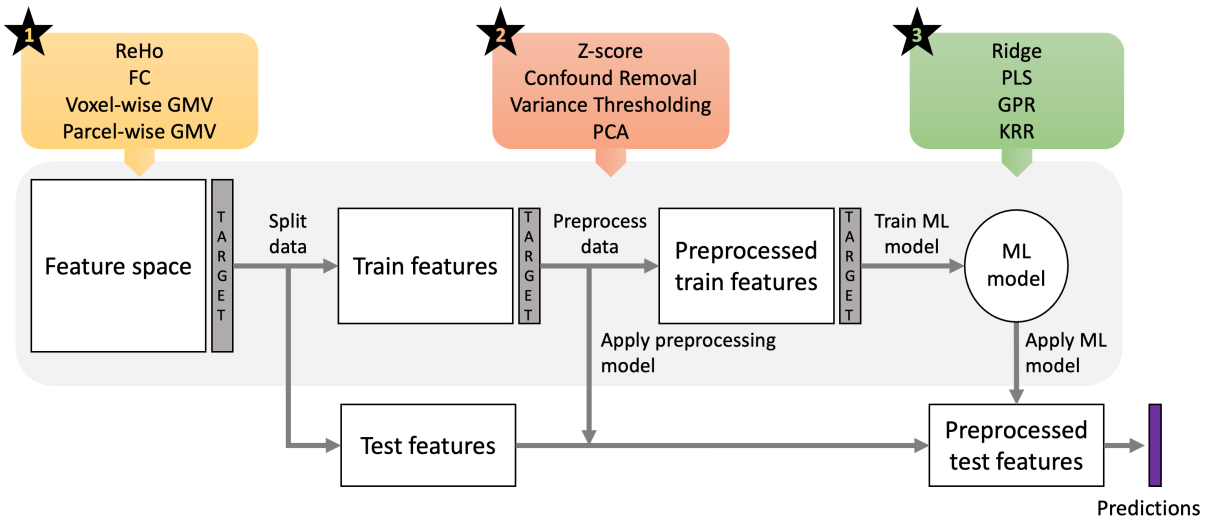

Figure 1: **Various steps in machine learning (ML) workflow design with some examples of (1) feature spaces, (2) preprocessing steps, and (3) ML algorithms.** First, the input data is split into training and test sets. Next, preprocessing steps are applied exclusively to the training features, and ML 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.

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.

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.

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.

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^3 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 Sundarrajan, 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., McMorrow, 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.

Chyzhyk, 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., Anatürk, 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., Guttmann, 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.

Kaczkurkin, 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 Sundarrajan, 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, Anatürk, 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., Antoniades, 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., Amlien, I.K., Baaré, W.F., Bartres-Faz, D., Bertram, L., Brandmaier, A.M., Drevon, 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.