1 2 9 0

UNIVERSIDADE Ð
COIMBRA

Maria de Fátima Machado Dias

Accurately Predicting Brain Age With Machine Learning: Implications For Biomarker Development

PhD Thesis in Informatics Engineering, Intelligent Systems, supervised by Professor Miguel Castelo Branco, Professor Paulo de Carvalho, and Doctor João Valente Duarte, and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra

January 2024

Accurately predicting brain age with machine learning: Implications for biomarker development

Maria de Fátima Machado Dias

January, 2024

The studies presented in this thesis were performed at the Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT) and Centre for Informatics and Systems of the University of Coimbra (CISUC) at the Informatics Department, Faculty of Sciences and Technology, University of Coimbra, Portugal.

This work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020; and CIBIT R&D Unit: UIDB/4050/2020 and UIDP4950/2020. FCT funded a doctoral grant to Maria de Fátima Dias (SFRH/BD/145798/2019).

Faculty of Sciences and Technology

University of Coimbra

Accurately predicting brain age with machine learning: Implications for biomarker development

Maria de Fátima Machado Dias

Supervisor:
Miguel Castelo-Branco

Co-supervisors:
Paulo de Carvalho
João Valente Duarte

Dissertation presented to obtain a Ph.D. degree in Informatics Engineering, Intelligent Systems, at the Faculty of Sciences and Technology of the University of Coimbra

Dissertação de Doutoramento apresentada à Faculdade de Ciências e Tecnologia da Universidade de Coimbra, para prestação de provas de Doutoramento em Engenharia Informática, Sistemas Inteligentes

January, 2024

Abstract

Brain age gap estimation (BrainAGE) is a putative ageing biomarker that aims to identify the onset of pathological ageing of the brain and monitor its progress. Diagnosing age-related brain conditions in a preclinical stage could enable early interventions, decreasing the disease burden. BrainAGE emerges as a promising ageing biomarker that seems to tackle biological ageing mechanisms. Its value is sensitive to lifestyle activities and numerous pathological conditions.

BrainAGE primarily models healthy brain ageing as a way to detect pathological deviations. Healthy brain ageing can be modelled by machine learning algorithms that learn ageing patterns from structural magnetic resonance imaging (MRI) data. The BrainAGE provides information about the difference between the predicted and the chronological age. A wide range of approaches have been considered to model brain age, from shallow to deep learning. However, current models lack generalisation when applied to data obtained in acquisition settings different from those used to train the model. Moreover, BrainAGE is sensitive to detecting changes in multiple diseases but lacks specificity, which might be essential for its adoption in clinical practice. This thesis aimed to improve the BrainAGE generalisability and specificity.

One hypothesis for the generalisability problem is that MRI data may contain scanner artefacts, which leads to a bias in machine learning models towards acquisition settings. Preprocessing has a preponderant role in reducing scanner artefacts. Nevertheless, there is no preprocessing gold standard in neuroimaging. FreeSurfer and SPM emerge as two frameworks that are extensively used in BrainAGE. The first study in this thesis assessed the reproducibility between the FreeSurfer and computational anatomy toolbox (CAT12), an SPM toolbox for structural data, and the reliability of each one. The results outlined that the reliability of the frameworks differs, and CAT12 outperformed FreeSurfer. Therefore, CAT12 was selected to preprocess the data.

Overfitting to training data could also cause the generalisability problem. Deep learning is extensively applied in the BrainAGE field with state-of-the-art performance. Nonetheless, the performance of these models often decreases when tested on an external test set. Transfer learning from a 3D-Convolutional Autoencoder (3D-CAE), an unsupervised model, was considered to overcome this limitation. The results outlined that reusing weights from pre-trained 3D-CAE improves generalisation on an external test set.

The specificity of BrainAGE might be improved by using multiple sources of information. The preprocessing of MRI data involves registering the image into a template. The transformations applied to overlap the two images are designated by deformation fields. Most studies use minimally preprocessed T_1-weighted images, or grey matter (GM) or white matter (WM) segmented images, to predict brain age. Few studies consider deformation fields to model brain age, and results are inconsistent. One study of this thesis focuses on comprehensively comparing the performance of the deformation fields with the GM, WM and cerebrospinal fluid (CSF). The results outlined that deformation fields yield better performance than WM and CSF. Furthermore, combining deformation fields with GM improves performance compared to GM alone.

Finally, the last study assessed sensitivity maps to explain the model predictions and increase BrainAGE specificity. All the previous analyses were combined. Five brain age models were trained, leveraging transfer learning from the 3D-CAE, using the following inputs: minimally processed, GM, WM, CSF and deformation results. In general, the BrainAGE was statistically significant in all models and conditions. The results evidenced model sensitivity, but lack specificity. The analysis of sensitivity maps revealed different patterns across the different diseases and input types, thus contributing to a biological explanatory framework. Explainability from multiple sources might provide insights on BrainAGE specificity.

This thesis contributed to the BrainAGE field on two axes: generalisability and specificity. The former was addressed using the preprocessing framework with higher reliability and transfer learning from an agnostic model. The latter was attained by including deformation fields in brain ageing modelling and exploring sensitivity maps for differential diagnosis.

Keywords: Ageing, Biomarker, Machine Learning, Neuroimaging

Resumo

Estimativa da diferença de idade cerebral (BrainAGE) é um possível biomarcador de envelhecimento que visa identificar o início do envelhecimento patológico e monitorar seu progresso. O diagnóstico de condições relacionadas com a idade em estágio pre-clínico poderá possibilitar intervenções precoces, diminuindo o impacto da doença. BrainAGE surge como um biomarcador de envelhecimento promissor que parece estar relacionado com os mecanismos biológicos de envelhecimento. O seu valor é sensível ao estilo de vida e a inúmeras condições patológicas.

BrainAGE modela o envelhecimento cerebral saudável como forma de detectar os desvios patológicos. O envelhecimento saudável do cérebro pode ser modelado por algoritmos de aprendizagem máquina, que aprendem padrões de envelhecimento a partir de dados estruturais de ressonância magnética. BrainAGE corresponde à diferença entre a idade prevista e a cronológica. Uma ampla gama de abordagens tem sido considerada para modelar a idade do cérebro, usando abordagens tradicionais (*shallow learning*) ou mais complexas (*deep learning*). No entanto, os modelos atuais carecem de generalização quando aplicados a dados obtidos em condições de aquisição diferentes daquelas utilizadas para treinar o modelo. Adicionalmente, o BrainAGE é sensível a múltiplas doenças e, por isso, tem uma reduzida especificidade, o que pode dificultar a sua adoção na prática clínica. Esta tese teve como objetivo melhorar a generalização e especificidade do BrainAGE.

O problema de generalização pode advir dos dados de ressonância magnética conterem artefatos específicos das condições da aquisição, o que leva a um viés nos modelos em relação às configurações de aquisição. O pre-processamento tem um papel preponderante na redução destes artefatos. No entanto, não existe uma abordagem padrão definida para o pre-processamento em neuroimagem. Duas ferramentas amplamente utilizadas em BrainAGE são o FreeSurfer e o SPM. O primeiro estudo desta tese avalia a reprodutibilidade entre o FreeSurfer e a CAT12, uma ferramenta do SPM para dados estruturais, e a fiabilidade de cada ferramenta. Os resultados mostram que a fiabilidade das ferramentas difere, a CAT12 tem um melhor desempenho que o FreeSurfer. Pelo que, a CAT12 foi selecionada para pre-processar os dados.

O *overfitting* aos dados de treino pode também ser a causa do problema de generalização. As abordagens de *deep learning* têm sido amplamente utilizadas em BrainAGE. No entanto, o desempenho destes modelos diminui quando testados em dados independentes. Para ultrapassar este entrave, foi avaliada a transferência de conhecimento de um *3D Convolucional Autoencoder (3D-CAE)*, um modelo não supervisionado. Os resultados revelam que a reutilização de pesos do 3D-CAE melhora a generalização do modelo.

A especificidade do BrainAGE pode ser melhorada usando múltiplas fontes de informação. O pre-processamento de imagens de ressonância magnética envolve o registo das imagens com uma imagem base. As transformações aplicadas para sobrepor as duas imagens são designadas por campos de deformação. A maioria dos estudos usa imagens minimamente pre-processadas, substância cinzenta (GM) ou substância branca (WM), para prever a idade cerebral. Poucos estudos consideram campos de deformação para modelar a idade cerebral e os resultados são inconsistentes. Um estudo desta tese concentra-se na comparação do desempenho dos campos de deformação com a GM, WM e líquido cefalorraquidiano (CSF). Os resultados demonstraram que os campos de deformação apresentam melhor desempenho que WM e CSF. Adicionalmente, a combinação dos campos de deformação com GM melhora o desempenho do modelo em comparação com a utilização de apenas GM.

Finalmente, o último estudo avaliou a utilização de mapas de sensibilidade para explicar as previsões do modelo e aumentar a especificidade do BrainAGE. Nesta análise, todas os estudos anteriores foram combinados. Foram treinados cinco modelos para prever a idade cerebral, utilizando os pesos do 3D-CAE, usando os seguintes tipos de dados: imagens com processamento mínimo, GM, WM, CSF e campos de deformação. De modo geral, o BrainAGE foi estatisticamente significativo em todos os modelos e condições. Os resultados evidenciam a sensibilidade do biomarcador e sua falta de especificidade. A análise dos mapas de sensibilidade revelou diferentes padrões nas diferentes doenças e tipo de dados. Desta forma, a explicação das previsões, utilizando múltiplos tipos de dados, pode aumentar a especificidade do BrainAGE.

Esta tese contribuiu para o BrainAGE em dois eixos: generalização e especificidade. O primeiro foi abordado usando a ferramenta de pre-processamento com maior fiabilidade e transferência de conhecimento de um modelo agnóstico. Em relação à especificidade foram incluídos os campos de deformação na modelação do envelhecimento cerebral e explorados os mapas de sensibilidade para o diagnóstico diferencial.

Palavras-chave: Envelhecimento, Biomarcador, Aprendizagem máquina, Neuroimagem.

Acknowledgements

To my supervisors, Miguel Castelo-Branco, Paulo de Carvalho, and João Duarte, I express my thanks for all the guidance and advice.

I gratefully acknowledge FCT for finance this work through the Ph.D. grant SFRH/BD/145798/2019.

I acknowledge the Coimbra Institute for Biomedical Imaging and Translational Research (CIBIT), Centre for Informatics and Systems of the University of Coimbra (CISUC) and Associate Laboratory for Intelligent Systems (LASI) for support my research.

I would also like to acknowledge the Laboratory for Advanced Computing at the University of Coimbra (LCA) for providing essential computing resources that contributed to the results presented in this thesis. I sincerely thank Tiago Cerqueira for his guidance in the supercomputer usage.

To all my friends, I express my gratitude for their encouraging words and support.

I express my deepest gratitude to my parents and sisters for their unwavering love and support throughout my life.

Special thanks to Bruno for his love, incredible support and insightful discussions during the development of this thesis.

Lastly, I express my profound gratitude to my beloved daughter, Alice, whose kind smile has brightened my days throughout this final year.

Contents

Acronyms

3D-CAE	3D-convolutional autoencoder 93, 94, 105, 118–120, 135–137
3DCAE-MRI	3D-convolutional autoencoder magnetic resonance imaging 92–97, 99–111, 211, 212
ABIDE	autism brain imaging data exchange 48, 50, 60, 93, 108, 114, 118, 120
AD	Alzheimer's disease 2, 7, 9, 11, 12, 38, 42, 93, 107, 109, 116–120, 122, 124, 127–129, 131, 132, 136, 211, 224, 225
ADNI	Alzheimer's Disease Neuroimaging Initiative 93, 108, 114, 118, 120
ANCOVA	analysis of covariance 52, 62, 66, 120–122, 124, 127
ASD	autism spectrum disorder 46, 50
BrainAGE	brain age gap estimation 1–5, 19, 21, 35, 38, 42, 44, 76, 93, 109, 116–118, 120–122, 124, 131–133, 135–137
Cam-CAN	Cambridge Centre for Ageing and Neuroscience dataset 119, 120
CAT12	computational anatomy toolbox 3, 4, 46–48, 50, 51, 54, 56, 58, 60, 63, 65, 68, 70–73, 78, 79, 109, 119, 135, 136
CCA	canonical correlation analysis 26, 29
CI	confidence interval 56, 62, 65
CNN	convolution neural networks 21, 32–34, 36, 93–95, 97, 99, 103, 106–108, 111, 120
CSF	cerebrospinal fluid 9, 12, 14, 20, 35, 38, 76, 78, 79, 84, 85, 87, 88, 117, 119, 121, 122, 124, 127, 129–132, 136, 224
DenseNet	dense convolutional network 33
DL	deep learning 92, 94, 95, 97, 103, 105–108, 113

FCP1000	1000 Functional Connectomes Project 108, 114, 118, 120
GH	Guy's Hospital 50, 66, 68, 72, 78, 82, 84, 109, 119, 120
GM	grey matter 7–12, 20, 34, 35, 76–79, 81, 84–88, 109, 110, 112, 117, 119, 121, 122, 124, 127, 129, 131, 132, 136, 137, 224
GPR	gaussian process regression 29
GSP	Brain Genomics Superstruct 108, 114, 118, 120
HH	Hammersmith Hospital 50, 66, 68, 72, 78, 82, 84, 109, 119–122, 124
ICC	intraclass correlation coefficient 51, 52, 54, 58, 60, 62, 63, 66, 68, 71
IOP	Institute of Psychiatry 50, 66, 68, 72, 78, 82, 84, 88, 105, 109, 112, 119–122
IXI	Information eXtraction from Images 48, 50, 63, 68, 78, 79, 109, 112, 114, 119–122
MAE	mean absolute error 22, 23, 25, 26, 29, 32, 34–38, 43, 81, 82, 84–86, 88, 95, 97, 105, 106, 108, 111, 113, 121, 122, 211
MCI	mild cognitive impairment 2, 38, 42
MR	magnetic resonance 13, 17, 19, 37, 50
MRI	magnetic resonance imaging 2, 4, 7, 12, 13, 15, 16, 20, 34–36, 43, 46–48, 50, 51, 56, 76–79, 81, 92–94, 108, 109, 116, 119, 137
MSE	mean squared error 34
NMF	non-negative matrix factorization 26, 29
OASIS	Open Access Series of Imaging Studies 48, 56, 63, 66, 68, 72, 93, 108, 114, 118, 120
PCA	principal component analyis 21, 26, 29, 79, 81, 84, 88, 94–97, 99, 101, 102, 105, 106, 112, 212
ReLU	rectified linear units 33
ResNet	residual network 33, 93
RF	radio frequency 13–17
rm-ANOVA	repeated measures analysis of variance 82, 84, 85, 95, 99, 102, 113
RMSE	root mean squared error 34

ROBEX	Robust Brain Extraction 4
ROI	region-of-interest 20, 23–25, 29, 46, 47, 51, 52, 54, 56, 58, 60, 63, 65, 66, 68, 70–73, 121, 127, 131
RVM	relevant vector machine 94, 96, 97, 99, 101, 102, 105, 106, 112, 212
RVR	relevant vector regression 29, 32, 76, 79, 81
SFCN	simple fully convolutional network 33, 94–97, 99–108, 110–112, 211, 212
SNR	signal-to-noise ratio 51, 52, 54, 62, 66, 71, 72
SVR	support vector regression 25, 29
T2D	type 2 diabetes 7, 9–12, 42, 116–120, 122, 124, 127–132, 136, 137
TE	time of echo 14–16, 78
TPM	tissue probability maps 50, 109
TR	time of repetition 16, 78
VGG	visual geometry group network 33, 93
WHO	World Health Organization 1, 10, 11
WM	white matter 7–12, 20, 35, 76, 78, 79, 84–88, 119, 121, 122, 124, 127, 129, 131, 132, 136, 224

List of Figures

List of Tables

Chapter 1

Introduction

Age-related diseases have been increasing over the past years [5, 6]. The search for biomarkers that can predict the healthy and pathological ageing of the brain has been gathering a lot of attention [7]. Brain age gap estimation (BrainAGE) is a putative biomarker which tackles brain ageing trajectory by computing the difference between the estimation of brain age and the chronological age [8]. BrainAGE has been shown to be increased in multiple pathological conditions suggesting an acceleration of the ageing process as a common mechanism [9]. This chapter delves into the motivation for an ageing biomarker, provides an overview of BrainAGE and its existing limitations, and outlines the objectives of this thesis.

1.1 Motivation

It is anticipated that by 2050 the number of individuals over the age of 65 years will surpass the number of both adolescents and young adults [10]. With the rising number of elderly individuals, the burden of deaths and disability caused by age-related conditions, such as neurodegenerative dementias, is increasing. The number of people diagnosed with dementia worldwide in 2019 was 57.4 million, and this number is expected to rise up to 152.8 million in 2050 [6].

The prevalence of age-related conditions is being recognised as a global public health problem. The World Health Organization (WHO) has designated the decade from 2020 to 2030 as the decade of healthy ageing. *Healthy ageing* is defined by WHO as "the process of developing and maintaining functional abilities that enable well-being in older age" [11]. Several factors seem to positively impact brain health. Education [12], physical exercise [13, 14], meditation [15], among others are associated with neuroprotective mechanisms and potentially slower ageing processes. Other lifestyle factors, such as tobacco [16] and alcohol consumption [17, 18], as well as age-related diseases [19] seem to accelerate the ageing process. Thus, a biomarker sensitive to deviations in brain ageing trajectory might detect, in an early stage, diseases related to pathological ageing. In a pre-clinical stage, the detection of age-related diseases could prove valuable in personalised medicine, opening a window of opportunity for early interventions designed to increase health span or even decelerate ageing.

1.1.1 BrainAGE as a putative brain ageing biomarker

BrainAGE aims to gauge brain ageing to identify early deviations from the typical healthy trajectory. Over the past few years, this putative biomarker has gathered significant attention within the scientific community [20, 21]. In a nutshell, BrainAGE (1) models healthy brain ageing and (2) compute the difference between the brain age estimation and chronological age. To model healthy brain ageing, a machine learning algorithm is trained using data from individuals who have been confirmed to be healthy. It is assumed that, in healthy individuals, the brain age is equal to the chronological age. Different sources of information have been explored in BrainAGE [21]; this thesis focuses on using structural brain images acquired with magnetic resonance imaging (MRI), a non-invasive neuroimaging technique. Structural images were selected due to their high resolution and current use in clinical practice [22].

The rationale behind this putative biomarker is that a delayed ageing process results in a negative BrainAGE. In contrast, accelerated ageing is manifested as a positive value. A negative BrainAGE has been reported in individuals with higher education level [23], as well as in individuals who practice physical exercise [24] and meditation [25] regularly. As discussed above, these activities are related to neuroprotective mechanisms and, consequently, might decelerate brain ageing. Furthermore, tobacco [26, 27] and alcohol consumption [27] yield a positive BrainAGE. Moreover, BrainAGE is positive in multiple pathological conditions, such as Alzheimer's disease (AD), mild cognitive impairment (MCI), schizophrenia and epilepsy [9]. Thus, BrainAGE seems to translate the biological mechanisms underlying the brain ageing process and might be able to capture the transition from healthy to pathological ageing.

1.1.2 Limitations on BrainAGE

The performance of brain age models on test sets acquired in the same conditions as training data is, in general, similar to the performance on the validation test set. However, the model's performance on data collected in different acquisition settings is lower; in some cases, the error is two- or three-fold higher [28–31]. Therefore, brain age models are not reliable. To be used on other sites, these models must be retrained and validated with local data from healthy individuals. Therefore, the usage of BrainAGE in clinical practice is compromised by this generalisation issue. Consequently, it is of uttermost importance to address this limitation.

BrainAGE has a high sensitivity for different pathologies, nevertheless, its specificity is low [9]. Currently, it solely identifies deviations without necessarily specifying the underlying disease. Therefore, approaches should be developed to detect the deviations and suggest the specific pathology causing the abnormal ageing. A recent research trend aims to overcome this problem by predicting local brain age rather than a global age, yet the performance of these models still needs to improve.

1.2 Goals and main contributions

This thesis aims to add insights to the BrainAGE as a ageing biomarker by improving its generalisability and specificity.

On the generalisability axis, two studies were performed; one focused on preprocessing, while the other focused on model training. Concerning the preprocessing, a comparative analysis of two commonly employed MRI preprocessing frameworks [20], computational anatomy toolbox (CAT12) and Freesurfer was performed, focusing on two critical aspects: reliability and reproducibility. The preprocessing phase plays a crucial role in reducing noise and removing non-brain tissue, which is essential to ensure the removal of data artefacts that could mislead the learning process [20]. The framework comparison was conducted using cortical thickness, a common feature in neuroimaging studies. This investigation selects the more reliable preprocessing framework, which is subsequently applied in the following studies. Another explored aspect on the generalisability axis was the usage of transfer learning as a training strategy for deep learning models. Deep learning models are data-driven, with remarkable performance on different tasks, namely object recognition and language processing [32]. Nevertheless, these models demand high amounts of data. This requirement is often a constraint in the neuroimaging field, where the data is relatively scarce compared to natural images, in which millions of labelled images are available. Therefore, we explore the potential of transfer learning from an autoencoder to train brain age models and assess the impact of this training strategy on performance and generalisability.

The BrainAGE specificity was addressed by proposing a new feature, the deformation fields, and exploring the explainability of brain age models. The deformation fields were considered to predict brain age and were compared to segmented images commonly used in this field. The exploration of novel features can enhance the model's specificity and help to understand the underlying biological processes. In this case, it sheds light on how morphology changes with age. Finally, the last study integrated all the previous analysis and assessed whether the explainability of brain age predictions could increase the BrainAGE specificity. The explainability highlights the regions most critical for age prediction and reveals the brain areas that exerted different influences when comparing the predictions in individuals with a particular disease with healthy controls. Identifying the regional pattern of accelerated ageing may unveil the underlying pathologies and add specificity to the model.

1.2.1 Scientific Outcomes

Throughout the course of this thesis, different achievements have been reached. Specifically, four articles were authored, with two already published and two currently undergoing the revision process. Most of the data used in the scope of this thesis are from open-source repositories. The code for each study is publicly available on GitHub (`https://mfmachado.github.io/brainage/`). Moreover, some contributions were made to the nipype library [33], which included the addition of CAT12 and Robust Brain Extraction (ROBEX) [34] to the supported preprocessing frameworks.

1.2.1.1 Peer-reviews journal articles

- Dias, Maria de Fátima Machado, Paulo Carvalho, Miguel Castelo-Branco, and João Valente Duarte. "Cortical thickness in brain imaging studies using freesurfer and cat12: A matter of reproducibility." Neuroimage: Reports 2, no. 4 (2022): 100137.
- Dias, Maria de Fátima Machado, Paulo Carvalho, João Valente Duarte, and Miguel Castelo-Branco. "Deformation fields: a new source of information to predict brain age." Journal of Neural Engineering 19, no. 3 (2022): 036025.
- Dias, Maria de Fátima Machado, Tiago FT Cerqueira, João Valente Duarte, Miguel Castelo-Branco, and Paulo Carvalho. "3DCAE-MRI: Overcoming Data Availability Limitations in Small Sample MRI Studies." Scientific Reports (2023). [under revision]
- Dias, Maria de Fátima Machado, João Valente Duarte, Paulo de Carvalho and Miguel Castelo-Branco. "Unravelling pathological ageing with brain age in Alzheimer's Disease, Diabetes, and Schizophrenia." Brain Communications (2023). [under revision]

1.3 Thesis outline

The structure of this thesis is organised as follows:
Chapter 2 introduces fundamental concepts related to neural ageing and MRI.
Chapter 3 provides an overview of the state-of-the-art in brain age models, starting with a review of traditional machine learning models and deep learning models. The application of BrainAGE in multiple diseases is also discussed.
Chapter 4 is a published paper that compares preprocessing results using two widely adopted frameworks in neuroimaging.

Chapter 5 proposes the use of deformation fields as a feature for predicting brain age, this study is a published paper.
Chapter 6 is a paper under review that analyses the implications of transfer learning regarding generality and performance in deep learning models.
Chapter 7 explores sensitivity maps to increase the specificity of BrainAGE. This chapter is a paper under review.
Chapter 8 concludes the thesis by summarising the essential findings and discussing potential future directions.

Chapter 2

Background Concepts

This chapter overviews fundamental concepts of neural ageing and magnetic resonance imaging (MRI). Section 2.1 delves into neural ageing and is divided into two subsections. Subsection 2.1.1 comments on the changes during healthy ageing. In contrast, subsection 2.1.2 focuses on the disruptions that specific pathologies, particularly Alzheimer's disease (AD), type 2 diabetes (T2D) and schizophrenia, cause in the brain compared to healthy controls. The assessment of brain structure is performed with MRI, a neuroimaging technique. Section 2.2 describes the principles of structural MRI acquisition.

2.1 Neural ageing

The brain ageing process can be categorised into seven general hallmarks: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient-sensing, mitochondrial dysfunction and altered intercellular communication [19]. These hallmarks play a crucial role in the brain's healthy and pathological ageing trajectory. Pathological ageing is characterised by an acceleration of the ageing process in some hallmarks and/or a disruption in the biological mechanisms compared to the healthy ageing process [19]. The following two subsections will delve into the healthy and pathological ageing changes.

2.1.1 Healthy ageing

Healthy brain ageing is characterised by cerebral atrophy that results in four main morphometric changes: volume loss, cortical thinning, sulcal widening and ventricular enlargement [1]. Figure 2.1 illustrates the hallmarks of cerebral atrophy. In early adulthood, the brain's total volume decreases by 0.2% per year; this value increases to 0.5% at the beginning of late adulthood, around 60 years [35]. Higher rates of brain atrophy are related to neurodegenerative diseases and cognitive decline. Both grey matter (GM) and white matter (WM) loss contribute to the global volume decrease.

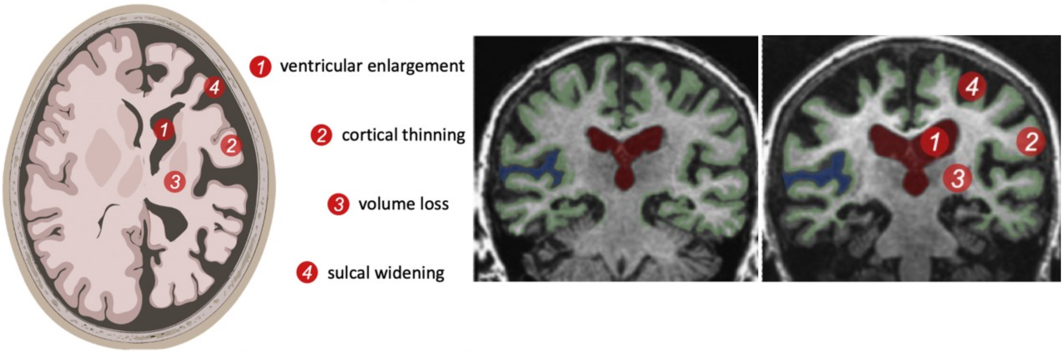


Figure 2.1: Hallmarks of brain atrophy: volume loss, cortical thinning, sulcal widening and ventricular enlargement. Adapted from [1] with permission of the authors.

GM comprises the outer layer of the cerebrum, known as the cerebral cortex, and the deep subcortical structures. The cortex is a highly folded structure that plays a central role in cognitive functions, processing sensory input, attention and decision-making, language comprehension and production, among others [36]. The deep subcortical structures encompass the thalamus, basal ganglia, hippocampus and amygdala. These structures have crucial functions in memory, emotions, hedonic pleasure, and neurohormone production [37, 38]. GM primarily consists of neuronal cell bodies; this tissue decreases, on average, around 0.5% per year [39]. The decline seems to be linked to the depletion of neurophil, the contraction of neurons, modifications in dendrite structure, and a decrease in synapse number [39, 40]. The cortical volume linearly declines non-uniformly across the brain; certain regions seem more susceptible to atrophy than others [41–43]. The same applies to subcortical structures, and each has an ageing pattern [44]. Beneath the cortex lies the cerebral WM, the distance between the outermost layer of the cortex, known as the pial surface, and the WM boundary is referred to as cortical thickness. Ageing results in a non-uniform gradual thinning of the cortex [42, 43, 45–48], and a non-uniformly gradually decline of the total area of the cortex outer layer [43, 46, 49]. The cortical and subcortical volume decline of GM and the cortical thinning have been suggested to be related to cognitive decline and memory impairments [40]. Moreover, as depicted in Figure 2.1, the grooves or fissures on the cortical surface [42, 50, 51] become wider and, the cortex becomes less folded and convoluted [46, 52, 53]. Two measures have been extensively considered to portray the convolution patterns: gyrification ratio and fractal dimension. The former correspondents to the ratio of the inner surface divided by the outer cortex surface, research suggests that there is a decrease in gyrification ratio with age [46,52,53]. The fractal dimension characterises the complexity of the cortical folding patterns by measuring the level of self-similarity in the cortical surface. Studies suggest that fractal dimension negatively correlates with age, indicating that ageing leads to a less convoluted and folded cortex [54–57].

WM mainly comprises myelinated axons; myelin protects the neurons and speeds the information flow along the axon [58]. The ageing process affects the WM volume and myelin integrity. Myelin is continuously lost and renewed throughout life, but the rate of remyelination is not constant [59–61]. The pattern of myelin distribution across the lifespan follows an inverted U-shape curve. Until mid-adulthood, the myelin production exceeds loss, increasing the total amount of WM. However, production slows down after reaching 40-50 years of age, and the loss of myelin surpasses its renewal [62, 63]. The loss of WM is not uniform across the brain [1], and in contrast to GM, it decreases significantly in late adulthood. The total volume loss of WM at 70 years is around 6% compared to the WM volume at 30 years; this value raises to values between 21.6% and 25.0% at 80 years [1]. Moreover, the ageing process also results in a decrease of WM integrity after 30 years, particularly in the anterior part of the brain [1]. A hypothesis proposed to explain the pronounced degradation in this region suggests that the myelin generated at older ages is thinner and weaker, leading to quicker deterioration [64–66]. Another notable change in ageing WM is the presence of WM lesions. These lesions result from ischemic damage in the brain, which can be attributed to the ageing process, hypertension, or vascular diseases [67]. Two distinct theories attempt to explain these lesions: one suggests that they are the result of continuous demyelination [68], whereas the other hypothesises that changes in the blood-brain barrier lead to reduced brain oxygenation and subsequent ischemia [69].

The loss of brain tissue culminates with the ventricle enlargement [41, 70–72], as shown in Figure 2.1. Ventricles are four interconnected cavities in the brain filled with cerebrospinal fluid (CSF). The loss of volume liberates space, which enables the increase in volume.

2.1.2 Pathological ageing

The transition from healthy to pathological ageing has yet to be understood. Certain medical conditions cause an acceleration of the ageing process [19]. Three pathologies are studied in the scope of this thesis: schizophrenia, T2D and AD. Each disease manifests in a distinct period of the lifetime, and all are related to atypical ageing of the brain. In the following subsections, a brief overview is provided along with the associated morphometric changes.

2.1.2.1 Schizophrenia

Schizophrenia is a chronic mental disorder of unknown aetiology characterised by psychotic episodes. The psychosis often involves paranoid delusions and auditory hallucinations. Other symptoms involve disorganised speech and/or behaviour and negative symptoms [73, 74]. The disease usually manifests in late adolescence or early adulthood. Although the pathogenesis and the pathophysiology remain to be uncovered, some evidence suggests the disease can be labelled as a neurodevelopmental disorder [73–75]. Schizophrenia affects nearly 24 million individuals globally [76], causing some incapacitation, with only approximately 14% of patients successfully reintegrating into society within a five-year period [77]. The disease was first identified in the nineteenth century by Kraepelin, who defined it as premature dementia [78]. This theory was further supported by postmortem studies, which suggest a higher atrophy of the brain [79]. Brain imaging studies enable the systematic investigation of neuropsychiatric disorders. Compared to healthy controls, the structural changes involve ventricle enlargement, wider sulci, GM and WM tissue loss [80–82]. Longitudinal studies also revealed that the first psychosis episode severely impacts the brain atrophy, but also highlight that brain atrophy begins before the first psychosis episode [74]. Moreover, episodes of psychosis have been suggested to have a neurotoxic nature [77]. The initial stages of schizophrenia are characterised by an increase in subcortical volume [83, 84], a non-uniform reduction of GM cortical volume and cortical thinning [85], along with a decline in the gyrification pattern [84]. Concerning the disease progression, multiple studies report an acceleration in brain atrophy, suggesting that schizophrenia is a degenerative progressing disease [77]. Recent studies challenge this hypothesis suggesting that the atrophy might be caused by the medication rather than the pathophysiology of the disease [77, 86]. Besides the morphological changes, the disease is also characterised by disruptions in brain connectivity, both at functional and anatomical level [87], a phenomenon often linked to a reduction in WM [88].

Schizophrenia shares symptomatology with bipolar disorder, and distinguishing between them in the early stages can be challenging [84]. Early and accurate diagnosis is essential for successful therapeutic interventions [84]. As a result, there has been a growing interest in discovering new neuroimaging biomarkers for the early identification of the condition, which can be implemented in clinical practice in recent years.

2.1.2.2 Type 2 Diabetes

T2D is a metabolic disorder associated with pathological ageing. The condition is marked by cellular resistance to insulin and/or decreased insulin production, which increases the glucose in the bloodstream [89]. According to World Health Organization (WHO), in 2014, diabetes affected 422 million individuals worldwide, with a prevalence of 8.5% among the adult population [90]. T2D accounts for 95% of the diabetes cases [90]. Some risk factors to develop T2D included age over 45 years, family history, overweight, smoking, and stress [90].

Chronic hyperglycemia negatively impacts various systems within the body, with primary emphasis on the endothelium [91, 92]. Endothelial dysfunction leads to a two- to four-fold predisposition to develop vascular complications [92]. Atherosclerosis emerges as the main vascular complication of the endothelium's abnormal functioning and is a risk factor for vascular disease. Vascular dysfunction is often associated with an increase in brain lesions (WM hyperintensities, lacunar infarcts, large infarcts, microbleeds, microinfarcts), which are associated with a decrease in cognitive capabilities and dementia [93]. Furthermore, small vessel disease, which affects the cerebral circulatory system, resulting in brain damage, has been related to the T2D [94]. Therefore, T2D may affect the ageing trajectory of the brain. The brain atrophy in T2D is more pronounced than in healthy ageing; compared to healthy controls, the total brain volume decreases 0.5-2.0% [93, 95]. The GM volume loss is non-uniform across the brain; the regions most affected are the medial temporal, anterior cingulate, and medial frontal lobes [93]. The subcortical structures which exhibit more noticeable atrophy appear to be the hippocampus and amygdala [93, 95]. The cortical thinning also seems more exacerbated in T2D patients compared to healthy controls [93,95]. Concerning WM, besides the increase in hyperintensities, studies also suggest an integrity loss of the WM fibers [95]. Since diabetes accelerates cerebral atrophy, the disease is also associated with an expansion of ventricles in comparison to healthy controls [96]. T2D has been associated with deterioration in cognitive performance and memory impairments [93, 95, 97] and a predisposition for dementia. AD and vascular dementia are the two most prevalent types of dementia in T2D; the relative risk of T2D patients develop vascular dementia and AD is 2.27 and 1.63, respectively [97].

2.1.2.3 Alzheimer Disease

AD is the most prevalent neurodegenerative disease; in total, it accounts for 60-80% of dementia cases [98, 99]. Dementia is characterised by a severe loss of cognitive function paired with difficulties in performing daily living activities. WHO estimated that, in March 2023, more than 55 million people worldwide lived with dementia, and this number increases by 10 million every year [100].

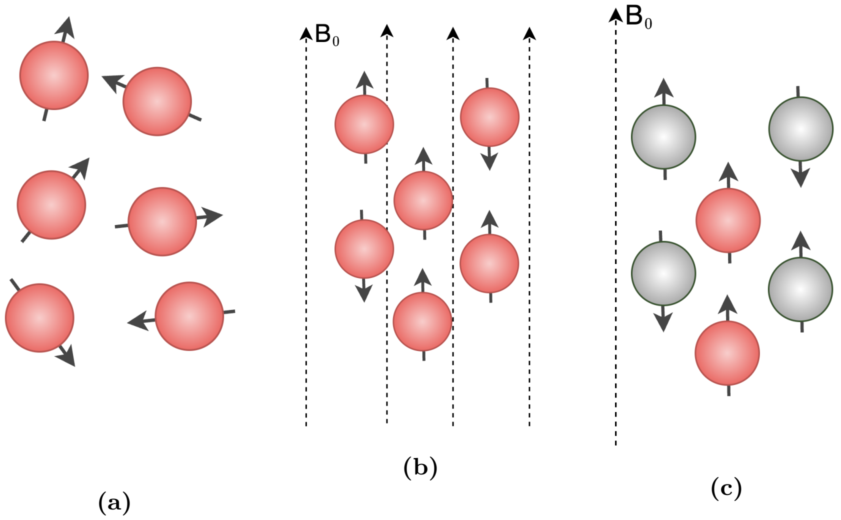
The aetiology of the AD seems to be a combination of factors, the highest risk factor for the disease is age [22,101]. AD was first described by psychiatrist Alois Alzheimer in 1907 [102]. Currently, it is the most commonly studied and prevalent neurodegenerative disease [102]. Nevertheless, pathogenesis and pathophysiology remain under debate, and no treatment is available. Unequivocally, the accumulation of the proteins β-amyloid and tau are two hallmarks of the disease [22, 101]. Currently, the hypothesis with more acceptance is the amyloid cascade. In this theory, the accumulation of β-amyloid in the extracellular space promotes tau's deposition and aggregation, forming neurofibrillary tangles. The overaccumulation of these composites seems to cause synaptic disruption and trigger neuronal loss [101, 103]. This brain tissue loss follows the accumulation of amyloid plaques and neurofibrillary tangles across the brain and is linked to cognitive impairments [22]. In its early stages, usually the area most affected is the medial temporal lobe, which includes subcortical structures such as the hippocampal region, subcortical regions with a vital role in memory and learning [22,71]. As the disease progresses, atrophy spreads to the neocortex [22, 101]. It should be highlighted that the disease progresses at different rates across individuals; therefore, the atrophy rates differ across studies. Nevertheless, the whole brain atrophy is around 0.5% and 1.9% on healthy controls and AD, respectively [22]. The hippocampus has a pronounced atrophy rate in AD, comparatively to healthy controls, the atrophy is at least two-fold higher [22, 71]. Concerning the WM, the impact of myelin degradation or WM lesions on dementia appears to be relatively minor compared to GM changes [22,104]. Finally, concerning CSF, studies report a three- to four-fold increase comparing the ventricle volume of AD patients with healthy controls [70–72].

The AD diagnosis comprises the identification of cognitive impairments, based on clinical and neuropsychological criteria, along with one of the three strategies: atrophy in the medial temporal region, using MRI; abnormal neuronal CSF markers, such as Aβ42 T-tau and P-tau, this screening involves a lumbar puncture; or temporoparietal hypometabolism using positron emission tomography [105]. While CSF markers are helpful for early detection of AD, their sensitivity diminishes as the disease advances. Conversely, cerebral atrophy, though less prominent in the early stages, is a sensitive indicator of disease progression [22].

Finally, AD has been proposed to be characterised by insulin resistance and hypometabolism, particularly in the temporoparietal region. The pathophysiology similarities between AD and T2D lead to the generation of the following hypothesis: AD is a type of diabetes that targets the brain, entitled "type 3 diabetes" [106]. Nevertheless, this theory is highly controversial in the scientific community, and the pathogenesis of AD remains under discussion.

2.2 Exploring the brain structure with magnetic resonance imaging

MRI is a high-resolution non-invasive neuroimaging technique that enables the in-vivo study of brain structure. It was first invented in the early 1970s by Paul C Lauterbur and Peter Mansfield [107], who were awarded with the Nobel Prize for Medicine in 2003 [108]. This neuroimaging technique leverages nuclei spins to produce detailed images of body tissues. Spin is a property of elemental particles, an intrinsic form of angular momentum that can be visualised as a rotation around the particle's axis. Composed particles like nuclei also possess spin. The nuclei containing an even number of particles have a zero net spin, whereas nuclei with an odd number of particles have an angular momentum. The nuclei, composed of protons and neutrons, are positively charged. The movement of nuclei with integral spin generates an intrinsic magnetic field. Consequently, the atom is sensitive to the presence of external magnetic fields [109].



(a) (b) (c)

Figure 2.2: On the left, the nuclei are stochastically orientated; in the centre, the nuclei are oriented in the presence of an external magnetic field B_0. On the right is represented the net magnetisation given the external magnetic field B_0.

Magnetic resonance (MR) images are generated by combining a strong magnetic field, B_0, with a sequence of radio frequency (RF) pulses that generate signals reflecting the intrinsic biological properties of tissues. The nuclei within the human body are stochastically oriented, as shown in Figure 2.2a. When a magnetic field is applied, the nuclei align in the direction of the generated field, see Figure 2.2b. The alignment can either be parallel or antiparallel to B_0, with the parallel alignment being more energetically favourable. The measured magnetisation is the sum of individual magnetisations, resulting in the net magnetisation, M_0, corresponding to the difference between the up and down nuclei, Figure 2.2c. At this stage the atom nuclei have been polarised. Assuming B_0 is applied along the z-axis (as is typically the case in imaging), resulting in longitudinal magnetisation, M_z. Each atom precesses at the frequency designated by *Larmor frequency* or ω_0, which is proportional to the strength of B_0, and can be computed using the equation 2.1, where γ_0 represents the gyromagnetic ratio which is atom specific [109].

$$\omega_0 = \gamma_0 \times B_0 \tag{2.1}$$

Larmor frequency is the backbone of the MRI. A nuclei will absorb energy from a RF pulse and transition to a higher energy state if its frequency matches the nuclei's *Larmor frequency*. The energy absorption will result in two phenomena: phase coherence and transverse magnetisation. Before the RF pulse, the nuclei are dephased; after the pulse absorption, their phase briefly aligns with the RF. However, the phase coherence is soon lost due to the influence of the intrinsic magnetic field of the neighbouring atoms, a mechanism designated by spin-spin relaxation. T$_2$ is a tissue property that translates the dephasing process; it corresponds to the time

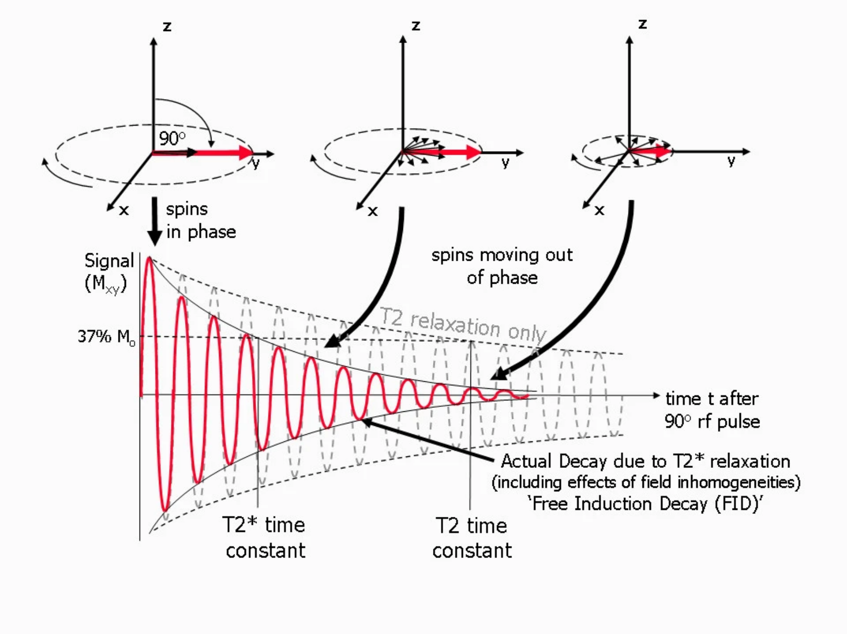


Figure 2.3: Representation of the dephasing process after the emission of a radio frequency pulse, and the corresponding "free induction signal decay" phenomenon. Adapted from [2].

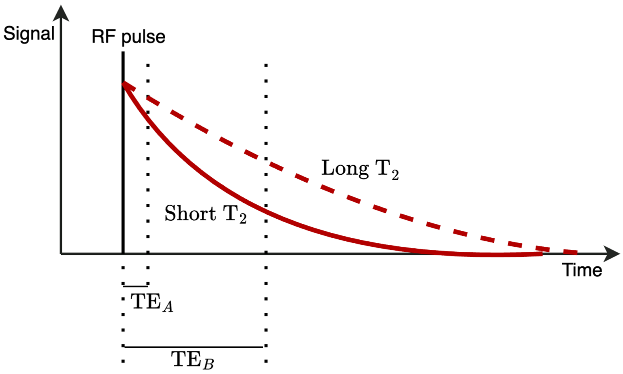


Figure 2.4: Representation of the signal acquisition into two-time points, TE_A and TE_B, for two tissues: one with a long T_2 and the other with a short T_2.

for 37% of the phase to be lost. Nonetheless, the dephasing process accelerates due to the interaction between spins and magnetic field inhomogeneities. This effect is designated by T_2* relaxation. The dephasing process leads to a decrease in the intensity of the magnetic signal and originates a phenomenon known as "free induction signal decay", as depicted in Figure 2.3. The mathematical expression describing signal decay is presented in Equation 2.2, where $S(t)$ describes the signal intensity, A the amplitude, t the time and T_2 the decaying rate of the signal [109].

$$S(t) = Acos(wt) \exp^{-\frac{t}{T_2}} \tag{2.2}$$

The dephasing process is fast on bones and lungs, ergo these tissues have a short T_2. On fluids such as water, blood and CSF, the phase coherence lasts longer; thus, these tissues have a long T_2. Time of echo (TE) is the amount of time between RF

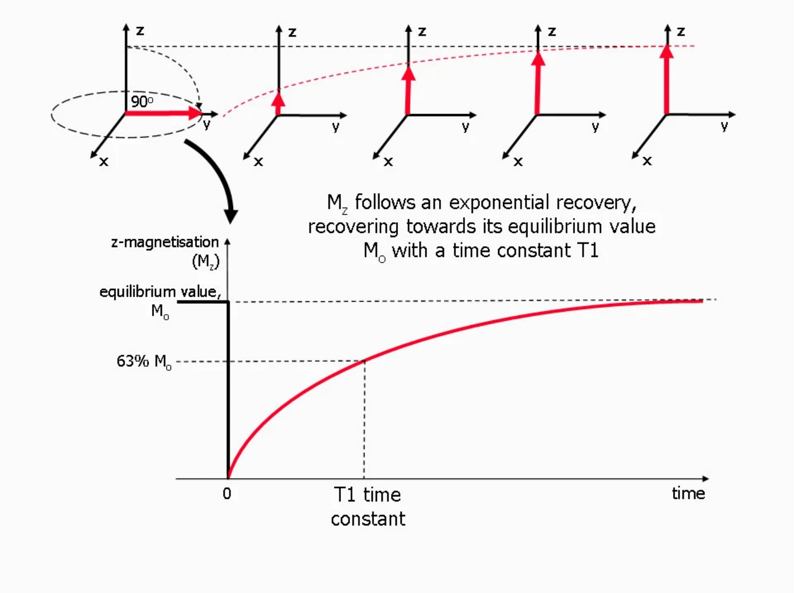

z
90º
y
x
M_z follows an exponential recovery, recovering towards its equilibrium value M_o with a time constant T1
z-magnetisation (M_z)
equilibrium value, M_o
63% M_o
0
T1 time constant
time

Figure 2.5: Longitudinal magnetisation recovery after the emission of a radio frequency pulse. Adapted from [2].

pulse being turned off and the signal detection. Figure 2.4 illustrates the impact on signal acquisition of different TE and T_2. If the TE is too short, for instance, the image is acquired right after the RF pulse is turned off (time point TE_A), the phase coherence across tissues is similar; therefore, the contrast between tissues would be negligible. On the other hand, the signal is captured at different dephasing phases (time point TE_B), the tissues would be distinguishable on the image. Tissues with a short T_2 will appear dark on images, given that the amplitude of its signal is lower at that point, and tissues with a long T_2 will appear bright. In a nutshell, T_2 is a source of contrast. Images that leverage T_2 to distinguish across tissues are entitled T_2-weighted [2].

Besides phase coherence, the RF pulse shifts the magnetisation from the z-plane to the xy-plane, resulting in transverse magnetisation (M_{xy}), as is depicted in Figure 2.5. When the RF pulse is turned off, the nuclei realign with the magnetic field B_0. Therefore, the nucleus loses transverse magnetisation and regains longitudinal magnetisation. The realignment with the B_0 involves energy dissipation; the energy is dissipated to the surroundings of the nuclei, also called *lattice*, and the dissipation energy rate depends on the interaction between the spin and the lattice. Each tissue in the brain has a different spin-lattice interaction and, consequently, a different dissipation rate. The time for the nuclei to achieve 66% of the original longitudinal orientation is called T_1. The equation which describes the regain of longitudinal magnetisation is given by Equation 2.3 [109].

$$M_z = M_0(1 - e^{-\frac{t}{T_1}}) \tag{2.3}$$

The acquisition of MRI images requires multiple excitations of the nuclei. The period

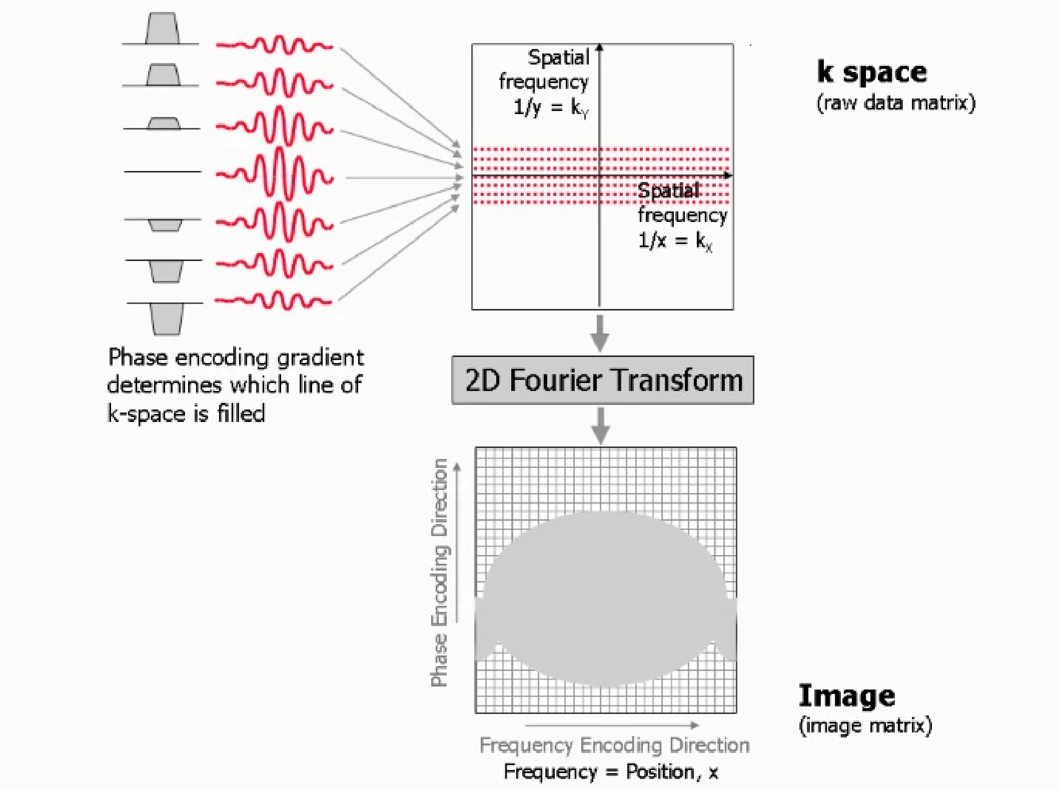

Spatial frequency 1/y = k_y
Spatial frequency 1/x = k_x
k space
(raw data matrix)
Phase encoding gradient determines which line of k-space is filled
2D Fourier Transform
Phase Encoding Direction
Image
(image matrix)
Frequency Encoding Direction
Frequency = Position, x

Figure 2.6: Image reconstruction from the *k*-space. The magnetic resonance signals obtained from each phase encoding step are preserved in a raw data matrix referred to as k-space. A two-dimensional Fourier transformation on the *k*-space leads to the generation of the reconstructed image. Adapted from [2].

between excitations is called the time of repetition (TR). The TR and T_1 parameters are intrinsically connected. For a long TR, the effects of T_1 become insignificant, but as the TR decreases, the T_1-effects gain more weight. When RF pulse is emitted, tissues with a shorter T_1 would have more longitudinal magnetisation than the ones with higher T_1. Therefore, these tissues will absorb more energy to realign with the RF pulse. Consequently, these tissues produce more induced current and appear brighter, whereas tissues with longer T_1 appear darker on images. Hence, images generated with a small period between the pulses are designated by T_1-weighted images because T_1 tissue parameter defines the contrast in these images. In T_1-weighted images, the TE is kept short, making the phase coherence effects insignificant [109].

The MRI scanner is composed of a coil which transmits a homogeneous constant magnetic field B_0 in the order of magnitude of one Tesla. As previously mentioned, the equation 2.1 is the backbone of the MRI, and the γ_0 is atom specific. MRI targets the hydrogen nuclei (^1H) in the body to acquire images; each hydrogen nucleus consists of a single proton. Hydrogen nuclei are abundant in the human body; for instance, each water molecule is composed of two hydrogen, and water constitutes approximately 70% of the human body's composition. The scanner aligns the ^1H-

nuclei with the magnetic field, then upon a RF pulse emission, all the ^{1}H-nuclei throughout the body will be excited and aligned with the RF [109].
Consequently, the signal measured will be a sum of all the body ^{1}H nuclei, and the spatial origin of each signal is unknown. The location of each signal is uncovered by encoding the location with three gradients, one across each plane (x, y and z). At first, the magnetic field is homogeneous; then, a gradient field is applied across the z-plane; consequently, by the Larmor frequency, spins will process at different rates. Thus, a specific plane in the z-axis can be targeted by emitting a frequency pulse that comprises the frequencies in the region of interest; this is designated by slice encoding. Afterwards, the phase encoding is performed, and a short gradient is applied in the y-axis; this gradient will change the spin precessing phases along the y-axis. Therefore, phase is the signature of these atoms. Finally, another gradient is applied along the x-axis. In this case, spins will precess slower and faster where this gradient is lower and higher, respectively. This is designated by frequency encoding and is depicted in Figure 2.6. This process is repeated multiple times, and each time, a different phase and frequency gradient is applied. In each repetition, the measurement is stored in the k-space. The image is decoded from the k-space via inverse 2D-Fourier transform [2, 109], Figure 2.6 summarises the MR acquisition procedure.

Chapter 3

State of the art

Brain age has been gathering much attention in the past decade [20, 21]. Machine learning strategies for predicting brain age encompass a range of approaches. This chapter reviews the state-of-the-art brain age models using T_1-weighted structural images. The first section, section 3.1, begins with an overview of the three basic blocks of a brain age pipeline. Section 3.2 details the state-of-the-art on shallow learning methods, whereas section 3.3 summarises the recent works using deep learning. Brain age models have been extensively applied to different neurological and psychiatric conditions. A comment on the application of brain age gap estimation (BrainAGE) on different diseases can be found in section 3.4. Finally, the last section wraps up with a discussion of the shortcomings of the current literature.

3.1 Brief introduction to a brain age pipeline

BrainAGE leverages machine learning models to capture brain ageing patterns from magnetic resonance (MR) images. Before diving into the state-of-the-art, an overview is provided concerning the three main steps of a brain age machine learning pipeline: preprocessing, modelling and evaluation, a scheme is shown in Figure 3.1.

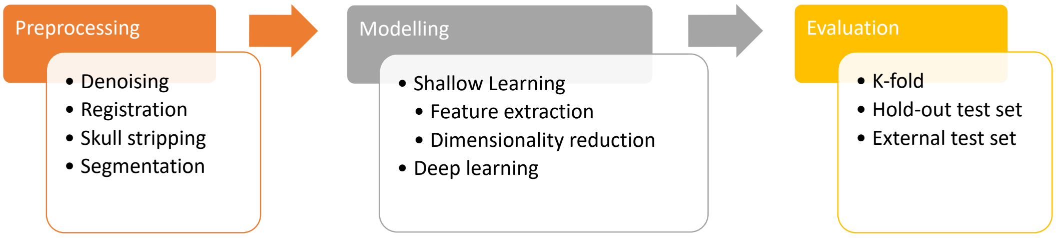


Figure 3.1: Representation of the three main steps of a brain age pipeline and the corresponding methodologies.

3.1.1 Preprocessing of magnetic resonance images

Preprocessing in neuroimaging is crucial in increasing the signal, reducing noise generated by the scanner and physiological effects, and removing undesirable information, such as the skull. There is no gold standard in neuroimaging for preprocessing T_1-weighted images; different frameworks and pipelines have been considered [20]. Nevertheless, the preprocessing generally encompasses the following steps: denoising, registration into a standardised template, skull stripping and segmentation. Different types of noise are produced during a magnetic resonance imaging (MRI) acquisition due to the scanner, physiological mechanisms and the interaction between the two. Multiple denoising techniques can be applied to reduce the noise and increase the image contrast [110]. The registration step aims to normalise the brain images and ensure the correct image orientation [20]. In the registration step, voxel-wise transformations are applied to the image to overlap it with a template. These transformations, designated by deformation fields, might retain valuable morphology information; Figure 3.2f portrays a deformation fields image. Finally, the non-tissue can be removed using skull stripping methods [111]. Taking the preprocessing a step further, images can be segmented into different tissues [112]. In this case, each voxel is assigned a probability of containing grey matter (GM), white matter (WM) or cerebrospinal fluid (CSF); an example of each tissue is in Figure 3.2c 3.2d and 3.2e, respectively. Another alternative is region-of-interest (ROI) parcellation; the preprocessed images are divided into different ROIs, according to a given template, and features are extracted per ROI.

Different levels of preprocessing have been considered to model brain age: raw, minimally processed, segmented images, and parcelled images. The raw images are the lowest preprocessed images. Despite the "raw" concept, the images are still registered into a template; a raw image is shown in Figure 3.2a. Minimally processed images, represented in Figure 3.2b, are denoised T_1-weighted images with only brain tissue. Thus, the preprocessing steps of minimally processed images encompass denoising, registration and skull stripping.

Most studies leverage frameworks that contain routines with all the preprocessing

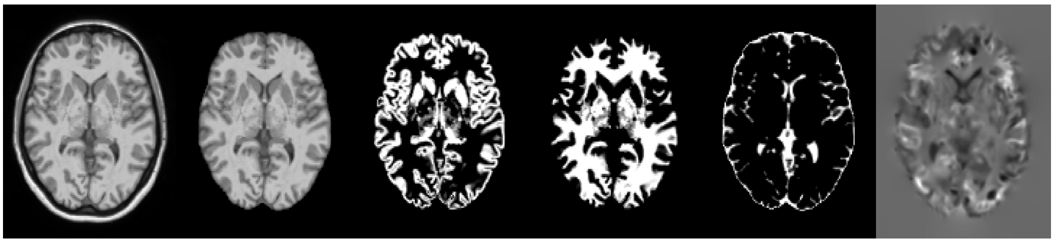

(a) Raw **(b)** MP **(c)** GM **(d)** WM **(e)** CSF **(f)** DF

Figure 3.2: Preprocessing byproducts of a T_1-weighted image: raw, minimally preprocessed (MP), grey matter (GM), white matter (WM), cerebrospinal fluid (CSF) and deformation fields (DF).

steps implemented. FreeSurfer and SPM emerge as popular frameworks extensively used in the brain age context [113,114]. FreeSurfer is the well-established framework in the neuroimaging field for preprocessing images and segmenting the different brain structures. Nevertheless, this tool is computationally demanding. Therefore, other alternatives, such as SPM, have emerged and have been adopted in BrainAGE field. The choice of the framework also depends on the pipeline considered. Studies focusing on feature extraction mainly use FreeSurfer to preprocess images [28,29,55,56,115–121], whereas studies which use the minimally processed or segmented images to train the model use both FreeSurfer [4,9,117,122–129] or SPM [3,29,118,130–133].

3.1.2 Modelling

Multiple machine learning methodologies have been considered to predict brain age; Figure 3.3 depicts an overview of the different approaches considered. The models can be divided into shallow and deep learning methodologies. The former, known as traditional machine learning, relies on regression models to predict age. In this case, brain age models either learn the age pattern from handcrafted features or a compressed version of the image. Shallow learning presents some weaknesses. Feature engineering requires a significant effort and domain knowledge to carefully design and select features that capture relevant information from the neuroimaging data. An option for feature engineering is to apply dimensionality reduction techniques, such as principal component analyis (PCA), to the image [134]. However, dimensionality reduction techniques are generally lossy strategies. Consequently, the compression involves losing information that might contain vital information to predict the ageing pattern.

Deep learning emerges as an alternative to shallow learning [32]. Concretely, convolution neural networks (CNN) are a subcategory of the deep learning field which attained remarkable performance on visual recognition tasks [135]. The image is the substrate of the CNN. Thus, these models learn the relevant problem-specific features directly from data. Nevertheless, deep learning also has disadvantages. Contrasting with shallow learning, which usually has dozens or hundreds of parameters, these models have millions or billions of parameters. Thus, these models generally require large amounts of data, are computationally demanding and are challenging to interpret.

3.1.3 Evaluation strategy

Different evaluation strategies have been considered to assess the performance of brain age models. Usually, the data are divided into three categories: training, validation and test set. The training set is used during model weights optimisation; the validation can be utilised in hyperparameter tuning; and finally, the model is evaluated on unseen data designated by the test set. The validation set is optional

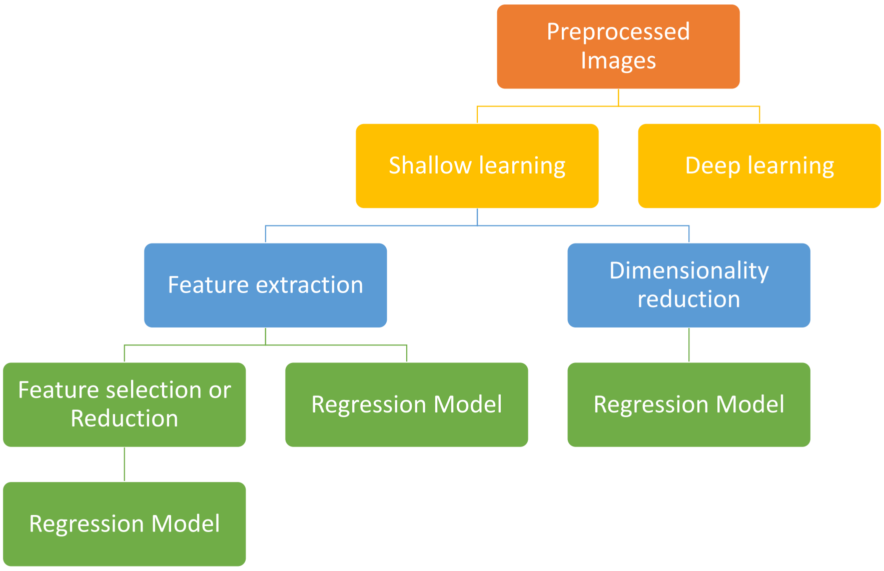


Figure 3.3: Different types of modelling considered in brain age.

and only required if hyperparameter tuning or model selection is performed.

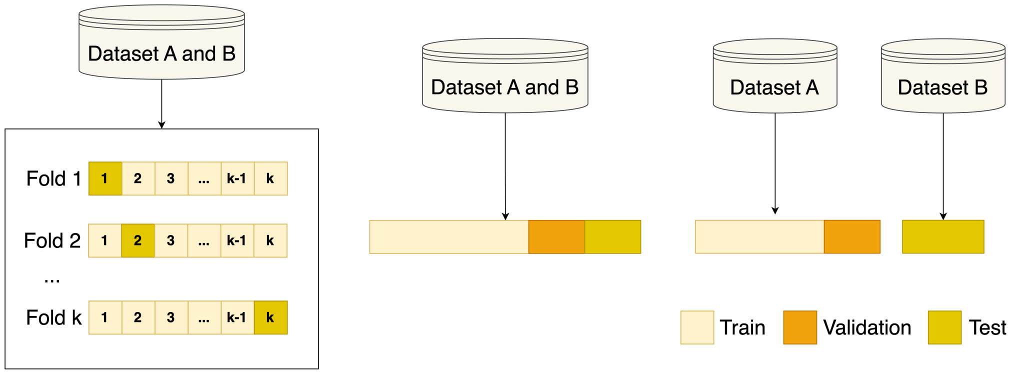
Three approaches are commonly used to evaluate the brain age model performance: holdout test set, k-fold and external test sets; Figure 3.4 portrays the three evaluation strategies. The holdout test set entails using a certain percentage or a specific number of instances from the dataset to train, validate and test the models. Alternatively, the k-fold approach involves dividing the dataset into k subgroups. Iteratively, a subgroup is used as the test set and the remaining $k-1$ subgroups are used for training. This process is repeated k times until all subgroups have been considered once as the test set. Thus, training, in total, k models. Concerning the external test set, this strategy involves evaluating the model on an independent/external dataset. This option enables the assessment of the model to variations of acquisition settings. Suppose that a given problem contains data from two datasets, A and B, in both the holdout sample and k-fold approaches; the training, validation, and test samples are derived from the two datasets; Figure 3.4a and 3.4b. In the external test set approach, dataset A is used to train and validate the model, whereas dataset B is used as a test set, Figure 3.4c.

The performance of brain age models is usually reported with the mean absolute error (MAE) and/or R^2.

MAE is given by equation 3.1, where N is the number of instances, y_i is the chronological age and $\hat{y}_i$ is the model prediction.

$$\text{MAE} = \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{N} \tag{3.1}$$

The R^2 metric measures the proportion of predicted age variance that is explained by the chronological age. The computation encompasses measuring the sum of



(a) *k*-fold **(b)** Holdout sample **(c)** External test set

Figure 3.4: Representation of the three evaluation strategies used to report the performance of brain age models.

squares of residuals SS_{res}, equation 3.2 where $\hat{y}_i$ is the predicted age; the total sum of squares SS_{tot}, equation 3.3 where the $\overline{y}$ represents the age's mean. Then one is subtracted to the ratio between the former and the latter, equation 3.4.

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.2}$$

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{3.3}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.4}$$

Finally, it must be outlined that performance results are not directly comparable across studies. The age range of both training and test sets tremendously impacts the reported performance. The MAE is proportional to the age range considered; studies with a small age range tend to have a lower MAE than those with a broader age range [136].

3.2 Shallow learning

In brain age prediction, two shallow learning strategies have been considered: handcrafted features and dimensionality reduction, subsections 3.2.1 and 3.2.2 debrief the state-of-the-art of the two approaches, respectively. Then, subsection 3.2.3 discusses the performance of different shallow pipelines on brain age prediction. Finally, the section finishes with an overview of the model evaluation across studies.

3.2.1 Handcrafted features

The handcrafted features approach is an ad-hoc strategy based on prior knowledge of morphological ageing patterns. In general, multiple features are extracted per ROI.

Cortical and subcortical brain measures have been extensively explored for brain age prediction [9, 28, 55, 56, 115–117, 137–140], a summary is presented in Tables 3.1 and 3.2. Regarding subcortical features, volume is one of the most commonly extracted features [9, 28, 117, 137–139, 141]. In terms of cortical measures, most of the studies focus on morphology features, but the complexity of the cortical ribbon also received some attention [55, 56]. Concerning morphology features, cortical thickness [28, 55, 56, 115–121, 137, 138] and cortical volume [28, 29, 116–121] have emerged as the most extensively explored features in the prediction of brain age. To a lesser extent, other morphology features have also been considered for age prediction, such as cortical area [28, 115, 116, 120, 137, 139], cortical curvatures [115, 120] and sulcal depth [120]. One common finding among studies is the overperformance of brain age models trained solely with cortical thickness compared with models trained only with cortical area or cortical curvatures [28, 115, 116]. This result is in line with other studies which point out that surface area [46, 49] and cortex curvatures [49] are less impacted by ageing than cortical thickness.

An intriguing aspect is the contradictory performance results observed between cortical thickness and cortical volume [116, 120]. Valizadeh *et al.* [116] suggest that cortical thickness outperforms cortical volume, whereas Liu *et al.* [120] have found the opposite. Similarly, contrasting performance results are reported between cortical thickness and subcortical volumes [28, 117, 119]. Becker *et al.* [117] and Rokicki *et al.* [119] report a higher predictive value of subcortical volumes whilst Liem *at al.* [28] report that cortical thickness yields better performance than subcortical structures. The inconsistencies observed may be influenced by several factors: the preprocessing pipeline, parcellation granularity and/or machine learning pipeline. Concerning the former, all the previous studies used Freesurfer to extract the cortical and subcortical measures; nevertheless, the version differs. Valizadeh *et al.* [116] and Liu *et al.* [120], who report dissonant results concerning cortical thickness and cortical volume, used the FreeSurfer version 5.3 and 6.0, respectively. The FreeSurfer version impacts the measures of cortical thickness and subcortical volumes [142]. Thus, the version of FreeSurfer might be a source of variability that helps explain some of the incongruencies of brain age models. Nonetheless, the FreeSurfer version is not enough to explain all the variability of the results. Liem *et al.* [28] and Becker *et al.* [117] used the same FreeSurfer (version 5.3), yet the authors report contrasting results. Moreover, Rokicki *et al.* [119] and Becker *et al.* [117] use different versions, and both report the superior performance of subcortical structures over cortical thickness. The parcellation scheme might explain the variability in the model's performance; the granularity of the template has been shown to impact the model's performance [55, 143]. In all the handcrafted features brain age models, the cortical granularity differed, the lowest and highest granularity used was 34 and 5124 ROIs, respectively. Finally, the machine learning pipeline, which differed across the previous studies, might also be the source of variability [138, 144]. Regarding complexity features, the gyrification and fractal dimension have been widely used to capture the folding pattern of the cortical ribbon. Brain age models trained with these features have dissonant results compared to models trained with cortical thickness. Some studies report that cortical thickness yields superior performance than gyrification [120], yet contrasting results have been reported by Madan *et al.* [55]. Analogous to the results previously discussed, this discrepancy might be due to different FreeSurfer versions, parcellation schemes, and/or machine learning pipelines, which are different in both studies. Regarding fractal dimension, Marzi *et al.* [56] performed a comprehensive analysis of the impact of fractal dimension computational methodologies on brain age models. The study highlights that the computation methodology of fractal dimension impacts the brain age model's performance. The authors report a MAE variation of 4.7 years in the model's performance for different approaches. Furthermore, the authors compared the performance of different fractal dimension computation strategies with cortical thickness and gyrification. The fractal dimension overperformed these two metrics in two strategies, whereas in the other two, it is the inverse.

Shallow brain age models often incorporate multiple features to improve predictive accuracy [28, 115, 116, 118, 120]. However, combining multiple features in some cases does not result in superior performance [115, 120]. The reasons behind these inconsistencies can be multifaceted. Cortical features are redundant [46, 54, 55, 145]. Therefore, regularisation, feature selection or reduction is crucial in preventing overfitting and enhancing generalisation. Improper regularisation might explain this dissonant result [137]. Liu *et al.* combined six cortical features extracted from 62 ROIs in a support vector regression (SVR) model. The results suggest that combining all these features yields lower performance than using only cortical thickness, volume, or folding index. The authors combine all the features and train the SVR without implementing a feature selection or reduction procedure. Furthermore, SVR [146] has hyperparameters that, if not properly tunned, might lead to erratic brain age performance [130]. Thus, inadequate regularisation and hyperparameter tuning can lead to unstable and unreliable results, influencing the comparative performance of features and feature fusion methods. Wang *et al.* assess the performance of brain age models on cortical thickness, cortical area, mean curvature and gaussian curvature. Different combinations of features were performed: 1) cortical thickness and cortical area; 2) cortical thickness and mean- and gaussian- curvatures; 3) all four features. All three combinations yield better performance than solely a single feature, yet the lowest MAE was attained by combining only the cortical thickness with the cortical area.

In conclusion, the results of different shallow learning approaches in age prediction exhibit intriguing patterns and inconsistencies. Cortical thickness consistently outperforms cortical area and cortical curvatures, while conflicting results emerge between cortical thickness and cortical volume. These inconsistencies may stem from the preprocessing procedure, parcellation granularity, and machine learning pipeline. Combining multiple features yields, in general, better results than using standalone features.

3.2.2 Dimensionality reduction

Applying dimensionality reduction techniques directly to the images is an alternative approach to handcrafted features. Dimensionality reduction at the voxel level has gained prominence as a strategy to model brain age. A summary of the results using different dimensionality reduction strategies in the context of brain age prediction is presented in Tables 3.3 and 3.4. Primary two dimensionality reduction procedures have been widely explored: similarity matrix [3, 29, 118, 124, 132, 147] and PCA [9, 117, 124, 130, 131]. Additionally, canonical correlation analysis (CCA) [124] and non-negative matrix factorization (NMF) [131] have also gathered some attention. The similarity matrix approach calculates pairwise similarities between subjects. PCA, CCA and NMF are factorisation-dependent methods that differ in the optimisation process. PCA extracts orthogonal components that capture the maximum variance in the data. CCA identifies linear combinations of variables that maximise the correlation between the data and the corresponding label. NMF decomposes data into non-negative components.

Concerning the similarity matrix approach, congruent results are reported across different studies. This consistency suggests that the similarity matrix technique is a robust and reliable approach to predict brain age. Nonetheless, it is essential to acknowledge that similarity matrix techniques heavily rely on the representativeness of the dataset used for training the brain age model. The accuracy and generalisability of this technique are contingent upon the diversity and size of the dataset. Therefore, it is crucial to ensure the inclusion of a broad range of subjects spanning different age groups and demographic characteristics to construct robust and accurate brain age prediction models. As with regard to PCA, at first sight, it seems this compression technique is more unreliable than the similarity metric given that Xifra-Porxas *et al.* [124] reported a MAE of 9.32 years, which is almost two-fold higher compared to the MAE reported by other studies [9, 117, 130, 131]. Nevertheless, this result is, in fact, in agreement with the reported results by other researchers. Xifra-Porxas *et al.* [124] reported the result considering using five PCA components only. Franke *et al.* [130] showed that the number of PCA components influence the MAE and that for a reduced number of components, the MAE is high. Therefore, the number of components has a preponderant factor in accurately predicting brain age. The number of PCA components used should be carefully selected to obtain a reliable brain age model. Xifra-Porxas *et al.* [124] compared the CCA dimensionality reduction strategy with similarity matrix and PCA. Despite the lack of proper statistical analysis, the authors report that CCA outperforms

Table 3.1: Overview of some manuscripts that predict brain age using a shallow learning feature-based approach.

Study	Preprocessing	Subjects	Age range [years]	Evaluation	Pipeline	Features	R^2	MAE [years]
Wang *et al.* [115]	FreeSurfer	360	20-82	10-fold	RVR	cortical thickness	n.s	$6,05 \pm 0,05$
						mean curvature	n.s	$7,88 \pm 0,07$
						gaussian curvature	n.s	$11.21 \pm 0,10$
						cortical area	n.s	$10,52 \pm 0,22$
					RFE-SVM and RVR	cortical thickness, mcurv, gcurv	n.s	$5,97 \pm 0,11$
						cortical thickness, cortical area	n.s	$4,57 \pm 0,04$
						cortical thickness, cortical area, mean curvature, gaussian curvature	n.s	$5,06 \pm 0,09$
Valizadeh *et al.* [116]	FreeSurfer	3144	7-96	Split (50-50)	NN	cortical volume	0,7	n.s
						cortical thickness	0,66	n.s
						cortical area	0,57	n.s
						cortical volume, cortical area, cortical thickness	0,84	n.s
Liem *et al.* [28]	FreeSurfer	2354	19-82	5-fold	SVR	cortical thickness	n.s	5.95 ± 4.69
					SVR	cortical area	n.s	7.29 ± 5.96
					SVR	subcortical volumes	n.s	6.44 ± 5.02
					SVR	cortical thickness, cortical area subcortical volumes	n.s	4.83 ± 4.01
Becker *et al.* [117]	FreeSurfer	1563	6-92	5-fold	GPR	subcortical volumes	0,87	5,52
					GPR	cortical thickness	0,8	6,5
Madan *et al.* [55]	FreeSurfer	1056	18-94	external test set	PCA and RVR	cortical thickness	0,59	n.s
						gyrification index	0,62	n.s
						fractal dimension	0,67	n.s
						cortical thickness, gyrification index, fractal dimension	0,81	n.s
				external test set	PCA and RVR	cortical thickness	0,35	n.s
						gyrification index	0,5	n.s
						fractal dimension	0,71	n.s
						cortical thickness,gyrification index, fractal dimension	0,71	n.s
Beacker *et al.* [9]	FreeSurfer	10824	47-73	holdout test set	SVR	cortical volume and subcortical volumes	0.42	4.06 ± 0.02
					RVR	cortical volume and subcortical volumes	0.42	4.10 ± 0.02
					GPR	cortical volume and subcortical volumes	0.42	4.08 ± 0.01
Lee *et al.* [137]	FreeSurfer	492	18-87	10-fold	ElasticNet	cortical thickness, cortical area, subcortical volumes	n.s	7.2

PCA: Principal Component Analysis; GPR: Gaussian Regression Processes; RVR: Relevance vector Regression; RFE: Recursive feature elimination; SVR: Support Vector Regression; n. s.: not specified.

Table 3.2: Overview of some manuscripts that predict brain age using a shallow learning feature-based approach.

Study	Preprocessing	Subjects	Age range [years]	Evaluation	Pipeline	Features	R^2	MAE [years]
Marzi *et al.* [56]	FreeSurfer	86	19-85	5-fold	Linear regression	cortical thickness	n.s	12
						gyrification index	n.s	14,2
						fractal dimension (Kiselev *et al.*, 2003)	n.s	15,8
						fractal dimension (Goñi *et al.*, 2013)	n.s	15,8
						fractal dimension (Marzi *et al.*, 2018)	n.s	12,3
						fractal dimension (Marzi, 2020)	n.s	11,1
Costa *et al.* [118]	SPM12	2640	17-90	5-fold	PCA and GPR	cortical volume	n.s	7.187
					PCA and GPR	mean curvature	n.s	7.30
					PCA and GPR	cortical thickness, cortical volume	n.s	6.385
					PCA and GPR	cortical thickness, cortical volume, mean curvatures	n.s	6.132
Rokicki *et al.* [119]	FreeSurfer	750	18-85.3	10-fold	Random Forest	Subcortical volumes	0.67	7.5
						Cortical thickness	0.59	8.4
						Cortical and subcortical	0.72	6.9
Liu *et al.* [120]	FreeSurfer	2501	20-94	holdout test set	SVR	cortical area	0.86	5.90
						cortical depth	0.84	6.34
						folding index	0.90	4.96
						local gyrification index	0.82	6.92
						cortical volume	0.91	5.03
						cortical thickness	0.90	4.79
						cortical area, cortical depth, folding index, local gyrification index, cortical volume, cortical thickness	0.88	5.30
Zhu *et al.* [121]	FreeSurfer	100	n.s.	5-fold	GPR	brain volume	n.s	6.86 ± 0.13
					GPR	cortical thickness	n.s	9.12 ± 0.15
Richard *et al.* [139]	FreeSurfer	612	20-88	external test set	xgboost	cortical thickness, cortical area, cortical volume, subcortical structures	n.s.	6.76
Aycheh *et al.* [140]	FreeSurfer	2705	45-91	10-fold	Lasso	cortical thickness	n.s	4.033
Dafflon *et al.* [138]	FreeSurfer	10307	18-89	10-fold	RVR	cortical thickness, subcortical volumes	n.s.	5.474 ± 0.140
					TPOT	cortical thickness, subcortical volumes	n.s.	4.612 ± 0.124

PCA: Principal Component Analysis; GPR: Gaussian Regression Processes; SVR: Support Vector Regression; xgboost: eXtreme Gradient Boosting; TPOT: tree-based pipeline optimization tool, n. s.: not specified.

PCA and the similarity matrix. Moreover, Varikuti *et al.* [131] compared NMF with PCA, the results evidence that NMF is an alternative approach to PCA. Nonetheless, the performance of CCA and NMF was not assessed on an external test set. Therefore, further analysis should be performed to validate these dimensionality reduction techniques in brain age prediction.

In conclusion, voxel-level strategies are a promising approach to brain age prediction. Further exploration of CCA and NMF's potential in predicting brain age could provide valuable insights into its efficacy compared to other methods.

3.2.3 The significance of shallow learning pipeline in brain age performance

The shallow learning pipeline can profoundly impact the model's performance. As previously discussed, the preprocessing pipeline, the parcellation scheme, the feature extraction and selection, or dimensionality reduction can lead to significantly different results. The comparison between a handcrafted feature-based model (using subcortical and cortical features) and dimensionality reduction with PCA was performed by Baecker *et al.* [9]. The findings evidence that the latter attained lower and stabler performance across three regression models (relevant vector regression (RVR), SVR and gaussian process regression (GPR)) than using ROI-wise features. Furthermore, the regression model and the tuning of the corresponding hyperparameters have also been suggested to influence the performance of shallow brain age pipeline [130, 137, 144]. Franke *et al.* [130] showed that the hyperparameter tuning of a SVR model influences its performance; improper tuning can lead to an increase of the MAE from 5 to 9 years. Regularisation techniques significantly impact the model performance; Lee *et al.* [137] compared several models and showed that the one without regularisation performed poorly. A comprehensive comparison across multiple regression models was performed by Beheshti *et al.* [144]. The findings reveal that the choice of the kernel and the regression model affect the MAE value. In this study, the models with the best and worst performance were the gaussian SVR and Gaussian regression (with a squared exponential kernel) with a MAE of 3.04 and 7.54 years, respectively. Given the infinite choices of regression models and their hyperparameters, Dafflon *et al.* [138] and Costa *et al.* [118] used genetic programming to select the best pipeline (feature type and regression model). The results evidence that no unanimous pipeline performs best for all cases [118, 138]. Different datasets benefit from different machine learning pipelines [118, 138]. Costa *et al* [118] showed that best pipeline differ across acquisition settings. Furthermore, the findings of Dafflon *et al.* [138] outline that there is an advantage in using genetic programming only when the age of the data is not uniformly distributed; otherwise, an RVR model yields similar performance to an optimised pipeline [138].

In conclusion, the pipeline choice affects the models' performance. Multiple compar-

Table 3.3: Overview of some manuscripts that predict brain age using a shallow learning dimensionality reduction based approach.

Study	**Preprocessing**	**Subjects**	**Age range** [years]	**Evaluation**	**Tissue**	**Feature reduction or compression**	**Regression model**	R^2	**MAE** [years]
Franke *et al.* [130]	SPM	410	20-86	external test set	GM	PCA	RVR	n.s	4,96
Becker *et al.* [117]	FreeSurfer	1563	6-92	5-fold	GM	PCA	GPR	0,86/-	5,65
Cole *et al.* [3]	SPM	2001	18-90	holdout test set	GM	similarity	GPR	0.89	4.66
					WM	similarity	GPR	0.84	5.88
					GM and WM	similarity	GPR	0.91	4.41
					raw	similarity	GPR	0.32	11.81
Cole *et al.* [147]	n.s.	2001	18-90	10-fold	GM and WM	similarity	GPR	n.s	5,02
		2001	18-90	external test set	GM and WM	similarity	GPR	n.s	7,08
Varikuti *et al.* [131]	VBM8	1084	18–81	10-fold	GM	PCA	RVR	n.s	5.7
						NMF	RVR	n.s	6.1
Jonsson *et al.* [29]	CAT12	1264	18-75	holdout test set	GM and WM	similarity	ridge	0.728	4.937
Costa *et al.* [118]	SPM12	2640	17-90	5-fold	GM	similarity	SVR	n.s	5.004
					WM	similarity	SVR	n.s	5.589
					GM and WM	similarity	SVR	n.s	4.571
					GM	PCA	Linear regression	n.s	13.609
					WM	PCA	Linear regression	n.s	13.613
Jiang *et al.* [132]	SPM8			holdout test set	FPN GM	similarity	GPR	0.70	7.74
Baecker *et al.* [9]	FreeSurfer	10824	47-73	holdout test set	GM	PCA	SVR	0.51 ± 0.02	3.77 ± 0.04
						PCA	RVR	0.51 ± 0.02	3.82 ± 0.03
						PCA	GPR	0.51 ± 0.02	3.81 ± 0.03

GM: Grey Matter; WM: White Matter; PCA: Principal Component Analysis; CCA: Canonical Correlation Analysis; NMF: Non-negative Matrix Factorization; FPN: FrontoParietal Network; GPR: Gaussian Regression Processes; RVR: Relevance vector Regression; SVR: Support Vector Regression; n. s.: not specified.

Table 3.4: Overview of some manuscripts that predict brain age using a shallow learning dimensionality reduction based approach.

Study	**Preprocessing**	**Subjects**	**Age range** [years]	**Evaluation**	**Tissue**	**Feature reduction or compression**	**Regression model**	R^2	**MAE** [years]
Xifra-Porxas *et al.* [124]	FreeSurfer	652	18-88	10-fold	cortical	PCA	GPR	n.s	10.05 ± 0.89
						similarity	GPR	n.s	7.14 ± 0.68
						CCA	GPR	n.s	7.01 ± 0.60
					subcortical	PCA	GPR	n.s	8.84 ± 0.76
						similarity	GPR	n.s	5.98 ± 0.63
						CCA	GPR	n.s	5.79 ± 0.53
					GM	PCA	GPR	n.s	9.32 ± 0.81
						similarity	GPR	n.s	6.20 ± 0.63
						CCA	GPR	n.s	6.02 ± 0.52
					WM	PCA	GPR	n.s	10.98 ± 1.07
						similarity	GPR	n.s	6.56 ± 0.66
						CCA	GPR	n.s	6.38 ± 0.54
Beheshti *et al.* [144]	SPM	788	788 (18-94)	10-fold	GM	PCA	Linear SVR	0.88	5.40
							Gaussian SVR	0.95	3.04
							GPR (kernel: exponential)	0.89	5.29
							GPR (kernal: squared potential)	0.81	7.54

GM: Grey Matter; WM: White Matter; PCA: Principal Component Analysis; CCA: Canonical Correlation Analysis; NMF: Non-negative Matrix Factorization; FPN: Fronto Parietal Network; GPR: Gaussian Regression Processes; RVR: Relevance vector Regression; SVR: Support Vector Regression; n. s.: not specified.

isons across different regression models were performed, yet different studies report distinct better models. There is not a gold-standard shallow pipeline that can be applied to multiple cases [118, 138]. Nevertheless, on a positive note, RVR yield a good performance across multiple studies [9, 130, 137, 138].

3.2.4 Evaluation

Brain age shallow learning models have been evaluated using the three strategies (*k*-fold, holdout test set and external test set). Concerning the shallow learning based on feature extraction, the most prominent evaluation methodology is the *k*-fold [28, 56, 115, 117, 118, 120, 121, 137, 138, 140]. To a smaller degree, the holdout [9, 116, 120] and external test [55, 55, 139] were also considered to evaluate the performance of brain age models. In dimensionality reduction, the three evaluation strategies have also been considered, but in this case the proportion of studies which considered *k*-fold [117,118,124,131] is equal to the holdout test set [3,9,29,132]. Two studies assess the model performance on an external test set [130, 147], but only Cole *et al.* [147] compared the model performance of a holdout with an external test set. The result showed a two-year increase in MAE on data acquired in different acquisition settings. This discrepancy across acquisition settings might be due to different acquisition parameters and/or scanner types, which lead to different contrast and noise levels on T_1-weighted images [148].

3.3 Deep Learning

This section summarises recent studies on brain age modelling using deep learning. The section begins with an introduction to CNN and the architectures used to predict brain age. The different inputs and ensembly strategies considered in this context are overviewed in section 3.3.2 and 3.3.3, respectively. CNN are considered black-box models; the recent strategies to uncover the model predictions are discussed in subsection 3.3.4. Finally, the section 3.3.5 finishes with a comment on the evaluation strategy used to assess the brain age model performance.

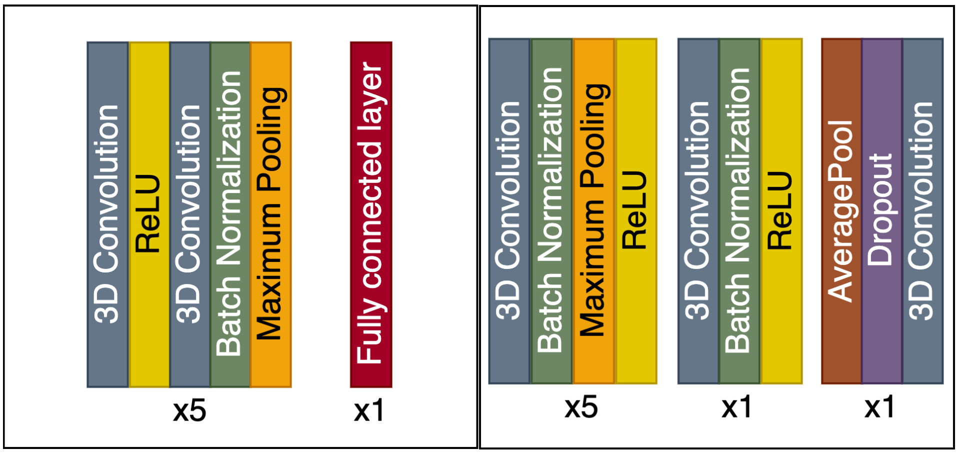
3.3.1 Deep learning architectures and optimisation strategies

CNN were developed in the context of visual recognition tasks [32]. Their outstanding performance led to their application in numerous fields, namely neuroimaging [135]. CNN typically consist of five fundamental layers: convolution, activation, pooling, batch normalisation, and dropout [32, 135]. In the convolution layer, multiple kernels, also known as filters or feature extractors, are convolved with the input image to extract relevant features. The activation layer applies a non-linear function to introduce non-linearity into the network, enabling it to learn complex relationships between features. Pooling is an operation that reduces the spatial dimensions of the feature maps by selecting representative values from a set of pixels and downsampling the data. Batch normalisation [149] is employed to normalise the weights of each layer, mitigating the vanishing gradient problem and preventing overfitting. Lastly, dropout [150] is a regularisation technique that prevents overfitting by randomly dropping a given proportion of connections between two layers during training, thus encouraging the network to learn more robust and generalisable features.

One of the early contributions to deep learning was made by Cole *et al.* [3]; the authors proposed a custom architecture, a scheme is shown in Figure 3.5a. The network contains five convolution block architectures; each convolution block comprises a 3D convolution, rectified linear units (ReLU), 3D convolution, batch normalisation, ReLU, and a maximum pooling layer. A fully connected layer succeeded the convolution blocks, followed by the output layer. In total, the network contains 889960 parameters. The authors compared the network's performance in age prediction with a shallow pipeline. The study revealed that CNN models were a promising alternative to shallow learning. Since then, deep learning has gained widespread application in brain age estimation [114]. A summary of some deep learning studies on brain age is presented in Tables 3.5 through 3.7.

Various CNN architectures have been explored in the context of age prediction. Some studies adapted to 3D established 2D networks renowned for their performance in visual recognition tasks, such as visual geometry group network (VGG) [127, 132], residual network (ResNet) [4, 29, 151], U-Net [133] or dense convolutional network (DenseNet) [152]. Nevertheless, these networks have a high number of parameters. 3D ResNet-18, 3D ResNet50, and 2D VGG contain 33.2 million, 46.2 million, and 133 million, respectively [4]. Thus, these networks are computationally demanding and might be prone to overfitting for small training sizes. To overcome this constraint, some studies tailored their architectures. Custom architectures vary in layer combinations [3, 4, 122]. Simple fully convolutional network (SFCN) [4], proposed by Peng *et al.*, is a lightweight network that won the 2019 Predictive Analysis Challenge for brain age prediction. The network contains approximately 3 million parameters, the architecture is depicted in Figure 3.5b. As can be observed, it consists of five convolution blocks, each encompassing a 3D convolution layer, a batch normalisation, maximum pooling and the ReLU activation function. The five-block is followed by a similar block without the pooling layer; finally, the layer finishes with a block containing average pooling, dropout and a convolution layer. The authors compared the performance of the SFCN with four ResNet architectures [153]. Despite the absence of statistical analysis, the results suggest that the proposed architecture outperforms all four ResNet architectures. Given the remarkable performance of SFCN and its simplicity, it has been applied to multiple brain age studies [30, 123, 125, 154].

The CNN training involves adjusting network weights using gradient descent-based algorithms. For the studies considered in this state-of-the-art, the optimisation



(a) Cole *et al.* [3] **(b)** SFCN [4]

Figure 3.5: Representation of custom CNN architectures widely used for age prediction. The architecture proposed by Cole *et al.* [3] marks the beginning of deep learning on brain age prediction. SFCN [4] is a lightweight network with formidable performance in age prediction.

methods considered were stochastic gradient descent [3,4,30,127,132] and Adam [4, 29,125,133,151,152,154,155]. The weights are updated to minimise a predefined loss function, which computes the error between the real and the predicted age. The three loss functions considered in this field are: MAE [3, 29, 30, 132], mean squared error (MSE) [122, 125, 127, 154], and root mean squared error (RMSE) [126]. The CNN weights are optimised through multiple iterations or epochs. Some researchers train their models for a fixed number of epochs and do not perform model selection [132, 154, 155]. Others select the model that performed the best over the last few epochs or since the beginning of training [4, 30, 125].

3.3.2 Comparison of the diverse input approaches

Images are the substrate of CNN models. As discussed in section 3.1.1, different preprocessing levels result in distinct image types. Each study employs one or more preprocessing byproducts of T_1-weighted images or other MRI modalities.

Two studies explored raw T_1-weighted images to predict the age [3,152]. The findings evidence that raw images generally lead to inferior [3, 152] and less reliable results compared to segmented images [3] or minimally processed images [152]. These results might be related to the higher noise, artefacts, or irrelevant information that can hinder accurate age estimation.

Most brain age models learn age patterns from minimally processed T_1-weighted [3,4,29,123,152,154–156] or segmented images [3,4,29,123,126,127,133,151]. Despite the lack of proper statistical comparison between minimally processed images and segmented images, the studies report congruent results; i.e. minimally processed images yield better results than any segmented image [4, 29, 123]. GM emerge as the segmented image most often considered for brain age prediction [3, 4, 29, 123, 126, 127, 132, 151], followed by WM [3, 4, 29, 123, 133]. Most of the studies suggest that GM yields better performance than WM [3, 4, 123, 133]. In contrast, Jonsson *et al.* [29] report that a model trained with WM outperforms GM. However, the MAE difference observed is 0.09 years, yet a statistical was not performed to compare the performance of different models.

Deformation fields have also been considered to estimate the brain age [29, 155]. Nonetheless, the results are dissonant. Jonsson *et al.* [29] reported that deformation fields yield a higher MAE than either minimally processed, GM or WM. On the contrary, He *et al.* [155] reported superior performance of deformation fields when compared to minimally processed. The distinct preprocessing pipelines or CNN architectures might explain the different conclusions.

3.3.3 Exploring ensemble strategies

Multiple image types and models can be combined to attain lower MAE in age prediction. Three fusion strategies are considered in brain age: input, layer and model level; a scheme is shown in Figure 3.6. The former combines diverse modalities for richer input representation. Layer fusion merges features from different networks. Model fusion combines predictions from multiple models using an aggregation function. Currently, there is no systematic comparison to determine the best fusion level in BrainAGE. He *et al.* [155] compared the three fusion levels. Despite the absence of statistical analysis, the findings suggest that, independently of the level, fusion yields better results than a single model. Furthermore, layer-level fusion yields the highest performance compared to input or model fusion. Nevertheless, dissonant results are reported by Cole *et al.* [3]. The authors combined GM, WM at the input level and assessed the model on a holdout dataset. The results highlight that the fusion of GM and WM yields better performance than training a model solely with WM, yet the fusion model underperforms comparatively to a GM model. The contradictory results concerning fusion at the input level might be explained by the type of inputs considered. Cole *et al.* [3] compared GM and WM, He *et al.* [155] combined minimally processed and deformation fields.

Most studies adopt the fusion at the model level [29, 125, 152, 154, 155]. The results are congruent, model fusion leads to superior performance compared to using a single image per model [4, 29, 125, 152, 154, 155]. Different combinations have been performed. Jonsson *et al.* [29] combined the predictions of minimally processed image, deformation fields and segmented images (GM, WM, and CSF). Peng *et al.* [4] combined the prediction of four models trained with minimally processed linear registered, minimally processed non-linear registered, GM and WM. Some studies combine models trained with other MRI modalities [125, 152, 154]. Wood *et al.* [152] merged the output predictions of T_1-weighted with Axial T_2. Despite the lack of proper statistical analysis, combining prediction from multiple modalities seems to yield better results. The same conclusion is suggested by Mouches *et al.* who evaluated the added value of magnetic resonance angiography [125, 154]. Shallow and deep learning models have also been combined [125], with the results suggesting a superior performance when both strategies are combined.

Another fusion strategy encompasses the aggregation of the predictions from multiple models trained with the same data on the same architecture. The fusion is performed at the output level either by averaging the multiple predictions [4, 123, 127], or performing a regression on the model predictions, thus assigning each model prediction a different weight [122, 128]. Combining the predictions of multiple models with different initial parameters achieves lower MAE than using a single model [4, 122, 123, 127, 128].

3.3.4 Explainability

Deciphering the reasons behind the predictions of CNN models is challenging due to the high number of parameters these models have. Two strategies have been considered to overcome this limitation: predicting the age at a lower level (slice or subvolume) or computing the sensitivity maps. Slice-based models enable the scrutiny of predictions at a slice level, facilitating the explainability of the model predictions. This approach has some advantages. Firstly, it reduces the computational complexity compared to using the whole 3D volume and increases the number of training instances. Ballester *et al.* [151] trained a 2D-CNN with MRI slices; the results outline that the model's performance depends on the slice and the MRI plane

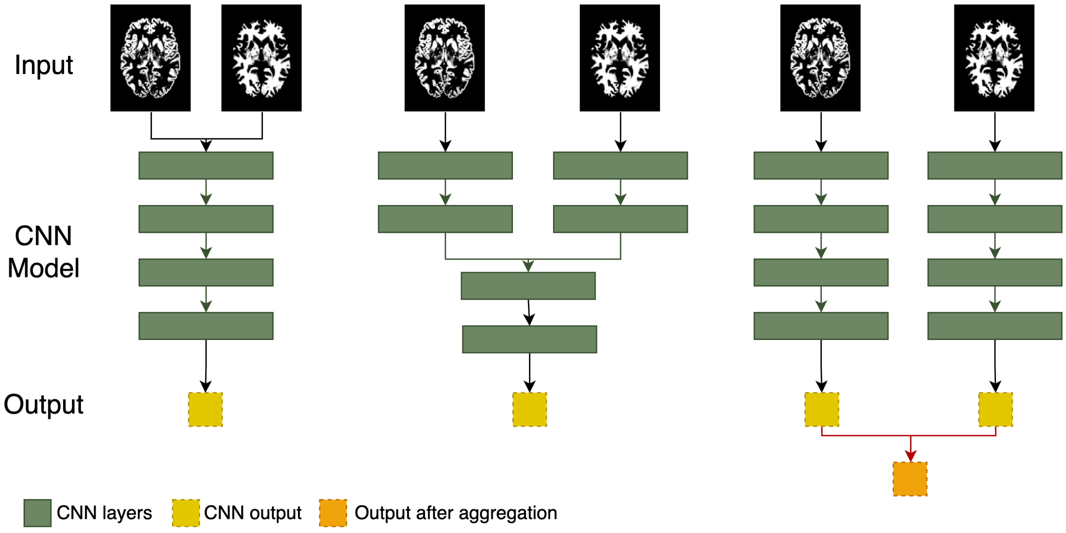


Figure 3.6: Representation of different fusion strategies used in brain age. An example of combining information from a grey- and white-matter. On the left, the images are combined at the input level; on the centre, the features extracted from each image are combined at the layer level; on the right, the predictions of each network are combined using an aggregation function at the model level.

(axial, sagittal and coronal). Therefore, valuable information might be lost with the transition from the volume to the slice level. Another option is to increase the explainability of the brain age per region using a patch-based approach [113,133,157]. A MR image is divided into smaller subvolumes or patches, and the age is predicted per patch. Thus, in this case, different brain regions may exhibit different age predictions. This approach is consistent with the current literature. It is believed that some brain regions undergo ageing faster than others, such as the hippocampus [41–43]. Insights on the abnormally ageing brain areas can be valuable in differentiating between diseases and understanding their progression. Nonetheless, currently, the prediction at the patch level yields the poorest performance compared with either the whole 3D volume or slice-level models. Such results might be explained by the reduced information these models contain to make a prediction compared to the slice or global brain age prediction.

The sensitivity maps emerge from the growing necessity to interpret models. These maps translate the influence that each pixel exerted on a prediction. Grad-CAM [158] and SmoothGrad [159] are two strategies that generate sensitivity maps by computing the influence of each pixel using the partial derivative of the input volume with respect to the prediction. These methods have been used to unveil the voxel's influence on age predictions [113, 122, 125, 126]. The results are congruent across studies, the regions with higher influence on age prediction include the ventricles and the subcortical regions [113, 122, 125, 126]. These results are consistent with the findings discussed in section 2.1.1. Additionally, Dinsdale *et al.* [127] shown that the registration process can affect the salience maps. In non-linear registered images, the maps highlight areas around the ventricles, while in linear registration, no specific region significantly contributes to the prediction. The explanations from sensitivity maps should be carefully analysed. A systematic comparison of eight salience map techniques, including Grad-CAM and SmoothGrad, based on four trustworthiness criteria (reproducibility, reliability, localisation utility, and sensitivity to model weights) revealed that all methods fail to meet at least one criterion [160]. Supporting this finding, Levakov *et al.* [122] concluded that the initial weights of the model impact on its explainability.

The integration of a patch-based model with sensitivity maps might help to increase the reliability of the explanations. Hepp *et al.* [113] showed that the regions with higher influence on predictions yield higher performance in the patch-based approach. Similarly to the other studies [122, 125, 126], the regions with more substantial influence were in the ventricles and subcortical regions. In these areas, MAE of the patch-based approach was around five years, whereas in the overall MAE was 11.97 ± 12.78.

3.3.5 Evaluation approaches

Similarly to shallow learning, the evaluation strategy differs across studies. Holdout test set [3,4,29,122,123,125–127,132,133,152,156] is the strategy extensively adopted to assess the performance of deep learning brain age models. Nevertheless, *k*-fold [154, 155] and external test set [29, 30] have also been considered.

The results obtained using the holdout sample and *k*-fold approach are consistent across studies that focused on the same age range. For studies that included participants in late adulthood only, the MAE typically ranged between 2 to 3 years [4, 123, 127]. However, studies that included subjects in early adulthood exhibited a higher MAE [3,29,122,125,126,132,152,154,155]. Unsurprisingly, models evaluated on external datasets tended to have higher MAE compared to those using a holdout sample [29, 30]. Jonsson *et al.* [29] reported a MAE of 3.39 years and 8.49 years for the holdout and external test set, respectively. Similarly, Leonardsen *et al.* [30] explored the model performance on multiple external test sets. The authors outline that the MAE varied across external test sets, with the lowest and highest reported MAEs being 2.82 and 6.94 years, respectively. These findings highlight that the performance on the holdout test set is not representative of the model performance on different acquisition settings. Although the holdout images were not used for training and validation, the model is trained with images acquired in the same acquisition settings as the holdout test set. Therefore, the model images might be biased towards the acquisition setting patterns. Jonsson *et al.* [29] showed that if the models were retrained with some images from the external test set, while the remaining samples were used as a holdout test set, the MAE decreased from 8.49 to 3.63 years. Furthermore, the MAE variability reported by Leonardsen *et al.* [30] on multiple external test sets is evidence that a single external test set is insufficient to represent the model performance on unseen data.

3.4 BrainAGE: A complex landscape

BrainAGE has been associated to multiple diseases [9], which suggest that this putative biomarker encodes an acceleration of the ageing process.

BrainAGE seem to capture the disease-related morphological brain changes in preclinical stages. Alzheimer's disease (AD), the most prevalent neurodegenerative disease, consistently exhibits statistically significant differences in BrainAGE when comparing prediction results between healthy controls and the AD group [117, 119, 131, 161–168]. Mild cognitive impairment (MCI), a prodromal stage of AD, is a condition associated with a decline in cognitive performance. Studies evidence that MCI is also related to an increased BrainAGE, albeit with lower magnitude [28, 119, 131, 161–163, 165, 168, 169]. BrainAGE yields higher accuracy rates in the prediction conversion of conversion from MCI than CSF markers, structural atrophy measures

Table 3.5: Overview of some manuscripts that predict brain age using deep learning.

Study	Subjects	Age range [years]	Evaluation	Input (fusion approach)	Architecture	R^2	MAE [years]
Cole *et al.* [3]	2001	18-90	holdout test set	GM	custom CNN	0.92	4.16
				WM	custom CNN	0.88	5.14
				GM and WM (input level fusion)	custom CNN	0.91	4.34
				raw	custom CNN	0.88	4.65
Jonsson *et al.* [29]	1264	18-75	holdout test set	MP	ResNet	4.006	0.829
				GM	ResNet	4.641	0.776
				WM	ResNet	4.189	0.812
				DF	ResNet	4.804	0.758
				MR, GM, WM and DF (model fusion level)	ResNet	3.388	0.872
			external test set	MP, GM, WM and DF (model fusion level)	ResNet	-0.630	8.494
Jiang *et al.* [132]	1303	18-90	holdout test set	GM FPN	VGG	0.76	5.55
Levakov *et al.* [122]	9158	n.s.	holdout test set	MP	custom CNN	n.s.	3.07
Peng *et al.* [4]	12949	44-80	holdout test set	MP, MP*, GM and WM (model fusion level)	SFCN	n.s.	2.14 ± 0.05
					ResNet	n.s.	2.50 ± 0.06
Gong *et al.* [123]	7896	n.s	holdout test set	MP, MP*, T1, GM and WM (model fusion level)	SFCN	n.s.	2.98
Leonardsen *et al.* [30]	53542	n.s.	external test set	MP	SFCN	n.s.	2.82
			external test set	MP	SFCN	n.s	6.94
Mouches *et al.* [125]	1340	n.s	holdout test set	MP	SFCN	0.872	4.01 ± 3.08
				TOF MRA	SFCN	0.805	4.11 ± 3.75
				MP and TOF MRA (model fusion level)	SFCN	0.882	3.85 ± 2.90
Mouches *et al.* [154]	1658	21-81	5-fold	MP	SFCN	n.s	4.20 ± 0.18
				cortical features	DL	n.s	5.54 ± 0.20
				MP/ cortical features (model fusion level)	SFCN/DL	n.s	4.11 ± 0.08

TOP MRA: Time of Flight Magnetic Resonance Angiography; MP: T_1-weighted Minimal Processed; GM: Grey Matter; WM: White Matter; DF: Deformation fields; FPN: Frontoparietal Network; CNN: Convolution Neural Network; SFCN: Simple Fully Convolutional Network; ResNet: Residual Neural Network; VGG: Visual Geometry Group Network; DL: Deep Learning; n. s.: not specified.

Table 3.6: Overview of some manuscripts that predict brain age using deep learning.

Study	Subjects	Age range [years]	Evaluation	Input (fusion approach)	Architecture	R^2	MAE [years]
Dinsdale *et al.* [127]	6223	n.s	holdout test set	male GM	VGG	n.s.	3.09 ± 2.37
He *et al.* [155]	6049	0-97	5-fold	DF	custom CNN	n.s.	4.09 ± 0.10
				minimally processed	custom CNN	n.s.	3.58 ± 0.26
				MP and DF (input fusion level)	custom CNN	n.s.	3.31 ± 0.13
				MP and DF (layer fusion level)	custom CNN	n.s.	3.12 ± 0.22
				MP and DF (decision fusion level)	custom CNN		3.38 ± 0.12
Hofmann *et al.* [128]	2016	18–82	10-fold	MP	custom CNN	n.s.	4.11
				FLAIR	custom CNN	n.s.	4.16
				SWI	custom CNN	n.s.	5.74
				MP, FLAIR, SWI (decision fusion level)	custom CNN	n.s.	3.86
Wood *et al.* [152]	1551	18-95	external test set	Axial T2-weight 3D CNN	DenseNet	n.s.	3.83
				MP	DenseNet	n.s.	3.86
				raw	DenseNet	n.s.	4.86
				MP + Axial T2	DenseNet	n.s.	3.35
Ballester *et al.* [151]	2639	n.s	holdout test set	GM slice sagittal	ResNet	n.s.	4.52
				GM slice coronal	ResNet	n.s.	5.04
				GM sliceaxial	ResNet	n.s.	5.09

FLAIR: Fluid-attenuated inversion recovery; SWI: Susceptibility weighted imaging; MP: T_1-weighted Minimal Processed; GM: Grey Matter; WM: White Matter; DF: Deformation fields; CNN: Convolution Neural Network; ResNet: Residual Neural Network; DenseNet: Dense Neural Network; VGG: Visual Geometry Group Network; DL: Deep Learning; n. s.: not specified.

Table 3.7: Overview of some manuscripts that predict brain age using deep learning.

Study	Subjects	Age range [years]	Evaluation	Input (fusion approach)	Architecture	R^2	MAE [years]
Lam *et al.* [156]	7312	45-81	holdout test set	MP slice - sagittal	CNN and LSTM	n.s.	2.86
				MP slice - coronal	CNN and LSTM	n.s.	2.98
				MP slice - axial	CNN and LSTM	n.s.	6.25
Popescu *et al.* [133]	3463	18-90	holdout test set OASIS3	patches (GM+WM)	U-Net	n.s.	8.09 ± 6.08
			holdout test set AIBL	patches (GM+WM)	U-Net	n.s.	10.23 ± 7.08
			holdout test set Wayne State	patches (GM+WM)	U-Net	n.s.	8.08 ± 6.40
Hepp *et al.* [113]	10691	20-72	5-fold	MP	ResNet	n.s.	3.21ś2.45
				MP - patch based	ResNet	n.s.	11.97 ± 12.78

MP: T_1-weighted Minimal Processed; GM: Grey Matter; WM: White Matter; CNN: Convolution Neural Network; ResNet: Residual Neural Network; DL: Deep Learning; LSTM: Long Short-Term Memory; U-Net: U Network; n. s.: not specified.

or cognitive scales [162]. Furthermore, type 2 diabetes (T2D), which is a risk factor for AD and is associated with higher brain atrophy rates [97], is also linked to higher BrainAGE [170,171]. Schizophrenia have consistently an increased BrainAGE compared to healthy controls [121, 137, 172–176]. The first episode of psychosis seems to be associated with higher brain atrophy [74]; increased BrainAGE has also been reported in individuals that experienced their first episode of psychosis [177, 178]. Moreover, individuals with a higher risk of abnormal neurodevelopment had increased BrainAGE compared to healthy controls [179]. Therefore, BrainAGE could be used as a screening tool for the early identification of clinical conditions at their pre-clinical stage.

BrainAGE seems to tackle brain atrophy and could also be used as a prognostic tool in some diseases. BrainAGE has been reported to be increased in patients with multiple sclerosis [180]. The results suggest that the BrainAGE is related to the degree of disability and the disease progression. Moreover, BrainAGE has been shown to increase in individuals who suffered a traumatic brain injury [181–183] and correlates with the amount of time since the traumatic brain injury occurred. Epilepsy also yields an increased BrainAGE [184,185]. Verma *et al.* [184] reported that epileptic subjects with frequent seizures, at least one weekly, yield higher BrainAGE than patients with less seizures. These findings suggest that BrainAGE might capture the trajectory of the underlying pathological process and be used as a prognosis tool.

However, discrepant results are reported in some disorders. Bipolar disorder, whose symptomatology is very similar to schizophrenia in its early stages, has yielded incongruent findings regarding its association with BrainAGE [172–174, 176]. Similarly, major depressive disorder has exhibited elevated BrainAGE in some research [186], while contradictory evidence has also been reported [187]. The reported discrepancies might be attributed, in part, to variations in the choice of models used to measure BrainAGE. Lee *et al.* [137] assessed the impact of different shallow learning pipelines on BrainAGE. The authors compared the BrainAGE on two schizophrenia datasets. The results highlight that the dataset and the model choice influence BrainAGE results. Despite BrainAGE being considered significant in almost all models on both datasets, except for one case, its value varied from 3.8 to 5.2 years on one dataset and between 4.5 to 11.7 on another.

In conclusion, consistent associations have been observed in some conditions such as AD, MCI, schizophrenia, multiple sclerosis and epilepsy. Nonetheless, dissonant results have been found in conditions such as bipolar disorder and major depressive disorder. Figure 3.7 summarises conditions considered in this state-of-the-art, which are reported to yield a superior BrainAGE. The choice of the model and dataset used to measure BrainAGE might be a potential factor contributing to the inconsistencies. Despite the complexities and discrepancies, exploring BrainAGE remains a promising avenue for early diagnosis in prodromal clinical stages and

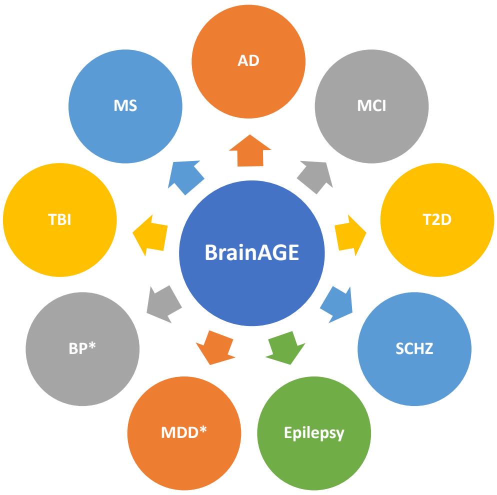


Figure 3.7: Pathologies in which the brain age gap estimation (BrainAGE) is suggested to yield significant differences with healthy controls. The asterisk (*) denotes the pathologies with inconsistent results in the literature. Abbreviations: Alzheimer's Disease (AD); Mild Cognitive Impairment (MCI); Type 2 Diabetes (T2D); Schizophrenia (SCHZ), Major Depressive Disorder (MDD), Bipolar disorder (BP), Traumatic Brain Injury (TBI); Multiple Sclerosis (MS).

disease prognosis.

3.5 Conclusion

Various preprocessing levels of T_1-weighted images have been considered in brain age context: raw, minimally processed, and segmented images. The results outline that raw images are unreliable across scanners. Furthermore, models tested on an external test set, in which the acquisition settings differ from the training and validation data, yield higher MAE compared to the performance on a holdout dataset. An explanation for this finding might be the fact that images contain information concerning the acquisition settings. Thus, MRI preprocessing is vital to reduce noise and remove undesirable information. Nevertheless, currently, there is no gold-standard strategy in preprocessing; mainly, two preprocessing frameworks are considered in the state-of-the-art: FreeSurfer and SPM.

Shallow and deep learning have been considered for modelling brain age. Two shallow learning strategies have been considered, one based on handcrafted features and the other on dimensionality reduction. The former approach yields multiple inconsistencies, which might result from differences in preprocessing or in the machine learning pipeline. On a positive note, the models based on a dimensionality reduction approach are more consistent. Nevertheless, the state-of-the-art highlighted that there is no gold-standard machine learning pipeline.

Deep learning emerged as an alternative to shallow learning; the results suggest that deep learning tends to outperform shallow learning in age prediction. Currently, these models encounter two challenges: explainability and generalisability. To solve the former, sensitivity maps were considered, and the results are congruent across studies. Thus, sensitivity maps might be a valuable tool to uncover the reasons behind a prediction. Concerning generalisability, deep learning models generalise poorly on data acquired in different acquisition settings. As previously mentioned, this might also be a problem of preprocessing, but it might also be due to a model bias towards the acquisition settings of the training data.

Finally, BrainAGE is a promising candidate for the early diagnosis and the prognosis of different conditions. This putative biomarker seems to capture the atrophy of the brain and could be used as a tracker of brain ageing healthy status. Nevertheless, BrainAGE is increased in multiple conditions; thus, to be considered in clinical practice for early diagnosis, the specificity of this putative biomarker should be addressed.

Chapter 4

Cortical thickness in brain imaging studies using FreeSurfer and CAT12: A matter of reproducibility

This chapter is based in the following article:

Dias, Maria de Fátima Machado, Paulo Carvalho, Miguel Castelo-Branco, and João Valente Duarte. "Cortical thickness in brain imaging studies using freesurfer and cat12: A matter of reproducibility." Neuroimage: Reports 2, no. 4 (2022): 100137.

4.1 Abstract

A reproducibility crisis has been reported across many research fields, including neuroimaging, reaching up to 70% of studies. Neuroimaging data, such as magnetic resonance imaging (MRI), requires preprocessing to allow for inter-subject comparison, increase signal contrast and noise reduction. As manual MRI preprocessing is time consuming and requires expertise, multiple automatic preprocessing frameworks have been proposed. However, neuroimaging studies often report divergent results, even for similar populations, thus it is important to determine whether this occurs as a result of different processing tools. Two of the most used tools are FreeSurfer and the computational anatomy toolbox (CAT12). In this study, we assessed the reproducibility between these two automatic preprocessing frameworks for structural MRI and the test-retest reliability within the framework on estimating cortical thickness. Our results show that the reproducibility between the frameworks is lower at the region-of-interest (ROI) level than at the individual level. Furthermore, we found that the reproducibility was lower in paediatric samples than in adults. Finally, an acquisition site effect was also identified. Given the widespread use of these frameworks in basic and clinical neuroscience, the results of multi-centric cross-sectional studies must be interpreted with caution, particularly with paediatric samples. The observed reproducibility issue might be one of the sources of discrepancies reported in neuroimaging studies. On a positive note, framework test-retest reliability within subject is high, suggesting that inconsistency of results may be less concerning in longitudinal studies. The code is available at: https://cibit-uc.github.io/fs-cat12-cortical-thickness-reproducibility.

4.2 Introduction

Structural MRI allows to estimate various morphometric features of the brain such as cortical thickness, which is widely used in basic and clinical research. The cortical thickness of healthy brains in humans measures, on average, 2.5mm [188], and is associated with cognitive function [189, 190]. Cortical thickness is reported to be correlated with disease progression in Alzheimer's disease [189, 191], schizophrenia [51, 192], and Parkinson's disease [193–195]. Nonetheless, inconsistent results have been reported regarding cortical thickness changes in neurodevelopment [196] and in multiple diseases such as autism spectrum disorder (ASD) [197–200] bipolar disorder [201], or major depression disorder [202].

Cortical thickness corresponds to the distance between the pial surface of the brain and the white matter boundary. Despite this straightforward definition, the evaluation of the cortical thickness is difficult given the inherent challenge of creating accurate thickness estimations due to the highly folded brain morphology. The manual preprocessing of magnetic resonance (MR) images is time consuming and requires field expertise, therefore different computational frameworks have been proposed to automatically process MR images and to compute morphometric measures [112, 203–205]. The automation has several advantages: it does not require deep user expertise, it can be parallelized, errors are systematic rather than random, and its reproducibility is amenable to explicit testing. Yet, currently, there is no gold standard preprocessing framework. Moreover, each framework uses its implementation to preprocess MRI data and extract brain tissue estimations such as cortical thickness. The different implementations can be a source of variability, which might explain inconsistencies reported in literature, as mentioned above.

This study assesses the cortical thickness estimation's reproducibility between two of the most widely used preprocessing frameworks: FreeSurfer and CAT12. The cortical thickness estimations' reproducibility between CAT12 and FreeSurfer was evaluated previously in very small sample size studies [206, 207], in which authors reported a mean R^2 and correlation between frameworks, for cortical thickness, of 0.89 and 0.92, respectively. However, the studies exhibit multiple limitations: the number of subjects used in the study was very low; the studies only focused on an adult population, and therefore did not consider the changes that occur in early life periods such as childhood and adolescence; and the studies did not assess whether the reproducibility depends upon the site of acquisition setting and individual age. Another study [208] also evaluated the preprocessing pipeline impact on cortical thickness estimation. Nevertheless, the study did not assess the reproducibility at the subject level. To overcome these shortcomings, the current work uses MRI data from three open data sharing initiatives with a large number of subjects. Furthermore, our study investigates whether the reproducibility between the two frameworks depends upon the acquisition setting, the subjects' age and brain region ROI. In particular, the objectives of the current work are (a) to compare the test-retest reliability within each framework, (b) to investigate whether the reproducibility of FreeSurfer and CAT12 frameworks depends on the age and neurodevelopmental stage, and (c) to assess these frameworks' reproducibility on different acquisition settings and verify whether different frameworks yield different results modelling age using cortical thickness values.

4.3 Material and Methods

This work evaluates the reliability of cortical thickness estimations within two frameworks: CAT12 and FreeSurfer, and the reproducibility of cortical thickness estimations between these two frameworks. The study was divided into three objectives:

- Objective (a) assesses and compares the test-retest reliability of each framework (FreeSurfer and CAT12). In this objective, we analysed two T_1-weighted images (Run_1 and Run_2) from each participant and cortical thickness is extracted for both using both frameworks. The reliability metrics are computed within framework using the cortical thickness estimations of the two images. Therefore, two groups were considered in this objective and the participants were the same in both groups. For this analysis the Open Access Series of Imaging Studies (OASIS)-3 dataset was used, all participants who had two consecutive images acquired within less than 60 minutes were included.

- Objective (b) compares the reproducibility metrics between two age-groups, the paediatrics group i.e., individuals with age lower or equal than 18 years old, and the adults group, i.e., individuals with age greater than 18 years old. For this objective the autism brain imaging data exchange (ABIDE) I dataset was used.
- Objective (c) assesses the reproducibility between FreeSurfer and CAT12 across different acquisition sites. In this objective two datasets were used: the Information eXtraction from Images (IXI) and OASIS-3 datasets, the former has three acquisition sites, whereas the latter has one acquisition site. Therefore, four groups were compared, one per acquisition site.

Figure 4.1 summarises the groups and data used for each objective.

4.3.1 Data

Three open sharing initiatives were used in this work: OASIS-3 [209], ABIDE I [209] and IXI [210]. The demographics statistics and the acquisition protocol are shown in Tables A.3, A.2, A.3 in the Appendix A. Each of the open sharing initiatives used in this work is described below.

4.3.1.1 OASIS-3

OASIS-3 is an online repository that comprises images of the same individuals captured from multiple modalities, such as MRI and positron emission tomography. Each subject has one or more session, and a session may have multiple runs of the same MRI modality. All images of the OASIS-3 dataset were acquired in the same research centre over the course of 30 years, using different scanners. In this work, only images acquired with the Siemens TIM Trio 3T scanner were used. For the objective (c) all the subjects considered cognitively normal at the first MRI session were used, which performed a total of 494 participants, three participants had preprocessing errors, in total 491 subjects were considered. For the test-retest analysis it is required that each subject has two T_1-weighted images acquired, at maximum, 60 minutes apart, 299 subjects fulfilled this requirement, three subjects were excluded from the subset due to preprocessing errors, thus in total 296 subjects were used in the analysis.

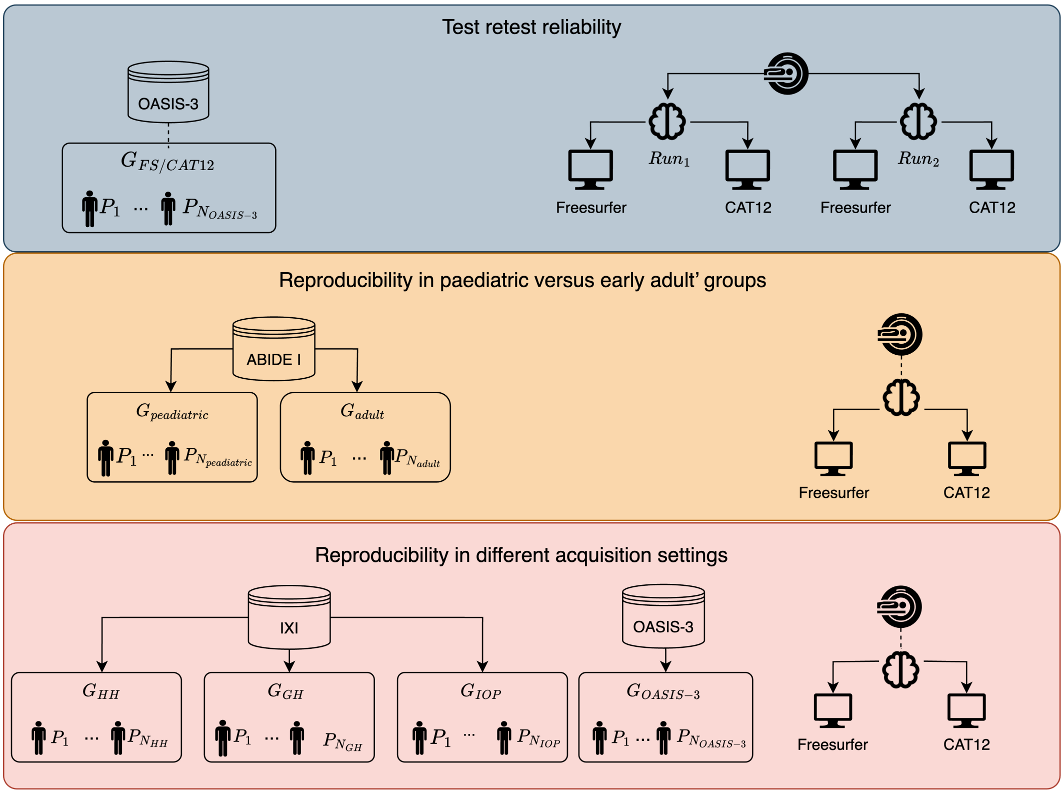


Figure 4.1: Depiction of the objectives of this work as well as the data, groups (G) and pipeline used in each objective. In this figure N represents the total number of participants of a given group.

4.3.1.2 ABIDE I

ABIDE I is a multi-site data sharing initiative (17 sites in total) with neurotypical controls and subjects diagnosed with ASD. In this study only the neurotypical controls were considered. Five subjects were discarded due to errors during the preprocessing step, in total 566 subjects were analysed.

4.3.1.3 IXI

IXI is a dataset collected from healthy adults in three distinct hospitals in London: the Hammersmith Hospital (HH), the Guy's Hospital (GH) and the Institute of Psychiatry (IOP). Philips MRI scanners with 3T and 1.5T were used to acquire brain images in HH and GH, respectively. The IOP images were acquired with a scanner from General Electrics with a field strength of 3T. In total, this dataset contains data collected from 562 subjects; from these a subset composed of 558 subjects were successfully preprocessed in this study, the demographics of the IXI dataset is in Table A.2 in the Appendix A.

4.3.2 Frameworks

4.3.2.1 CAT12

CAT12 (http://www.neuro.uni-jena.de/cat/) is a computational framework to automatically preprocess T_1-weigthed MR images. Its routines are implemented in MATLAB. In this work the default preprocessing routine (*Segment*) of CAT12 was used to preprocess the T_1-weighted images. The CAT12, SPM12 and MATLAB version were version 1742, v7771 and R2020a (9.8.0.1323502), respectively. The backbone of the segmentation routine is the unified segmentation algorithm [203]. The segmentation step relied on tissue probability maps (TPM) to differentiate the brain tissues. The TPM must be properly adjusted for morphological changes that occur in the human brain during neurodevelopment [211]. CAT12 integrates the toolbox template-o-matic to generate customised TPM based on the age and gender of the subjects in the dataset. A paediatric TPM was generated for the ABIDE I dataset, using all the individuals with age lower or equal to 18 years old. In the CAT12 preprocessing, in the paediatric group only, the customised TPM was then used.

4.3.2.2 FreeSurfer

FreeSurfer [204] is an established computational framework in the neuroscience community [212]. This framework contains a routine called *recon-all* which implements all the required steps to preprocess T_1-weighted MR images and extract cortical thickness estimations. In this work the images where preprocessed using the recon-all routine of the FreeSurfer version 7.1.1.

4.3.2.3 MRIQC

The MRI quality control tool (MRIQC) [213] is an open-source framework which extracts different quality metrics from the T_1-weighted images. This framework was used to extract the quality metrics from the MRI images. The signal-to-noise ratio (SNR) is one of the metrics extracted and was the metric chosen to represent the quality of an image.

4.3.3 Cortical Parcellation

A standard approach in cortical thickness analysis is the division of the cortical ribbon into several ROI according to a given template. In this work, the cortical template considered was the Destrieux template [214]. This template divides each hemisphere into 74 ROIs based on a sulco-gyral parcellation. The ROIs are then grouped into five lobes: frontal, insula, temporal and occipital, parietal and limbic. A ROI may be in more than one lobe, for instance the a*nterior part of the cingulate gyrus and sulcus* is in the frontal and limbic lobe.

4.3.3.1 Reproducibility and Reliability Metrics

The reproducibility and reliability between and within CAT12 and FreeSurfer, respectively, is assessed in this work. The metrics extracted were the same for both: the coefficient of determination (R^2) of a regression and the intraclass correlation coefficient (ICC), more specifically the ICC(3,1) proposed by [215]. The computation methodology is explained below.

Test-retest reliability

The test-retest reliability aimed to assess the reliability of cortical thickness estimations within framework over a short period of time (Run_1 and Run_2). A fitting was performed in which the Run_1 and Run_2 cortical thickness estimations were the independent and dependent variables, respectively, from the fitting the R^2 was extracted. The ICC was computed comparing the results of the two runs.

Reproducibility

To obtain reproducibility metrics between CAT12 and FreeSurfer a linear regression was fitted to the data in which the independent and dependent variables were the cortical thickness estimations extracted by FreeSurfer and CAT12, respectively. The R^2 of the fitting was one of the reproducibility metrics, the other was the ICC which was computed comparing the cortical thickness estimations extracted by both frameworks.

4.3.4 Statistics

Two analyses were carried out per objective: a "participant analysis" and a "ROI analysis". In the former reproducibility/reliability metrics were extracted per individual whereas in the second metrics were computed per ROI.

For each objective an overall analysis, participant and ROI analysis was performed as described in subsections 4.3.4.1, 4.3.4.2 and 4.3.4.3, respectively.

A significance threshold of 0.05 was considered for statistical analyses in this chapter.

4.3.4.1 Overall analysis

In this analysis all the cortical thickness estimations for all ROIs and individuals were considered. Two plots were used to show the relationship of two estimates of cortical thickness: a regression plot and a Bland-Altman plot. The former shows the relation of the two estimates, whereas the latter depicts the mean and mean difference of these estimations.

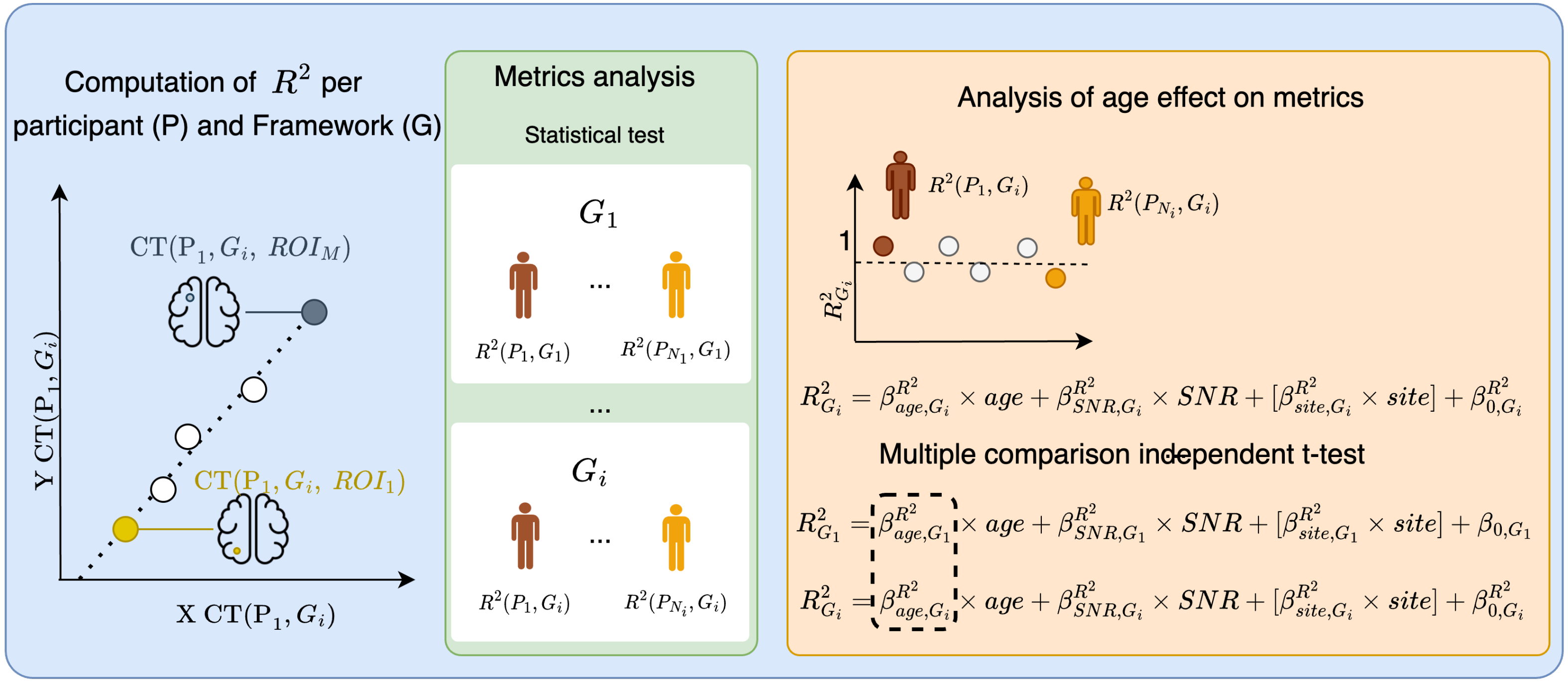
4.3.4.2 Participant Analysis

The workflow of the participant analysis for the reliability/reproducibility study is depicted in Figure 4.2. In this analysis the reproducibility/reliability metrics are computed per participant using all participant' ROIs. Therefore, each participant has two reproducibility/reliability metrics: R^2 and ICC. Then, two studies are carried out: a *metric analysis* and an *analysis of age effect on reproducibility/test-retest reliability metrics.*

In the metric analysis a statistical test was performed per reproducibility/reliability metric to compare the metric means of different groups. The statistical test performed depends upon the objective as follows:

- Objective (a) – to compare the test-retest reliability metric means of different frameworks a paired t-test was conducted;
- Objective (b) - to compare the reproducibility metric mean of paediatric and adults' groups an analysis of covariance (ANCOVA) was performed, in which the dependent variable was the metric, whereas the factors were the age, SNR, the age-group and the acquisition setting;
- Objective (c) – to test the acquisition setting effect on the reproducibility metric an ANCOVA was performed in which the independent variables were the acquisition setting, age and SNR, whereas the dependent variable was the reproducibility metric in consideration.

The assessment of the age impact on the reproducibility/reliability metrics is designated in this work by *Analysis of age effect on reproducibility/test-retest reliability metrics.* The values from the metrics analysis were grouped by metric and group.


Computation of R^2 per participant (P) and Framework (G)
$\mathrm{CT}(\mathrm{P}_1, G_i,\ ROI_M)$
$\mathrm{CT}(\mathrm{P}_1, G_i,\ ROI_1)$
$\mathrm{Y\ CT}(\mathrm{P}_1, G_i)$
$\mathrm{X\ CT}(\mathrm{P}_1, G_i)$
Metrics analysis
Statistical test
G_1
...
$R^2(P_1, G_1)$
$R^2(P_{N_1}, G_1)$
...
G_i
...
$R^2(P_1, G_i)$
$R^2(P_{N_i}, G_i)$
Analysis of age effect on metrics
$R^2(P_1, G_i)$
$R^2(P_{N_i}, G_i)$
1
$R^2_{G_i}$
$R^2_{G_i} = \beta^{R^2}_{age,G_i} \times age + \beta^{R^2}_{SNR,G_i} \times SNR + [\beta^{R^2}_{site,G_i} \times site] + \beta^{R^2}_{0,G_i}$
Multiple comparison independent t-test
$R^2_{G_1} = \beta^{R^2}_{age,G_1} \times age + \beta^{R^2}_{SNR,G_1} \times SNR + [\beta^{R^2}_{site,G_1} \times site] + \beta_{0,G_1}$
$R^2_{G_i} = \beta^{R^2}_{age,G_i} \times age + \beta^{R^2}_{SNR,G_i} \times SNR + [\beta^{R^2}_{site,G_i} \times site] + \beta^{R^2}_{0,G_i}$

objective (a) | $i \in \{FS, CAT12\}$
$\mathrm{X} = \mathrm{Run}_1, \mathrm{Y} = \mathrm{Run}_2$

objective (b) | $i \in \{peadiatric, adult\}$
$\mathrm{X} = \mathrm{FS}, \mathrm{Y} = \mathrm{CAT12}$

objective (c) | $i \in \{HH, Guys, IOP, OASIS-3\}$
$\mathrm{X} = \mathrm{FS}, \mathrm{Y} = \mathrm{CAT12}$

Figure 4.2: Generic representation of the workflow of the Participant Analysis for the objectives proposed, considering the R^2 as the reliability/reproducibility metric. The first step of this analysis is the computation of the R^2, per participant (P), using the cortical thickness (CT) estimations for that participant (blue block). In the objective (a) the R^2 is computed comparing the CT estimations of Run_1 and Run_2, whereas for the objective (b) and (c) the R^2 is computed using the CT estimations of FreeSurfer (FS) and CAT12. Then, a metrics analysis is performed: the R^2 means of different groups (G_i) are compared using the appropriate statistical test (green block). Finally, the analysis of age effect on metrics is done by performing a fit per group (orange block). The dependent variable is the R^2 and the independent variables are the age, SNR and site. It should be noted the site is only used in the objective (b) since it is the only one that has multiple sites within a group. The slopes of the age for the different groups, $\beta^{R^2}_{age,G_i}$, are compared using independent t-tests corrected for multiple comparisons.

Then another linear fit was performed in which the dependent variable was the metric whereas the independent variables were the age and SNR. An additional term, the acquisition setting, was used for the objective (b) which is the only objective that has multiple acquisition settings within a group (see Figure 4.2). It was considered that age influenced the metric if the p-value of the second fit was lower than the significance value ($p\text{-}value<0.05$).

The relation of age and a metric is given by the age' slope (β_{age}) of the second fit. In this study, the β_{age} of the R^2 and ICC of the second fit is given by $\beta_{age}^{R^2}$, and β_{age}^{ICC}, respectively. To further verify whether the relationship between age and reproducibility/reliability metrics was the same in every group, a *independent samples t-test* corrected for multiple comparisons was performed, per metric, comparing the β_{age} of the different groups (see Figure 4.2).

4.3.4.3 ROI analysis

This analysis aimed to understand the reproducibility/reliability across ROIs. To achieve this two studies were conducted: a *metric analysis* and paired t-test for each ROI.

In the metric analysis a linear regression was fitted for each ROI in which the dependent and independent variables were the CAT12 and FreeSurfer estimations of cortical thickness for all individuals (see Figure 4.3). To compare whether the reproducibility metrics across ROIs were identical in every group, the appropriate statistical test was performed:

- Objective (a) a paired-test was performed, per metric, to compare the ROI reliability metrics of the two frameworks.
- Objectives (b) and (c) a pairwise *independent samples t-test* was performed, and corrected with the Bonferroni method [216].

Moreover, for the all objectives a paired t-test comparing the cortical thickness estimations of CAT12 and FreeSurfer was performed per ROI and group/framework.

4.3.4.4 Modelling age using cortical thickness

The reproducibility problem arises when different studies report different findings. The objective (c) aims to verify whether the use of different frameworks may have an impact in the same analysis. For this analysis, we considered the brain age problem [21], in which a person's age is decoded from cortical thickness estimations, in our case. A multiple linear regression was fitted for each framework (CAT12 and FreeSurfer), in which the independent variables were the cortical thickness values for each ROI, and the dependent variable was the age. To test whether the reproducibility metric, in this case the mean reproducibility R^2, had an impact on the reported results we compared the ROIs which were considered significant in each framework

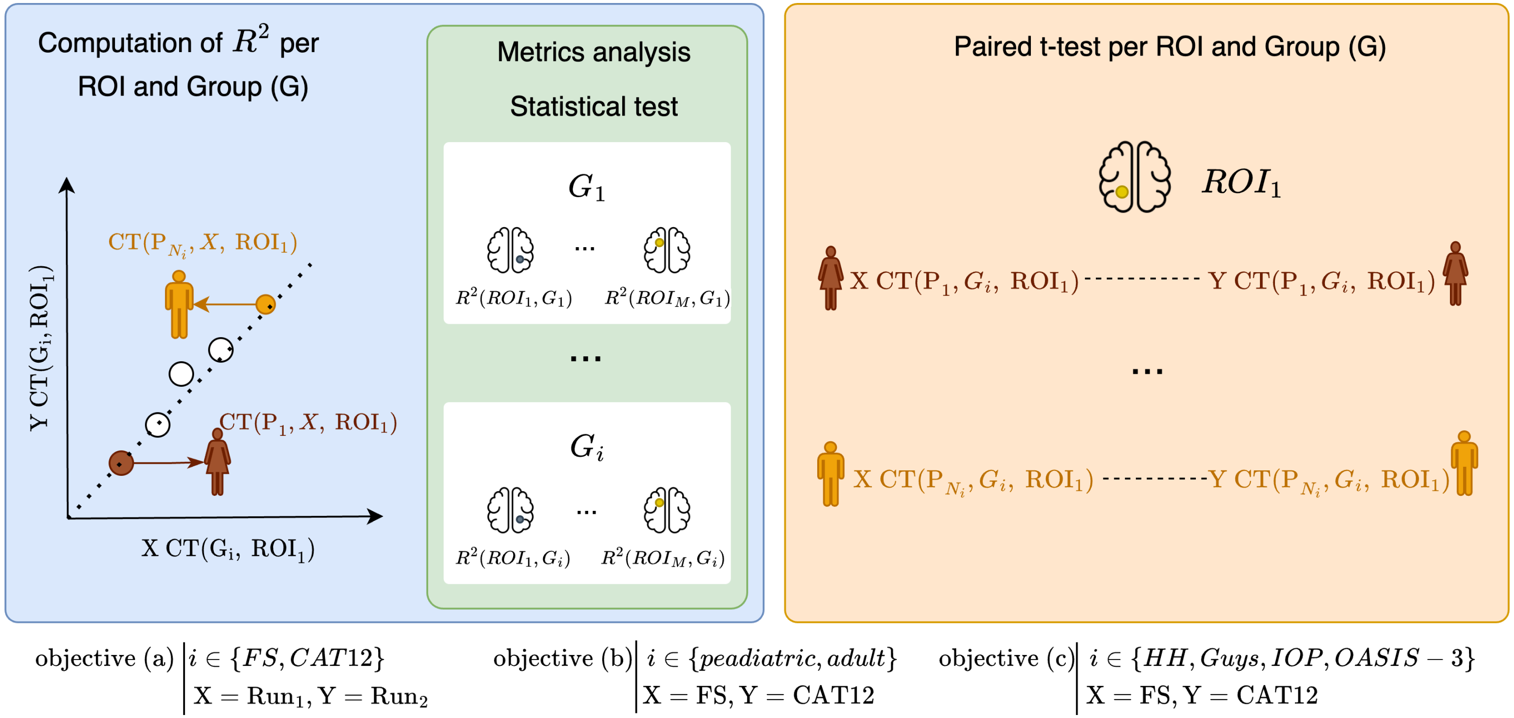

Computation of R^2 per ROI and Group (G)
CT(P_{N_i}, X, ROI_1)
CT(P_1, X, ROI_1)
Y CT(G_i, ROI_1)
X CT(G_i, ROI_1)
Metrics analysis
Statistical test
G_1
$R^2(ROI_1, G_1)$
$R^2(ROI_M, G_1)$
G_i
$R^2(ROI_1, G_i)$
$R^2(ROI_M, G_i)$
Paired t-test per ROI and Group (G)
ROI_1
X CT(P_1, G_i, ROI_1)
Y CT(P_1, G_i, ROI_1)
X CT(P_{N_i}, G_i, ROI_1)
Y CT(P_{N_i}, G_i, ROI_1)
objective (a) $i \in \{FS, CAT12\}$ $X = Run_1, Y = Run_2$
objective (b) $i \in \{peadiatric, adult\}$ X = FS, Y = CAT12
objective (c) $i \in \{HH, Guys, IOP, OASIS-3\}$ X = FS, Y = CAT12

Figure 4.3: Generic representation of the workflow of the ROI Analysis for the three objectives. In this figure the is used as the example for the reproducibility/reliability metric. A fit is performed using all the subject's cortical thickness estimations of a given ROI and from this fit the R^2 is extracted (blue block). The R^2 values of the different groups (G_i) are then compared using an appropriate statistical test (green block). Besides the fitting, a paired t-test (orange block) is performed, per ROI and group, comparing the cortical thickness estimations for different participants of the two runs, in the objective (a), and the two frameworks in the objectives (b) and (c).

model. The ROIs of each model were divided into overlapping and non-overlapping, the former are the ROIs which were significant in both framework models whereas the non-overlapping were the ROIs significant in only one framework model. An independent t-test was performed to verify whether the mean reproducibility R^2 was different in both groups (overlapping and non-overlapping).

4.4 Results

The results section is divided into three sub-sections, one for each objective:

- Objective (a) - Test-retest reliability
- Objective (b) - Reproducibility in paediatric versus early adults groups
- Objective (c) - Reproducibility in different acquisition settings

Each subsection begins with an overview analysis of the data, then the individual analysis is presented, which is divided into the *metrics analysis* and *analysis of age and effect on the reproducibility/reliability metrics.* Finally, the results of the ROI analysis are described.

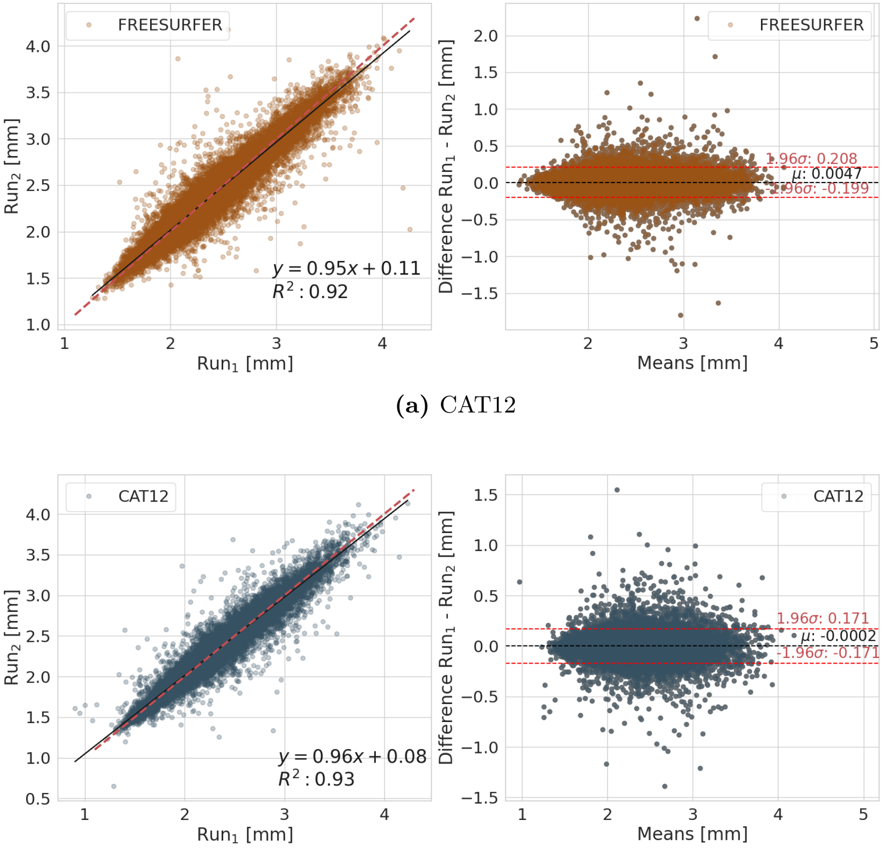
4.4.1 Test-retest reliability

In this section, the goal is to assess the test-retest reliability of each framework. For this, we used 296 subjects from the OASIS-3 dataset. The protocol followed in the OASIS-3 dataset implied that at least two MRI images per subject were acquired in the same day within a 60-minute period. The two images per subject were preprocessed by both frameworks and the test-retest reliability within framework was assessed.

4.4.1.1 Overall Analysis

The relationship between the cortical thickness extracted by the same framework in different images, independently of the framework is depicted in Figure 4.4. The equation is given by $y = 0.95x + 0.11$ and $y = 0.96x + 0.08$, and the R^2 of the fitting is 0.92 and 0.93 for FreeSurfer and CAT12, respectively.

The results of the Bland-Altman plot show that the cortical thickness mean computed with each framework is 2.38 and 2.36 for FreeSurfer and CAT12, respectively. Concerning the bias, the value is 0.005 mm and -0.0002 mm for FreeSurfer and CAT12, respectively. The former is significantly different from 0 ($t_{43807} = 9.39$, $Cohen'd : 0.04$) and the latter is not statistically significant. The confidence interval (CI) is [-0.20; 0.20] mm and [-0.17, 0.17] mm, therefore the CI width is 0.40 mm and 0.34 mm for FreeSurfer and CAT12, respectively.



(a) CAT12

(b) FreeSurfer

Figure 4.4: Regression and Bland-Altman plot for CAT12 and FreeSurfer: On the left the distribution of cortical thickness extracted in Run$_1$ and Run$_2$, respectively, each point corresponds to a subject's ROI. The dashed red line on the left plot represents the equation $y = x$. On the right is the cortical thickness difference between runs.

Table 4.1: Test-retest metrics analysis, showing the mean and the standard deviation of and ICC, for the participant analysis, per framework.

Framework	$\boldsymbol{R^2}$	**ICC**
FreeSurfer	0.93 ± 0.07	0.97 ± 0.04
CAT12	0.94 ± 0.06	0.96 ± 0.04

Table 4.2: Analysis of age effect on test-retest metrics: results of the regression with each of the reproducibility metrics: R^2 ($\beta_{age}^{R^2}$) and ICC (β_{age}^{ICC})

	Framework	$\beta_{age}^{R^2}$	β_{age}^{ICC}
Age	FreeSurfer	3.2×10^{-4} ($p = 0.485$)	2.2×10^{-4} ($p = 0.445$)
	CAT12	6.4×10^{-4} ($p = 0.1$)	3.9×10^{-4} ($p = 0.121$)

4.4.1.2 Participant analysis

Test-retest metrics analysis

The reliability test-retest results for the participant analysis, per framework, are shown in Table 4.1.

The difference between the FreeSurfer and CAT12 test-retest reliability metrics' means of the frameworks is -0.01 for both R^2 and ICC. To verify whether the means are statistically different a paired t-test comparison was performed. The results suggest that the mean of FreeSurfer and mean of CAT12 is different for the R^2 ($t_{295} = 4.14,\ Cohen'd = 0.15$) and ICC ($t_{295} = 3.6,\ Cohen'd = 0.13$).

Analysis of age effect on test-retest metrics

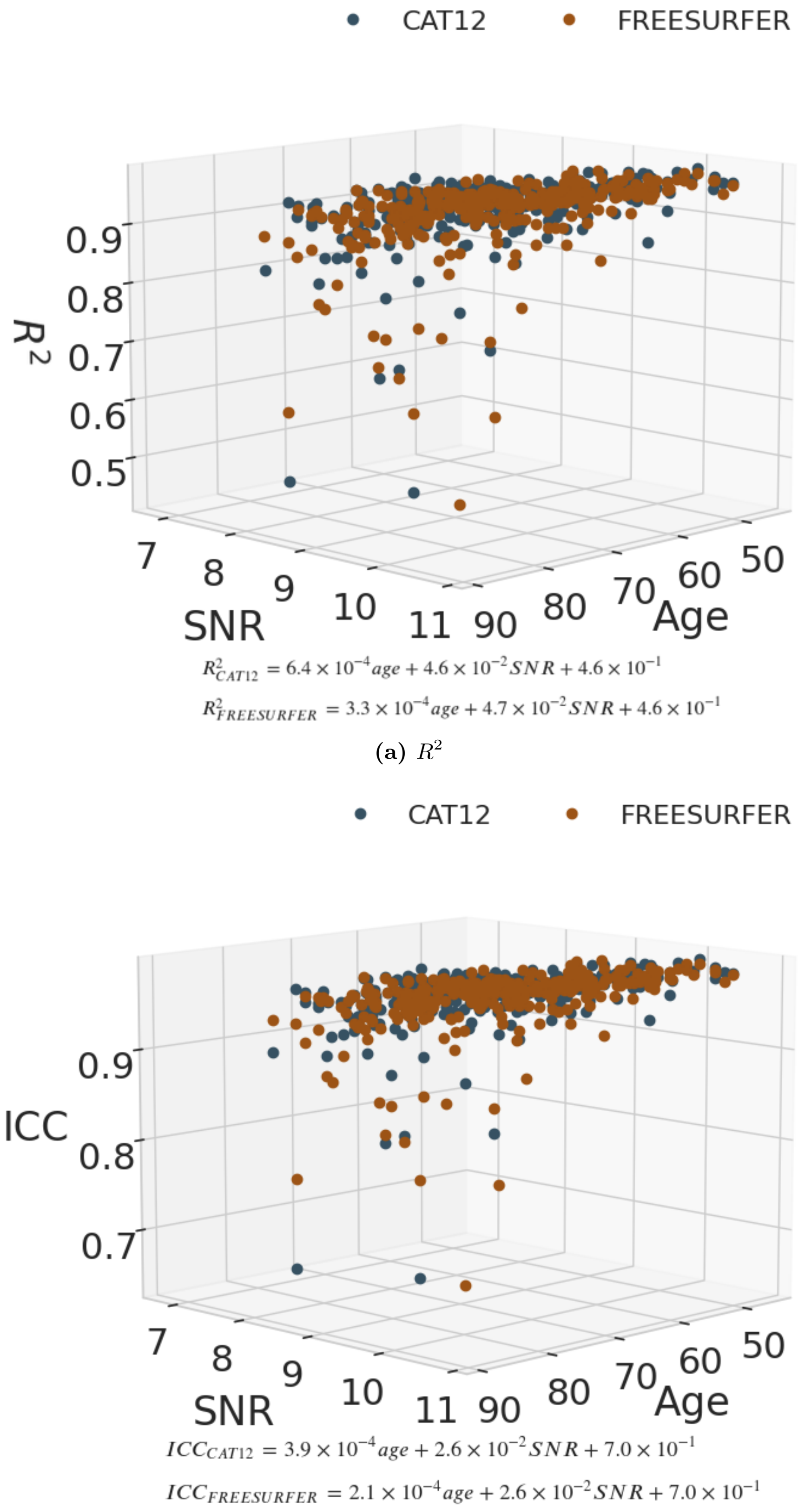
The distribution of the test-retest reliability metrics with age is depicted in Figure 4.5 and the statistical analysis results are shown in Table 4.2. The results show that age had no effect on CAT12 and FreeSurfer on test-retest reliability metrics.

4.4.1.3 ROI Level Analysis

The reliability test-retest results, per framework, are presented in Table 4.3 and in Figure 4.6, the detailed results of the R^2 grouped by lobe is in Table A.4 in the Appendix A whereas the R^2 and ICC by ROI are presented in Tables A.5-A.6 of the Appendix A. The test-retest reliability metrics are lower for the ROI analysis than for the participant analysis. The comparison between the ROI test-retest reliability metric's mean values of FreeSurfer and CAT12 (paired t-test) showed a significant difference between the frameworks for R^2 (mean difference -0.08, $t_{147} =$

Table 4.3: Mean and standard deviation of the R^2 and ICC values for the test-retest reliability metrics ROI analysis.

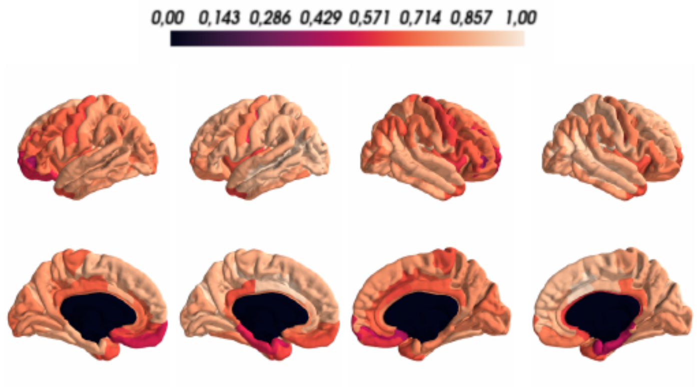
Framework	$\boldsymbol{R^2}$	**ICC**
Freesurfer	0.70 ± 0.10	0.84 ± 0.06
CAT12	0.78 ± 0.09	0.88 ± 0.05



(a) R^2

(b) *ICC*

Figure 4.5: Distribution of the R^2 and ICC with age for each of the frameworks. Each point represents the R^2 or the ICC value, at the participant level, of the cortical thickness test-retest reliability of each framework.



(a) Left FreeSurfer (b) Left CAT12 (c) Right Freesurfer (d) Right CAT12

Figure 4.6: R^2 values distribution between the cortical thickness extracted by FreeSurfer or CAT12 values per ROI, per hemisphere.

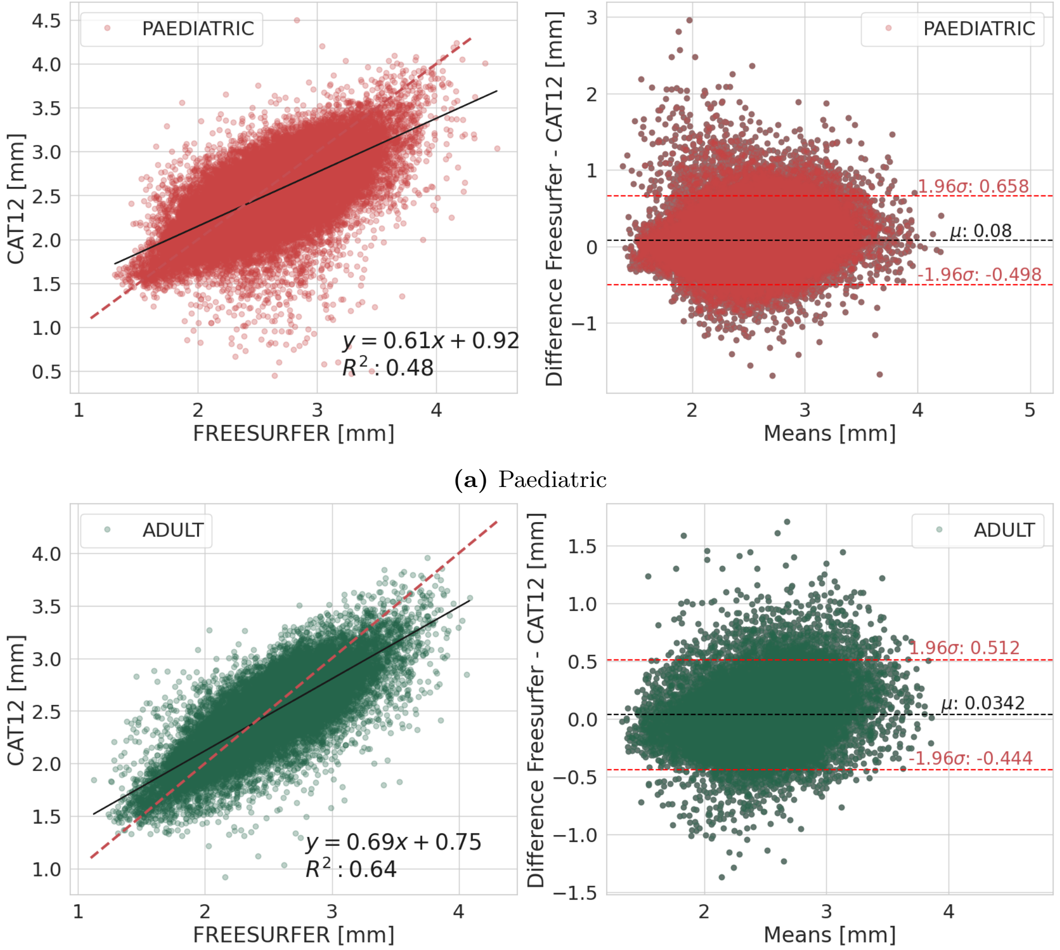
9.83, $Cohen'd = 0.83$) and ICC (mean difference -0.06, $t_{147} = 9.28$, $Cohen^{'}d =$ 0.81). The individual ROI analysis revealed that the lobe with higher R^2 for CAT12 is the Parietal lobe (0.83) whereas for FreeSurfer is the temporal and occipital lobes (0.77). The lobe with lower R^2 is the frontal lobe for both CAT12 (0.76) and FreeSurfer (0.63). The ROIs with higher difference between R^2 in test-retest reliability are the *parahippocampal gyrus, inferior occipital gyrus and sulcus, posterior-ventral part of the cingulate gyrus,* and the *planum polare of the superior temporal gyrus.* Concerning the *paired t-tests* performed for each ROI, the results suggest that the cortical thickness estimations are similar between runs within each ROI, for all ROIs except the *Left Short insular gyri* in FreeSurfer, detailed results are in Table A.7 in the Appendix A.

4.4.2 Reproducibility in paediatric versus early adults' groups

This section investigates whether different age stages (paediatric versus adults) have different reproducibility behaviour. The dataset used was the ABIDE I dataset, and the age range of the participants is from 6.5 to 56.2 years, and the analysis was carried out for two age groups: paediatric and adults.

4.4.2.1 Overall Analysis

All the cortical thickness values extracted by both frameworks are depicted in Figure 4.7. The equation that translates the relationship between FreeSurfer and CAT12 values is $CAT12 = 0.61 \times FREESURFER + 0.92$ and $CAT12 = 0.69 \times FREESURFER + 0.75$, for paediatric and adult samples, respectively, and the R^2 is 0.48 and 0.64. The mean difference between the estimations is 0.065 ± 0.28 mm. The Bland-Altman plot shows a FreeSurfer (CAT12) mean of cortical thickness of 2.60 (2.52) mm and 2.49 (2.45) mm for paediatric and adult samples, respectively.



(a) Paediatric

(b) Adult

Figure 4.7: Regression and Bland-Altman plot for the paediatric and adults' analysis. On the left the distribution of cortical thickness extracted with FreeSurfer and CAT12, respectively, each point corresponds to a subject's ROI. The dashed yellow line on the left plot represents the equation $y = x$. On the right the cortical thickness means and difference (FreeSurfer – CAT12) of the frameworks.

Table 4.4: Reproducibility metrics analysis: Mean +/- standard deviation of R^2 and ICC, for the individual analysis, grouped by age (paediatric versus adults) for the neurodevelopment analysis

Framework	$\boldsymbol{R^2}$	**ICC**
paediatric	0.51 ± 0.14	0.69 ± 0.11
adult	0.64 ± 0.09	0.79 ± 0.06

Concerning the cortical thickness difference, the CI is between [-0.50, 0.66] mm and [-0.44,0.51] mm, thus a CI width of 1.16 mm and 0.95 mm for paediatric and adult samples, respectively. The bias mean is 0.080 mm and 0.034 mm, for paediatric and adult samples, respectively. The t-test suggests that the bias is statistically different from zero in both cases (paediatric , $t_{56831} = 6.47$, $cohen'd : 0.27$ and adult , $t_{26935} = 2.3$, $Cohen'd : 0.14$*)*.

4.4.2.2 Participant Analysis

Reproducibility metrics analysis

The results observed when analysing the two age groups are summarised in Table 4.4. The difference between the paediatric and adult group is -0.13 and -0.10 for R^2 and ICC, respectively. Two ANCOVAs were performed, one per reproducibility metric. The ANCOVA results are shown in Tables A.8 and A.9 in the Appendix A. In summary, in the two metrics all the independent factors (age, SNR, age-group and acquisition setting) are suggested to influence the reproducibility metrics. The SNR was the factor with higher dispersion, followed by age, acquisition setting and age-group.

Analysis of age effect on reproducibility metrics

The evolution of reproducibility metrics with age and SNR per acquisition setting, is depicted in Figure 4.8. It shows an interesting and unexpected pattern: in children and adolescents all reproducibility metrics improve with age. Moreover, the variability decreases with age as well as the worst attained value, for each reproducibility metric.

Table 4.5: Reproducibility metrics and age relation analysis: regression between age and SNR and each of the reproducibility metrics R^2 ($\beta_{age}^{R^2}$) and ICC (β_{age}^{ICC}) per age group. The significant values are shown in bold

Framework	$\beta_{age}^{R^2}$	β_{age}^{ICC}
paediatric	$\mathbf{1.2 \times 10^{-2}}$ $(\mathbf{p = 8.6 \times 10^{-7}})$	$\mathbf{8.9 \times 10^{-3}}$ $(\mathbf{p = 1.3 \times 10^{-5}})$
adult	$\mathbf{2.9 \times 10^{-3}}$ $(\mathbf{p = 1.2 \times 10^{-3}})$	$\mathbf{2.9 \times 10^{-3}}$ $(\mathbf{p = 1.2 \times 10^{-3}})$

Table 4.5 shows the relation between the age for each of the reproducibility metrics. In the paediatric group the relation was found to be significant, for both metrics (R^2 and ICC) and age. Regarding the β_{age} comparisons, the *independent t-test* suggested that the difference between paediatric and adult group ($\beta_{age,\ R^2}$: 9.1×10^{-3} and

Table 4.6: Mean and standard deviation of R^2 and ICC, for the ROI analysis, grouped by age (greater or less than 18 years) for the neurodevelopment analysis.

Age group	R^2	ICC
Paedriatic	0.30 ± 0.14	0.52 ± 0.14
Adult	0.38 ± 0.16	0.59 ± 0.15

$\beta_{age,\ ICC} : 6.0 \times 10^{-3}$) of relation between age and reproducibility metrics is different between the two age groups (paediatric vs adult) for the two reproducibility metrics: $\beta_{R^2}(t_{564} = 25.96,\ \textit{Cohen'd: 2.34})$ and $\beta_{ICC}(t_{564} = 23.72,\ \textit{Cohen'd: 0.69})$.

4.4.2.3 ROI Level Analysis

The reproducibility metrics for the ROI analysis are depicted in Figure 4.9. Table 4.6 presents a summary of the metrics, per age group (the detailed lobes results is in Table A.10 and the results per ROI for R^2 and ICC is in Tables A.11 and A.12, respectively, in the Appendix A).

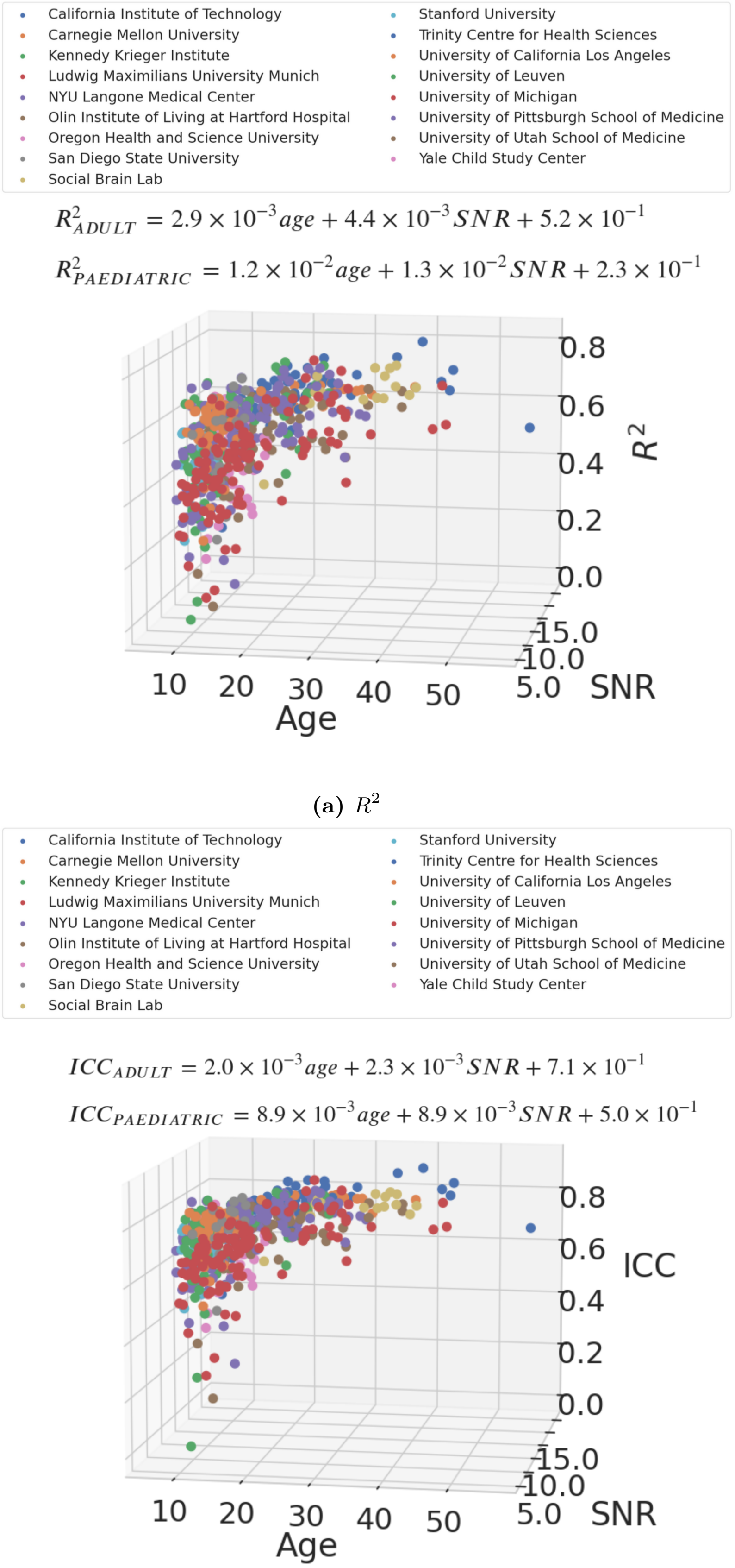
To verify whether the difference between the paediatric and adult age groups was significant for the R^2 (-0.08) and ICC (-0.07) an *independent samples t-test* was performed. The results showed differences between age groups, for the two metrics: R^2 ($t_{294} = -4.79,\ Cohen'd:\ 0.56$) and ICC ($t_{294} = -3.98,\ Cohen'd:\ 0.46$).

The lobe with higher difference between the age groups is the temporal and occipital lobe whereas with the lowest difference is the frontal lobe. The parietal lobe is the one with higher mean R^2, 0.49 and 0.41 for adults and paediatric, respectively. Whereas insula lobe is the one with lower mean R^2, 0.24 and 0.16 for adults and paediatric, respectively. The ROI with higher reproducibility in children and adolescents is the *postcentral sulcus*, which is also the one of the ROIs with lowest difference between the age groups.

The ROI *paired t-tests* showed that 137 ROIs had at least one age-group with differences in the cortical thickness estimations of both frameworks. In the children and adult group, 130 and 109 ROIs had differences between the cortical thickness estimations extracted by Freesurfer and CAT12, respectively. The p-values were corrected for the 296 comparisons, detailed results in Table A.13 in the Appendix A.

4.4.3 Reproducibility in different acquisition settings

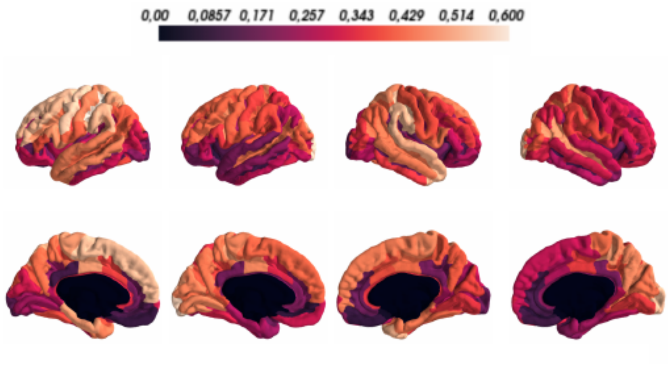
This section aims to understand whether the acquisition setting and age has impact on reproducibility metrics on mature brains. In this analysis two open-data sharing initiatives were used: IXI and OASIS-3. The analysis was conducted per acquisition setting. Therefore, four groups were analysed: three groups for the IXI dataset (one per site/acquisition setting) and one group for the OASIS-3. Furthermore, the reproducibility between CAT12 and FreeSurfer metrics was also computed for a subset of OASIS-3. The subset was the one used in test-retest reliability analysis,



(a) R^2

(b) ICC

Figure 4.8: R^2 and ICC by age and SNR for each site of the ABIDE I.



(a) Left Paediatric
(b) Left Adult
(c) Right Paediatric
(d) Right Adult

Figure 4.9: R^2 values distribution between the estimations of cortical thickness extracted by Freesurfer and CAT12 values per ROI, hemisphere, and per age group.

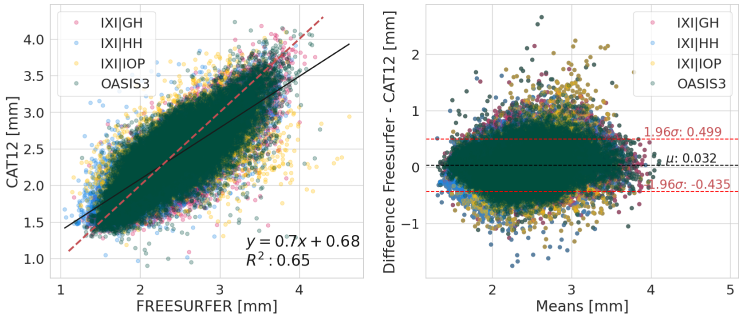


Figure 4.10: Regression and Bland-Altman for the adult dataset with different acquisition settings. On the left the distribution of cortical thickness extracted with FreeSurfer and CAT12. The dashed red line on the left plot represents the equation $y = x$. On the right is the plot with the mean and difference (FreeSurfer – CAT12) between the cortical thickness estimates of the frameworks. In this plot each point corresponds to a participant's ROI.

objective (a), which includes subjects who have two T_1-weighted images. In this case the cortical thickness was averaged over the two runs and the reproducibility metrics were computed.

4.4.3.1 Overall analysis

In this case, given that there are four acquisitions settings we decided to report the Bland-Altman results independently of the acquisition setting, the detailed results per acquisition setting are in Figure A.1 in the Appendix A. Considering all the adult individual ROIs, the overall reproducibility metrics were extracted, which are shown in Figure 4.10 with a Bland-Altman plot. The linear regression has the following equation: $CAT12 = 0.70 \times FREESURFER + 0.68$, with a R^2 of 0.65. Regarding the Bland-Altman analysis the mean cortical thickness value is 2.39 mm and 2.36 mm for FreeSurfer and CAT12, respectively. Regarding the cortical thickness difference between the frameworks the mean difference is 0.032 which is statistically different from zero ($t_{155251} = 5.30,\ Cohen'd : 0.13$). Regarding the CI the values range from

[-0.44;0.50] mm which is a width of 0.93 mm.

4.4.3.2 Participant Analysis

Reproducibility metrics analysis

The overall R^2 for the adult datasets is 0.65, for the participant level analysis, and Table 4.7 summarises the reproducibility metrics per acquisition setting. The acquisition setting with higher reproducibility metrics is the GH, followed by OASIS-3, HH and IOP.

Table 4.7: Reproducibility metrics analysis: mean and standard deviation of R^2 and ICC for the participant analysis, grouped by acquisition setting for the IXI and OASIS-3 datasets.

Acquisition setting	R^2	ICC
GH	0.69 ± 0.06	0.82 ± 0.04
HH	0.65 ± 0.09	0.79 ± 0.07
IOP	0.38 ± 0.10	0.60 ± 0.08
OASIS3	0.65 ± 0.08	0.80 ± 0.06

The ANCOVA suggested an acquisition setting and SNR effect in both metrics, yet the acquisition setting has the highest factor of dispersion independent of the metric. The age had effect on the R^2 metric. The post-hoc comparisons revealed differences between all acquisition settings, detailed results are in Tables A.14-A.17 in the Appendix A. In Figure 4.11 the acquisition setting effect is quite visible, i.e., the IOP reproducibility metrics are clustered apart from the GH, HH and OASIS-3. The reproducibility results for the OASIS-3 test-retest subset were 0.67 ± 0.07 and 0.81 ± 0.06 for the R^2 and ICC, respectively.

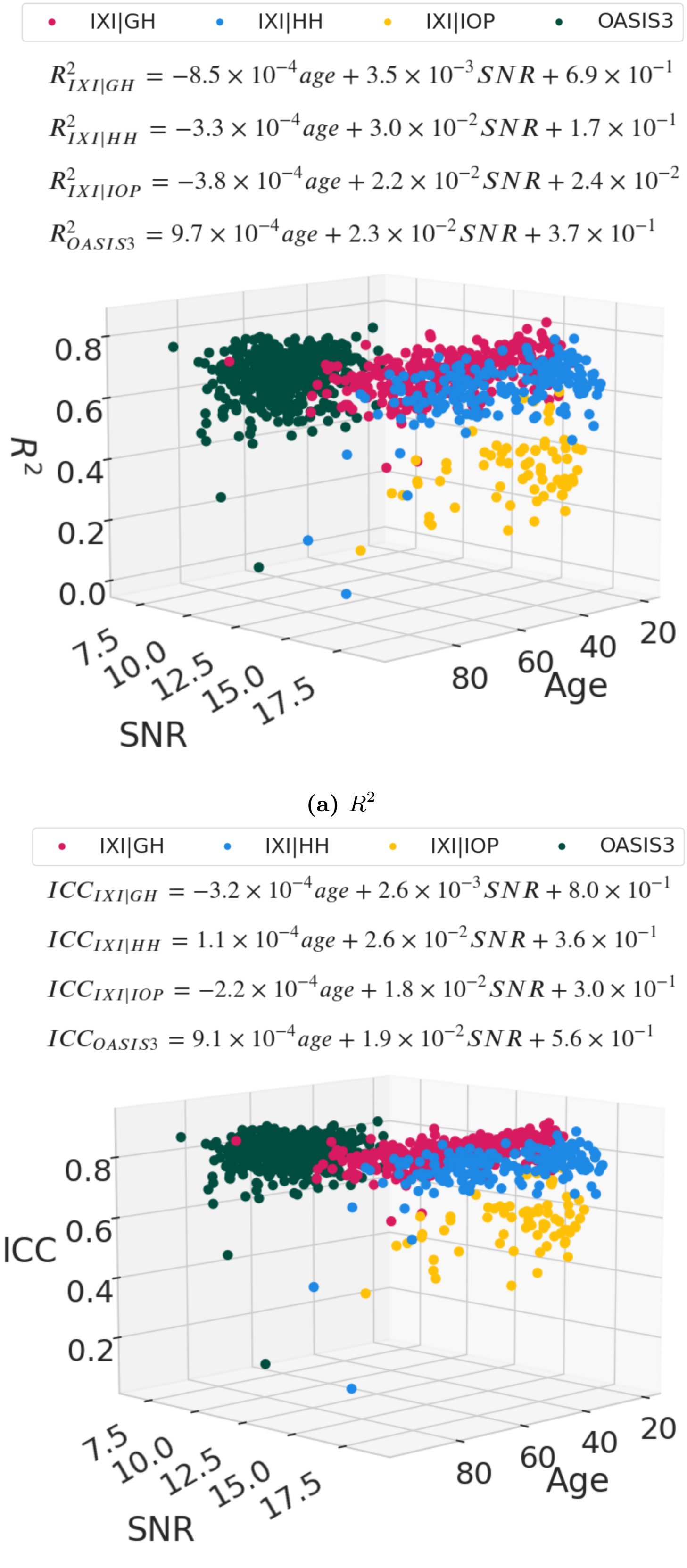
Analysis of age effect on reproducibility metrics

Figure 4.11 and Table 4.8 summarise the assessment of the impact of age and acquisition setting on reproducibility metrics. The results are consistent independently of the metric. The statistical analysis revealed an age effect on GH and OASIS-3 sites.

The comparative analysis of the β_{age} ($\beta_{age}^{R^2}$,β_{age}^{ICC}) of each acquisition setting points out that the two β_{age} are indeed different between all acquisition settings, except between HH and IOP, detailed results are in Tables A.18-A.19 in the Appendix A. Therefore, the relation between age and the reproducibility metrics is dependent on the acquisition setting.

4.4.3.3 ROI Level Analysis

A summary of the reproducibility metrics results for the ROI analysis, per acquisition setting, is shown in Table 4.9. The results show that the HH, GH and OASIS-3 have better reproducibility metrics than the IOP.



$$R^2_{IXI|GH} = -8.5 \times 10^{-4} age + 3.5 \times 10^{-3} SNR + 6.9 \times 10^{-1}$$

$$R^2_{IXI|HH} = -3.3 \times 10^{-4} age + 3.0 \times 10^{-2} SNR + 1.7 \times 10^{-1}$$

$$R^2_{IXI|IOP} = -3.8 \times 10^{-4} age + 2.2 \times 10^{-2} SNR + 2.4 \times 10^{-2}$$

$$R^2_{OASIS3} = 9.7 \times 10^{-4} age + 2.3 \times 10^{-2} SNR + 3.7 \times 10^{-1}$$

(a) R^2

$$ICC_{IXI|GH} = -3.2 \times 10^{-4} age + 2.6 \times 10^{-3} SNR + 8.0 \times 10^{-1}$$

$$ICC_{IXI|HH} = 1.1 \times 10^{-4} age + 2.6 \times 10^{-2} SNR + 3.6 \times 10^{-1}$$

$$ICC_{IXI|IOP} = -2.2 \times 10^{-4} age + 1.8 \times 10^{-2} SNR + 3.0 \times 10^{-1}$$

$$ICC_{OASIS3} = 9.1 \times 10^{-4} age + 1.9 \times 10^{-2} SNR + 5.6 \times 10^{-1}$$

(b) ICC

Figure 4.11: Reproducibility metrics (R^2 and ICC) at the participant level and its relationship with age and SNR for the different sites of the IXI and the OASIS-3 datasets.

Table 4.8: Reproducibility metrics and age relation analysis of the regression of age with each of the reproducibility metrics R^2 ($\beta_{age}^{R^2}$) and ICC (β_{age}^{ICC}), per acquisition setting. The significant values are shown in bold.

Acquisition setting	$\beta_{age}^{R^2}$	β_{age}^{ICC}
IXI|GH	$\mathbf{-8.5 \times 10^{-4}}$ **(p = 1.3 × 10⁻⁴)**	$\mathbf{-3.2 \times 10^{-4}}$ $\mathbf{(p = 2.3 \times 10^{-2})}$
IXI|HH	-3.3×10^{-4} ($p = 4.2 \times 10^{-1}$)	1.1×10^{-4} ($p = 7.5 \times 10^{-1}$)
IXI|IOP	-3.8×10^{-4} ($p = 6.1 \times 10^{-1}$)	-2.2×10^{-4} ($p = 7.1 \times 10^{-1}$)
OASIS3	$\mathbf{9.7 \times 10^{-4}}$ $\mathbf{(p = 1.8 \times 10^{-2})}$	$\mathbf{9.1 \times 10^{-4}}$ $\mathbf{(p = 2.9 \times 10^{-3})}$

Regarding the reproducibility results for ROI analysis and using the OASIS-3 test-retest dataset the R^2 and ICC were 0.40 ± 0.21 and 0.60 ± 0.21, respectively.
The R^2 values per ROI for each open sharing initiative are depicted in Figure 4.12 and the ROI and lobe R^2 values are in Tables A.20-A.21 and the ICC values per ROI are in Tables A.22-A.23 in the Appendix A. The R^2 differs throughout the brain, it is clear that there are regions with higher reproducibility metrics than others. Yet, the ROI reproducibility pattern is similar across acquisition settings. In general, while the parietal lobe has the higher mean R^2, the limbic lobe has the lower mean R^2. The subcallosal area and anterior transverse collateral sulcus (both ROIs are near the corpus callosum) are the ROIs with the lowest R^2.

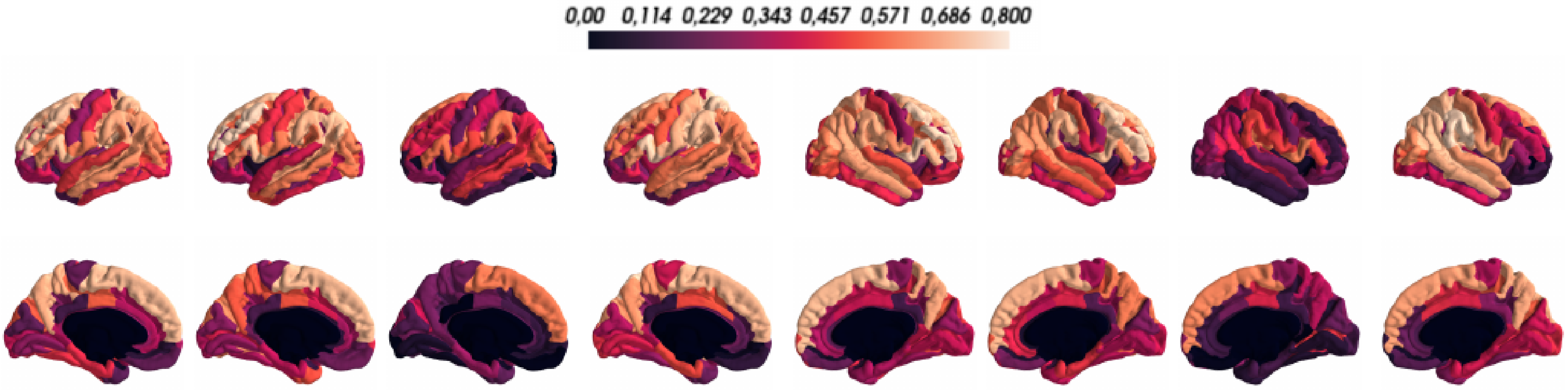
Table 4.9: Mean and standard deviation of R^2 and ICC, for the ROI analysis, per acquisition setting for the IXI and OASIS-3 datasets.

Acquisition Setting	$\boldsymbol{R^2}$	**ICC**
GH	0.40 ± 0.19	0.60 ± 0.18
HH	0.39 ± 0.19	0.59 ± 0.18
IOP	0.24 ± 0.16	0.44 ± 0.18
OASIS3	0.39 ± 0.21	0.59 ± 0.20

Most ROIs have a similar R^2 value for the right and left hemisphere, yet there are discrepancies. The *transverse temporal sulcus*, *middle occipital sulcus, lunatus sulcus* and *orbital gyri* yield a difference in R^2 between hemispheres higher than 0.5, independently of the acquisition setting. Nevertheless, in some ROIs, such as the *middle frontal gyrus*, the difference between hemispheres is only observed in the OASIS-3 dataset but not in others.
Four paired t-test were performed per ROI, one per acquisition setting, the p-values were corrected for the 576 comparisons. The results revealed differences between the frameworks on 145 ROIs, in at least one acquisition setting. Briefly, 132, 116, 94 and 130 ROIs were significantly different in GH, HH, IOP and OASIS-3, respectively. Detailed results are presented in Table A.24 in the Appendix A.

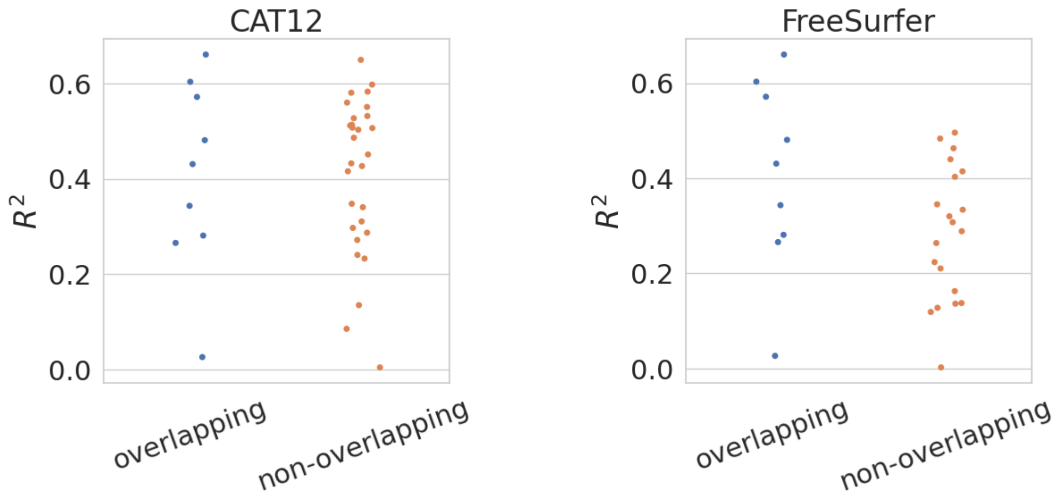
Modelling age using cortical thickness values

Two models were trained, one per framework, with adult data from the IXI and OASIS-3 datasets. The adjusted R^2 for FreeSurfer and CAT12 models is 0.71 and 0.73, respectively. The number of ROIs considered significant for age prediction is



(a) Left GH (b) Left HH (c) Left IOP (d) Left OASIS3 (e) Right GH (f) Right HH (g) Right IOP (h) Right OASIS3

Figure 4.12: R^2 values distribution between the cortical thickness extracted by FreeSurfer and CAT12 values per ROI, hemisphere and acquisition setting.



(a) CAT12

(b) FreeSurfer

Figure 4.13: Distribution of the mean reproducibility, R^2, of the overlapping and non-overlapping ROIs for CAT12 and FreeSurfer.

29 and 38 for FreeSurfer and CAT12, respectively, the details are in Tables A.25 and A.26 in the Appendix A. An analysis was carried out to verify which ROIs were significant in age prediction in both frameworks. The results point out that there are 9 overlapping ROIs, regions are significant in age prediction using both frameworks for estimation of cortical thickness. The overlapping ROIs and their mean reproducibility, R^2, across sites (which is represented between the parentheses) are: *left superior frontal gyrus (0.66), left triangular part of the inferior frontal gyrus (0.60), right angular gyrus (0.57), right superior occipital sulcus and transverse occipital sulcus (0.48), right superior part of the precentral sulcus (0.43), right occipital pole (0.34), left lingual gyrus (0.27), right anterior transverse temporal gyrus (0.28)* and the *left medial orbital sulcus (0.03).*

In Figure 4.13 is depicted the distribution of the reproducibility metric values, R^2, for each framework for the two groups: overlapping and non-overlapping ROIs. The independent t-test performed, per framework, revealed no difference between the groups for neither framework.

4.5 Discussion

This study reveals that the test-rest reliability of cortical thickness estimations within two computation frameworks (CAT12 and FreeSurfer) attains higher values than the reproducibility between frameworks, as expected. Concerning the test-retest reliability the Bland-Altman analysis suggested that the bias is only significant in the FreeSurfer framework, yet the effect is small. Furthermore, the dispersion for CAT12 is lower than for FreeSurfer. The cortical thickness estimation mean was identical between frameworks (2.36mm and 2.38mm), these mean values are close to previous report mean cortical thickness [188]. The confidence interval was 0.34

mm and 0.40 mm, respectively, which corresponds to an error of 14% and 17% for a mean cortical thickness of 2.37mm.

At the individual level, the CAT12 attains statistically higher test-retest metrics than FreeSurfer, despite a very small 0.01 difference, which may be negligible from the biological point of view. An age effect was not detected in neither CAT12 nor FreeSurfer test-retest reliability metrics. Nevertheless, a SNR effect was detected in the two test-retest reliability metrics, therefore the performance within the framework is sensitive to the image quality. The positive slope between the SNR and the metrics reveals that images with higher quality have higher reliability. An analysis of the relation between age and SNR showed that older individuals have lower SNR values. This is in agreement with the literature reporting that motion artifacts are higher in older individuals and, which in turn affect cortical thickness estimation [55]. At the ROI level the test-retest difference between frameworks is higher (0.08 for R^2) and was found to be statistically significantly with a strong effect on ICC and R^2. Once again, the CAT12 has higher test-retest metrics than FreeSurfer. Regarding test-retest reliability per lobes the parietal lobe and the temporal and occipital lobes have higher test-retest reliability. In contrast the frontal lobe has lower reproducibility in both frameworks.

This study reveals a significant discrepancy between the cortical thickness reproducibility metrics extracted by FreeSurfer and CAT12 in adults and paediatric populations. The reproducibility metrics are significantly lower in children and adolescents' than in adults, which may impact on the interpretation of neurodevelopmental studies. Furthermore, the reproducibility metrics increase with age until the age of 18 years old and stabilise in adulthood (Figure 4.8). The reason behind this result might be nonlinear brain maturation, which changes brain morphology in children/adolescents [217]. These changes are a source of variability that might increase the difficulty to accurately estimate cortical thickness. Besides maturation, another cause for the low reproducibility in paediatric group might be motion, which tend to be higher in younger individuals [218] and causes artifacts that affect the accurate extraction of cortical thickness [219]. It should be noted that age had an effect in children's cortical thickness reproducibility even when our models introduce SNR as a covariate. Thus, we believe some of the reproducibility issues might be inherent to biological variability introduced by the brain maturation process. These results are pursuant with published findings where it is shown that different preprocessing pipelines, in early childhood, yield different results [220]. The low reproducibility between pipelines might be an explanation for cortical thickness inconsistencies found in the neurodevelopment phase in health and disease [196–200].

Concerning the ROI analysis, the overall pattern of reproducibility metrics is the same in both age groups: the parietal lobe is the region with higher reproducibility whereas the insula is the one with lower reproducibility. There is previous evidence that the insula has high annual changes in cortical thickness during childhood and adolescence whereas in parietal cortex the annual changes are less exacerbated [221]. Therefore, the reproducibility might be related to the annual change rate of cortical thickness, the higher is the rate of annual changes the lower is the reproducibility.

In adults, the reproducibility between frameworks slowly decreases with age, i.e., older adult individuals exhibit lower reproducibility metrics than younger adult individuals, in some acquisition settings. An age effect was detected, for both reproducibility metrics, in GH and OASIS-3 acquisition settings, even though we controlled for the image SNR. The reproducibility decrease with age may be related to the fact that older individuals move more in the scanner than younger individuals, which adds noise to the images and, consequently, the methods have more difficulties in extracting accurate metrics [218, 222]. Additionally, the changes in brain morphometry that occur naturally with the ageing process might also underly the decrease in reproducibility of these measures with increasing age. The lack of an age effect on the IOP might be related to its sample size (the IOP has 67 participants, whereas the GH, HH and OASIS-3 have 312, 179 and 491 participants, respectively).

Besides the age effect, an acquisition setting effect was detected on the reproducibility metrics, which translates that sites with a similar SNR distribution had different reproducibility distributions. Furthermore, the relationship between age and reproducibility metrics is dependent on the acquisition setting. Therefore, the low reproducibility may be related to the acquisition setting that affected all the images in the same way, yet each acquisition setting affects the images differently, and the frameworks were unable to overcome this limitation. Moreover, the test-retest results also suggest that the reproducibility metrics are similar even when the cortical thickness estimation is averaged across two images of the same subject. This was as expected given the high values of the test-retest reliability of each frameworks. The current results highlight the challenge of multicentric studies. The conclusions of these large (typically consortia) studies might be vulnerable to acquisition settings imprints on the data and must be interpreted with caution. Nonetheless, multicentre studies are crucial to hypothesis generalisation. Therefore, data harmonisation techniques should ideally be available in preprocessing frameworks to reduce the acquisition setting impact on the images.

The reproducibility at the individual level is higher than at the ROI level. This is expected since at the individual level there is only variance related to the framework whereas at the ROI level analysis there is variance related to both the participant and framework. This finding highlights the fact that the frameworks are not capable to overcome the individual variability and cross-sectional studies are more susceptible to inconsistencies leading to inaccurate conclusions. Nevertheless, the reproducibility at the ROI level is consistent across the four acquisition settings. This finding corroborates the results published in [208] in which it is shown that the ROI reproducibility is similar across datasets.

The analysis of the outcome of the age prediction problem showed that CAT12 and FreeSurfer yield different results. Nonetheless, despite the differences, the results also show that some ROIs with high reproducibility between FreeSurfer and CAT12 overlap across frameworks. Yet, a high reproducibility value does not guarantee the same result in both models, as there are ROIs with high values of reproducibility across frameworks that are only significant in predicting age in one of the frameworks.

4.6 Conclusion

This work shows that the reproducibility of cortical thickness estimations with CAT12 and FreeSurfer is weaker at the ROI-level than at the subject-level. These findings show that the inference from cross-sectional studies, in which conclusions at the ROI level are made from distinct sites, groups and frameworks, might explain inconsistencies reported in the literature. Furthermore, conclusions from metanalyses in which the preprocessing protocol was not the same might be compromised. Thus, the preprocessing framework should be considered when different studies are being compared. Our results show that the test-retest reliability is reassuringly high at the individual level for images obtained using the same acquisition protocol. Consequently, the conclusions from longitudinal studies, with the same acquisition and preprocessing protocol are less sensitive to the described processing framework effects.

Chapter 5

Deformation Fields: A new source of information to predict brain age

This chapter is based in the following article:

Dias, Maria de Fátima Machado, Paulo Carvalho, João Valente Duarte, and Miguel Castelo-Branco. "Deformation fields: a new source of information to predict brain age." Journal of Neural Engineering 19, no. 3 (2022): 036025.

5.1 Abstract

Objective: The modelling of healthy ageing critically requires the identification of methods that detect subtle changes in this process. In the last few years multiple machine learning models have been proposed that learn age patterns from magnetic resonance imaging (MRI). Current standard information sources rely on local volumetric information of brain tissues, namely grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF). Information about patterns of brain deformation remains underexplored. In this paper an assessment is performed to understand better the predictive value of the deformation fields.

Approach: A shallow approach was used to compare the predictive value of deformation fields with the brain tissues (GM, WM and CSF). Images were compressed into a lower dimension space using Principal Components Analysis and then, a relevant vector regression (RVR) learned the age patterns from the components. A model was trained per modality (deformation fields, GM, WM and CSF) and the performance between the models was compared. To evaluate whether the deformation fields increased the predictive power of GM, a model fusion approach was explored in which the final estimator was a RVR. Each model was validated using a cross-validation approach and was also evaluated on an external dataset.

Main results: We found that models trained with deformation patterns have higher predictive value than the ones trained with WM or CSF. Furthermore, deformation fields had a significantly better performance on the test set and also yield the lower difference between the validation and test set. Moreover, the predictions based on the combination of deformation patterns with GM volume yields better results than GM volumetric information alone.

Significance: These findings suggest that deformation fields have a higher predictive power than WM and CSF and are robustly invariant across a set of confounding variables. Therefore, deformation fields should be considered in brain age models.

Keywords: Deformation Fields, BrainAge, Ageing; Machine Learning

5.2 Introduction

The brain undergoes multiple changes across distinct brain regions during the lifespan. In early adulthood, GM volume slowly decreases [223–225] and in late adulthood (>45 years) the WM volume also starts to show signs of decline [59, 226]. These brain tissues' fluctuations lead to alterations of the global and local brain morphologies [39, 227, 228] that create different deformation patterns, which can be used to train machine learning models to predict the chronological age (commonly designated by brain age models) [21].

The brain age research field aims to accurately model healthy ageing using data from healthy participants. The brain age gap estimation (BrainAGE), the difference between the chronological age and the predicted age, is has been shown to be higher in multiple diseases [8, 21], namely Alzheimer's [161, 162], Schizophrenia [173, 229] and has also been related to mortality risk [147]. Therefore, these models may then be clinically important to detect pathological brain age deviations in neurodegenerative diseases [21,161,162,230]. Current state-of-the-art brain age models, leveraging from MRI data, mainly use GM information only (volumetry, cortical thickness, area and depth of sulci in the brain) [8]. Differential morphology features capable of capturing the non uniformity of brain tissue loss are underexplored in brain age prediction. Previous studies have attempted to improve brain age prediction by using morphology features such as gyrification, which corresponds to the ratio between the GM surface length (including the sulci) and the outer GM surface contour length (excluding the sulci) [54, 55], and fractal dimension, which is a measure of brain complexity [145]. However, gyrification might be not representative of the brain morphology changes, as a decrease of its value might be either because of an increase in the outer surface or a decrease in the surface length, or a combination of both. On the other hand, fractal dimensions obey self-similarity rules, which may render them less sensitive to detect differential spatial patterns. Accordingly, the self-similarity of a region may not change significantly despite shape modifications. Typically, the models trained with these morphology metrics have higher errors [55] than the ones trained with volumetry features [116]. The cause for such poor performance might be the extent into which these features are representative of the morphology changes. Moreover, these morphology features have lower spatial resolution, as they are typically computed for larger regions than volumetric features (e.g., GM voxel-wise volume), which might lead to the vanishment of local deformations that happen only in a small portion of the region(s).

In this paper a new morphology feature is explored in the context of brain age prediction: the deformation fields. Deformation fields consist of the voxel-wise nonlinear transformations (e.g., contraction or elongation) that each voxel must go through to match a template. This template fitting is a procedure in MRI preprocessing that aims to reduce interindividual variability and to ensure that features are computed for the same brain location in every image [231]. However, its potential in the context of brain age prediction has not been fully explored.

Different tissue atrophy rates across brain regions lead to a different voxel relative position in relation to the adjacent voxels. Consequently, the repositioning of voxels with a different neighbourhood requires non-linear transformations. Furthermore, the regional structural deformation varies with age [232]. Thus, we raise the hypothesis that deformation fields might carry novel information about brain morphology with predictive value in the context of brain age.

The demonstration that brain deformation is closely associated with age was provided in a cross sectional study [232], although only in a relatively young age range (younger than 51 years). Surprisingly, this finding has so far not been fully exploited to predict the brain age. In fact, the above-mentioned study did not evaluate the prediction error of the deformation fields to predict chronological age. More recently, deformation fields were used to predict the brain age in rodents [233] and in humans along with WM and GM [29], although performance was not investigated because the study's focus was the deep learning architecture rather than the source of information. In conclusion, there are no studies evaluating the gain of information of this feature or comparing it to conventional volumetric features. Thus, this study aims to (a) compare the predictive value of deformation fields with that of GM, WM and CSF, (b) evaluate the added value of the deformation fields and (c) evaluate its robustness against confounding factors such as scanning site and the deformation fields quantitative relationships with age. To achieve these objectives a published brain age pipeline was used [230].

5.3 Methods

5.3.1 Data

The data used in this study are from the open-source repository Information eXtraction from Images (IXI) [210]. IXI contains data on healthy individuals acquired in three sites of London: the Hammersmith Hospital (HH), the Guy's Hospital (GH) and the Institute of Psychiatry (IOP).

The T_1-weighted images of the HH were acquired using a Philips Medical Systems Intera, with a magnetic field strength of 3T, and the acquisition parameters were 9.6 ms, 4.6 ms, and 8° for time of repetition (TR), time of echo (TE) and flip angle, respectively. The GH acquired the data using a Philips Medical Systems Gyroscan Intera with a field strength of 1.5T, and the acquisition parameters were 9.8 ms, 4.6 ms, and 8° for TR, TE and flip angle, respectively. The IOP acquired the data using a General Electrics of 1.5T, yet the acquisition parameters of the images are not provided for this site.

The IXI repository contains 581 T_1-weighted images, 19 images do not have age information and computational anatomy toolbox (CAT12) was unable to preprocess two other images, therefore, in total, 560 T_1-weighted images were used. The different sites have an equivalent distribution of the age and gender (Table 5.1).

In this work, the GH and HH sites, 492 instances, were used to train and validate the models using a cross-validation approach of 30-fold. The IOP was used to test the model's robustness to data from a new site, therefore 68 instances in total.

5.3.2 Image Preprocessing

T_1-weighted MRI images were preprocessed with the CAT12 toolbox (`http://www.neuro.uni-jena.de/cat/`, r1742) within SPM12 (`https://www.fil.ion.ucl.ac.uk/spm/software/spm12/`, *v7771*) using MATLAB (v9.8). CAT12 requires images

Table 5.1: Distribution of the female/male, the mean and standard deviation of the age (years) and the minimum (min) and maximum (max) age in years of the subjects of IXI dataset used in this work.

Site	GH	HH	IOP
Male/female	140/173	86/93	24/44
Mean age ± standard deviation [years]	50.7 ± 16.0	47.0 ± 16.7	42.4 ± 16.6
Minimum/Maximum age [years]	20.1/86.2	20.2/81.9	20.0/86.3

to be aligned with the Anterior Commissure – Posterior Commissure (AC-PC) plane. For this step the ATRA registration toolbox (https://www.nitrc.org/projects/art,v2.0) [234] was used. Then, the CAT12 default segmentation procedure was executed, which includes the standard preprocessing steps such as non-brain tissue removal, noise reduction, image registration and tissue segmentation into GM, WM and CSF. During the registration each voxel goes through different nonlinear transformations, so the native image matches a template in a standard space. These voxel-wise transformations are designated by deformation fields, and they are used in this work to predict the brain age. The registration template used was the CAT12 default template, which was derived from 555 healthy subjects of the IXI dataset.

The data preprocessing and the training of the models were carried in the computing cluster of Laboratory for Advanced Computing (https://www.uc.pt/lca) at the University of Coimbra.

5.3.3 Pipeline overview

The workflow is illustrated in Figure 5.1. The deformation fields, GM, WM, and CSF segmented images that resulted from the preprocessing of the MRI images were used to model chronological age. These images are high dimensional 3D volumes with millions of voxels. All the images were spatially registered in the Montreal Neurological Institute space with dimensionsof 121 x 145 x 121 voxels.

The number of voxels per image greatly outnumbers the number of instances. If all voxels are used without a careful regularisation the model will overfit the training data and have a poor performance on the test data. To overcome this limitation and given the high correlation between adjacent voxels, a data reduction approach using principal component analyis (PCA) was applied to encode the data into a low-dimensional space. The rule of thumb for regression problems is to have at least 10 instances per feature [235]. Thus, given the training and validation dataset size (492 instances) the maximum number of components used to train the models and avoid overfitting and generalisation problems is approximately 40 components.

The gender [236, 237] might have an impact on the component's value. Therefore, the gender was added as a feature, as a binary variable. The standardised PCA components along with gender are the input of a RVR model that predicts the chronological age.

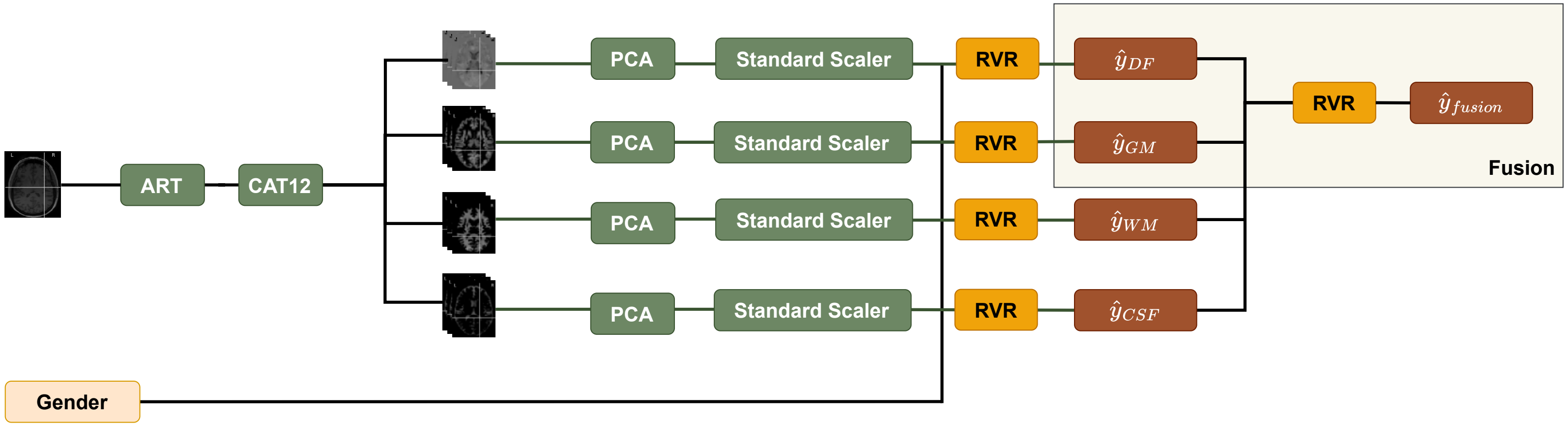


Figure 5.1: Pipeline Overview: A T_1-weighted MRI image is aligned with the AC-PC plane using the ART toolbox, and then the image is preprocessed with the CAT12 software, resulting in four images: the deformation field, and the GM, WM and CSF tissue maps. Then each image is compressed using PCA, the components are standardised and the Relevance Vector Regression model learns the age from the patterns of each feature components and from the gender. The prediction of the age from the deformation fields and GM models are combined using a Relevance Vector Regression model.

The model training and test, except for the image preprocessing, was performed in Python 3.8 using the Scikit-Learn library (V0.24.0) [238] and the statistical analysis were performed in R [239]. The code is available at: `https://cibit-uc.github.io/brainage-deformation-fields`.

5.3.3.1 Dimension reduction

PCA as a dimension reduction tool is a lossy compression algorithm that finds an orthogonal linear transformation such that the data in the new projected space have maximum variance. The first projection is the one with highest variance explained, the second explains less variance than the first one, and so on. If the total ratio of variance explained, of the selected components, is less than 1, then information is lost during the compression, the more PCA components are used the more information is explained.

5.3.3.2 Model

A model is a mapping function that converts the input data, x, into an estimate of the output, $\widehat{y}$. The regression model selected to decode age information from low dimensional MRI-derived images was the RVR model. An architecture combining PCA and RVR was reported as a suitable framework for the brain age using GM information only [230]. The main advantages are the low customisation requirements and its robustness to overfitting. Therefore, this model was the one selected to evaluate the predictive value of deformation fields and compare it to other brain tissue types.

5.3.3.3 Fusion

A model fusion approach was used to evaluate whether deformation fields increased the performance of brain age prediction when combined with GM volumetric information, relative to the prediction based on information from GM volume alone. Succinctly, the predicted age of the models trained only with GM volume and deformation fields alone were the input of a RVR model which in turn predicted an age by weighting the two predictions.

5.3.3.4 Evaluation

To evaluate the goodness of fit, two metrics were used: the mean absolute error (MAE) and R^2.

MAE summarises the mean predictive error and is given by equation 5.1, where N is the number of instances, y_i is the chronological age and $\hat{y}_i$ is the model estimate of the chronological age.

$$MAE = \sum_{i=1}^{N} \frac{|\hat{y_i} - y_i|}{N} \tag{5.1}$$

The R^2 metric encodes the variance explained of the real age value by the predictions, and is computed with the equations (5.2, 5.3, 5.4) below, where the $\overline{y}$ represents the age's mean.

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{5.2}$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{5.3}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5.4}$$

To validate and statistically compare the models a 30-fold cross-validation procedure was performed. The dataset was divided into 30 non-overlapping subsets and in each iteration a different subset was used as a validation set and the remaining ones were used to train the model. The reasoning behind the 30-fold was to have 30 independent values of MAE and R^2, which, according to the central theorem, is the minimum number of independent variables to assure the normal distribution of the sample.

Besides the model validation, performed with data from HH and GH hospitals, the models were also tested using the data from the IOP site, a scheme is shown in Figure 5.2.

5.3.4 Statistical analysis

The statistical test performed depends on the objective, and they are described below. It should be noted that all the MAE and R^2 validation values are independent, thus, we were able to perform statistical analysis using parametric methods.

5.3.4.1 Comparison of deformation fields predictive value with GM, WM, and CSF

In each fold iteration, four different models (one per modality) were trained and evaluated on the same data. In the end each model yielded 30 values of MAE, one per fold. The appropriate test to compare these results is the one-way repeated measures analysis of variance (rm-ANOVA) with modality as a factor and the MAE as the dependent variable. Then, if the p-value is significant, for the modality, a post hoc analysis is performed using Tukey's Honest Significant Difference (HSD) [216], to verify which modalities yield a different predictive value. The results were considered significant if the tests yielded a p-value less than 0.05.

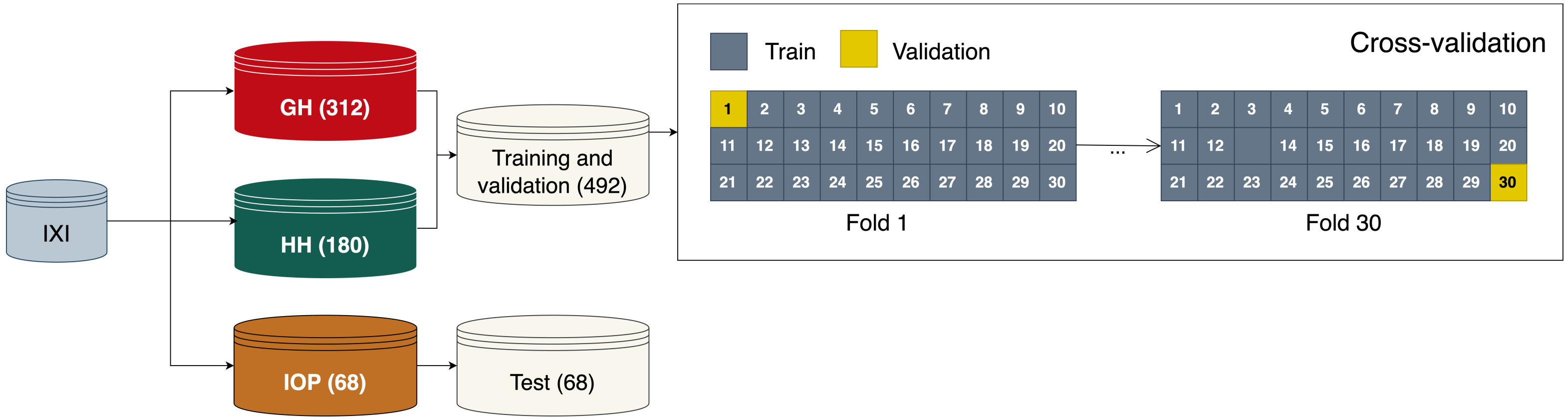


Figure 5.2: Evaluation overview: The IXI dataset is composed by data acquired in three sites of London (GH, HH and IOP). The data from GH and HH were used to train and validate the models whereas data from the IOP site was used to test the models. The model validation is performed using a 30-fold cross-validation approach.

5.3.4.2 Brain age prediction with GM only and GM plus deformation fields

In this case the objective was to test whether there is a performance improvement when the prediction of deformation fields is combined with GM predictions. The fusion was performed at the model level, thus for each data instance there is a value of GM prediction and a fusion prediction (GM + deformation fields). To test the hypothesis a paired t-test was performed. The null hypothesis was that the performance of the model was the same either with GM only or GM and deformation fields predictions combined. The alternative hypothesis was that fusion model have lower (higher) value for the MAE (R^2).

5.3.5 Site impact on components value

To investigate the site impact on the components, the PCA was trained with all the data instances from the training and validation set (GH and HH) and then all the dataset instances (GH, HH and IOP) were compressed using PCA. Then the first two components with higher explained variance were analysed. The objective was to verify whether the site had any effect on the components value. A model was fitted in which the dependent variable was the component and the independent variables (factors) were the age, site, gender and interaction between gender and site. A quadratic age factor was added if a component followed a quadratic trend with age (by visual inspection). The site was one hot encoded and was considered to have an impact on the component if there was a site weight value statistically different from zero.

5.4 Results

5.4.1 Comparison of deformation fields predictive value with GM, WM, and CSF

The results of each model performance on the validation and test set are shown in Table 5.2. The rm-ANOVA, for validation results, point out that different inputs have a significantly different MAE ($F_{3,87} = 5.7$, $p = 0.011$). A follow-up analysis using HSD was performed to identify which pair of modalities yield a significant difference in MAE. We have found that the error in brain age prediction based on deformation fields is statistically lower than using WM ($p = 0.011$) or CSF ($p = 0.0007$). Thus, deformation fields, for the proposed approach, yielded better results than WM and CSF volumetric patterns in brain age prediction. Despite the non-significant result between deformation fields and GM, the former had a better performance than the latter, the MAE difference between both is 0.3 years difference (deformation fields yielded a lower MAE, although not significant). The

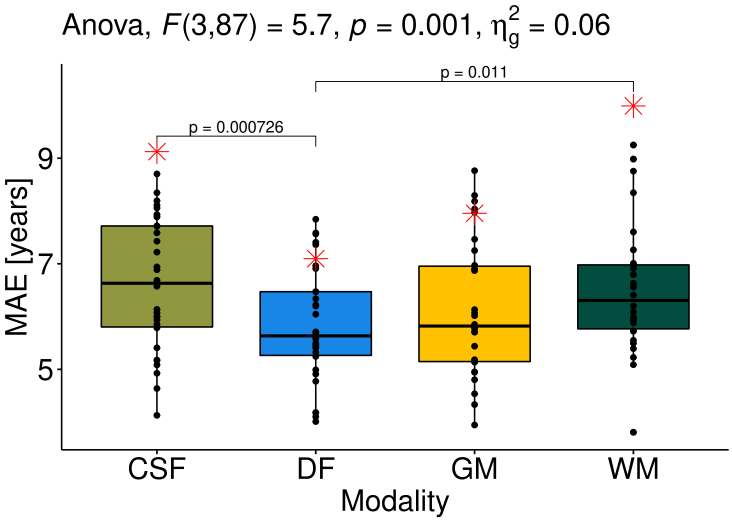


Figure 5.3: Boxplot of the MAE per modality: The statistical results for the one-way repeated measures ANOVA and post-hoc analysis. Each point in the boxplot is the MAE for a cross-validation fold. The horizontal black line represents the median value, and the limits of the boxes represent the 1^{st} and 3^{rd} quartiles. The red star identifies the MAE of the test set.

Table 5.2: Mean and standard deviations of the MAE (in years) for the validation set, 30-fold cross-validation, and MAE for the test, for each input type.

MAE [years]	Deformation fields	GM	WM	CSF
Validation	5.84 ± 1.09	6.14 ± 1.29	6.50 ± 1.20	6.61 ± 1.26
Test	7.10	7.96	9.99	9.12

detailed results for the rm-ANOVA with the MAE as the dependent variable and the modality as the factor are shown in Figure 5.3.

Concerning the results for the test set, the modality with lower MAE is the deformation fields with 7.10 years. The difference between the test set and mean MAE is 1.26, 1.82, 3.49 and 2.51 years for the deformation fields, GM, WM and CSF, respectively. Therefore, the deformation fields yields the lower difference between validation and test set. In the Appendix B.2 the Figure B.1 depicts the relationship between the chronological age and predicted age and tables B.1-B.4 show the models' performance results per age group, considering bins of 10 years. It can be seen that in early adulthood the model overestimates the age whereas in late adulthood the model underestimates the age, as previously reported in literature [240–242]. A study suggests that the bias is neither specific to the model nor the age distribution of the datasets [240].

5.4.2 Brain age prediction with GM only and GM plus deformation fields

Despite the model with deformation fields having a lower MAE value than the GM model (not significant, though), we were interested in investigating whether the

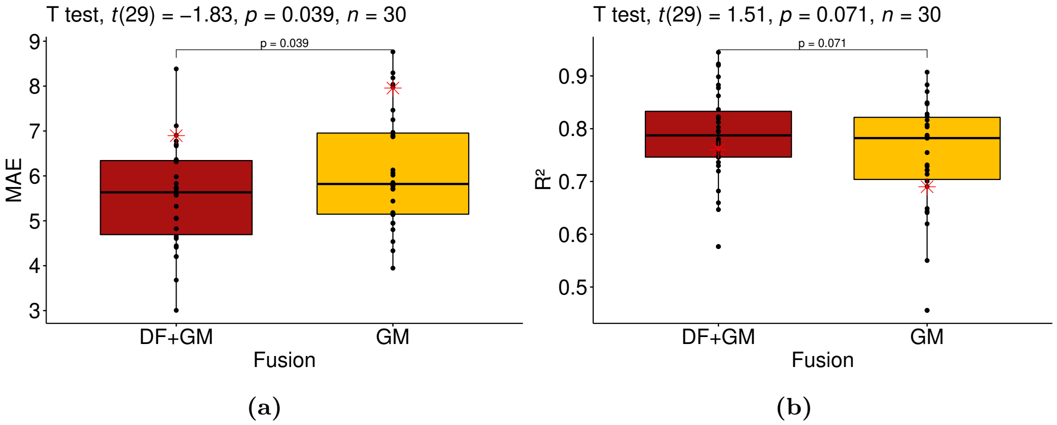


Figure 5.4: Performance comparison of GM model versus a fusion model: Each point in the boxplot is the MAE (a) or R^2 (b) for a cross-validation fold. The horizontal black line represents the median value, and the limits of the boxes represent the 1^{st} and 3^{rd} quartiles. The p-value is the result of a paired t-test. The red star identifies the MAE and R^2 of the test set.

deformation fields add information to the brain age prediction using the volume of GM alone. A paired t-test was performed, and the result is shown in Figure 5.4 and Table 5.3. Indeed, we observed that the MAE of the model fusion is significantly lower ($t_{29} = -1.83$, p=0.039), and that the R^2 is higher, yet not significantly, than those originated by the model trained with the volumetric information of GM only. Concerning the results on the test set, the fusion model yields lower MAE and higher R^2 than the GM model. Furthermore, the difference between the results of the test and validation set is lower on the fusion model, i.e., 1.35 years and 0.06 for MAE and R^2, respectively.

Table 5.3: Mean and standard deviations of the MAE (in years) and R^2 for the validation set, 30-fold cross-validation, and MAE and R^2 for the test of the fusion and GM model.

	MAE [years]		$\boldsymbol{R^2}$	
	Validation	Test	Validation	Test
GM	6.14 ± 1.29	7.96	0.75 ± 0.10	0.69
Deformation fields and GM	5.55 ± 1.14	6.90	0.79 ± 0.09	0.76

An analysis was also performed to compare the performance of deformation fields combined with GM versus GM combined with WM, the results are in the Figure B.3 and Table B.6 in the Appendix B. The paired t-test did not reveal any differences between the performance of the two fusion models on the validation. Nonetheless, on the test set, we could verify the better combination is deformation fields combined with GM rather than GM combined with WM for both MAE (a difference of 1.02 years) and R^2 (a difference of 0.12).

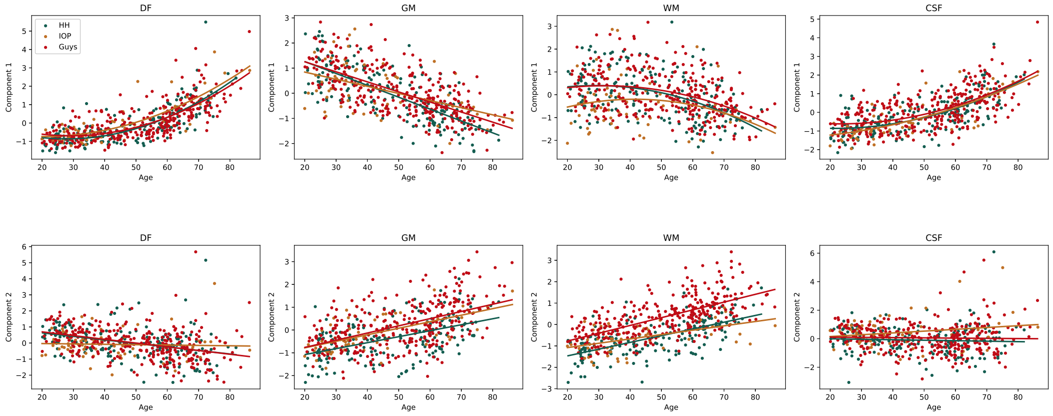


Figure 5.5: First component (first row) and second component (second row) of deformation fields (DF) and each tissue type (GM, WM, and CSF), per image acquisition site. Each point represents a subject and the lines shows the quadratic (linear), linear (linear), quadradic (linear) and quadradic (linear) fitting for DF, GM, WM, and CSF for the first (second) component, respectively, in each site.

5.4.3 Site impact on the components value

The deformation fields were the only modality in which a confounding site effect was not observed (a sign of robustness) on the first two components, the ones which explain more variance. The transformation into a 40-dimension space encodes 46%, 33%, 36% and 52% for deformation fields, GM, WM and CSF, respectively. In Figure B.2 in the Appendix B is shown slices of the original image and the reconstruction from the 40 components. One can see that, although more than 48% of the variance is lost, the images are similar, and the overall information prevails. The variance explained for the first (second) component is 15% (3%), 15% (1%), 18% (2%), 23% (7%) for deformation fields, GM, WM and CSF, respectively. It should be noted that the first two components explain, approximately half of the variance explained by the 40 components. The evolution of the first two components' value with age is shown in Figure 5.5. The first component follows a quadratic evolution for the deformation fields, monotonically linear decreases for GM, an inverted u-shape for WM and a quadratic increase for CSF, respectively. Regarding the second component, all the components follow a linear trend with age. The site effect in GM and WM the second component is more evident than in any other image, suggesting loss of invariance to this confounding variable.

5.5 Discussion

The major finding of this study is that voxel-wise local deformation provides novel valuable information to predict brain age. The fusion, at the model level, of the GM and deformation field components, yields better results than a model trained with GM volume only. Furthermore, we have found that the model trained only with deformation fields outperforms models trained with WM or CSF volume information and has similar performance to the model trained with common GM volumetric maps. These results show the deformation fields have complementary information to GM images and should be considered as an independent feature in brain age models. Deformation based morphometry studies have shown differences between healthy controls and patients with neurological diseases (schizophrenia [243] and Alzheimer's disease [244] although they had not been used for prediction. Brain ageing deviations have been related to different diseases [8, 161, 162, 173, 245], the inclusion of this feature in the models not only improves its performance but might also be valuable for a differential diagnosis. Here we show that deformation fields capture deviations from the healthy ageing that have predictive value. Further. It remains however a question whether this information is useful in the clinical setting. Therefore, further research is required to understand whether any of these models can be used as a biomarker for neurological diseases such as schizophrenia and Alzheimer's disease.

The analysis of the first component of deformation fields (Figure 5.5) revealed an acceleration of the ageing process from 50-60 years onwards, which arises from differences in regional tissue loss across the brain and, consequently, different deformation patterns. The first PCA component has a positive quadratic relation of the deformation fields with age, which means that more non-linear transformations must be made to match a template, suggesting that older brains are more deformed. The deformation fields dependency at the voxel and region-of-interest level with age was studied in Pieperhoff at [232], which showed a linear positive relation of the deformation fields with age. However, the study only included participants up to 51 years old. Accordingly, by inspecting Figure 5.5, before 50 years old the first component of DF seems to have a positive linear relation with age. Yet, from 50-60 years old onwards the component values dramatically increase with age, at least quadratically. We believe that the first component of deformation fields represents the overall deformation fields pattern in the brain since the first component of GM, WM and CSF followed the expected reported trend with age. Previous reports suggest that GM volume monotonically decreases with age [246], that the relation of WM volume with age has an inverted u-shape pattern [59, 61] and that the volume of CSF [246, 247] increases with age patterns, which one can observe in the first component of these tissues.

Notably, our results reveal that the deformations fields appear to be more robust to variations elicited by the site of image acquisition. The statistical analysis of the first two PCA components to the site effect revealed an invariance to the site. On the contrary, on the first two components of GM, WM and CSF volume maps there was a significant site effect. This effect is notorious in the second component in Figure 5: the IOP components are completely clustered apart for the WM and partially for GM volume maps. Currently, the MAE doubles when the test set has data from a different site of the training set. Data harmonisation algorithms have been developed to reduce site effects and remove the scanner imprints, in an attempt to overcome the current observation and achieve generalised models [248]. However, in most of the cases models must be retrained to learn the scanner patterns. Our results suggest that the deformations fields might be intrinsically more robust to scanner artifacts and site dependencies, even without data harmonisation.

5.6 Conclusion

The deformations fields show a great potential to predict the brain age, i.e., they should be considered as an independent feature in brain age models and seem to be robust against confounding factors such as image acquisition sites. In this work a simple fusion approach was employed. In the future further approaches should explore how to optimally combine the information from the deformation fields with that of grey matter volumetric maps. Moreover, the robustness of deformations fields regarding site of image acquisition makes this feature potentially more suitable to achieve more generalisable models.

Chapter 6

3DCAE-MRI: Overcoming Data Availability Limitations in Small Sample MRI Studies

This chapter is based in the following article:

Dias, Maria de Fátima Machado, Tiago FT Cerqueira, João Valente Duarte, Miguel Castelo-Branco, and Paulo Carvalho. "3DCAE-MRI: Overcoming Data Availability Limitations in Small Sample MRI Studies." Scientific Reports (2023).

6.1 Abstract

Deep learning (DL) methods are data-driven models that learn abstract, hierarchical features from data. These models perform exceptionally well in image recognition tasks when large amounts of data are used. The availability of open-source large-scale annotated datasets is crucial for developing high-performance models. Nevertheless, the use of DL models in neuroimaging is restricted due to data availability constraints. Transfer learning has been used to overcome this issue. Nonetheless, most related studies have focused on transferring knowledge from/to an Alzheimer's disease context. In this paper, we propose a 3D-convolutional autoencoder magnetic resonance imaging (3DCAE-MRI) model to improve upon the performance of supervised DL models in small-sample magnetic resonance imaging (MRI) studies. We exploited multiple open-source MRI datasets to train 3DCAE-MRI. Transfer learning was applied to two model architectures and benchmark problems: age prediction and sex classification. With respect to age prediction, conducting off-the-shelf learning with 3DCAE-MRI while using only 300 training samples yielded remarkable performance on an external dataset; this performance was superior to that of state-of-the-art methods that use thousands of images. In terms of both age prediction and sex classification, off-the-shelf learning using 3DCAE-MRI led to better performance, higher generalisability and more stable models than fine-tuning or training the network from scratch.

6.2 Introduction

Supervised DL models are over-reliant on large annotated datasets to attain optimal performance and generalisability [32]. ImageNet [249], a widely used open-source dataset comprising more than one million natural images classified into more than 1200 categories, has been indispensable for developing multiple state-of-the-art image segmentation and object recognition models. Unfortunately, no similar MRI dataset is available, mainly because the acquisition process is time-consuming, expensive and requires technical expertise. MRI studies often involve a reduced number of participants (from dozens up to hundreds), which is insufficient for training DL models. Despite the efforts of various initiatives [209,250–256] aimed at collecting hundreds or thousands of MR images, the data collection process still poses challenges in specific contexts. Collecting millions of MRI images per context may be impractical, and for certain rare pathologies, it may be impossible to gather such large amounts of data. To overcome this constraint, studies with small, annotated datasets employ transfer learning to improve the performance of their models.

Transfer learning reuses the weights trained in a particular context (source domain) to facilitate the learning process for similar problems (target domain) [257–264]. Two main types of images have been used as source domains in MRI studies:

natural images and MRI data [265]. Regarding natural images, the most commonly used dataset is ImageNet [257–259, 265, 266]; related studies have focused on using architectures known from visual recognition tasks such as the residual network (ResNet) [153], the visual geometry group network (VGG) [267], and AlexNet [268] pretrained on the ImageNet dataset. These networks were then retrained to address, for example, Alzheimer's disease (AD) related problems [257–259]. However, performing transfer learning based on this dataset and using these architectures lead to two main disadvantages. First, natural and medical images are morphologically different; thus, specific medical patterns may not be present in such images. Second, natural images are two-dimensional, whereas structural MR images are three-dimensional. Consequently, the information encoded between slices is lost, compromising the performance of the constructed model [126].

Despite the lack of an analogous dataset to ImageNet containing MRI data, several popular MRI databases have been used to pretrain 3D-convolution neural networks (CNN) models. Most of the related studies have focused on transfer learning from and to AD contexts using only the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset as a source domain [260–262, 265, 269]. Currently, most studies pretrain 3D-CNN models to distinguish between healthy controls and AD patients. These pretrained 3D-CNN models are subsequently fine-tuned in other contexts [260–262]. Another strategy proposed by Oh *et al.* [262] involves performing transfer learning from a 3D-convolutional autoencoder (3D-CAE). The authors trained a 3D-CAE solely with data acquired from healthy controls and AD patients. Then, the weights were transferred to a classification model, which aimed to distinguish between AD patients and two mild cognitive impairment conditions. This approach presents an opportunity for improvement: the features learned in the source domain might be too specific for AD, which might restrain the transfer learning process for different problems; furthermore, the generalisability of the method to other problems was not assessed. Despite the publication of multiple open-source MRI databases, such as the autism brain imaging data exchange (ABIDE) [209, 253], the ADNI [250], Open Access Series of Imaging Studies (OASIS) [254–256], among others, to our knowledge, there is still no approach that can exploit all these datasets as source domains. This paper proposes 3DCAE-MRI, a 3D-CAE architecture trained with multiple open-source MRI datasets. Transfer learning is applied to 3DCAE-MRI to optimise a regression model for predicting brain age and a classification model aimed at sex prediction.

Brain age gap estimation (BrainAGE) is a putative biomarker that aims to detect atypical brain ageing [21]. This biomarker leverages machine learning models to evaluate healthy brain ageing. The difference between the estimated brain age and the corresponding chronological age is used to measure the acceleration or deceleration of brain ageing. Studies suggest that the BrainAGE is increased in patients with multiple pathologies, such as AD [9,162,230], schizophrenia [9,229,270] and multiple sclerosis [9,180,271]. The development of accurate brain ageing models might be essential for assisting in disease diagnosis tasks at an early stage and for monitoring disease progression. Deep learning models trained using thousands of training instances have achieved remarkable results. However, the backbones of these studies are either local data or the UKBiobank dataset [272], a paid repository; thus, these data are not easily accessible by the scientific community. Moreover, most brain age models are evaluated on a test set comprising data acquired with the same acquisition settings used for the training data (a holdout test set). Studies that have evaluated brain age models on external test sets have reported decreased performance on external test sets compared to that attained on the holdout test set [28–31].

The sex prediction problem is considered in this study to evaluate the generalisability of transfer learning from 3DCAE-MRI for classification tasks. The brain structure and function of males and females are dissimilar [273–275]. It is being increasingly recognised that sex is a relevant source of bias in many neuropsychiatric conditions and that it affects disease processes in very distinct ways [276–278]. Machine learning models, which successfully classify sex based on MRI images [4, 279, 280], might help uncover the trajectory of sexual dimorphism in the brain and uncover the relationships between medical conditions and sex.

The current work proposes an unsupervised model, 3DCAE-MRI, to learn abstract and unspecific features from already-available MRI data. We assess whether conducting transfer learning from 3DCAE-MRI (in an off-the-shelf and fine-tuning fashion) to a problem-specific CNN yields better performance and is more stable than training the CNN from scratch. The achieved performance is assessed on a holdout set and an external set for the regression and classification problems. Furthermore, the performances of different deep learning training strategies under different training sizes are compared to those of a shallow learning approach.

6.3 Results

This work assesses whether conducting transfer learning from a 3D-CAE improves the performance attained for two benchmark problems: age prediction and sex classification. Two architectures are considered: LiteNet, which is a custom architecture, and the simple fully convolutional network (SFCN) [4], which is a contest-winning network. The performance of the DL models is compared to that of a shallow model. In the shallow model, the input MRI volume is compressed using principal component analyis (PCA), and a relevant vector machine (RVM) is subsequently used for classification or regression.

Transfer learning from 3DCAE-MRI is performed using an off-the-shelf and fine-tuning approach. In the former method, the reused weights are not optimised, whereas in the latter strategy, the weights are updated during training. Then, on

LiteNet, these two approaches are compared to two training-from-scratch strategies: one with augmentation and one without augmentation. The transformations considered for training from scratch with augmentation are random translation and flipping across the sagittal plane. For the SFCN, transfer learning from 3DCAE-MRI is compared to transfer learning from pretrained models (based on sex and age). Further details are provided in the Methods section.

6.3.1 Brain age prediction

6.3.1.1 Holdout test set

The mean absolute error (MAE) produced with different training sample sizes, training strategies and CNN architectures on the holdout test set are presented in Table 6.1. Figure C.1 in the Appendix C depicts the relation between the training sample size and the MAE for different training strategies. The repeated measures analysis of variance (rm-ANOVA) and post hoc analysis results obtained for LiteNet (SFCN) are shown in Tables C.3 (C.5) and C.4 (C.6) in the Appendix C. The results suggest that the training strategy plays a crucial role in determining the MAE under all training sample sizes. Concerning the LiteNet architecture, transfer learning (via off-the-shelf training or fine-tuning) from 3DCAE-MRI attains better performance on the holdout test set. Compared with other training strategies, off-the-shelf learning from 3DCAE-MRI significantly stands out as the best training strategy for small training sets (25, 50) with very large or large differences. Under 100 training samples, the fine-tuning and off-the-shelf methods have similar performances, and under 200 and 300 training samples, fine-tuning yields superior performance. In the case of the SFCN architecture, transfer learning from 3DCAE-MRI using fine-tuning attains better performance under 25 and 50 training samples. In contrast, transfer learning from an age-pretrained model performs better under 200 and 300 training samples. Under 100 training samples, the performances of transfer learning from 3DCAE-MRI and the pretrained model are similar. Nevertheless, the difference between these two training strategies (3DCAE-MRI fine-tuning and the age-pretrained model) is negligible for all training sample sizes. Concerning PCA versus DL models, the results highlight that both CNN architectures yield better performance than PCA for all the cases except for the LiteNet architecture trained from scratch with 25 samples. The performance differences between the DL and PCA models are usually very large.

6.3.1.2 External test set

The means and standard deviations of the MAEs obtained on the external test set are shown in Table 6.2 and Figure 6.1. The statistical analysis results obtained for LiteNet (SFCN) are shown in Tables C.7 (C.9) and C.8 (C.10) in the Appendix C. The rm-ANOVA suggests that the MAE differs across various training strategies

Table 6.1: MAEs and standard deviations obtained for the age prediction task on the holdout test set by PCA and two CNN architectures under different training strategies and training set sizes. The lowest value obtained by each CNN architecture under each training size is represented in bold; the symbols * and † indicate that PCA achieves better performance than do the LiteNet and SFCN architectures, respectively.

Training Strategy	PCA-RVM	LiteNet				SFCN			
		Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
		No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	8.80 ± 0.53	9.26 ± 1.32	9.84 ± 1.12	7.90 ± 1.26	$\mathbf{6.65 \pm 0.57}$	7.05 ± 0.66	7.27 ± 0.65	$\mathbf{6.98 \pm 0.54}$	7.07 ± 0.51
50	7.89 ± 0.48	6.69 ± 0.60	7.78 ± 0.96	6.02 ± 0.50	$\mathbf{5.76 \pm 0.40}$	6.45 ± 0.42	6.66 ± 0.62	$\mathbf{6.34 \pm 0.46}$	6.61 ± 0.39
100	7.37 ± 0.35	5.6 ± 0.46	6.78 ± 1.16	$\mathbf{5.20 \pm 0.41}$	5.28 ± 0.42	$\mathbf{6.14 \pm 0.53}$	6.58 ± 0.48	$\mathbf{6.14 \pm 0.41}$	6.42 ± 0.33
200	6.16 ± 0.56	5.06 ± 0.43	5.43 ± 0.81	$\mathbf{4.80 \pm 0.30}$	4.99 ± 0.36	$\mathbf{5.38 \pm 0.42}$	5.73 ± 0.5	5.46 ± 0.45	6.05 ± 0.41
300	5.76 ± 0.48	4.75 ± 0.36	6.3 ± 1.26	$\mathbf{4.42 \pm 0.35}$	4.67 ± 0.35	$\mathbf{4.96 \pm 0.45}$	5.32 ± 0.53	4.99 ± 0.49	5.59 ± 0.4

Table 6.2: The MAEs and standard deviations of the age prediction results produced for PCA-RVM and two CNN architectures under different training strategies and training set sizes are evaluated on the external test set. The lowest values obtained for each CNN architecture and training size are represented in bold; the symbols * and † indicate that PCA-RVM achieves better performance than do the LiteNet and SFCN architectures, respectively.

Training Strategy	PCA-RVM	LiteNet				SFCN			
		Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
		No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	11.26 ± 0.56	11.89 ± 1.43	11.97 ± 1.44	9.96 ± 0.84	$\mathbf{8.88 \pm 0.6}$	8.18 ± 0.58	8.55 ± 0.85	8.12 ± 0.71	$\mathbf{8.01 \pm 0.6}$
50	10.81 ± 0.6	10.85 ± 1.63	11.57 ± 1.42	8.95 ± 0.81	$\mathbf{8.11 \pm 0.43}$	7.88 ± 0.58	7.77 ± 0.72	$\mathbf{7.32 \pm 0.53}$	7.47 ± 0.42
100	9.79 ± 0.42	8.93 ± 1.36	10.64 ± 1.55	8.01 ± 0.71	$\mathbf{7.40 \pm 0.32}$	7.27 ± 0.48	7.55 ± 0.63	7.12 ± 0.44	$\mathbf{6.94 \pm 0.29}$
200	8.2 ± 0.32	9.02 ± 1.14	9.41 ± 1.37	7.36 ± 0.53	$\mathbf{6.95 \pm 0.35}$	7.05 ± 0.62	7.05 ± 0.61	6.73 ± 0.45	$\mathbf{6.44 \pm 0.35}$
300	7.42 ± 0.21	8.73 ± 1.06	10.02 ± 1.47	7.22 ± 0.54	$\mathbf{6.71 \pm 0.25}$	7.46 ± 0.98	6.71 ± 0.72	6.95 ± 0.6	$\mathbf{5.93 \pm 0.36}$

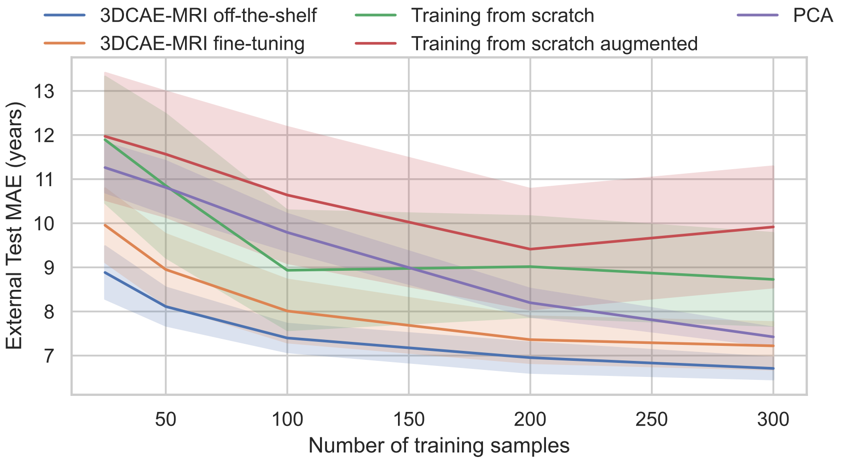
under all training dataset sizes. For both CNN architectures, transfer learning from 3DCAE-MRI yields lower MAEs than the other training strategies on the external test set. LiteNet attains lower mean MAEs using off-the-shelf transfer learning from 3DCAE-MRI. The post hoc results also provide evidence that the effect of training with an off-the-shelf strategy is very large compared to those of training from scratch and PCA-RVM for all training dataset sizes. Fine-tuning outperforms both PCA-RVM and training from scratch (with and without augmentation).

Similarly, for the SFCN architecture, the off-the-shelf training strategy yields the best performance for all training cases except for that with 50 training samples. Under 50 training samples, the best training strategy is fine-tuning for 3DCAE-MRI. Nevertheless, under both 50 and 100 training samples, the MAEs of off-the-shelf training and fine-tuning with 3DCAE-MRI are equivalent. Furthermore, the post hoc results reveal that under 25 training samples, the performances of off-the-shelf training and fine-tuning with 3DCAE-MRI are similar to that of the age-pretrained model. Under 100 training samples, the difference between the fine-tuning results obtained with 3DCAE-MRI and the age-pretrained model is not significant. Similarly, under 200 training samples, fine-tuning a model using weights derived from a sex-pretrained model and utilising 3DCAE-MRI are equivalent. Finally, under 300 training samples, a shallow approach (PCA-RVM) performs similarly to fine-tuning an age-pretrained model.

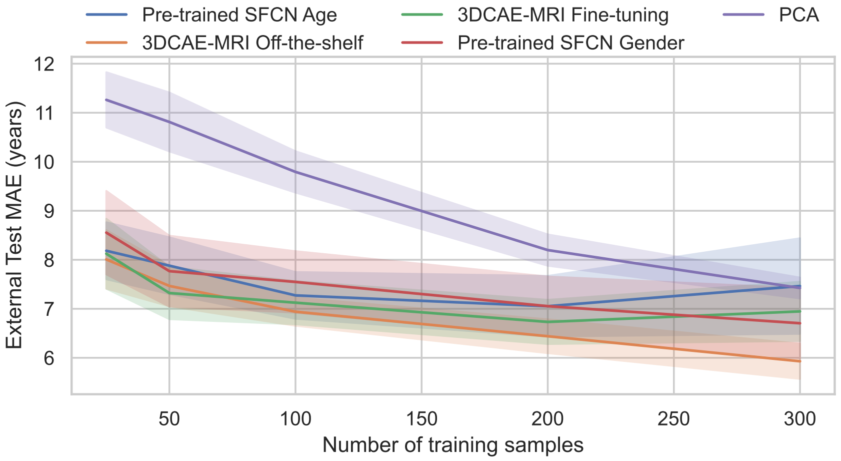
6.3.1.3 Stability

An analysis of the training stability of both CNN architectures over the last 30 epochs with different DL training strategies is exhibited in Table 6.3. A statistical analysis comparing the validation variability produced across different training strategies for the LiteNet (SFCN) architecture is shown in Table C.11 (C.13), and the post hoc analysis is shown in Table C.12 (C.14) in the Appendix C. For LiteNet, the results reveal that off-the-shelf learning with 3DCAE-MRI has lower validation variability than do the other methods, followed by 3DCAE-MRI fine-tuning and training from scratch (with and without augmentation). The statistical analysis of the validation variability produced across different training strategies suggests that the mean of the stability metric significantly differs across training strategies for all the training dataset sizes. The differences between the off-the-shelf approach and the other DL training strategies are very large for all the training dataset sizes.

The findings for the SFCN suggest that fine-tuning from an age-pretrained model is more stable than any other training strategy. The validation variability difference between fine-tuning an age-pretrained model and the other DL training strategies are very large. Fine-tuning and off-the-shelf processing from 3DCAE-MRI yield equivalent results for 25 and 50 training samples, respectively. Transfer learning from 3DCAE-MRI (using fine-tuning) or a sex-pretrained model, in terms of stability, is



(a) LiteNet



(b) SFCN

Figure 6.1: External test set MAE results for age prediction problem. Training evolution for the age prediction problem with different dataset training sizes for the four training strategies and convolution neural network architectures (LiteNet and SFCN). The shaded band represents to the standard deviation.

similar under 25, 100, 200 and 300 training samples. Finally, training from 3DCAE-MRI using an off-the-shelf approach yields the same stability as that attained by performing training with a sex-pretrained model.

6.3.2 Sex classification

6.3.2.1 Holdout test set

The results obtained on the holdout test set are shown in Table 6.4 and Figure C.2b in the Appendix Cl. The results show that the accuracy monotonically increases with the number of training instances for all the training strategies and both CNN architectures. The statistical analysis conducted for LiteNet (SFCN) is shown in Table C.15 (C.17), and the post hoc results are shown in Table C.16 (C.18) in the Appendix C.

For the LiteNet architecture, the results show that off-the-shelf transfer learning from 3DCAE-MRI has significantly greater accuracy than fine-tuning and training from scratch (with and without augmentation) under all dataset sizes. The accuracy differences between off-the-shelf training and training from scratch (with and without augmentation) are either large or very large. Similarly, the differences between the off-the-shelf and fine-tuning methods are large across all set training sizes. PCA-RVM outperforms the off-the-shelf method under 200 and 300 training samples, but the differences are not significant. PCA-RVM achieves significantly superior performance to that of training from scratch (with and without augmentation) and fine-tuning across all numbers of training samples except for 50.

The off-the-shelf training strategy with 3DCAE-MRI attains better performance on the holdout test set than do the other training strategies with the SFCN architecture. Nonetheless, the accuracy differences are not significant in some cases. Transfer learning from 3DCAE-MRI using an off-the-shelf method and the fine-tuning approach are equivalent for all training cases except for that with 300 training instances. Moreover, the difference between the accuracies of a shallow model (PCA-RVM) and the SFCN trained using an off-the-shelf approach is not significant. The results of the comparisons between the pretrained models are inconsistent. The sex-pretrained model attains better accuracy for 25 training samples. Nevertheless, the difference is not significant. The pretrained models yield the same accuracy for 100 training samples, whereas the age-pretrained model yields significantly greater accuracy under 50, 200 and 300 training samples.

6.3.2.2 External test set

The results obtained on the external test set are shown in Table 6.5 and Figure 6.2. The rm-ANOVA and post hoc analysis results produced for the LiteNet (SFCN) architectures are shown in Tables C.19 (C.21) and C.20 (C.22), respectively, in the Appendix C.

Table 6.3: Means and standard deviations of the validation variability across different training set sizes and training strategies for the age prediction problem. The lowest value attained for each training set size is represented in bold.

Training Strategy	LiteNet				SFCN			
	Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
	No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	5.57 ± 0.81	4.8 ± 0.74	4.52 ± 0.57	**1.74 ± 0.49**	**8.31 ± 1.73**	12.44 ± 3.86	12.91 ± 3.21	11.42 ± 3.1
50	2.42 ± 0.81	3.15 ± 0.97	1.95 ± 0.45	**0.79 ± 0.22**	**6.31 ± 1.79**	9.84 ± 2.0	11.83 ± 2.37	11.61 ± 2.64
100	2.65 ± 0.98	3.57 ± 0.95	2.01 ± 0.58	**0.86 ± 0.24**	**6.56 ± 1.51**	9.16 ± 1.18	9.5 ± 1.37	11.19 ± 2.9
200	2.12 ± 0.75	3.02 ± 0.71	1.53 ± 0.33	**0.89 ± 0.29**	**3.74 ± 0.91**	6.41 ± 1.01	6.6 ± 1.09	9.08 ± 2.41
300	1.93 ± 0.63	2.34 ± 0.57	1.17 ± 0.27	**0.63 ± 0.18**	**2.52 ± 0.62**	4.49 ± 0.91	4.25 ± 1.01	7.21 ± 1.46

Table 6.4: The means and standard deviations of the sex classification accuracies produced by PCA-RVM and two CNN architectures under different training strategies and training set sizes on the holdout test set. The lowest values for each CNN architecture and training size are represented in bold; the symbols * and † indicate that PCA-RVM achieves better performance than do the LiteNet and SFCN architectures, respectively.

Training Strategy	PCA-RVM	LiteNet				SFCN			
		Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
		No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	0.73 ± 0.05	0.69 ± 0.08	0.70 ± 0.06	0.68 ± 0.08	**0.76 ± 0.05**	0.77 ± 0.04	0.78 ± 0.04	0.79 ± 0.05	**0.80 ± 0.03**
50	0.78 ± 0.03	0.74 ± 0.05	0.73 ± 0.05	0.77 ± 0.05	**0.80 ± 0.03**	0.79 ± 0.04	0.78 ± 0.03	0.81 ± 0.03	**0.82 ± 0.03**
100	0.81 ± 0.04	0.79 ± 0.03	0.78 ± 0.05	0.79 ± 0.04	**0.83 ± 0.04**	0.80 ± 0.03	0.80 ± 0.04	0.82 ± 0.04	**0.84 ± 0.04**
200	0.86 ± 0.03*	0.81 ± 0.05	0.82 ± 0.04	0.82 ± 0.04	**0.85 ± 0.03**	0.84 ± 0.04	0.81 ± 0.04	0.85 ± 0.03	**0.86 ± 0.03**
300	0.88 ± 0.03*	0.83 ± 0.05	0.82 ± 0.05	0.84 ± 0.05	**0.87 ± 0.04**	0.85 ± 0.04	0.83 ± 0.04	0.86 ± 0.04	**0.89 ± 0.03**

Table 6.5: The means and standard deviations of the sex classification accuracies produced by PCA-RVM and two CNN architectures under different training strategies and training set sizes on the external test set. The lowest values for each CNN architecture and training size are represented in bold; the symbols * and † indicate that PCA-RVM performs better than do the LiteNet and SFCN architectures, respectively.

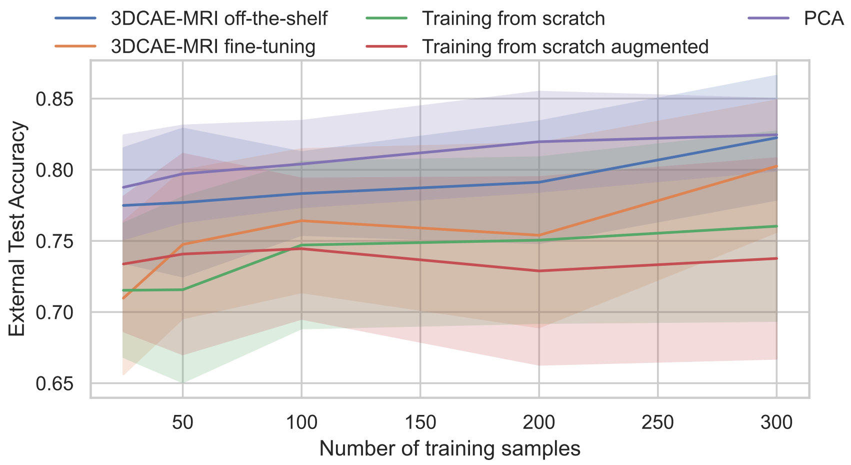
Training Strategy	PCA-RVM	LiteNet				SFCN			
		Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
		No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	0.79 ± 0.04 * †	0.72 ± 0.05	0.73 ± 0.05	0.71 ± 0.05	**0.78 ± 0.04**	0.77 ± 0.04	0.77 ± 0.04	0.79 ± 0.03	**0.80 ± 0.03**
50	0.80 ± 0.03*	0.72 ± 0.07	0.74 ± 0.07	0.75 ± 0.05	**0.78 ± 0.05**	0.77 ± 0.04	0.76 ± 0.04	**0.81 ± 0.02**	0.80 ± 0.02
100	0.80 ± 0.03 * †	0.75 ± 0.06	0.74 ± 0.05	0.76 ± 0.05	**0.78 ± 0.03**	0.77 ± 0.04	0.78 ± 0.05	0.79 ± 0.03	**0.80 ± 0.02**
200	0.82 ± 0.04 * †	0.75 ± 0.06	0.73 ± 0.07	0.75 ± 0.06	**0.79 ± 0.04**	0.77 ± 0.04	0.76 ± 0.04	0.80 ± 0.04	**0.81 ± 0.02**
300	0.82 ± 0.03 * †	0.76 ± 0.07	0.74 ± 0.07	0.80 ± 0.05	**0.82 ± 0.04**	0.79 ± 0.05	0.78 ± 0.07	**0.81 ± 0.03**	**0.81 ± 0.02**

A higher accuracy of 0.82 is attained for LiteNet by off-the-shelf training with 3DCAE-MRI and PCA-RVM. These training strategies have superior performance to that of fine-tuning or training from scratch (with and without augmentation). In this case, the accuracies of the off-the-shelf, PCA-RVM and training-from-scratch strategies are maintained or increase with the training size. The same conclusion applies to fine-tuning except for the case with 200 training samples, for which a small decrease is observed. These findings suggest that the selected training strategy impacts the model performance achieved with all dataset training set sizes. The off-the-shelf and PCA-RVM methods perform significantly better than do the other three training strategies on the external dataset, except under the training dataset size of 300, for which the difference between the off-the-shelf and fine-tuning methods is not significant. PCA-RVM significantly outperforms the off-the-shelf method under 100 and 200 training samples. The Cohen's d value differences between PCA-RVM and training from scratch are as follows (with the corresponding number of training samples in parentheses): very large (25, 200), large (50, 100) and moderate (300). The difference between the off-the-shelf and training-from-scratch strategies is very large for 25 training samples; large for 50, 200 and 300 training samples; and moderate for 100 samples. Fine-tuning yields better accuracy than training from scratch, but the difference is significant only for 50 and 300 training samples (with a moderate effect).

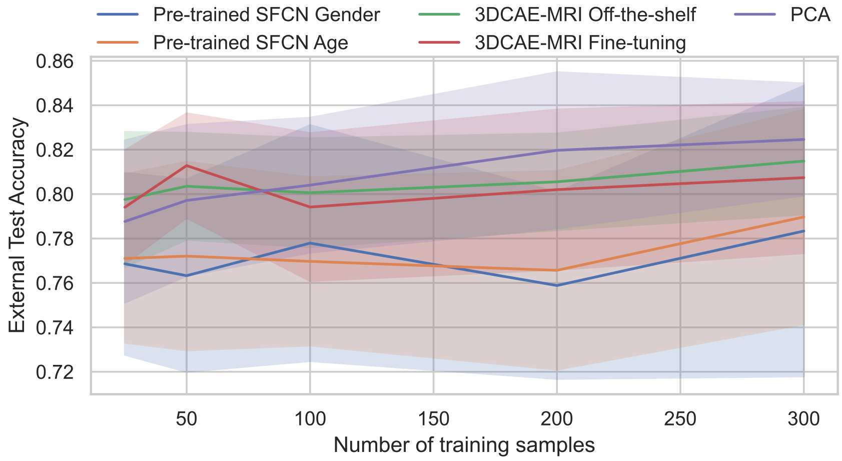
For the SFCN, transfer learning from 3DCAE-MRI attains better results than do the age- or sex-pretrained models. The models trained using the off-the-shelf approach have higher accuracies under 25, 100 and 300 training samples. Fine-tuning from 3DCAE-MRI yields better performance for 50 samples, and both strategies yield the same accuracy (0.81) for 200 training samples. Nonetheless, the accuracy differences between the off-the-shelf and fine-tuning training strategies with 3DCAE-MRI are not significant for any of the training samples. Concerning the comparison between the SFCN and PCA-RVM, the results highlight that the SFCN yields better performance for 25 and 50 training samples, but the accuracy differences are not significant. PCA-RVM and the SFCN yields the same performance for 100 training samples, and PCA-RVM outperformed the SFCN that transfer learning from an age-pretrained model and doing so from a sex-pretrained model are equivalent to sex classification.

6.3.2.3 Stability

The model stability results are shown in Table 6.6. The rm-ANOVA and post hoc results obtained for LiteNet (SFCN) are shown in Tables C.23 (C.25) and C.24 (C.26) in the Appendix C. The findings reveal that the selected training strategy impacts the validation variability levels of both architectures under all training samples, except for LiteNet with 200 training samples. Moreover, off-the-shelf



(a) LiteNet



(b) SFCN

Figure 6.2: Accuracy results obtained for the sex classification problem on the external test set. The training evolution trends yielded for the sex classification problem with different training dataset sizes by the different training strategies and CNN architectures (LiteNet and SFCN) are shown. The shaded band represents the standard deviation.

yields lower validation variability on LiteNet (i.e., greater stability) than do fine-tuning and training from scratch (with and without augmentation) under all the training sample sizes except 300. In this exception, training from scratch with augmentation yields better performance than off-the-shelf training; however, the difference is not significant. In the case of the SFCN, in general, the sex-pretrained model is equivalent to or more stable than any other DL-based training strategy.

Table 6.6: Means and standard deviations of the validation variability produced across different training set sizes and training strategies for the sex prediction problem. The lowest value obtained for each training set is represented in bold.

Training Strategy	LiteNet				SFCN			
	Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
	No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	0.08 ± 0.01	0.08 ± 0.01	0.09 ± 0.01	$\mathbf{0.060 \pm 0.004}$	$\mathbf{0.10 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	0.11 ± 0.01	0.12 ± 0.01
50	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.10 ± 0.01	$\mathbf{0.08 \pm 0.02}$	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.11 \pm 0.01}$	0.12 ± 0.01
100	0.10 ± 0.01	0.10 ± 0.01	0.11 ± 0.01	$\mathbf{0.09 \pm 0.01}$	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.11 \pm 0.02}$	0.13 ± 0.01
200	0.12 ± 0.01	0.12 ± 0.01	0.12 ± 0.02	$\mathbf{0.11 \pm 0.02}$	0.13 ± 0.01	$\mathbf{0.12 \pm 0.01}$	0.13 ± 0.01	0.14 ± 0.01
300	0.12 ± 0.01	$\mathbf{0.11 \pm 0.01}$	0.13 ± 0.01	0.12 ± 0.02	0.13 ± 0.01	$\mathbf{0.12 \pm 0.01}$	0.14 ± 0.01	0.15 ± 0.01

6.4 Discussion

The findings of this work elucidate how conducting transfer learning from 3DCAE-MRI to two CNN architectures (LiteNet and the SFCN) might improve the performance of these models.

Compared with the other training strategies, i.e., transfer learning from pretrained models, training from scratch with and without augmentation, and PCA-RVM, off-the-shelf learning from 3DCAE-MRI performs remarkably well in terms of age prediction on an external test set for both architectures. In general, it performs significantly better than the other training strategies. Moreover, in the most cases, the effect sizes are very large, which provides significant evidence for the superior performance of this training strategy. Most published works on age prediction using DL have reported that the MAE doubles on external datasets [28, 29, 31]. In our case, we verify the same trend for both architectures. For LiteNet, training a model from scratch without (with) augmentation yields an MAE of 4.75 (6.3) years on the holdout test set. The performance of these models on an external test set is 8.73 (10.02) years; thus, the difference between the test sets is 3.98 (3.90) years. When utilising the off-the-shelf training strategy on LiteNet, the MAE induced on the holdout test set is 4.67 years, whereas the MAE induced on the external dataset is 6.71 years; this is a difference of 2.04 years, which is almost half of the difference relative to that of a model trained from scratch. Furthermore, the performance of the SFCN fine-tuned with an age-pretrained model on the holdout test set is 4.96 years, while on the external test set, it is 7.46, which is a difference of 2.5 years. In this case, training the SFCN with an off-the-shelf approach from 3DCAE-MRI yields MAEs of 5.59 years on the holdout test set and 5.93 years on the external test set, producing a difference of 0.34 years. The difference observed across various acquisition settings might explain the higher MAEs attained on the external test sets. Images acquired with different scanners and acquisition parameters might result in different noise levels [29]. The scanner artefacts might not be removed entirely during preprocessing. Therefore, machine learning models might be biased towards the acquisition settings of the training data and might generalise poorly. This bias might be more problematic for models trained with data possessing low diversity with respect to the acquisition settings. In our case, age prediction and sex classification models are trained with data acquired under two acquisition site settings, whereas 3DCAE-MRI is trained with data acquired from 75 different settings. Thus, training via transfer learning from a 3D-CAE with the off-the-shelf approach for age prediction leads to more models for use with external datasets.

The Institute of Psychiatry (IOP) data are used to evaluate the performance of different training strategies on an external test set. Similarly, Leonardsen *et al.* [30] also used these data to assess the performance of the SFCN on an external test set.

Their dataset comprised 53542 instances, 42829 and 10713 of which were used to train and validate the models, respectively. The authors reported an MAE of 6.96 years for the IOP data. In this work, LiteNet and the SFCN, trained with only 300 training samples by using transfer learning from 3DCAE-MRI, achieved higher performances than did the SFCN trained with 42829 instances. LiteNet and the SFCN attained MAEs of 6.71 and 5.93 years, which were improvements of 0.25 and 1.03 years, respectively. This finding highlights that transfer learning from 3DCAE-MRI via off-the-shelf training requires fewer training instances to attain equivalent results to those produced by training models from scratch using thousands of training instances.

In general, reusing the 3DCAE-MRI weights for the two CNNs outperforms the other training strategies for both age prediction and sex classification tasks conducted on the holdout test set. With respect to age prediction, the LiteNet architecture trained via transfer learning from 3DCAE-MRI achieves better results under all the training samples. Compared with training from scratch (with and without augmentation), the off-the-shelf and fine-tuning methods achieve superior performance in most cases. Concerning sex classification, the off-the-shelf approach significantly outperforms the other training strategies for sample sizes less than 200. For dataset consisting of 200 and 300 training samples, the off-the-shelf and PCA-RVM methods have equivalent performances. With respect to the SFCN architecture, for age prediction, the fine-tuning weights acquired from 3DCAE-MRI and from an age-pretrained model are equivalent. For sex prediction, the off-the-shelf method yields greater performance on the holdout dataset than does using age- and sex-pretrained models. Comparing the SFCN with the shallow approach, the SFCN attains higher accuracy for all training sample sizes except for 200, in which the performances of both models are similar.

For sex classification, shallow learning achieves equivalent performance to that of the two CNNs. Training both CNN architectures using off-the-shelf transfer learning from 3DCAE-MRI yields similar performances to that of a shallow learning model. On LiteNet, the shallow approach outperforms 3DCAE-MRI fine-tuning and training the model from scratch (with and without augmentation). On the SFCN, the shallow approach yields superior accuracy to that of the fine-tuned age- or sex-pretrained models. Concerning age prediction, the results show that shallow learning outperforms the CNNs trained from scratch (with and without augmentation) on an external test set. These results are on par with those of previous reports highlighting the fact that DL does not consistently outperform shallow learning for small training set sizes [281, 282].

Regarding stability, the results diverge across LiteNet and the SFCN. On LiteNet, in general, for both sex classification and age prediction, models trained using the off-the-shelf approach are more stable. On the SFCN architecture, the results show that, in general, the pretrained models are more stable in the last training epochs. For the sex classification problem, the sex-pretrained model is more stable, whereas for age prediction, transfer learning from an age-pretrained model yields greater stability. The superior performance of the off-the-shelf model compared to that of other DL-based training strategies might be related to the number of trainable parameters. In this strategy, the weights are frozen and not updated during the training process. The number of parameters reused from 3DCAE-MRI is 1.16 million. LiteNet and the SFCN contain totals of 1.31 and 2.95 parameters, respectively; thus, in the off-the-shelf strategy, only 12% and 61% of the total parameters are trainable compared to those of the other training strategies. These results are in accordance with the results published by Ghafoorian *et al.* [264], in which freezing more layers yielded better white matter hyperintensity detection results. Therefore, the off-the-shelf method might act as a regularisation strategy, preventing overfitting on the training data.

Given the unsupervised nature of 3DCAE-MRI, the models learn intrinsic morphological patterns that are inherent to the brain's structure; thus, their features are agnostic. Our findings suggest that the features extracted by 3DCAE-MRI might be valuable in other neuroimaging contexts with few training instances. Moreover, the weights are transferred to two CNN architectures, highlighting the finding that a single 3DCAE-MRI model might be successfully applicable to multiple CNN architectures. The current results are consistent with those of self-supervised learning strategies, such as Model Genesis. Model Genesis [283], proposed by Zhou *et al.*, is a self-supervised approach that is trained to reconstruct an original image from a transformed version of the image. Akin to our results, Zhou et al. showed that their approach outperformed models trained from scratch and pretrained models on segmentation and classification tasks. Model Genesis and our approach differ in terms of their training strategies. Our approach involves training 3DCAE-MRI to compress and reconstruct the original image from a low-dimensional space. Model Genesis reconstructs original images from their transformations; by doing so, the models learn the spatial relations between different structures and might be able to better learn morphological features. Therefore, our approach is simpler, but even though Model Genesis is more complex, it might learn more robust features and be more generic than our approach.

The present study acknowledges certain limitations. First, our methodology is restricted to the analysis of two specific problems, namely, sex classification and age prediction. A complementary analysis is performed on an AD classification problem, as detailed in the Appendix C. The AD classification results are similar to the sex classification findings. With respect to the LiteNet architecture, transfer learning from the 3DCAE-MRI using an off-the-shelf strategy yields better results than training the network from scratch on either the holdout dataset or the external test set. Regarding the SFCN architecture, in general, equivalent results are attained using a model developed by transfer learning from 3DCAE-MRI or age- or sex-pretrained models. Although the results indicate the potential success of the proposed methodology in other contexts, importantly, its generalisation to different problem domains cannot be guaranteed. Finally, attaining interpretability in DL is challenging. While various methods have been proposed to demystify predictions, a systematic comparison of eight salience map techniques concluded that all methods failed to meet at least one trustworthiness criterion (reproducibility, reliability, localisation utility, and sensitivity to model weights) [160].

6.5 Conclusion

The current work outlines the importance of transfer learning to DL models in MRI studies with small sample sizes. Reusing weights from the 3DCAE-MRI model with an off-the-shelf approach outperforms the other DL-based training strategies and uses fewer training samples. Furthermore, the models are doubtlessly more stable than models trained from scratch. Moreover, the remarkable age and sex prediction results obtained by conducting transfer learning from the same 3DCAE-MRI model suggests the versatility of the proposed approach. 3DCAE-MRI is trained with thousands of multisite images. With respect to age prediction, off-the-shelf transfer learning from 3DCAE-MRI achieves greater performance than does training with thousands of images. Thus, problems with reduced training set sizes might be solved by using transfer learning from 3DCAE-MRI. Nevertheless, DL does not outperform shallow learning in all tasks. Shallow learning has equivalent performance compared to that of off-the-shelf DL during a sex prediction task and a lower MAE during an age prediction task conducted on an external dataset with respect to DL models trained from scratch.

6.6 Methods

This work assessed whether transferring weights from 3DCAE-MRI could yield better performance than training a model from scratch or using pretrained models. To achieve this goal, two CNN architectures were considered: LiteNet and the SFCN. 3DCAE-MRI was trained with images acquired from multiple repositories. The pretrained weights of the encoder were reused by two CNNs: LiteNet and the SFCN. In addition to DL, a shallow pipeline was also considered in this work. Different training dataset sizes were used to demonstrate the potential of the proposed solution. In particular, training set sizes of 25, 50, 100, 200 and 300 samples were applied. All methods were carried out in accordance with the relevant guidelines and regulations.

6.6.1 Data

Eight open data-sharing initiatives were used to train 3DCAE-MRI: ADNI [250], ABIDE I [209], ABIDE II [253], 1000 Functional Connectomes Project

(FCP1000) [251], Brain Genomics Superstruct (GSP) [252], OASIS-1 [254], OASIS-2 [255] and OASIS-3 [256]. All images in the datasets were used to train the autoencoder, and no inclusion or exclusion criteria were applied. In total, the 3DCAE-MRI model was trained with a total of 29478 T1-weighted images from 75 sites. Figure 6.3 depicts the repositories considered to train and validate the 3DCAE-MRI model; detailed information about the repositories and sites is provided in Table C.1 in the Appendix C. For training and evaluating the age and sex models, the Information eXtraction from Images (IXI) [210] dataset was used (Figure 6.3). The IXI dataset comprises MRI images acquired from healthy participants in three different hospitals in London, i.e., Hammersmith Hospital (HH), Guy's Hospital (GH) and the IOP. The GH and HH data were employed for training, validating, and testing the models. However, the IOP data were exclusively utilised to evaluate the performance of each model on an external dataset. Consequently, the data from the IOP site were not considered during the training or validation phases. Detailed information about the number of images used per dataset and the corresponding demographics are presented in Table C.2 of the Appendix C. All the databases considered were derived from open-sharing repositories. Consequently, the data are subject to international data protection regulations. Thus, for each database, all the experimental protocols were approved by an institutional and/or licensing committee. Moreover, informed consent was obtained from all subjects and/or their legal guardian(s).

6.6.2 Image preprocessing

T_1-weighted images were aligned in the anterior and posterior commissures plane using the ATRA toolbox [234] (v2.0). Then, the images were preprocessed using the default preprocessing pipeline (*Segment*) of the computational anatomy toolbox (CAT12) toolbox (http://www.neuro.uni-jena.de/cat/). The CAT12, SPM12 and MATLAB versions used were 1742, v7771 and R2020a (9.8.0.1323502), respectively. The CAT12 was the selected framework due to its low preprocessing time and high reliability [284]. *Segment* is based on a unified segmentation algorithm [203] and includes different steps: image denoising, nonbrain tissue removal, registration and image segmentation. All the images were registered into a template using nonlinear transformations. The images acquired from adult participants were registered to the default CAT12 template, and the images of children were registered to a custom template created with the Cerebromatic toolbox [285]. After the registration step, the images had shapes of 121x145x121. The images were segmented using tissue probability maps (TPM); the adult images were segmented using the default CAT12 TPM, and a personalised template was created for children and adolescents using the TOM8 toolbox [211]. Age prediction was one of the benchmark problems considered in this study. Brain ages seem to be increased in patients with neurode-

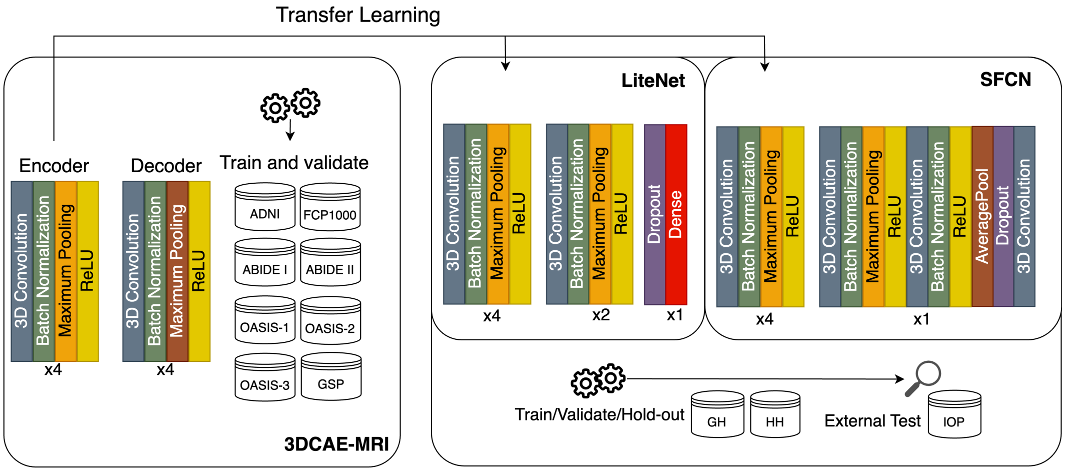


Figure 6.3: Scheme of the proposed methodology. The architectures of 3DCAE-MRI, LiteNet and the SFCN are shown. The 3DCAE-MRI comprised encoder and decoder blocks and was trained and validated on data acquired from eight MRI repositories (ADNI, ABIDE I, ABIDE II, FCP1000, GSP, OASIS-1, OASIS-2, OASIS-3). The arrows represent weight transfers from the encoder to the first four convolution blocks of LiteNet and the SFCN. Both CNN architectures (LiteNet and SFCN) were trained, validated and tested on data from the GH and HH sites, and the IOP site was used to assess the performance achieved on an external test set.

generative diseases, such as AD [9, 162, 230], schizophrenia [9, 229, 270] and multiple sclerosis [9, 180, 271]. The grey grey matter (GM) is severely impacted by these diseases [22, 84, 286, 287]. Thus, given the potential of the BrainAGE biomarker, only GM-segmented images were considered.

6.6.3 Deep learning

6.6.3.1 3DCAE-MRI

3DCAE-MRI comprises an encoder that reduces the input to a low-dimensional space and a decoder that reconstructs the input from a low-dimensional space. The 3DCAE-MRI network architecture is depicted in Figure 6.3. This is a subarchitecture of the SFCN proposed by Peng *et al.* [4]. The SFCN outperforms other popular network architectures, such as ResNet [153] , in brain age prediction tasks. The encoder has four encoder blocks, and each block is composed of a convolution layer, followed by a batch normalisation layer, a maximum pooling layer with a size of two in every dimension and, finally, a rectified linear unit (ReLU) activation layer. The numbers of kernels in the four encoding blocks are 32, 64, 128 and 256, from the first encoding block to the last block. The decoder is identical to the encoder, but the numbers of kernels in different blocks are in the reverse order (256, 128, 64 and 32), and the maximum pooling layer is replaced by an upsampling layer with a size of two in every dimension. Moreover, the decoder has a final inception convolution layer that concatenates the 32 feature maps into a single image, followed by sigmoid activation, which converts the output into zeros and ones. The convolution layers are composed of kernels of size three; the convolution process is performed with a step size of one in every dimension, and zero padding is performed before each convolution block, so the input and output sizes of the convolution layer are the same. Therefore, at the end of each encoding (decoding) block, the input dimensionality is reduced (increased) by two in every dimension. To ensure that the input is accurately recreated, the image shape should be a power of two. Thus, the volumes were reshaped to 128x128x128. 3DCAE-MRI was trained over 25 epochs using the adaptive moment estimation (Adam) optimiser [288] and the mean squared error as the loss function.

6.6.3.2 Convolutional neural network architectures

This work used two CNN architectures: LiteNet and the SFCN. The scheme of each model architecture is shown in Figure 6.3. LiteNet is a custom compact network with a reduced number of parameters; the model is composed of six convolution blocks that are similar to the convolution blocks of the encoder (a convolution layer, batch normalisation, maximum pooling and a ReLU function). The first four convolution blocks are equal to those of the encoder; in the last two blocks, the convolution layers have 64 and 32 filters, respectively, with a kernel size of two in every dimension, but no padding is performed in this case. The convolution blocks are subsequently transformed into a vector, a dropout layer (with a probability of 0.25) and a dense layer with a single neuron. In the classification model, a sigmoid activation function is added after the dense layer. The SFCN was described in detail by Peng *et al.* [4], and the architecture of the CNN is depicted in Figure 6.3. Nevertheless, the SFCN architecture was proposed for age classification rather than age prediction (regression). The network was adapted by removing the softmax layer and replacing the last layer with a single neuron, as proposed by Leonardsen *et al.* [30]. The numbers of parameters used for LiteNet and the SFCN were approximately 1.31 million and 2.95 million, respectively.

The optimisation processes of both networks were identical. The networks were optimised using Adam, and the loss functions were the MAE and the binary-cross entropy losses for the age prediction and sex classification problems, respectively. To demonstrate the applicability of the proposed transfer learning approach for different available data sample sizes, the training sample sizes considered were 25, 50, 100, 200 and 300. The models were trained over 150 epochs for all training sample sizes except for 25 and 50. For the cases with 25 and 50 training instances, the number of epochs was increased to 250. The selected model was the one that attained the lowest loss on the validation set over the last 30 training epochs.

6.6.3.3 Deep learning-based training strategies

Different training strategies were considered in this work: transfer learning from 3DCAE-MRI or pretrained models and training from scratch with and without augmentation. In the transfer learning strategy with 3DCAE-MRI, the weights of the 3DCAE-MRI encoder were transferred to the first four convolution blocks of both the LiteNet and SFCN models (Figure 6.3). Two transfer learning strategies were considered: off-the-shelf training and fine-tuning. In the former, the transferred weights were not updated during training, whereas in the latter, all the weights were updated during training. The number of encoder weights reused in the two architectures was approximately 1.16 million, representing 88% and 39% of the total number of parameters of the LiteNet and SFCN models, respectively. Therefore, in the off-the-shelf approach, in which the weights were not updated during training, the numbers of trainable parameters in LiteNet and the SFCN were 150 thousand and 1.79 million; thus, only 12% and 61% of the total number of weights were trainable in the LiteNet and SFCN models, respectively. Concerning the pretrained models, Peng *et al.* [4] constructed pretrained models for age and sex classification. These models were reused, and the fine-tuning strategy was applied; therefore, all the model weights were updated during training.

Finally, when training from scratch, two scenarios were considered: without and with augmentation. In the former, no transformations were applied to the images; in the latter, during every epoch, two random transformations could be applied: a translation (of 0, 1, or 2 voxels) and a flip (with a probability of 50% across the sagittal plane) [4].

6.6.4 Shallow learning

The age prediction and sex classification performance was assessed for a shallow pipeline. To accomplish this goal, a widely used age prediction framework was considered [117,130,162,289]; this framework combines PCA to encode GM-segmented images with the RVM to compute predictions. The number of components considered influences the performance of the model [130]. This study examined various numbers of components: 2, 10, 15, 20, 25, 30, 40, 50, 75, 100, 150, 200, 250, and 300. Importantly, for a given number of training samples, the number of evaluated components was either equal to or less than the number of training samples. For each training set size, the pipeline selected for evaluation was the one that demonstrated superior performance on the validation set.

6.6.5 Model evaluation

The performances of LiteNet and the SFCN were assessed by training thirty models for each training condition (PCA-RVM, off-the-shelf training, fine-tuning or training from scratch) and training size (25, 50, 100, 200, 300). The HH and GH data from the IXI dataset contained 492 instances, 92 of which were used to validate the model, and 100 instances were used as a holdout set to test the model. The 492 instances were shuffled in each training iteration, and the training, validation and holdout test data were randomly drawn using stratified sampling. The same training, validation and test data were used across the training conditions for a given iteration but were different across various training iterations. The IOP data were used as an external test set to assess the performance of the regression and classification models on an independent test set.

The regression model was evaluated using the MAE, and the computation of the MAE is described in Equation 6.1, where N is the number of instances and y_i and $\hat{y_i}$ are the actual and predicted age values of instance i, respectively.

Accuracy was the metric considered to assess the performance of the classification model. Its calculation formula is described in Equation 6.2, where TN, FN, TP and FP are the numbers of true negatives, false negatives, true positives and false positives, respectively.

$$MAE = \sum_{i=1}^{N} \frac{|y_i - \hat{y_i}|}{N} \qquad (6.1)$$

$$\text{accuracy} = \frac{TN + TP}{TN + FN + TP + FP} \qquad (6.2)$$

6.6.6 Statistical analysis

In total, 30 metric performance values (MAE or accuracy) were produced for both the test and external test sets for each training dataset size (25, 50, 100, 200 and 300 samples) and training condition. The analysis performed to compare the performance and stability of the model across the different training strategies is described below. The significance value considered for the statistical analysis was 0.05.

6.6.6.1 Performance comparison

This work aimed to verify which training strategy yielded better performance results on the holdout and external test sets. Since the same training, validation and test data were used in each iteration, the appropriate statistical test for comparing the performance values across different training strategies was a repeated-measures analysis of variance (rm-ANOVA). If the results were significant, a pairwise Tukey t test66 was performed to verify which groups yielded significant metric performance differences. To evaluate the magnitude of each difference, we used Cohen's d test [290]. Thresholds of 0.2, 0.5, 0.8 and 1.2, which were introduced in [290, 291], were used to assess whether the effect was small, medium, moderate or large, respectively. Cohen's d values above 1.2 were considered very large, as defined by Sawilowsky *et al.* [291].

6.6.6.2 Stability comparison

Model stability was assessed only for the DL models. The utilised stability metric was the standard deviation of the validation set results (MAE or accuracy) over the last 30 training epochs, which was designated in this work as the variation variability. To conduct the stability comparison across different training strategies, rm-ANOVA was carried out to compare the resulting variation variability metrics.

6.7 Code availability

The code and the 3DCAE-MRI model is available at: https://github.com/mfmachado/transfer-learning-age-sex-prediction.

6.8 Data availability

The data used in this work is from the following repositories: ADNI [250], ABIDE I [209], ABIDE II [253], FCP1000 [251], GSP [252] and OASIS-1 [254], OASIS-2 [255], OASIS-3 [256] and IXI [210]; which are open data repositories.

Chapter 7

Unravelling pathological ageing with BrainAGE on Alzheimer's Disease, Diabetes, and Schizophrenia

This chapter is based in the following article:

Dias, Maria de Fátima Machado, João Valente Duarte, Paulo de Carvalho and Miguel Castelo-Branco. "Unravelling pathological ageing with brain age in Alzheimer's Disease, Diabetes, and Schizophrenia." Brain Communications (2023).

7.1 Abstract

Brain age gap estimation (BrainAGE) represents a putative biomarker aiming to detect the transition from healthy to pathological brain ageing. The biomarker primarily models healthy ageing with machine learning models trained with structural magnetic resonance imaging (MRI) data. The difference between predicted brain age and chronological age is expected to translate the deviations in neural ageing trajectory. BrainAGE is increased in multiple pathologies, for instance, in Alzheimer's disease (AD), schizophrenia and type 2 diabetes (T2D). Accelerated ageing seems to be a general feature of neuropathological processes. However, neurobiological constraints remain to be identified to provide specificity to this biomarker. Explainability might be the key to uncovering age predictions and understanding which brain regions lead to an elevated predicted age on a given pathology compared to healthy controls. This is highly relevant to understanding the similarities and differences in neurodegeneration in AD and T2D, which remains an outstanding biological question. Sensitivity maps explain models by computing the importance of each voxel on the final prediction, thereby contributing to the interpretability of deep learning approaches. This paper assesses whether sensitivity maps yield different results across three conditions related to pathological neural ageing: AD, schizophrenia and T2D. Five deep learning models were considered, each model trained with a different MRI data: minimally processed T_1-weighted, grey matter, white matter, cerebrospinal fluid segments and deformation fields (after spatial normalisation). The results outlined an increased BrainAGE in all pathologies, with a different mean, which is the smallest in schizophrenia; this is in line with the observation that neural loss is secondary in this early-onset condition. Importantly, our findings suggest that the sensitivity, indexing regional weights, for all models varies with age. A set of regions were shown to yield statistical differences across conditions. These sensitivity results suggest that mechanisms of neurodegeneration are quite distinct in AD and T2D. For further validation, the sensitivity and the morphometric maps were compared. The findings outlined a high congruence between the sensitivity and morphometry maps for age and clinical group conditions. Our evidence outlines that the biological explanation of model predictions is vital in adding specificity to the BrainAGE and understanding the pathophysiology of chronic conditions affecting the brain.

7.2 Introduction

Ageing is a dynamic biological process intrinsically connected to the natural history of multiple diseases [292, 293]. The biological mechanisms behind the transition from healthy to pathological ageing remain to be uncovered in a broad range of conditions. Pathological ageing may share biological pathways with healthy ageing, suggesting that, in some cases, the former might represent an acceleration of the latter. Nevertheless, pathological ageing might also be characterised by disease-specific mechanisms [19]. The incidence of age-related diseases has been increasing in the last decade. Thus, developing an ageing biomarker that enables healthy ageing and detecting the transition to pathological ageing is paramount. BrainAGE, which corresponds to the difference between predicted brain age and chronological age [21], is such a potential biomarker. BrainAGE is increased in multiple pathological conditions [8], suggesting that its sensitivity is high, although these prior approaches suffer from a lack of specificity. The guidelines of the American Federation for Aging Research state that an ageing biomarker should be able to detect and specify the pathology in its early stages [294]. Thus, adding specificity to the prediction is essential to validate the BrainAGE as an ageing valuable biomarker in clinical practice. In this work, we aim to uncover the brain age models' predictions and verify whether the origin of an elevated BrainAGE is different across three chronic pathologies: AD, schizophrenia and T2D. These diseases are associated to specific periods of lifetime, AD manifests in late adulthood, T2D in mid to later adulthood, and schizophrenia in early adulthood or late adolescence. Furthermore, two morphological hallmarks are reflected in all the three pathologies: higher brain atrophy and ventricle enlargement. Notably, BrainAGE has been shown to reflect the atypical healthy trajectory ageing of the brain in these conditions, in distinct previous studies [117, 119, 121, 131, 137, 170], while biological specificity remained to be uncovered. AD is the most prevalent neurodegenerative disease, and ageing is a major risk factor for developing the disease. The structural brain changes in AD involve a pronounced loss of grey matter (GM) and an exacerbated increase in cerebrospinal fluid (CSF), particularly noticeable in the ventricles [22, 71]. T2D is a metabolic disorder characterised by decreased insulin production and/or increased insulin resistance which impacts glucose level regulation. Uncontrolled glucose levels have been associated with abnormal loss of GM, an increase in the CSF, and changes in the gyrification patterns compared to healthy controls [93, 95, 295]. Both AD and T2D have been associated with the controversial concept of brain insulin resistance, which may result in regional hypometabolism. While the pathogenesis of AD remains to be uncovered, some researchers suggest that insulin resistance prompts the accumulation of both beta-amyloid and tau, therefore suggesting that AD is "the diabetes of the brain" or Type 3 Diabetes (T3D) [106]. Nonetheless, it should be highlighted this account is still highly disputed. Lastly, schizophrenia is a late neurodevelopmental mental disorder characterised by psychotic manifestations. This disease is characterised by atypical neural connectivity both at anatomical and functional levels, as well as an enlargement of the subcortical structures alongside a reduction in cortical volume and thickness [74, 84, 87, 88]. Currently, state-of-the-art models on brain age belong to the category of deep learning, which are considered black-box models [114]. Understanding the contribution of different brain regions on a given prediction is essential to accepting deep learning models in clinical practice, and it might be the key to adding specificity to this putative biomarker - BrainAGE. This study has considered two strategies to address this issue: to predict age locally and to compute the sensitivity maps. Local brain age models predict the age per patch or region rather than using the entire volume [113, 133, 151, 156]. Thus, this strategy assigns an age per region, which enables the identification of accelerated ageing areas. However, the performance of such models is poor compared to global brain age models, which might be caused by insufficient information to derive accurate predictions. Another approach towards explainability is entailed by sensitivity or saliency maps [122,125,154]. Sensitivity maps unveil the influence that each voxel may exert on a prediction. These maps have been explored in the brain age context, and the results are congruent: the regions with a higher contribution to the predictions are located around the ventricles [122,125,154]. Nevertheless, to the best of our knowledge, no study has assessed these maps to discern the differentially elevated regional BrainAGE values in different pathologies. In brief, this work aims to:

1. Compare the BrainAGE in different pathologies across different input image modalities as a tool to understand the respective underlying neurobiology;

2. Assess sensitivity maps across different pathologies to test the hypothesis that biological explainability is distinct - this is critical for the hypothesised relation between T2D and AD;

3. Compare the results obtained with the computation of sensitivity maps with those obtained with the morphological maps as a validation approach.

7.3 Materials and methods

7.3.1 Data

This study considered thirteen datasets: eleven open-source data repositories and two local datasets. Table 7.1 provides a summary of demographic information for each of the datasets used in this study, except for training of the 3D-convolutional autoencoder (3D-CAE), for this case, the demographic information is provided in Table D.1 in the Appendix D. The first step of this methodology is the training of the 3D-CAE. To perform this step, eight of the datasets (autism brain imaging data exchange (ABIDE) I [209], ABIDE II [253], Brain Genomics Superstruct (GSP) [252], Open Access Series of Imaging Studies (OASIS)-1 [254], OASIS-2 [255], OASIS-3 [256], 1000 Functional Connectomes Project (FCP1000) [251], and Alzheimer's Disease Neuroimaging Initiative (ADNI) [250]) were considered, which comprised 29478 images from 75 sites. Two open-source repositories were considered in the model tuning phase: Cambridge Centre for Ageing and Neuroscience dataset (CamCAN) [296, 297] and the Information eXtraction from Images (IXI) [210]. IXI are multi-site repositories with data collected independently from three hospitals in London: Guy's Hospital (GH), Hammersmith Hospital (HH) and Institute of Psychiatry (IOP). Finally, to evaluate the performance of the models on the three diseases of interest in this study, three distinct datasets were employed, one dataset per pathology. For schizophrenia, the open-source repository COBRE dataset was used, whereas local datasets for T2D [298] and AD [287] were considered.

Table 7.1: Demographics of the datasets considered in the model tuning phase and to assess the BAG and sensitivity maps on healthy controls versus clinical group.

Dataset	Total	Number of males	Mean and standard deviation [years]	Min age [years]	Max age [years]	Number of Control
CamCAN	642	318	54.23 ± 18.6	18	88	642
IXI–GH	312	139	50.73 ± 15.98	20.07	86.20	312
IXI–HH	179	85	47.63 ± 16.61	20.17	81.94	179
IXI–IOP	67	24	42.13 ± 16.60	19.98	86.32	67
AD	38	19	66.08 ± 6.66	52	76	18
Diamarker	152	73	54.68 ± 9.61	40	76	82
Cobre	144	107	37.11 ± 12.82	18	65	72

7.3.2 Preprocessing

MRI structural T_1-weighted scans were considered in this study. The images were preprocessed using the computational anatomy toolbox (CAT12) default preprocessing pipeline (Segment) and SPM12 in MATLAB environment, due to the low preprocessing time and high reliability they provide [284]. The CAT12, SPM12 [299], and MATLAB versions used were version 1742, v7771 and R2020a (9.8.0.1323502), respectively. The CAT12 framework requires images to be aligned in the anterior and posterior commissures planes, which was performed using the ATRA toolbox (v2.0. Furthermore, CAT12 preprocessing guidelines recommend that, for images of individuals aged 18 years or less, a personalised template should be created for the registration and segmentation steps. Therefore, these templates were created using the Cerebromatic toolbox [285] and TOM8 toolbox [211], respectively. The default CAT12 template was considered for images of adult participants, whereas childrens' images were registered to a personalised template. The images considered in this work were registered in the MNI space, with the dimensions 121x145x121. Five types of information extracted from the MRI structural images were considered per subject, i.e., the minimally processed images; the T_1-weighted segmented in GM, white matter (WM) and CSF; and the deformation fields resulting from the normalisation (to MNI space) procedure. These different information types, extracted from the same MRI images, will be designated throughout this paper as modalities.

7.3.3 Brain age model tuning

Brain age models were developed leveraging off-the-shelf transfer learning from 3D-CAE, which has been suggested to yield superior results compared to training a convolution neural networks (CNN) from scratch [300]. Therefore, the creation of each brain age model encompasses the training of two models, the 3D-CAE and then the brain age CNN. The architectures considered for the 3D-CAE and regression CNN were the ones considered in a previous study [300]. Regarding the 3D-CAE, as described above, all data from the ABIDE I, ABIDE II, GSP, OASIS-1, OASIS-2, OASIS-3, FCP1000, and ADNI repositories were considered to train and validate the models. All images from these datasets were included, regardless of the condition of the participants, which comprised a total of 29478 images, out of which 250 instances were used to validate and select the best model. The 3D-CAE models were trained for over 50 epochs with a batch size of 16, except for the deformation fields modality. For the deformation fields, the loss of the 3D-CAE model diverged after 20 epochs, requiring an increase in batch size to 56 and an extension of the training to 150 epochs. The autoencoders were optimised using the Mean Squared Error as the loss. The 3D-CAE model selected was the one that exhibited the highest performance on the validation set during the last 30 training epochs. The 3D-CAE encoder weights were reused on the corresponding brain age CNN model. The transferred weights were frozen, and only the layers in which the weights were randomly initialised were updated during training, as described by Dias *et al.* [300]. The Cam-CAN and the IXI repository were considered to train, select, and evaluate the brain age CNN models. The brain age CNN models were trained with 954 images from the Cam-CAN repository and using GH data from the IXI repository. All the brain age models were trained over 150 epochs, and the data from the IOP of the IXI dataset were used to select the best model of the last 30 training epochs. Finally, to correct the model bias of age models, the data from the HH site were used to adjust the bias using the approach proposed by Beheshti *et al.* [129].

7.3.4 BrainAGE on different pathologies

The BrainAGE was investigated per pathology dataset (schizophrenia, T2D and AD) and modality. For each case, the BrainAGE was compared between clinical conditions (disease status versus healthy controls) using analysis of covariance (ANCOVA) controlled for age. Age should be accounted for in the analysis since brain age models often report an age bias, and models tend to underestimate and overestimate the age of young and older subjects, respectively [129]. Despite bias correction being performed, the bias might not be completely removed and, therefore, the age was still accounted for in statistical comparisons. An ANCOVA was performed with and without bias correction.

7.3.5 Sensitivity maps generation

Sensitivity maps were computed using the SmoothGrad approach [159]. Sensitivity maps tend to be noisy, thus to overcome this issue SmoothGrad applies Gaussian noise to the input and computes the corresponding sensitivity maps. The procedure is performed multiple times, and the final sensitivity map is the average of the multiple noisy sensitivity maps. The level of noise that should be added depends on the input type. In this work, five inputs were considered, i.e., minimally processed, GM, WM, CSF, and deformation fields. Thus, the appropriate noise level should be selected for each one. To select the best noise structure, an assumption was made to ensure the one that maximised the correlation between the age and sensitivity maps. To perform this analysis, the data from the HH of the IXI dataset were considered. The sensitivity maps were computed for each subject and parcelled according to the neuromorphometrics atlas. For each region-of-interest (ROI), the mean value was computed. Then, for each ROI, the Pearson correlation was computed between age and the ROI sensitivity. Different noise levels were evaluated for the five input types, with values ranging from 0% to 50%, with a step of 2%.

7.3.6 Sensitivity maps and morphometry on different pathologies

The morphometric and sensitivity maps were assessed individually (stage 1) and compared with each other (stage 2). Both analyses were performed at an ROI level. Thus, the modality images and sensitivity maps were parcelled according to the neuromorphometrics template and for each ROI the mean was computed. Stage 1 aims to understand the relation of the ROI value with age and health condition. In the morphometric case, the analysis portrays the regions that yield significant volume changes with age and clinical conditions. To achieve this an ANCOVA was performed with the ROI value (sensitivity or morphometric measure) as the dependent variable, the clinical condition as the group and age as a covariate. The p-values were corrected for multiple comparisons using the False Discovery Rate [301]. This analysis enables the assessment of the regions in which the sensitivity and/or or the morphometry correlates with age and which regions were sensitive to clinical conditions. Stage 2 aims to assess whether the results from the morphometric and sensitivity maps were congruent. To achieve this the Jaccard coefficient was considered. Jaccard index is the ratio between the number of significant regions in the analyses and the total number of significant regions in both analyses, thus measuring the overlap between the two maps.

7.4 Results

7.4.1 Brain age models performance

The results regarding mean absolute error (MAE) for the IOP and HH data and the correlation between MAE and BrainAGE with and without bias correction are displayed in Table D.2 in the Appendix D. The model selection process was based on the IOP dataset, and the MAE on minimally processed, GM, WM, CSF and deformation fields was 4.66, 5.36, 6.10, 5.61 and 5.48 years, respectively. Subsequently, the model was evaluated on an external dataset, the data from the HH of the IXI dataset. The results, on HH data, yield a lower MAE of 4.46, 5.75, 5.26, 5.44 and 6.53 years for minimally processed, GM, WM, CSF and deformation fields, respectively. Concerning the age bias, the findings reveal that, without any bias correction, a significant correlation exists between BrainAGE and age for all image types in both sites. Regarding the corrected age predictions, the bias is corrected in all modalities on the HH set, this finding is expected since this dataset was the one used for the correlation. In the case of the IOP data, the correlation decreased in all cases and is completely corrected for three modalities (minimally processed, CSF and deformation fields).

7.4.1.1 Schizophrenia

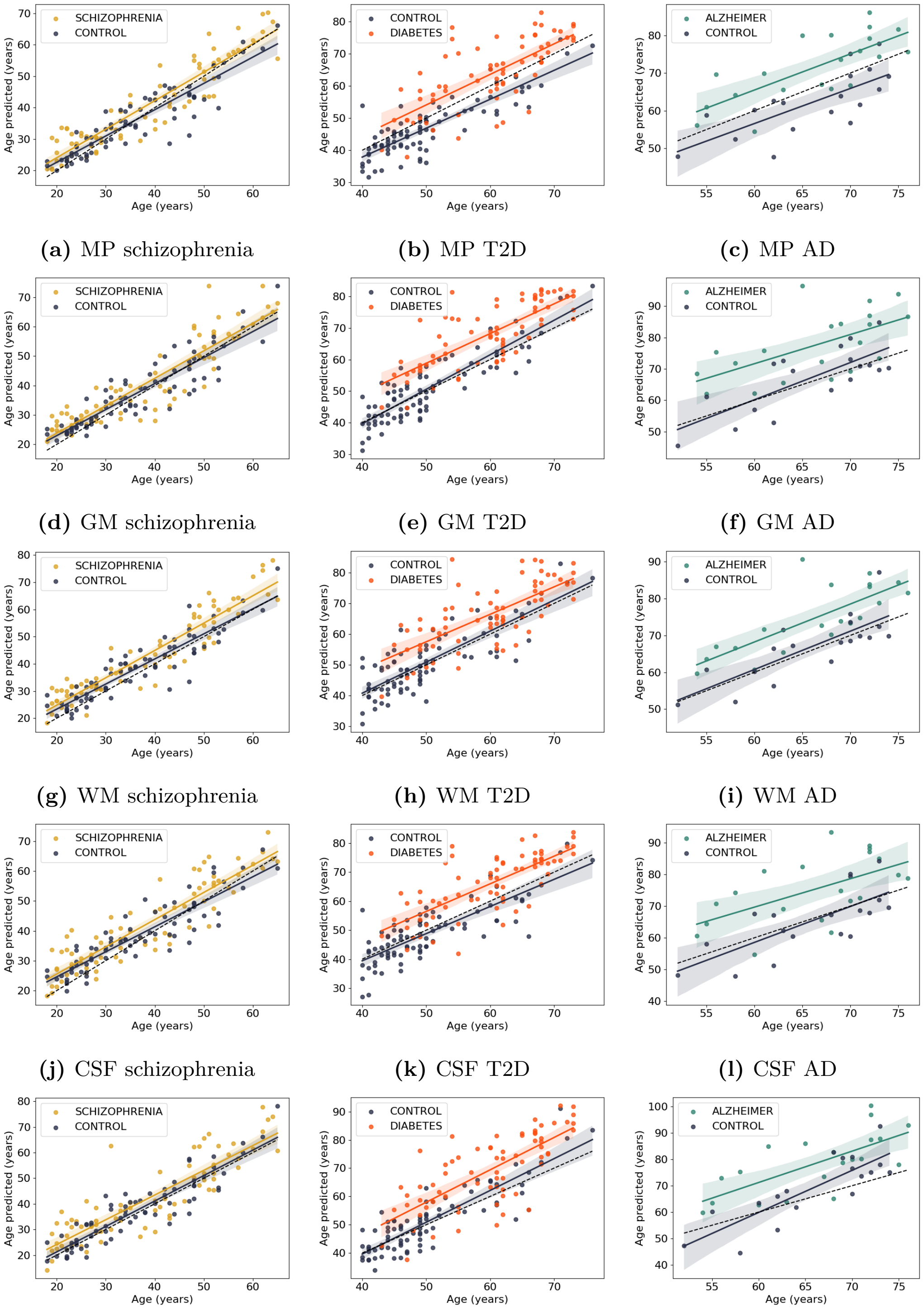
The corrected age prediction results for the schizophrenia and healthy control subjects are presented in Figure 7.1. A summary of the MAE and BrainAGE for different modalities can be found in Table D.3 in the Appendix D. The BrainAGE difference across groups is 2.40, 1.47, 2.82, 1.92, 2.37 years on minimally processed image, GM, WM, CSF and deformation fields, respectively. Thus, in this chronic mental disorder, the BrainAGE for the schizophrenia group is higher than in the control group. The ANCOVA results for the BrainAGE are in Table 7.2. The results indicate the significant BrainAGE on all modalities except GM. Furthermore, the bias on BrainAGE value was significant on minimally processed, GM and WM.

7.4.1.2 Type 2 Diabetes

The age prediction results for the T2D dataset are depicted in Figure 7.1. A summary of the results achieved for the mean MAE and BrainAGE for the health group versus T2D is in Table D.4 in the Appendix D. We found that the predicted age in the T2D group is higher than healthy controls. The mean BrainAGE difference between groups is 6.75, 7.76, 5.59, 7.2 and 8.47 years for minimally processed image, GM, WM, CSF, and deformation fields, respectively. The statistical analysis, for the BrainAGE is in Table 7.2. The results outline that the statistical difference is significant for all modalities. Concerning the age bias, the ANCOVA results exhibited no age bias on BrainAGE.

7.4.1.3 Alzheimer's disease

Figure 7.1 shows the relation between true age and corrected age predictions across various modalities within the AD dataset. Table D.5 in the Appendix D details the mean MAE and BrainAGE for the corrected predictions. The results reveal that, on average, the BrainAGE is higher in AD patients. Specifically, the mean BrainAGE difference between AD patients and healthy controls is 9.04, 9.96, 7.43,



(a) MP schizophrenia **(b)** MP T2D **(c)** MP AD

(d) GM schizophrenia **(e)** GM T2D **(f)** GM AD

(g) WM schizophrenia **(h)** WM T2D **(i)** WM AD

(j) CSF schizophrenia **(k)** CSF T2D **(l)** CSF AD

(m) DF schizophrenia **(n)** DF T2D **(o)** DF AD

Figure 7.1: Brain age predictions. Chronological age versus brain age prediction for the different pathologies considered in this work: schizophrenia, Type 2 Diabetes (T2D) and Alzheimer's Disease (AD). For the different brain age models: minimally processed (MP), grey matter (GM), white matter (WM), Cerebrospinal fluid (CSF) and deformation fields (DF).

9.69, and 9.04 years for minimally processed, GM, WM, CSF, and deformation fields, respectively. Thus, for all modalities except WM, the BrainAGE in AD patients is, on average, 9+ years higher compared to healthy controls. The ANCOVA results in Table 7.2 confirm the significant differences in BrainAGE between the two groups on all modalities. Regarding age bias, the ANCOVA analysis of the BrainAGE suggests no bias in this dataset.

7.4.2 Age

7.4.2.1 Relationship of ROI-based morphometry with age

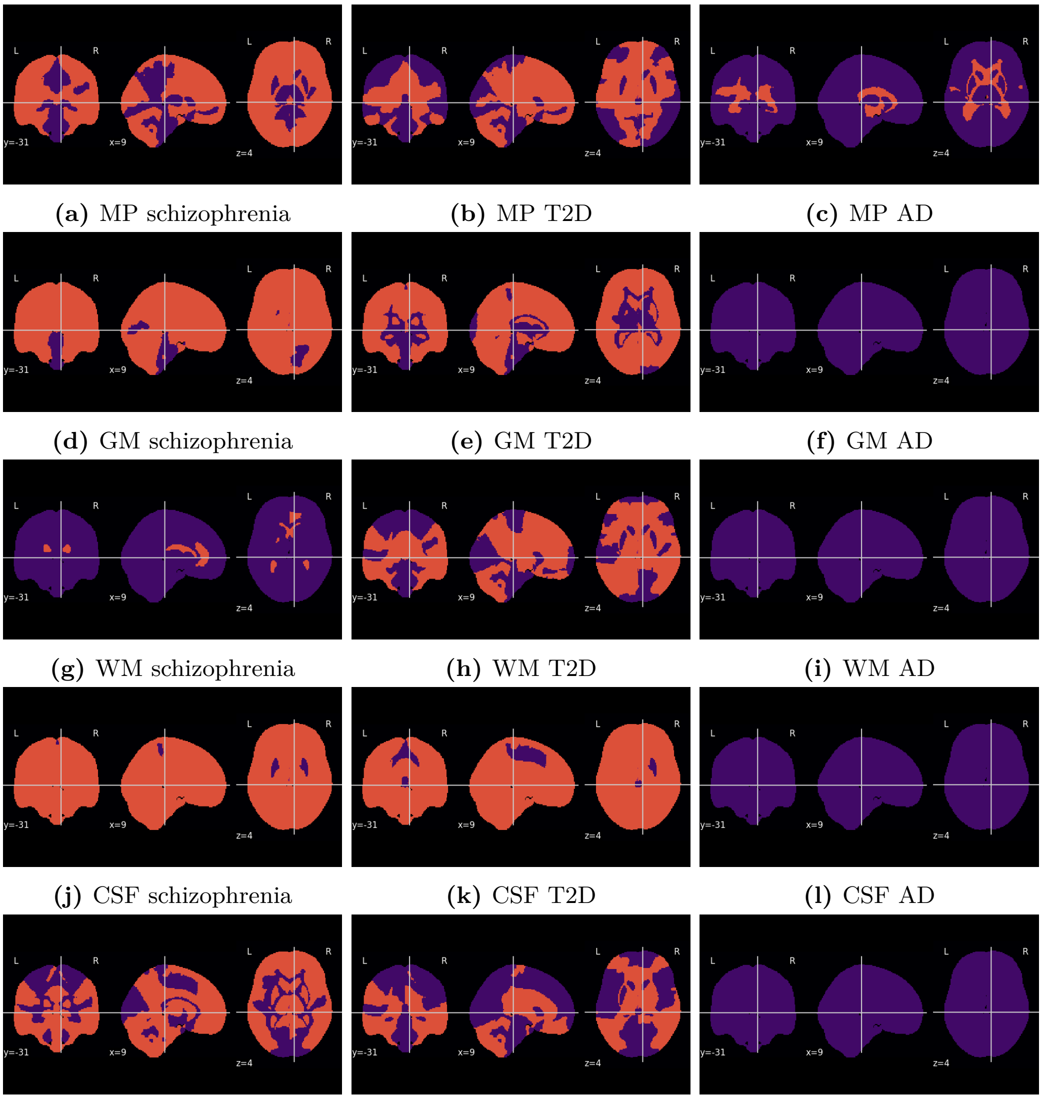
The regions showing significant regression coefficients for the age factor, or, in other words, regions whose value correlates with age in a significant manner are depicted in Figure 7.2. The mean morphometric maps for each dataset are shown in Figure D.1 (in Appendix D). The proportion of significant regions in each modality and dataset is in Table D.6 in the Appendix D. For schizophrenia and T2D, the results indicate that almost all regions are significant for the GM and CSF modalities. On minimally processed modality, between one-third and half of the regions are considered significant, whereas on the deformation fields, around half of the regions yield a significant correlation with age. The WM yield higher morphometry differences across the schizophrenia and T2D cases, while on T2D, only the ventricles are considered to yield volume differences with age, and on the schizophrenia cases, more than half of the regions are considered significant. The AD dataset yields the lowest number of regions significantly correlated with age. The modality with the highest number of significant regions is minimally processed images with 8.57% being significant regions. In all other modalities, no region is considered to be significant. These results might be explained by the narrow age range of the dataset as well as the number of instances in the dataset. The dataset contains only 38 subjects aged between 52 and 76 years. Thus, the statistical power of the data to detect variations might be smaller. Finally, it may also be possible that AD pathology overrides age-related changes.

7.4.2.2 Sensitivity maps on age prediction

The analysis of the relation between sensitivity maps and noise is discussed in Section D.2.3 in the Appendix D. The results reveal that each modality yields the maximum correlation at a different noise level. Thus, the noise level considered was different across modalities and corresponded to the noise yielding the highest correlation value between sensitivity maps and age on the HH data. The mean sensitivity maps for each dataset are presented in Figure D.2 (in Appendix D). The results indicate that regions with higher sensitivity are around the ventricles in all modalities, possibly due to their fast change in shape during ageing. This finding

Table 7.2: ANCOVA results comparing the BrainAGE and controlling for age. An ANCOVA was performed per disease and modality (minimally processed image, GM, WM, CSF and deformation fields) of healthy controls versus subjects diagnosed with schizophrenia, T2D, and AD. The significant values are represented in bold.

	Tissue	**Source**	**SS**	**F**	***p*-value**	**np2**
Schizophrenia	Minimal processed	clinical condition	251.0	10.56	**0.0014**	0.07
		age	350.98	14.77	**0.00018**	0.095
	GM	clinical condition	97.4	3.13	0.079	0.022
		age	172.79	5.55	**0.02**	0.038
	WM	clinical condition	295.76	11.40	**0.00094**	0.075
		age	20.18	0.78	0.38	0.0055
	CSF	clinical condition	169.14	5.79	**0.017**	0.039
		age	338.0	11.58	**0.00087**	0.076
	Deformation fields	clinical condition	209.37	4.99	**0.027**	0.034
		age	13.0	0.30	0.58	0.0022
T2D	Minimal processed	clinical condition	1602.92	42.04	1.24×10^{-9}	0.22
		age	65.32	1.7	0.19	0.011
	GM	clinical condition	1556.45	46.6	2.05×10^{-10}	0.24
		age	8.26	0.25	0.62	0.0017
	WM	clinical condition	1015.0	23.36	3.31×10^{-6}	0.14
		age	16.65	0.38	0.54	0.0026
	CSF	clinical condition	1705.30	41.33	1.64×10^{-9}	0.22
		age	33.57	0.80	0.37	0.0054
	Deformation fields	clinical condition	1473.89	28.53	3.38×10^{-7}	0.16
		age	185.43	3.59	0.06	0.024
AD	Minimal processed	clinical condition	777.54	22.05	4.01×10^{-5}	0.39
		age	4.79	0.14	0.71	0.0039
	GM	clinical condition	935.68	15.47	**0.00038**	0.31
		age	3.48	0.058	0.81	0.0016
	WM	clinical condition	519.94	15.95	**0.00032**	0.31
		age	2.01	0.062	0.81	0.0018
	CSF	clinical condition	888.08	13.59	**0.00076**	0.28
		age	0.010	0.000150	0.99	
	Deformation fields	clinical condition	750.61	11.62	**0.0017**	0.25
		age	229.55	3.55	0.068	0.092



(a) MP schizophrenia **(b)** MP T2D **(c)** MP AD

(d) GM schizophrenia **(e)** GM T2D **(f)** GM AD

(g) WM schizophrenia **(h)** WM T2D **(i)** WM AD

(j) CSF schizophrenia **(k)** CSF T2D **(l)** CSF AD

(m) DF schizophrenia **(n)** DF T2D **(o)** DF AD

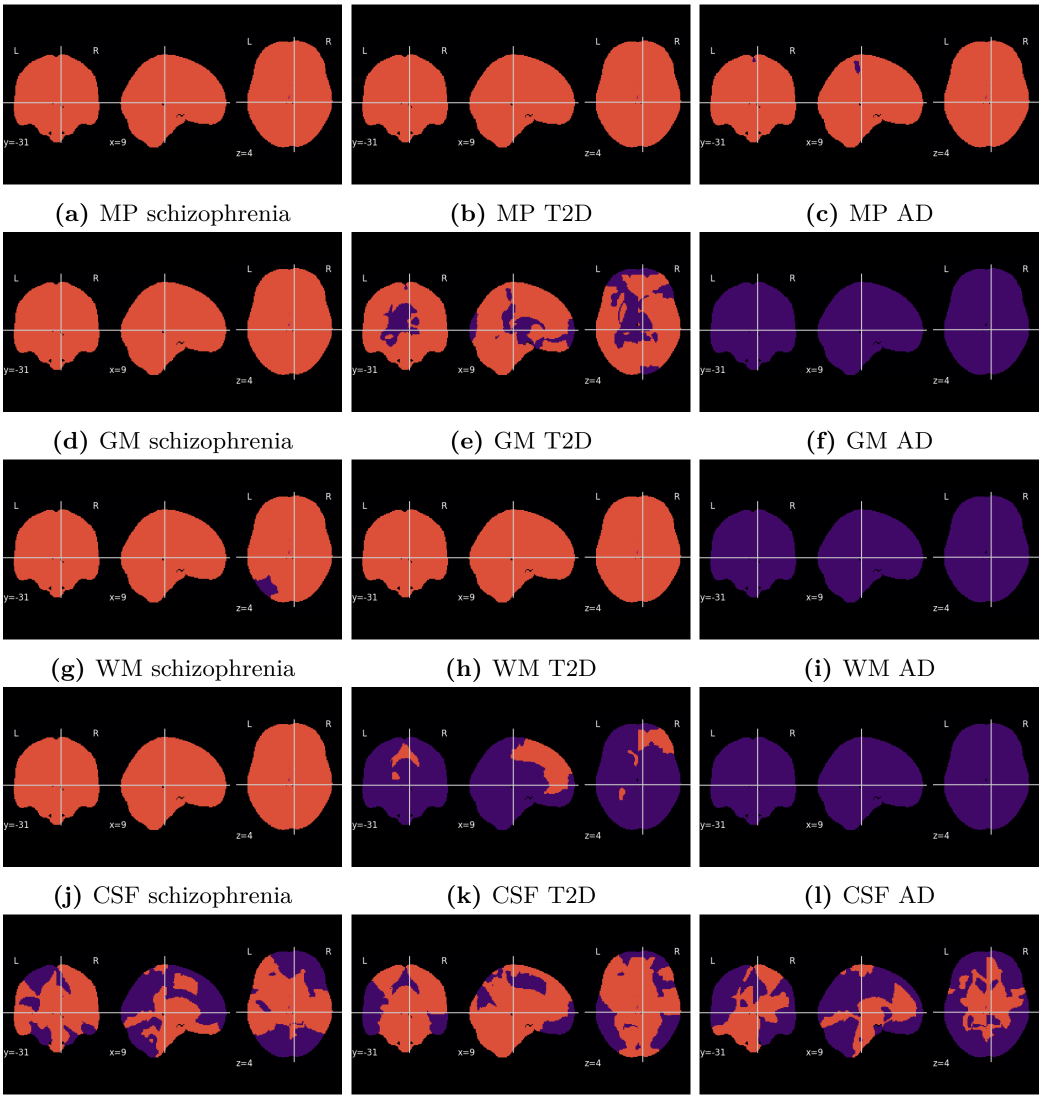
Figure 7.2: Significant region-of-interest (ROI) for age on morphometric maps. Statistically significant ROIs, represented in orange, exhibited a significant p-value for the age factor in an ANCOVA. The ANCOVA compared the morphometric map ROI mean of clinical conditions (health controls versus pathology) and controlling for age. Different morphometric maps assessed were minimally processed (MP), grey matter (GM), white matter (WM), Cerebrospinal fluid (CSF) and deformation fields (DF).

is congruent with other published works [122, 125, 154]. The significant regions concerning the sensitivity for the age factor are shown in Figure 7.3, and the percentage of the significant regions is in Table D.7 in the Appendix D. The results suggest that the sensitivity of a region correlates with age on all regions in the minimally processed image modality, in all datasets except the AD dataset, in which 99.29% of the regions were considered significant. In the case of GM and CSF modalities, all regions were significant for the schizophrenia dataset. On the T2D, the correlation of sensitivity with age was statistically significant in 72.14% and 10.00% of the regions for GM and CSF modalities, respectively. Similarly, to the morphometric results, on the AD, no region is correlated with age on GM, WM and CSF. However, in this case, almost all and 30% of regions are considered significant on minimally processed and deformation fields, respectively. Sensitivity vs morphometry maps on age prediction Some ROIs yield a significant age correlation both on morphometry and sensitivity maps analysis, Table D.10 in the Appendix D shows the Jaccard index between significant regions in both analyses. The findings reveal that the overlap depends upon the dataset and modality. An almost perfect agreement in both analyses is reported on the GM and CSF modalities on the schizophrenia. GM is also the modality with the highest overlap, a Jaccard index of 0.72, on the T2D. Concerning the minimally processed image, the Jaccard of 0.75 and 0.58 is reported for the schizophrenia and T2D cases, respectively. WM has a good agreement comparing the morphometry and sensitivity map results on T2D but not on schizophrenia. Around one-third of the regions are significant in both analyses regarding the deformation fields. Lastly, consistent with the preceding results, the AD dataset yields a reduced overlap between the regions. Nevertheless, it should be highlighted that the agreement is perfect on CSF, WM, and GM, and no region is considered significant in either analysis.

7.4.3 Clinical group comparisons

7.4.3.1 Morphometry

The statistically significant ROIs across groups in the ANCOVA are depicted in Figure 7.4, the percentage of significant regions is detailed in Table D.8 in Appendix D. Compared with the equivalent age results, in general, fewer regions are considered significant for the health condition. The exceptions include the WM on the T2D which yielded more than 10% of significant ROIs for the clinical group condition than for the age. On the AD dataset, concerning the age factor, no region is significant on CSF and deformation fields and on minimally processed only 8.57%, whereas on the clinical group condition factor 30%, 85%, and 30% are considered significant regions, respectively.



(a) MP schizophrenia **(b)** MP T2D **(c)** MP AD

(d) GM schizophrenia **(e)** GM T2D **(f)** GM AD

(g) WM schizophrenia **(h)** WM T2D **(i)** WM AD

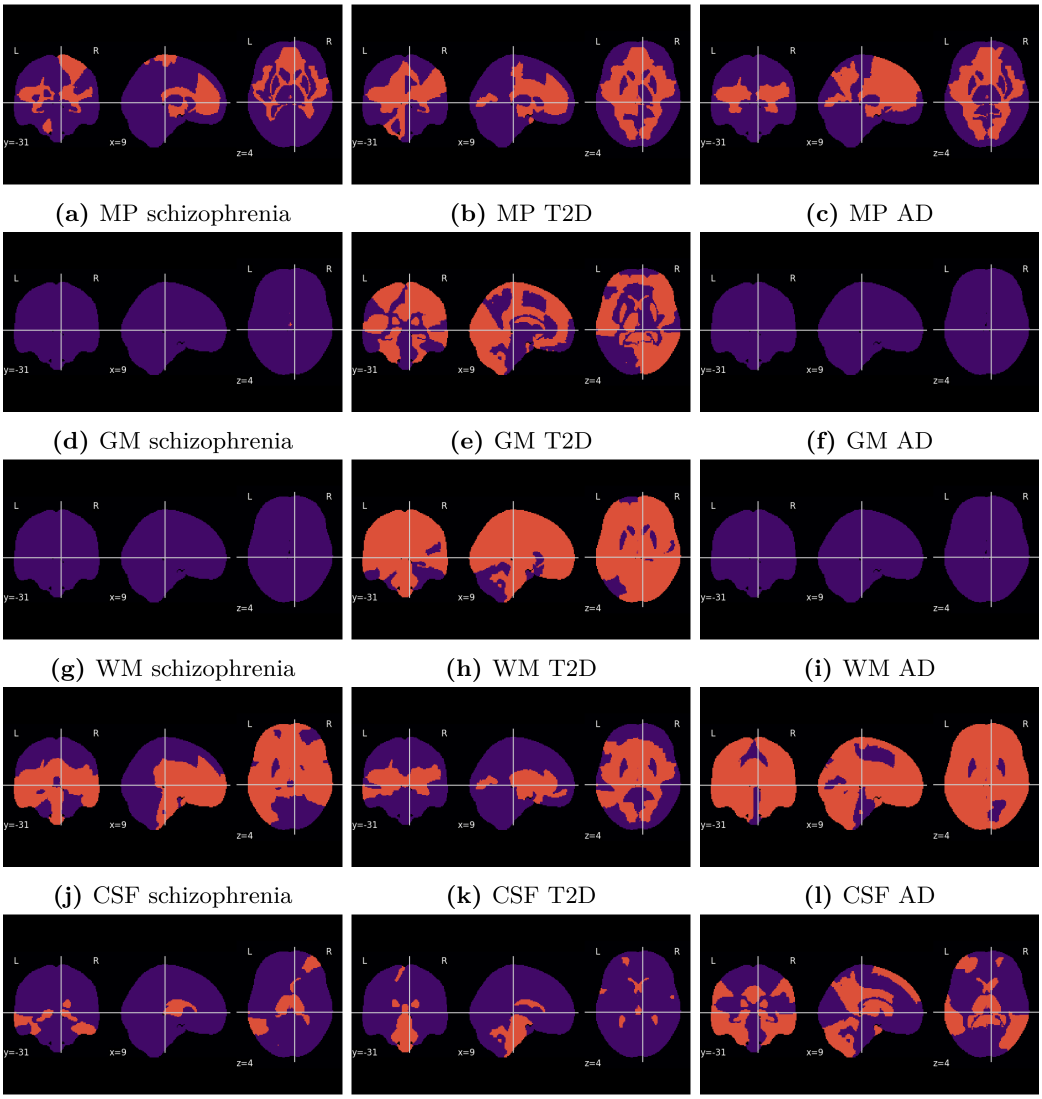
(j) CSF schizophrenia **(k)** CSF T2D **(l)** CSF AD

(m) DF schizophrenia **(n)** DF T2D **(o)** DF AD

Figure 7.3: Significant region-of-interest (ROI) for age on sensitivity maps. Statistically significant ROIs, represented in orange, exhibited a significant p-value for age factor in an ANCOVA. The ANCOVA compared the morphometric map ROI mean of clinical conditions (health controls versus pathology) and controlling for age. Different sensitivity maps derived from brain-age models trained with minimally processed (MP), grey matter (GM), white matter (WM), Cerebrospinal fluid (CSF) and deformation fields (DF) were assessed.

7.4.3.2 Sensitivity maps

Regions whose sensitivity varies with the condition are depicted in Figure 7.5 for each dataset, whereas Table D.9 in Appendix D summarises the percentage of significant regions. The analysis suggests that clinical conditions significantly influence sensitivity maps. Minimally processed consistently has a high number of significant regions on all datasets; the percentage of significant regions is 80%, 97.86%, and 69.29% for schizophrenia, T2D, and AD, respectively. Consequently, the overlap between the significant regions on this modality is also high. Concerning the



(a) MP schizophrenia **(b)** MP T2D **(c)** MP AD

(d) GM schizophrenia **(e)** GM T2D **(f)** GM AD

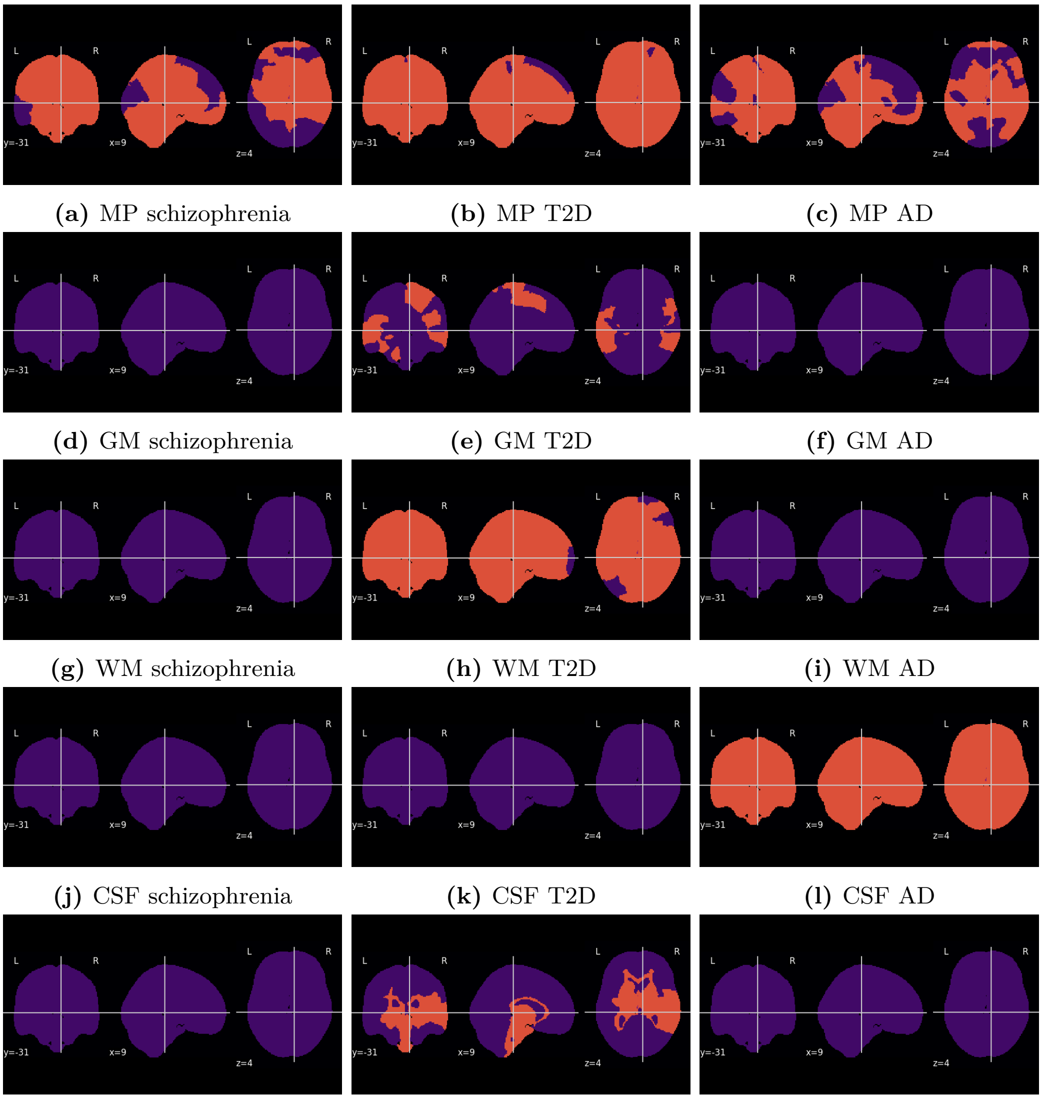
(g) WM schizophrenia **(h)** WM T2D **(i)** WM AD

(j) CSF schizophrenia **(k)** CSF T2D **(l)** CSF AD

(m) DF schizophrenia **(n)** DF T2D **(o)** DF AD

Figure 7.4: Significant region-of-interest (ROI) for clinical condition on morphometric maps. Statistically significant ROIs, represented in orange, exhibited a significant p-value for the clinical condition factor in an ANCOVA. The ANCOVA compared the morphometric map ROI mean of clinical conditions (health controls versus pathology) and controlling for age. Different morphometric maps assessed were minimally processed (MP), grey matter (GM), white matter (WM), Cerebrospinal fluid (CSF) and deformation fields (DF).

other modalities, the results across pathologies evidence much larger dissimilarities. GM modality only yields significant differences comparing healthy controls with T2D; in this case, 18.57% of regions are considered significant, encompassing the cortical ribbon. Regarding WM, almost all regions (94.29%) demonstrate sensitivity differences between healthy controls and the T2D group. In contrast, no WM region is considered significant when comparing schizophrenia or AD with healthy controls. The CSF results suggest that all regions have a significant role in predicting a higher age when comparing the results of the AD group with healthy controls. Nevertheless, comparing the sensitivity of T2D or schizophrenia with healthy controls no significant regions are significant for CSF. Lastly, the deformation fields modality yields significant differences only in sensitivity comparing T2D with healthy controls.



(a) MP schizophrenia **(b)** MP T2D **(c)** MP AD

(d) GM schizophrenia **(e)** GM T2D **(f)** GM AD

(g) WM schizophrenia **(h)** WM T2D **(i)** WM AD

(j) CSF schizophrenia **(k)** CSF T2D **(l)** CSF AD

(m) DF schizophrenia **(n)** DF T2D **(o)** DF AD

Figure 7.5: Significant region-of-interest (ROI) for clinical condition on sensitivity maps. Statistically significant ROIs, represented in orange, exhibited a significant p-value for the clinical condition factor in an ANCOVA. The ANCOVA compared the sensitivity map ROI mean of clinical conditions (health controls versus pathology) and controlling for age. Different sensitivity maps derived from brain-age models trained with minimally processed (MP), grey matter (GM), white matter (WM), Cerebrospinal fluid (CSF) and deformation fields (DF) were assessed.

7.4.3.3 Morphometric results versus sensitivity maps on pathologies

The overlap between morphometric and sensitivity maps analyses is shown in Table D.10 in Appendix D. In general, the findings outline a low agreement between the morphometry and sensitivity maps, with some exceptions. The two highest Jaccard scores are the CSF and WM on the AD and T2D, respectively. Notably, both analyses have a high number of significant regions. The minimally processed images yield a similar overlap score across all datasets, around one-third of the significant ROIs on both analyses. Moreover, although the Jaccard index is zero on the AD dataset on GM and WM and the WM on the schizophrenia dataset, in both analyses, no region is considered significant; consequently, the match is perfect. Similarly, on the GM on the schizophrenia dataset, only 1.43% of the regions are considered significant, while on the sensitivity maps, no region yielded significant results; thus, in this case, the match is almost perfect.

7.5 Discussion

The main finding of this article is that the explanation of brain age predictions, based on sensitivity maps, allows the identification of regional specificity of BrainAGE across pathologies. Furthermore, sensitivity maps provide a pathophysiological differentiation between AD and T2D. BrainAGE is significant for all the three pathologies considered (AD, schizophrenia and T2D) compared to healthy controls, yet the mean BrainAGE is different across pathologies. AD yields the highest mean BrainAGE (around 9 years), followed by T2D (around 5 years) and finally by schizophrenia (around 2 years). This result might be explained by the degree of structural changes in each one of the pathologies. Although no prior existing studies compare the structural changes of the three diseases, our data are consistent with the notion that schizophrenia has less direct neural loss compared to T2D, and T2D has, in turn, less structural changes and distinct regional pathology as compared to AD. AD and T2D are characterised by brain neurodegeneration [22, 93, 95], which is not the case for schizophrenia [77, 83]. This hypothesis is corroborated by the morphometry analysis conducted in this study, the results outline that T2D yields, on average, more significant ROIs than schizophrenia when comparing the ROIs of the segmented images. Therefore, the BrainAGE might reflect the degree of pathological ageing of the brain. The morphometric and sensitivity maps yield congruent results on the regions that are age-sensitive. The morphometric results outline that for all modalities (except the WM on the schizophrenia) more than 50% of the regions have significant changes with age on the T2D and schizophrenia. The sensitivity maps observe the same tendency (except for the CSF on the T2D). The overlap between the explainability and morphometry significant regions for age is almost perfect in some cases, for instance, on GM and CSF modalities of schizophrenia, and in other cases, it is around half or more of the overlap between the regions. Furthermore, on the AD dataset, no region is considered significant on morphometric and sensitivity maps on the GM, WM and CSF, which is also congruent.

Morphometry results endorse sensitivity maps concerning the regions with different importance across health conditions. On WM tissue, almost all the regions seem to exert a different importance on the prediction comparing the T2D group with the healthy group, a similar finding is outlined with the CSF when analysing the sensitivity differences between AD and healthy controls. Despite this intriguing result, the same trend is observed in the morphometric analysis, in which most regions are also considered significant in both cases. Furthermore, the overlap between the significant regions on morphometry versus explainability is high in both cases, specifically, the Jaccard index is 0.74 and 0.85 for WM and the CSF, respectively. Moreover, there is also a high agreement concerning the non-significant regions, as no region was considered significant for either GM and WM modalities when comparing the healthy group with the schizophrenia group and when comparing healthy controls with AD patients. In conclusion, the high agreement between the morphometric and sensitivity maps, validate the results of sensitivity analysis maps.

BrainAGE encodes disease specificity patterns and sensitivity maps may disclose structural and morphological differences driven by pathological ageing. Brain-age models were trained to tackle the healthy ageing process, and the predicted age of these models yielded statistical differences when comparing healthy controls and clinical groups across three diseases. This finding corroborates the hypothesis that diseases might cause an acceleration of the ageing process [19]. Sensitivity maps reveal the region's influence on a prediction, which is different across pathologies and modalities. The comparison of T2D with healthy controls reveals that almost all WM regions exerted different influences on both groups. This result suggests that T2D might be a diffuse pathology in WM, which is consistent with the pathophysiology of T2D, which causes generalised dysfunction of the endothelium and vascular damage [91]. Regarding AD, the sensitivity results evidence that all regions were considered significant regarding the CSF modality, but no region is considered significant on GM, WM, and deformation field modalities. These findings suggest that the model explanations on AD and T2D were distinct, suggesting that the pathophysiology of both conditions is quite distinct.

Sensitivity maps yield complementary information to morphometry maps. Despite the high agreement between the two approaches, there are also some differences in particular around ventricle regions. Multiple models retain complementary information to decode the pathology from the age prediction. The minimally processed yields better performance and generalisation. Nevertheless, its specificity to detect disease processes is reduced. The results reveal that the pattern is more dissimilar on the other modalities than on minimally processed images. Therefore, the minimally processed model can be used to obtain an accurate measure of the predicted age, yet other modalities might be essential to specify disease mechanisms.

7.6 Conclusions

This work highlights the potential of sensitivity maps to uncover the pathological ageing. The results reveal a high agreement between the morphometry and the sensitivity maps, which validates the sensitivity maps as a decoding tool. Furthermore, sensitivity maps yielded distinct patterns across different brain pathologies, highlighting that those predictions encode disease-specific information, and sensitivity maps might be the key to adding specificity to the BrainAGE biomarker. Finally, sensitivity maps can also be used as a complementary strategy to comprehend the biological mechanisms of age-related diseases.

Chapter 8

General Discussion and Future Work

This chapter begins with a section highlighting the main contributions of this thesis. Succinctly, the results outline that:

- Computational anatomy toolbox (CAT12) seems more reliable than Freesurfer;
- Deformation fields contain valuable information to predict brain age;
- Transfer learning from 3D-convolutional autoencoder (3D-CAE) improves the brain age model generalisability;
- Explainability of brain age prediction seems a promising avenue to increase the brain age gap estimation (BrainAGE) specificity.

The chapter closes with a section discussing possible future directions on BrainAGE field.

8.1 Main contributions

This work aimed to address two shortcomings in BrainAGE: generalisability and specificity.

Concerning generalisability, two studies were conducted at the preprocessing and modelling level. FreeSurfer and CAT12, two widely used preprocessing frameworks in BrainAGE, were compared regarding reproducibility and reliability in chapter 4. The study assessed the reproducibility between the frameworks and the reliability of each framework, comparing the cortical thickness measurements. The findings showed that reproducibility was dependent on the acquisition settings and was lower in children and adolescents than in adults. In terms of reliability, the results evidenced that CAT12 yields higher reliability than FreeSurfer. Therefore, in the subsequent studies, CAT12 was considered to preprocess data. The generalisability was also improved at the modelling level by leveraging transfer learning. The performance of deep learning models decreases when the models are applied to unseen data acquired in different acquisition settings. One of the reasons for this outcome might be model overfitting, i.e. the model might learn intrinsic scanner patterns that are in the training data. Therefore, transfer learning from a self-supervised 3D-CAE was assessed in chapter 6. The 3D-CAE was trained on data from multiple repositories. In total, data acquired in 75 different acquisition settings was considered. The transfer learning using an off-the-shelf strategy leads to higher performance on an external test set compared to training a model from scratch or using age-pretrained models.

BrainAGE is increased in multiple pathological conditions, as described in section 3.4. Thus, the ageing biomarker is sensitive to pathological ageing but lacks specificity. Two studies were performed to address this issue. The first one, described in chapter 5, assessed the predictive power of deformation fields comparatively to other T_1 byproducts. The result outlines that deformation fields yield better performance than white matter (WM) and cerebrospinal fluid (CSF) and are equivalent to grey matter (GM). Furthermore, combining GM and deformation fields at the fusion level yields better performance than solely GM. This study outlines the importance of including deformation fields in brain age models; this feature retains complementary brain morphology that might improve the model's performance. Furthermore, different types of information might play a vital role in increasing the specificity of this hypothetical biomarker. Finally, the last study, introduced in chapter 7, evaluated whether explainability could improve the BrainAGE sensitivity maps. Three diseases with distinct age dependence were assessed: Alzheimer's disease (AD), type 2 diabetes (T2D) and schizophrenia. The results outline that the regions that drive a BrainAGE increase differ across diseases and byproducts of T_1-weighted images. These findings are consistent with distinct pathophysiological processes affecting brain ageing in these conditions. Therefore, based on these disease-specific signatures, multimodal sensitivity maps might help in the differential diagnosis.

8.2 Future Research Directions

The preprocessing level impacts the model generalisability [3]. In this thesis, the preprocessing pipeline was selected using the reliability of cortical thickness. Nevertheless, this metric does not assess whether preprocessing improved the generalisability of the brain age model across different acquisition settings. Moreover, only FreeSurfer and CAT12 were compared. Therefore, further investigation should be performed at the preprocessing level. State-of-the-art harmonising techniques could improve generalisability. Multiple harmonisation algorithms have been proposed to either uniformise features or images [302], and feature harmonisation has been suggested to improve the performance of brain age models [303]. Moreover, the differences in images caused by different acquisition settings could be mitigated by providing more strict standard operating procedures also at the hardware acquisition level [148]. Thus, harmonisation at the imaging level and acquisition standardisation might result in better generalisation across different acquisition settings.

The results of this work outlined that transfer learning from 3D-CAE improve the model generalisability. Nevertheless, the brain age models evaluated with this training strategy were trained with solely GM, using a reduced dataset (300 training instances) and a single external test set. State-of-the-art brain age models use thousands of images. Therefore, transfer learning from 3D-CAE should be evaluated on a larger sample size, more external test sets and other image types. Furthermore, self-supervised models, such as Model Genesis [283], should be assessed. The 3D-CAE learns magnetic resonance imaging (MRI) patterns by a dimensionality reduction approach. Model Genesis has an equivalent approach, but rather than minimising the difference between the input and output, it applies transformations to the images and tries to reconstruct the original image from the transformed image. The authors reported a performance increase on multiple segmentation problems [283]. Therefore, transferring learning to the brain age model from self-supervised models could improve the model's generalisability.

On the specificity axis, this thesis focuses solely on T_1-weighted images. Integrating multiple sources of information could help in the differential diagnosis. In chapter 5, it is reported that the fusion of various types of information improves the model performance. Additionally, in chapter 7, it is outlined different image types yield a different pattern across pathologies. Currently, most studies focus on predicting brain age using solely structural imaging. Thus, aggregating other MRI modalities might help improve the model's performance and increase its specificity. For instance, schizophrenia is characterised by abnormal structural and functional connectivity [87, 88]. Therefore, including functional MRI, effective connectivity metrics, or diffusion tensor imaging on brain age models could increase the BrainAGE specificity and be helpful for the disease prognosis. MRI images, that are within the scope of radiomics, only capture macro-level brain ageing changes. Nonetheless, brain ageing is a complex biological process which affects multiple systems [19]. Therefore, adding other types of sources of information, such as proteomics and genomics, could improve the model's sensitivity. For instance, the combination of structural imaging with the dynamic profile of glucose levels could also be proven valuable in T2D to predict the conversion to dementia. Nonetheless, effectively combining multiple types of information can be challenging. Currently, there is no comprehensive comparison of fusion strategies. Thus, a systematic analysis of data fusion should also be addressed. Finally, the findings of this work show that sensitivity maps help to explain the prediction and might be relevant for the BrainAGE specificity. Combining local brain age prediction with sensitivity maps seems a promising candidate in increasing the explainability of brain age models [113]. Therefore, further investigation into combining these two strategies across different diseases could provide valuable information.

References

[1] Yana Blinkouskaya, Andreia Caçoilo, Trisha Gollamudi, Shima Jalalian, and Johannes Weickenmeier. **Brain aging mechanisms with mechanical manifestations**. *Mechanisms of Ageing and Development*, **200**:111575, 2021. xxv, 7, 8, 9

[2] John P. Ridgway. **Cardiovascular magnetic resonance physics for clinicians: part I**. *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, **12**, 2010. xxv, 14, 15, 16, 17

[3] James H. Cole, Rudra P.K. Poudel, Dimosthenis Tsagkrasoulis, Matthan W.A. Caan, et al. **Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker**. *NeuroImage*, **163**:115 – 124, 2017. xxvi, 21, 26, 30, 32, 33, 34, 35, 38, 39, 136

[4] Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. **Accurate brain age prediction with lightweight deep neural networks**. *Medical image analysis*, **68**, 2021. xxvi, 21, 33, 34, 35, 36, 38, 39, 94, 110, 111, 112

[5] Theo Vos, Abraham D. Flaxman, Mohsen Naghavi, Rafael Lozano, et al. **Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010**. *The Lancet*, **380**(9859):2163–2196, 2012. 1

[6] Emma Nichols, Jaimie D. Steinmetz, Stein Emil Vollset, Kai Fukutaki, et al. **Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019**. *The Lancet Public Health*, **7**:e105–e125, 2022. 1

[7] Albert T. Higgins-Chen, Kyra L. Thrush, and Morgan E. Levine. **Aging Biomarkers and the Brain**. *Seminars in cell & developmental biology*, **116**:180, 2021. 1

[8] Cole JH and Franke K. **Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers**. *Trends in neurosciences*, **40**(12):681–690, 2017. 1, 77, 88, 117

[9] Lea Baecker, Rafael Garcia-Dias, Sandra Vieira, Cristina Scarpazza, and Andrea Mechelli. **Machine learning for brain age prediction: Introduction to methods and clinical applications**. *eBioMedicine*, **72**:103600, 2021. 1, 2, 3, 21, 24, 26, 27, 29, 30, 32, 38, 93, 94, 110

[10] Geoffrey McNicoll. **World Population Ageing 1950-2050.** *Population and development Review*, **28**(4):814–816, 2002. 1

[11] Ewa Rudnicka, Paulina Napierała, Agnieszka Podfigurna, Błażej Męczekalski, et al. **The World Health Organization (WHO) approach to healthy ageing**. *Maturitas*, **139**:6, 2020. 1

[12] Norton W. Milgram, Christina T. Siwak-Tapp, Joseph Araujo, and Elizabeth Head. **Neuroprotective effects of cognitive enrichment**. *Ageing Research Reviews*, **5**:354–369, 2006. 1

[13] J. Eric Ahlskog, Yonas E. Geda, Neill R. Graff-Radford, and Ronald C. Petersen. **Physical Exercise as a Preventive or Disease-Modifying Treatment of Dementia and Brain Aging**. *Mayo Clinic Proceedings*, **86**:876–884, 2011. 1

[14] Miranka Wirth, Claudia M. Haase, Sylvia Villeneuve, Jacob Vogel, and William J. Jagust. **Neuroprotective pathways: lifestyle activity, brain pathology, and cognition in cognitively normal older adults**. *Neurobiology of Aging*, **35**:1873–1882, 2014. 1

[15] Chantal Villemure, Marta Čeko, Valerie A. Cotton, and M. Catherine Bushnell. **Neuroprotective effects of yoga practice: Age-, experience-, and frequency-dependent plasticity**. *Frontiers in Human Neuroscience*, **9**:136221, 2015. 1

[16] Gary E. Swan and Christina N. Lessov-Schlaggar. **The effects of tobacco smoke and nicotine on cognition and the brain**. *Neuropsychology Review*, **17**:259–273, 2007. 1

[17] Peter W. Vik, Tony Cellucci, Amy Jarchow, and Jill Hedt. **Cognitive impairment in substance abuse**. *Psychiatric Clinics of North America*, **27**:97–109, 2004. 1

[18] Robert L. Spencer and Kent E. Hutchison. **Alcohol, Aging, and the Stress Response**. *Alcohol Research & Health*, **23**:272, 1999. 1

[19] Menglong Jin and Shi Qing Cai. **Mechanisms Underlying Brain Aging Under Normal and Pathological Conditions**. *Neuroscience bulletin*, **39**:303–314, 2023. 1, 7, 9, 117, 132, 137

[20] Hedieh Sajedi and Nastaran Pardakhti. **Age Prediction Based on Brain MRI Image: A Survey**. *Journal of medical systems*, **43**(8), 2019. 2, 3, 19, 20

[21] Katja Franke and Christian Gaser. **Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained?** *Frontiers in Neurology*, **0**(JUL):789, 2019. 2, 19, 54, 76, 77, 93, 117

[22] Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. **The clinical use of structural MRI in Alzheimer disease**. *Nature reviews. Neurology*, **6**(2):67–77, 2010. 2, 11, 12, 110, 117, 131

[23] Jason Steffener, Christian Habeck, Deirdre O'Shea, Qolamreza Razlighi, et al. **Differences between chronological and brain age are related to education and self-reported physical activity**. *Neurobiology of aging*, **40**:138–144, 2016. 2

[24] Anna Maria Matziorinis, Christian Gaser, and Stefan Koelsch. **Is musical engagement enough to keep the brain young?** *Brain Structure and Function*, **228**:577–588, 2023. 2

[25] Eileen Luders, Nicolas Cherbuin, and Christian Gaser. **Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners**. *NeuroImage*, 2016. 2

[26] Zeqiang Linli, Jianfeng Feng, Wei Zhao, and Shuixia Guo. **Associations between smoking and accelerated brain ageing**. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, **113**:110471, 2022. 2

[27] Kaida Ning, Lu Zhao, Will Matloff, Fengzhu Sun, and Arthur W. Toga. **Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants**. *Scientific Reports 2020 10:1*, **10**:1–10, 2020. 2

[28] Franziskus Liem, Gaël Varoquaux, Jana Kynast, Frauke Beyer, et al. **Predicting brain-age from multimodal imaging data captures cognitive impairment**. *NeuroImage*, **148**:179 – 188, 2017. 2, 21, 24, 25, 27, 32, 38, 94, 105

[29] B. A. Jonsson, G. Bjornsdottir, T. E. Thorgeirsson, L. M. Ellingsen, et al. **Brain age prediction using deep learning uncovers**

associated sequence variants. *Nature Communications 2019 10:1*, **10**(1):1–10, 2019. 2, 21, 24, 26, 30, 32, 33, 34, 35, 38, 39, 78, 94, 105

[30] Esten H. Leonardsen, Han Peng, Tobias Kaufmann, Ingrid Agartz, et al. **Deep neural networks learn general and clinically relevant representations of the ageing brain**. *NeuroImage*, **256**, 2022. 2, 33, 34, 38, 39, 94, 105, 111

[31] Jian Zhai and Ke Li. **Predicting brain age based on spatial and temporal features of human brain functional networks**. *Frontiers in Human Neuroscience*, 2019. 2, 94, 105

[32] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. **Deep learning**, 2015. 3, 21, 32, 92

[33] Krzysztof Gorgolewski, Christopher D. Burns, Cindee Madison, Dav Clark, et al. **Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python**. *Frontiers in neuroinformatics*, **5**, 2011. 4

[34] Juan Eugenio Iglesias, Cheng Yi Liu, Paul M. Thompson, and Zhuowen Tu. **Robust brain extraction across datasets and comparison with publicly available methods**. *IEEE Transactions on Medical Imaging*, 2011. 4

[35] Anna M. Hedman, Neeltje E.M. van Haren, Hugo G. Schnack, René S. Kahn, and Hilleke E. Hulshoff Pol. **Human brain changes across the life span: a review of 56 longitudinal magnetic resonance imaging studies**. *Human brain mapping*, **33**:1987–2002, 2012. 7

[36] Edmund T. Rolls. **Cerebral Cortex: Principles of Operation**. *Cerebral Cortex*, 2016. 8

[37] Mark G. Packard and Barbara J. Knowlton. **Learning and Memory Functions of the Basal Ganglia**. *https://doi.org/10.1146/annurev.neuro.25.112701.142937*, **25**:563–593, 2003. 8

[38] Elizabeth A. Phelps. **Emotion and Cognition: Insights from Studies of the Human Amygdala**. *https://doi.org/10.1146/annurev.psych.56.091103.070234*, **57**:27–53, 2005. 8

[39] Fjell AM and Walhovd KB. **Structural brain changes in aging: courses, causes and cognitive consequences**. *Reviews in the neurosciences*, **21**(3):187–221, 2010. 8, 76

[40] Karen M. Rodrigue and Kristen M. Kennedy. **The Cognitive Consequences of Structural Changes to the Aging Brain**. *Handbook of the Psychology of Aging*, pages 73–91, 2011. 8

[41] Susan M Resnick, Dzung L Pham, Michael A Kraut, Alan B Zonderman, and Christos Davatzikos. **Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain**. *Journal of Neuroscience*, **23**(8):3295–3301, 2003. 8, 9, 37

[42] Maryam E. Rettmann, Michael A. Kraut, Jerry L. Prince, and Susan M. Resnick. **Cross-sectional and longitudinal analyses of anatomical sulcal changes associated with aging**. *Cerebral cortex (New York, N.Y. : 1991)*, **16**:1584–1594, 2006. 8, 37

[43] Herve Lemaitre, Aaron L. Goldman, Fabio Sambataro, Beth A. Verchinski, et al. **Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume?** *Neurobiology of aging*, **33**:617.e1–617.e9, 2012. 8, 37

[44] Anders M. Fjell, Lars T. Westlye, Inge Amlien, Thomas Espeseth, et al. **Minute effects of sex on the aging brain: a multisample magnetic resonance imaging study of healthy aging and Alzheimer's disease**. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **29**(27):8774–8783, 2009. 8

[45] David H. Salat, Randy L. Buckner, Abraham Z. Snyder, Douglas N. Greve, et al. **Thinning of the cerebral cortex in aging**. *Cerebral Cortex*, 2004. 8

[46] Larson J. Hogstrom, Lars T. Westlye, Kristine B. Walhovd, and Anders M. Fjell. **The structure of the cerebral cortex across adult life: Age-related patterns of surface area, thickness, and gyrification**. *Cerebral Cortex*, 2013. 8, 24, 25

[47] Andreas Berg Storsve, Anders M. Fjell, Christian K. Tamnes, Lars T. Westlye, et al. **Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: Regions of accelerating and decelerating change**. *Journal of Neuroscience*, 2014. 8

[48] Anders M. Fjell, Håkon Grydeland, Stine K. Krogsrud, Inge Amlien, et al. **Development and aging of cortical thickness correspond to genetic organization patterns**. *Proceedings of the National Academy of Sciences of the United States of America*, 2015. 8

[49] Xiaojing Long, Weiqi Liao, Chunxiang Jiang, Dong Liang, et al. **Healthy aging: an automatic analysis of global and regional morphological alterations of human brain**. *Academic radiology*, **19**:785–793, 2012. 8, 24

[50] Tao Liu, Perminder S. Sachdev, Darren M. Lipnicki, Jiyang Jiang, et al. **Longitudinal changes in sulcal morphology associated with late-life aging and MCI**. *NeuroImage*, **74**:337–342, 2013. 8

[51] Jing Yan, Yue Cui, Qianqian Li, Lin Tian, et al. **Cortical thinning and flattening in schizophrenia and their unaffected parents**. *Neuropsychiatric Disease and Treatment*, **15**:935, 2019. 8, 46

[52] Bo Cao, Benson Mwangi, Ives Cavalcante Passos, Mon Ju Wu, et al. **Lifespan Gyrification Trajectories of Human Brain in Healthy Individuals and Patients with Major Psychiatric Disorders**. *Scientific Reports*, 2017. 8

[53] Sander Lamballais, Elisabeth J. Vinke, Meike W. Vernooij, M. Arfan Ikram, and Ryan L. Muetzel. **Cortical gyrification in relation to age and cognition in older adults**. *NeuroImage*, 2020. 8

[54] Christopher R. Madan and Elizabeth A. Kensinger. **Cortical complexity as a measure of age-related brain atrophy**. *NeuroImage*, **134**:617–629, 2016. 8, 25, 77

[55] Christopher R. Madan and Elizabeth A. Kensinger. **Predicting age from cortical structure across the lifespan**. *European Journal of Neuroscience*, **47**(5):399–416, 2018. 8, 21, 24, 25, 27, 32, 71, 77

[56] Chiara Marzi, Marco Giannelli, Carlo Tessa, Mario Mascalchi, and Stefano Diciotti. **Toward a more reliable characterization of fractal properties of the cerebral cortex of healthy subjects during the lifespan**. *Scientific Reports*, **10**(1):16957, 2020. 8, 21, 24, 25, 28, 32

[57] Valentina Meregalli, Francesco Alberti, Christopher R. Madan, Paolo Meneguzzo, et al. **Cortical complexity estimation using fractal dimension: A systematic review of the literature on clinical and nonclinical samples**. *European Journal of Neuroscience*, **55**:1547–1583, 2022. 8

[58] K. B. Walhovd, H. Johansen-Berg, and R. T. Káradóttir. **Unraveling the secrets of white matter - Bridging the gap between cellular, animal and human imaging studies**, 2014. 8

[59] Alan Peters. **The effects of normal aging on myelin and nerve fibers: A review**, 2002. 9, 76, 88

[60] Alan Peters and Claire Sethares. **Is there remyelination during aging of the primate central nervous system?** *Journal of Comparative Neurology*, 2003. 9

[61] S. Wang and K. M. Young. **White matter plasticity in adulthood**, 2014. 9, 88

[62] Huan Liu, Yuanyuan Yang, Yuguo Xia, Wen Zhu, et al. **Aging of cerebral white matter**. *Ageing Research Reviews*, **34**:64–76, 2017. 9

[63] Lars T. Westlye, Kristine B. Walhovd, Anders M. Dale, Atle Bjørnerud, et al. **Life-span changes of the human brain white matter: Diffusion tensor imaging (DTI) and volumetry**. *Cerebral Cortex*, 2010. 9

[64] H. Braak and E. Braak. **Development of Alzheimer-related neurofibrillary changes in the neocortex inversely recapitulates cortical myelogenesis**. *Acta Neuropathologica*, 1996. 9

[65] Barry Reisberg, Emile H. Franssen, Liduïn E.M. Souren, Stefanie R. Auer, et al. **Evidence and mechanisms of retrogenesis in Alzheimer's and other dementias: Management and treatment import**. *American Journal of Alzheimer's Disease and other Dementias*, 2002. 9

[66] Adam M. Brickman, Irene B. Meier, Mayuresh S. Korgaonkar, Frank A. Provenzano, et al. **Testing the white matter retrogenesis hypothesis of cognitive aging**. *Neurobiology of Aging*, 2012. 9

[67] D. H. Salat. **Imaging small vessel-associated white matter changes in aging**, 2014. 9

[68] Niels D. Prins and Philip Scheltens. **White matter hyperintensities, cognitive impairment and dementia: An update**, 2015. 9

[69] Alida A. Gouw, Alexandra Seewann, Wiesje M. Van Der Flier, Frederik Barkhof, et al. **Heterogeneity of small vessel disease: A systematic review of MRI and histopathology correlations**, 2011. 9

[70] Sarah K. Madsen, Boris A. Gutman, Shantanu H. Joshi, Arthur W. Toga, et al. **Mapping ventricular expansion onto cortical gray matter in older adults**. *Neurobiology of Aging*, **36**:S32 – S41, 2015. Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD). 9, 12

[71] CR Jack, MM Shiung, JL Gunter, PC O'brien, et al. **Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD**. *Neurology*, **62**(4):591–600, 2004. 9, 12, 117

[72] Sean M Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, et al. **Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database**. *Brain*, **131**(9):2443–2454, 2008. 9, 12

[73] Thomas R. Insel. **Rethinking schizophrenia**. *Nature*, **468**:187–193, 2010. 9

[74] Robert A. McCutcheon, Tiago Reis Marques, and Oliver D. Howes. **Schizophrenia-An Overview**. *JAMA psychiatry*, **77**:201–210, 2020. 9, 10, 42, 117

[75] Robin M. Murray, Shon W. Lewis, and L. Lecturer. **Is schizophrenia a neurodevelopmental disorder?** *British medical journal (Clinical research ed.)*, **295**:681–682, 1987. 9

[76] World Health Organization. **Schizophrenia**. https://www.who.int/news-room/fact-sheets/detail/schizophrenia, online at 2023-12-31. 10

[77] Robert B. Zipursky, Thomas J. Reilly, and Robin M. Murray. **The myth of schizophrenia as a progressive brain disease**. *Schizophrenia bulletin*, **39**:1363–1372, 2013. 10, 131

[78] Assen Jablensky. **Living in a Kraepelinian world: Kraepelin's impact on modern psychiatry**. *History of psychiatry*, **18**:381–388, 2007. 10

[79] Paul J. Harrison. **Postmortem studies in schizophrenia**. *Dialogues in Clinical Neuroscience*, **2**:349, 2000. 10

[80] Eve C. Johnstone, C. D. Frith, T. J. Crow, Janet Husband, and L. Kreel. **Cerebral ventricular size and cognitive impairment in chronic schizophrenia**. *Lancet (London, England)*, **2**:924–926, 1976. 10

[81] Adrianne M. Reveley, Christine A. Clifford, Michael A. Reveley, and Robin M. Murray. **Cerebral ventricular size in twins discordant for schizophrenia**. *Lancet (London, England)*, **1**:540–541, 1982. 10

[82] Daniel R. Weinberger, E. Fuller Torrey, Andreas N. Neophytides, and Richard Jed Wyatt. **Structural abnormalities in the cerebral cortex of chronic schizophrenic patients**. *Archives of general psychiatry*, **36**:935–939, 1979. 10

[83] Ian Ellison-Wright, David C. Glahn, Angela R. Laird, Sarah M. Thelen, and Ed Bullmore. **The anatomy of first-episode and chronic schizophrenia: an anatomical likelihood estimation meta-analysis**. *The American journal of psychiatry*, **165**:1015–1023, 2008. 10, 131

[84] Nuno Madeira, João Valente Duarte, Ricardo Martins, Gabriel Nascimento Costa, et al. **Morphometry and gyrification in bipolar disorder and schizophrenia: A comparative MRI study**. *NeuroImage: Clinical*, 2020. 10, 110, 117

[85] Neeltje E.M. Van Haren, Hugo G. Schnack, Wiepke Cahn, Martijn P. Van Den Heuvel, et al. **Changes in cortical thickness during the course of illness in schizophrenia**. *Archives of general psychiatry*, **68**:871–880, 2011. 10

[86] P. Fusar-Poli, R. Smieskova, M. J. Kempton, B. C. Ho, et al. **Progressive brain changes in schizophrenia related to antipsychotic treatment? A meta-analysis of longitudinal MRI studies**. *Neuroscience and biobehavioral reviews*, **37**:1680–1691, 2013. 10

[87] Mary Ellen Lynall, Danielle S. Bassett, Robert Kerwin, Peter J. McKenna, et al. **Functional connectivity and brain networks in schizophrenia**. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **30**:9477–9487, 2010. 10, 117, 137

[88] Ian Ellison-Wright and Ed Bullmore. **Meta-analysis of diffusion tensor imaging studies in schizophrenia**. *Schizophrenia research*, **108**:3–10, 2009. 10, 117, 137

[89] Mujeeb Z Banday, Aga S Sameer, and Saniya Nissar. **Pathophysiology of diabetes: An overview**. *Avicenna journal of medicine*, **10**:174–188, 2020. 10

[90] World Health Organization. **Diabetes**. https://www.who.int/newsroom/fact-sheets/detail/diabetes, online at 2023-12-31. 10

[91] An S. De Vriese, Tony J. Verbeuren, Johan Van De Voorde, Norbert H. Lameire, and Paul M. Vanhoutte. **Endothelial dysfunction in diabetes**. *British journal of pharmacology*, **130**:963–974, 2000. 10, 132

[92] Avogaro A, Albiero M, Menegazzo L, de Kreutzenberg S, and Fadini GP. **Endothelial dysfunction in diabetes: the role of reparatory mechanisms**. *Diabetes care*, **34 Suppl 2**:823–826, 2011. 10

[93] Amir Moheet, Silvia Mangia, and Elizabeth R. Seaquist. **Impact of diabetes on cognitive function and brain structure**. *Annals of the New York Academy of Sciences*, **1353**:60–71, 2015. 11, 117, 131

[94] Junfeng Liu, Loes Rutten-Jacobs, Ming Liu, Hugh S. Markus, and Matthew Traylor. **Causal Impact of Type 2 Diabetes Mellitus on Cerebral Small Vessel Disease: A Mendelian Randomization Analysis**. *Stroke*, **49**:1325–1331, 2018. 11

[95] Manon Brundel, L. Jaap Kappelle, and Geert Jan Biessels. **Brain imaging in type 2 diabetes**. *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology*, **24**:1967–1981, 2014. 11, 117, 131

[96] Stefan L.C. Geijselaers, Simone J.S. Sep, Coen D.A. Stehouwer, and Geert Jan Biessels. **Glucose regulation, cognition, and brain MRI in type 2 diabetes: a systematic review**. *The lancet. Diabetes & endocrinology*, **3**(1):75–89, 2015. 11

[97] Geert Jan Biessels, Flavio Nobili, Charlotte E. Teunissen, Rafael Simó, and Philip Scheltens. **Understanding multifactorial brain changes in type 2 diabetes: a biomarker perspective**. *The Lancet. Neurology*, **19**:699–710, 2020. 11, 42

[98] Authors Martin Prince, Anders Wimo, Maëlenn Guerchet, Miss Gemma-Claire Ali, et al. **World Alzheimer Report 2015: The global impact of dementia: An analysis of prevalence, incidence, cost and trends**, 2015. 11

[99] Warren W. Barker, Cheryl A. Luis, Alice Kashuba, Mercy Luis, et al. **Relative frequencies of Alzheimer disease, Lewy body, vascular and frontotemporal dementia, and hippocampal sclerosis in the State of Florida Brain Bank**. *Alzheimer disease and associated disorders*, **16**(4):203–212, 2002. 11

[100] World Health Organization. **Dementia**. https://www.who.int/news-room/fact-sheets/detail/dementia, online at 2023-12-31. 11

[101] Colin L. Masters, Randall Bateman, Kaj Blennow, Christopher C. Rowe, et al. **Alzheimer's disease**. *Nature reviews. Disease primers*, **1**, 2015. 11, 12

[102] Camilla Ferrari and Sandro Sorbi. **The complexity of Alzheimer's disease: an evolving puzzle**. *Physiological reviews*, **101**:1047–1081, 2021. 11

[103] Charles R. Harrington. **The molecular pathology of Alzheimer's disease**. *Neuroimaging clinics of North America*, **22**(1):11–22, 2012. 11

[104] Sara E. Nasrabady, Batool Rizvi, James E. Goldman, and Adam M. Brickman. **White matter changes in Alzheimer's disease: a focus on myelin and oligodendrocytes**. *Acta neuropathologica communications*, **6**(1):22, 2018. 12

[105] Bruno Dubois, Howard H. Feldman, Claudia Jacova, Steven T. DeKosky, et al. **Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria**. *The Lancet. Neurology*, **6**:734–746, 2007. 12

[106] Suzanne M. De La Monte and Ming Tong. **Brain metabolic dysfunction at the core of Alzheimer's disease**. *Biochemical pharmacology*, **88**:548–559, 2014. 12, 117

[107] P. C. Lauterbur. **Image formation by induced local interactions: Examples employing nuclear magnetic resonance**. *Nature*, **242**(5394):190–191, 1973. 12

[108] **The Nobel Prize in Physiology or Medicine 2003 - NobelPrize.org**. 12

[109] D. Weishaupt, V.D. Kochli, B. Marincek, and E. E. Kim. **How Does MRI Work? An Introduction to the Physics and Function of Magnetic Resonance Imaging**. *Journal of Nuclear Medicine*, 2007. 12, 13, 14, 15, 16, 17

[110] J. Mohan, V. Krishnaveni, and Yanhui Guo. **A survey on the magnetic resonance image denoising methods**, 2014. 20

[111] P. Kalavathi and V. B.Surya Prasath. **Methods on Skull Stripping of MRI Head Scan Images—a Review**. *Journal of Digital Imaging*, **29**(3):365–379, 2016. 20

[112] Lingraj Dora, Sanjay Agrawal, Rutuparna Panda, and Ajith Abraham. **State-of-the-Art Methods for Brain Tissue Segmentation: A Review**, 2017. 20, 47

[113] Tobias Hepp, Dominik Blum, Karim Armanious, Bernhard Schölkopf, et al. **Uncertainty estimation and explainability in deep learning-based age estimation of the human brain: Results from the German National Cohort MRI study**. *Computerized Medical Imaging and Graphics*, **92**:101967, 2021. 21, 37, 41, 118, 137

[114] M. Tanveer, M. A. Ganaie, Iman Beheshti, Tripti Goel, et al. **Deep learning for brain age estimation: A systematic review**. *Information Fusion*, **96**:130–143, 2023. 21, 33, 117

[115] Jieqiong Wang, Wenjing Li, Wen Miao, Dai Dai, et al. **Age estimation using cortical surface pattern combining thickness with curvatures**. *Medical and Biological Engineering and Computing*, **52**(4):331–341, 2014. 21, 24, 25, 27, 32

[116] S. A. Valizadeh, J. Hänggi, S. Mérillat, and L. Jäncke. **Age prediction on the basis of brain anatomical measures**. *Human Brain Mapping*, 2017. 21, 24, 25, 27, 32, 77

[117] Benjamin Gutierrez Becker, Tassilo Klein, and Christian Wachinger. **Gaussian process uncertainty in age estimation as a measure of brain abnormality**. *NeuroImage*, 2018. 21, 24, 26, 27, 30, 32, 38, 112, 117

[118] Pedro F. Da Costa, Jessica Dafflon, and Walter H.L. Pinaya. **Brain-Age Prediction Using Shallow Machine Learning: Predictive Analytics Competition 2019**. *Frontiers in Psychiatry*, **11**:604478, 2020. 21, 24, 25, 26, 28, 29, 30, 32

[119] Jaroslav Rokicki, Thomas Wolfers, Wibeke Nordhøy, Natalia Tesli, et al. **Multimodal imaging improves brain age prediction and reveals distinct abnormalities in patients with psychiatric and neurological disorders**. *Human brain mapping*, **42**:1714–1726, 2021. 21, 24, 28, 38, 117

[120] Xia Liu, Iman Beheshti, Weihao Zheng, Yongchao Li, et al. **Brain age estimation using multi-feature-based networks**. *Computers in Biology and Medicine*, **143**:105285, 2022. 21, 24, 25, 28, 32

[121] Jun Ding Zhu, Shih Jen Tsai, Ching Po Lin, Yi Ju Lee, and Albert C. Yang. **Predicting aging trajectories of decline in brain volume, cortical thickness and fractional anisotropy in schizophrenia**. *Schizophrenia (Heidelberg, Germany)*, **9**, 2023. 21, 24, 28, 32, 42, 117

[122] Gidon Levakov, Gideon Rosenthal, Ilan Shelef, Tammy Riklin Raviv, and Galia Avidan. **From a deep learning model back to the brain-Identifying regional predictors and their relation to aging**. *Human brain mapping*, **41**:3235–3252, 2020. 21, 33, 34, 36, 37, 38, 39, 118, 127

[123] Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, Stephen M. Smith, and Han Peng. **Optimising a Simple Fully**

Convolutional Network for Accurate Brain Age Prediction in the PAC 2019 Challenge. *Frontiers in psychiatry*, **12**, 2021. 21, 33, 34, 35, 36, 38, 39

[124] Alba Xifra-Porxas, Arna Ghosh, Georgios D. Mitsis, and Marie Hélène Boudrias. **Estimating brain age from structural MRI and MEG data: Insights from dimensionality reduction techniques**. *NeuroImage*, **231**:117822, 2021. 21, 26, 31, 32

[125] Pauline Mouches, Matthias Wilms, Deepthi Rajashekar, Sönke Langner, and Nils D. Forkert. **Multimodal biological brain age prediction using magnetic resonance imaging and angiography with the identification of predictive regions**. *Human brain mapping*, **43**:2554–2566, 2022. 21, 33, 34, 35, 36, 37, 38, 39, 118, 127

[126] Johnny Wang, Maria J. Knol, Aleksei Tiulpin, Florian Dubost, et al. **Gray Matter Age Prediction as a Biomarker for Risk of Dementia**. *Proceedings of the National Academy of Sciences of the United States of America*, **116**:21213–21218, 2019. 21, 34, 37, 38, 93

[127] Nicola K. Dinsdale, Emma Bluemke, Stephen M. Smith, Zobair Arya, et al. **Learning patterns of the ageing brain in MRI using deep convolutional networks**. *NeuroImage*, **224**, 2021. 21, 33, 34, 36, 37, 38, 40

[128] Simon M. Hofmann, Frauke Beyer, Sebastian Lapuschkin, Ole Goltermann, et al. **Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain**. *NeuroImage*, **261**, 2022. 21, 36, 40

[129] Iman Beheshti, Scott Nugent, Olivier Potvin, and Simon Duchesne. **Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme**. *NeuroImage. Clinical*, **24**, 2019. 21, 120

[130] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. **Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters**. *NeuroImage*, **50**(3):883 – 892, 2010. 21, 25, 26, 29, 30, 32, 112

[131] Deepthi P. Varikuti, Sarah Genon, Aristeidis Sotiras, Holger Schwender, et al. **Evaluation of non-negative matrix factorization of grey matter in age prediction**. *NeuroImage*, **173**:394–410, 2018. 21, 26, 29, 30, 32, 38, 117

[132] Huiting Jiang, Na Lu, Kewei Chen, Li Yao, et al. **Predicting Brain Age of Healthy Adults Based on Structural MRI Parcellation Using**

Convolutional Neural Networks. *Frontiers in Neurology*, 2020. 21, 26, 30, 32, 33, 34, 38, 39

[133] Sebastian G. Popescu, Ben Glocker, David J. Sharp, and James H. Cole. **Local Brain-Age: A U-Net Model**. *Frontiers in aging neuroscience*, **13**, 2021. 21, 33, 34, 35, 37, 38, 41, 118

[134] Benson Mwangi, Tian Siva Tian, and Jair C. Soares. **A review of feature reduction techniques in Neuroimaging**. *Neuroinformatics*, **12**:229–244, 2014. 21

[135] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, et al. **Recent advances in convolutional neural networks**. *Pattern Recognition*, **77**:354–377, 2018. 21, 32

[136] Ann Marie G. de Lange, Melis Anatürk, Jaroslav Rokicki, Laura K.M. Han, et al. **Mind the gap: Performance metric evaluation in brain-age prediction**. *Human brain mapping*, **43**:3113–3129, 2022. 23

[137] Won Hee Lee, Mathilde Antoniades, Hugo G. Schnack, Rene S. Kahn, and Sophia Frangou. **Brain age prediction in schizophrenia: Does the choice of machine learning algorithm matter?** *Psychiatry research. Neuroimaging*, **310**, 2021. 24, 25, 27, 29, 32, 42, 117

[138] Jessica Dafflon, Walter H.L. Pinaya, Federico Turkheimer, James H. Cole, et al. **An automated machine learning approach to predict brain age from cortical anatomical measures**. *Human Brain Mapping*, **41**:3555–3566, 2020. 24, 25, 28, 29, 32

[139] Geneviève Richard, Knut Kolskår, Anne Marthe Sanders, Tobias Kaufmann, et al. **Assessing distinct patterns of cognitive aging using tissue-specific brain age prediction based on diffusion tensor imaging and brain morphometry**. *PeerJ*, 2018. 24, 28, 32

[140] Habtamu M. Aycheh, Joon-Kyung Seong, Jeong-Hyeon Shin, Duk L. Na, et al. **Biological Brain Age Prediction Using Cortical Thickness Data: A Large Scale Cohort Study**. *Frontiers in Aging Neuroscience*, **10**(AUG):252, 2018. 24, 28, 32

[141] Timothy T. Brown, Joshua M. Kuperman, Yoonho Chung, Matthew Erhart, et al. **Neuroanatomical assessment of biological maturity**. *Current Biology*, 2012. 24

[142] Ed H.B.M. Gronenschild, Petra Habets, Heidi I.L. Jacobs, Ron Mengelers, et al. **The effects of FreeSurfer version, workstation**

type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one*, **7**, 2012. 24

[143] Budhachandra S. Khundrakpam, Jussi Tohka, Alan C. Evans, William S. Ball, et al. **Prediction of brain maturity based on cortical thickness at different spatial resolutions**. *NeuroImage*, 2015. 24

[144] Iman Beheshti, M. A. Ganaie, Vardhan Paliwal, Aryan Rastogi, et al. **Predicting Brain Age Using Machine Learning Algorithms: A Comprehensive Evaluation**. *IEEE journal of biomedical and health informatics*, **26**:1432–1440, 2022. 25, 29, 31

[145] Kiho Im, Jong Min Lee, Uicheul Yoon, Yong Wook Shin, et al. **Fractal dimension in human cortical surface: Multiple regression analysis with cortical thickness, sulcal depth, and folding area**. *Human Brain Mapping*, **27**(12):994–1003, 2006. 25, 77

[146] Asa Ben-Hur and Jason Weston. **A user's guide to support vector machines**. *Methods in molecular biology (Clifton, N.J.)*, **609**:223–239, 2010. 25

[147] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. C. Valdés Hernández, et al. **Brain age predicts mortality**. *Molecular Psychiatry*, 2018. 26, 30, 32, 77

[148] Vincent A. Magnotta and Lee Friedman. **Measurement of signal-to-noise and contrast-to-noise in the fBIRN multicenter imaging study**. *Journal of Digital Imaging*, **19**(2):140–147, 2006. 32, 137

[149] Sergey Ioffe and Christian Szegedy. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. In *32nd International Conference on Machine Learning, ICML 2015*, 2015. 33

[150] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. **Dropout: A simple way to prevent neural networks from overfitting**. *Journal of Machine Learning Research*, 2014. 33

[151] Pedro L. Ballester, Laura Tomaz da Silva, Matheus Marcon, Nathalia Bianchini Esper, et al. **Predicting Brain Age at Slice Level: Convolutional Neural Networks and Consequences for Interpretability**. *Frontiers in psychiatry*, **12**, 2021. 33, 34, 36, 40, 118

[152] David A. Wood, Sina Kafiabadi, Ayisha Al Busaidi, Emily Guilhem, et al. **Accurate brain-age models for routine clinical MRI examinations**. *NeuroImage*, **249**, 2022. 33, 34, 35, 38, 40

[153] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Deep residual learning for image recognition**. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. 33, 93, 110

[154] Pauline Mouches, Matthias Wilms, Agampreet Aulakh, Sönke Langner, and Nils D. Forkert. **Multimodal brain age prediction fusing morphometric and imaging data and association with cardiovascular risk factors**. *Frontiers in neurology*, **13**, 2022. 33, 34, 35, 36, 38, 39, 118, 127

[155] Sheng He, Diana Pereira, Juan David Perez, Randy L. Gollub, et al. **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**, 2021. 34, 35, 38, 40

[156] Pradeep K. Lam, Vigneshwaran Santhalingam, Parth Suresh, Rahul Baboota, et al. **Accurate brain age prediction using recurrent slice-based networks**. *bioRxiv*, page 2020.08.04.235069, 2020. 34, 38, 41, 118

[157] Kyriaki Margarita Bintsi, Vasileios Baltatzis, Arinbjörn Kolbeinsson, Alexander Hammers, and Daniel Rueckert. **Patch-based Brain Age Estimation from MR Images**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **12449 LNCS**:98–107, 2020. 37

[158] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, et al. **Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization**. *Proceedings of the IEEE International Conference on Computer Vision*, **2017-October**:618–626, 2017. 37

[159] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. **SmoothGrad: removing noise by adding noise**. 2017. 37, 120

[160] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, et al. **Assessing the Trustworthiness of Saliency Maps for Localizing**

Abnormalities in Medical Imaging. *Radiology. Artificial intelligence*, **3**, 2021. 37, 108

[161] Katja Franke and Christian Gaser. **Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's Disease**. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, **25**(4):235–245, 2012. 38, 77, 88

[162] Christian Gaser, Katja Franke, Stefan Klöppel, Nikolaos Koutsouleris, et al. **BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease**. *PLoS ONE*, **8**(6), 2013. 38, 42, 77, 88, 93, 110, 112

[163] Annie M. Racine, Michael Brickhouse, David A. Wolk, and Bradford C. Dickerson. **The personalized Alzheimer's disease cortical thickness index predicts likely pathology and clinical progression in mild cognitive impairment**. *Alzheimer's & dementia (Amsterdam, Netherlands)*, **10**:301–310, 2018. 38

[164] Maria Ly, Gary Z. Yu, Helmet T. Karim, Nishita R. Muppidi, et al. **Improving brain age prediction models: incorporation of amyloid status in Alzheimer's disease**. *Neurobiology of aging*, **87**:44–48, 2020. 38

[165] Mengxue Wang, Qingguo Ren, Yachen Shi, Hao Shu, et al. **The effect of Alzheimer's disease risk factors on brain aging in normal Chineses: Cognitive aging and cognitive reserve**. *Neuroscience Letters*, **771**:136398, 2022. 38

[166] Jing Li and Hanna Lu. **MRI-informed cortical features for brain age prediction in age-specific adulthoods**. *Human Brain Mapping*, **44**:301–303, 2023. 38

[167] Irene Cumplido-Mayoral, Marina García-Prat, Grégory Operto, Carles Falcon, et al. **Biological brain age prediction using machine learning on structural neuroimaging data: multi-cohort validation against biomarkers of Alzheimer's disease and neurodegeneration stratified by sex**. *eLife*, **12**, 2023. 38

[168] Shammi More, Georgios Antonopoulos, Felix Hoffstaedter, Julian Caspers, et al. **Brain-age prediction: A systematic comparison of machine learning workflows**. *NeuroImage*, **270**, 2023. 38

[169] McKenna E. Williams, Jeremy A. Elman, Linda K. Mcevoy, Ole A. Andreassen, et al. **12-year prediction of mild cognitive impairment aided by Alzheimer's brain signatures at mean age 56**. *Brain Communications*, **3**, 2021. 38

[170] Katja Franke, Christian Gaser, Brad Manor, and Vera Novak. **Advanced BrainAGE in older adults with type 2 diabetes mellitus**. *Frontiers in aging neuroscience*, **5**, 2013. 42, 117

[171] James H. Cole. **Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors**. *Neurobiology of aging*, **92**:34–42, 2020. 42

[172] Nikolaos Koutsouleris, Christos Davatzikos, Stefan Borgwardt, Christian Gaser, et al. **Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders**. *Schizophrenia bulletin*, **40**:1140–1153, 2014. 42

[173] Igor Nenadić, Maren Dietzek, Kerstin Langbein, Heinrich Sauer, and Christian Gaser. **BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder**. *Psychiatry Research: Neuroimaging*, **266**:86 – 89, 2017. 42, 77, 88

[174] Saba Shahab, Benoit H. Mulsant, Melissa L. Levesque, Navona Calarco, et al. **Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls**. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, **44**:898–906, 2019. 42

[175] Shalaila S. Haas, Ruiyang Ge, Nicole Sanford, Amirhossein Modabbernia, et al. **Accelerated Global and Local Brain Aging Differentiate Cognitively Impaired From Cognitively Spared Patients With Schizophrenia**. *Frontiers in psychiatry*, **13**, 2022. 42

[176] Tomas Hajek, Katja Franke, Marian Kolenic, Jana Capkova, et al. **Brain Age in Early Stages of Bipolar Disorders or Schizophrenia**. *Schizophrenia bulletin*, **45**:191–198, 2019. 42

[177] Marian Kolenic, Katja Franke, Jaroslav Hlinka, Martin Matejka, et al. **Obesity, dyslipidemia and brain age in first-episode psychosis**. *Journal of psychiatric research*, **99**:151–158, 2018. 42

[178] Sean Mcwhinney, Marian Kolenic, Katja Franke, Marketa Fialova, et al. **Obesity as a Risk Factor for Accelerated Brain Ageing in First-Episode Psychosis-A Longitudinal Study**. *Schizophrenia bulletin*, **47**:1772–1781, 2021. 42

[179] Yoonho Chung, Jean Addington, Carrie E. Bearden, Kristin Cadenhead, et al. **Use of Machine Learning to Determine Deviance in Neuroanatomical Maturity Associated With Future Psychosis in Youths at Clinically High Risk**. *JAMA psychiatry*, **75**:960–968, 2018. 42

[180] James H. Cole PhD, Joel Raffel MD, Tim Friede PhD, PhD Arman Eshaghi MD, et al. **Longitudinal Assessment of Multiple Sclerosis with the Brain-Age Paradigm**. *Annals of neurology*, **88**:93–105, 2020. 42, 94, 110

[181] James H. Cole, Robert Leech, and David J. Sharp. **Prediction of brain age suggests accelerated atrophy after traumatic brain injury**. *Annals of neurology*, **77**:571–581, 2015. 42

[182] Ricky R. Savjani, Brian A. Taylor, Laura Acion, Elisabeth A. Wilde, and Ricardo E. Jorge. **Accelerated Changes in Cortical Thickness Measurements with Age in Military Service Members with Traumatic Brain Injury**. *Journal of neurotrauma*, **34**:3107–3116, 2017. 42

[183] Virginia F.J. Newcombe, Nicholas J. Ashton, Jussi P. Posti, Ben Glocker, et al. **Post-acute blood biomarkers and disease progression in traumatic brain injury**. *Brain : a journal of neurology*, **145**:2064–2076, 2022. 42

[184] Gaurav Verma, Yael Jacob, Manish Jha, Laurel S. Morris, et al. **Quantification of brain age using high-resolution 7 tesla MR imaging and implications for patients with epilepsy**. *Epilepsy & behavior reports*, **18**, 2022. 42

[185] Gyujoon Hwang, Bruce Hermann, Veena A. Nair, Lisa L. Conant, et al. **Brain aging in temporal lobe epilepsy: Chronological, structural, and functional**. *NeuroImage. Clinical*, **25**, 2020. 42

[186] Laura K.M. Han, Richard Dinga, Tim Hahn, Christopher R.K. Ching, et al. **Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group**. *Molecular psychiatry*, **26**:5124–5139, 2021. 42

[187] Bianca Besteher, Christian Gaser, and Igor Nenadić. **Machine-learning based brain age estimation in major depression showing no evidence of accelerated aging**. *Psychiatry Research: Neuroimaging*, **290**:1–4, 2019. 42

[188] Chloe Hutton, Enrico De Vita, John Ashburner, Ralf Deichmann, and Robert Turner. **Voxel-based cortical thickness measurements in MRI**. *Neuroimage*, **40**(4):1701, 2008. 46, 70

[189] B. C. Dickerson, E. Fenstermacher, D.H. Salat, D.A. Wolk, et al. **Detection of cortical thickness correlates of cognitive performance:**

Reliability across MRI scan sessions, scanners, and field strengths. *NeuroImage*, **39**(1):10, 2008. 46

[190] Narr KL, Woods RP, Thompson PM, Szeszko P, et al. **Relationships between IQ and regional cortical gray matter thickness in healthy adults**. *Cerebral cortex (New York, N.Y. : 1991)*, **17**(9):2163–2171, 2007. 46

[191] Rik Ossenkoppele, Ruben Smith, Tomas Ohlsson, Olof Strandberg, et al. **Associations between tau, Aβ, and cortical thickness with cognition in Alzheimer disease**. *Neurology*, **92**(6):e601–e612, 2019. 46

[192] van Haren NE, Schnack HG, Cahn W, van den Heuvel MP, et al. **Changes in cortical thickness during the course of illness in schizophrenia**. *Archives of general psychiatry*, **68**(9):871–880, 2011. 46

[193] Emma J. Burton, Ian G. McKeith, David J. Burn, E. David Williams, and John T. O'Brien. **Cerebral atrophy in Parkinson's disease with and without dementia: a comparison with Alzheimer's disease, dementia with Lewy bodies and controls**. *Brain : a journal of neurology*, **127**(Pt 4):791–800, 2004. 46

[194] Naroa Ibarretxe-Bilbao, Carme Junque, Barbara Segura, Hugo C. Baggio, et al. **Progression of cortical thinning in early Parkinson's disease**. *Movement disorders : official journal of the Movement Disorder Society*, **27**(14):1746–1753, 2012. 46

[195] Thomas Jubault, Jean François Gagnon, Sherif Karama, Alain Ptito, et al. **Patterns of cortical thickness and surface area in early Parkinson's disease**. *NeuroImage*, **55**(2):462–467, 2011. 46

[196] Kristine B. Walhovd, Anders M. Fjell, Jay Giedd, Anders M. Dale, and Timothy T. Brown. **Through Thick and Thin: a Need to Reconcile Contradictory Results on Trajectories in Human Cortical Development**. *Cerebral Cortex (New York, NY)*, **27**(2):1–10, 2016. 46, 71

[197] Saashi A. Bedford, Min Tae M. Park, Gabriel A. Devenyi, Stephanie Tullo, et al. **Greater cortical thickness in individuals with ASD**. *Molecular Psychiatry 2020 25:3*, **25**(3):507–508, 2020. 46, 71

[198] Budhachandra S. Khundrakpam, John D. Lewis, Penelope Kostopoulos, Felix Carbonell, and Alan C. Evans. **Cortical Thickness Abnormalities in Autism Spectrum Disorders Through Late Childhood, Adolescence, and Adulthood: A Large-Scale MRI Study**. *Cerebral cortex (New York, N.Y. : 1991)*, **27**(3):1721–1731, 2017. 46, 71

[199] Daan Van Rooij, Evdokia Anagnostou, Celso Arango, Guillaume Auzias, et al. **Cortical and Subcortical Brain Morphometry Differences Between Patients With Autism Spectrum Disorder and Healthy Individuals Across the Lifespan: Results From the ENIGMA ASD Working Group**. *The American journal of psychiatry*, **175**(4):359, 2018. 46, 71

[200] Gregory L. Wallace, Nathan Dankner, Lauren Kenworthy, Jay N. Giedd, and Alex Martin. **Age-related temporal and parietal cortical thinning in autism spectrum disorders**. *Brain*, **133**(12):3745, 2010. 46, 71

[201] D. P. Hibar, L. T. Westlye, N. T. Doan, N. Jahanshad, et al. **Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group**. *Molecular Psychiatry*, **23**(4):932, 2018. 46

[202] Qian Li, Youjin Zhao, Ziqi Chen, Jingyi Long, et al. **Meta-analysis of cortical thickness abnormalities in medication-free patients with major depressive disorder**. *Neuropsychopharmacology 2019 45:4*, **45**(4):703–712, 2019. 46

[203] John Ashburner and Karl J. Friston. **Unified segmentation**. *NeuroImage*, **26**(3):839–851, 2005. 47, 50, 109

[204] Bruce Fischl. **FreeSurfer**, 2012. 47, 50

[205] Uroš Vovk, Franjo Pernuš, and Boštjan Likar. **A review of methods for correction of intensity inhomogeneity in MRI**, 2007. 47

[206] Rene Seiger, Sebastian Ganger, Georg S. Kranz, Andreas Hahn, and Rupert Lanzenberger. **Cortical Thickness Estimations of FreeSurfer and the CAT12 Toolbox in Patients with Alzheimer's Disease and Healthy Controls**. *Journal of Neuroimaging*, **28**(5):515–523, 2018. 47

[207] Juan Velázquez, Julieta Mateos, Erick H. Pasaye, Fernando A. Barrios, and Jorge A. Marquez-Flores. **Cortical Thickness Estimation: A Comparison of FreeSurfer and Three Voxel-Based Methods in a Test-Retest Analysis and a Clinical Application**. *Brain topography*, **34**(4):430–441, 2021. 47

[208] Shahrzad Kharabian Masouleh, Simon B. Eickhoff, Yashar Zeighami, Lindsay B. Lewis, et al. **Influence of Processing Pipeline on Cortical Thickness Measurement**. *Cerebral Cortex*, **30**(9):5014–5027, 2020. 47, 72

[209] A. Di Martino, C. G. Yan, Q. Li, E. Denio, et al. **The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism**. *Molecular Psychiatry*, 2014. 48, 92, 93, 108, 114, 118

[210] **IXI Dataset – Brain Development**. https://brain-development.org/ixi-dataset/, online at 2023-12-31. 48, 78, 109, 114, 119

[211] Marko Wilke, Scott K. Holland, Mekibib Altaye, and Christian Gaser. **Template-O-Matic: A toolbox for creating customized pediatric templates**. *NeuroImage*, **41**:903–913, 2008. 50, 109, 119

[212] Tayyabah Yousaf, George Dervenoulas, and Marios Politis. **Advances in MRI Methodology**. *International Review of Neurobiology*, **141**:31–76, 2018. 50

[213] Oscar Esteban, Daniel Birman, Marie Schaer, Oluwasanmi O. Koyejo, et al. **MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites**. *PLOS ONE*, **12**(9):e0184661, 2017. 51

[214] Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. **Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature**. *NeuroImage*, **53**(1):1–15, 2010. 51

[215] Patrick E. Shrout and Joseph L. Fleiss. **Intraclass correlations: uses in assessing rater reliability**. *Psychological bulletin*, **86**:420–428, 1979. 51

[216] Winston Haynes. *Tukey's Test*, pages 2303–2304. Springer New York, 2013. 54, 82

[217] Robbin Gibb and Anna Kovalchuk. **Brain Development**. *The Neurobiology of Brain and Behavioral Development*, pages 3–27, 2018. 71

[218] Heath R. Pardoe, Rebecca Kucharsky Hiess, and Ruben Kuzniecky. **Motion and morphometry in clinical and nonclinical populations**. *NeuroImage*, **135**:177–185, 2016. 71, 72

[219] Alysha D. Gilmore, Nicholas J. Buser, and Jamie L. Hanson. **Variations in structural MRI quality significantly impact commonly used measures of brain anatomy**. *Brain Informatics 2021 8:1*, **8**:1–15, 2021. 71

[220] Logan Haynes, Amanda Ip, Ivy Y.K. Cho, Dennis Dimond, et al. **Grey and white matter volumes in early childhood: A comparison**

of voxel-based morphometry pipelines. *Developmental Cognitive Neuroscience*, **46**:100875, 2020. 71

[221] Elizabeth R. Sowell, Paul M. Thompson, Christiana M. Leonard, Suzanne E. Welcome, et al. **Longitudinal Mapping of Cortical Thickness and Brain Growth in Normal Children**. *The Journal of Neuroscience*, **24**:8223, 2004. 72

[222] Martin Reuter, M. Dylan Tisdall, Abid Qureshi, Randy L. Buckner, et al. **Head motion during MRI acquisition reduces gray matter volume and thickness estimates**. *NeuroImage*, **107**:107–115, 2015. 72

[223] Yasuyuki Taki, Ryoi Goto, Alan Evans, Alex Zijdenbos, et al. **Voxel-based morphometry of human brain with age and cerebrovascular risk factors**. *Neurobiology of Aging*, **25**(4):455–463, 2004. 76

[224] Elizabeth R. Sowell, Bradley S. Peterson, Paul M. Thompson, Suzanne E. Welcome, et al. **Mapping cortical change across the human life span**. *Nature Neuroscience 2003 6:3*, **6**(3):309–315, 2003. 76

[225] Catriona D Good, Ingrid S Johnsrude, John Ashburner, Richard N Henson, et al. **A voxel-based morphometric study of ageing in 465 normal adult human brains**. *NeuroImage*, **14**(1 Pt 1):21–36, 2001. 76

[226] George Bartzokis. **Age-related myelin breakdown: A developmental model of cognitive decline and Alzheimer's disease**, 2004. 76

[227] Raz N, Ghisletta P, Rodrigue KM, Kennedy KM, and Lindenberger U. **Trajectories of brain aging in middle-aged and older adults: regional and individual differences**. *NeuroImage*, **51**(2):501–511, 2010. 76

[228] Eric Courchesne, Heather J. Chisum, Jeanne Townsend, Angilene Cowles, et al. **Normal Brain Development and Aging: Quantitative Analysis at in Vivo MR Imaging in Healthy Volunteers1**. *https://doi.org/10.1148/radiology.216.3.r00au37672*, **216**(3):672–682, 2000. 76

[229] Schnack HG, van Haren NE, Nieuwenhuis M, Hulshoff Pol HE, et al. **Accelerated Brain Aging in Schizophrenia: A Longitudinal Pattern Recognition Study**. *The American journal of psychiatry*, **173**(6):607–616, 2016. 77, 93, 110

[230] Franke K, Ziegler G, Klöppel S, and Gaser C. **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, 2010. 77, 78, 81, 93, 110

[231] Paul M Thompson and Arthur W Toga. **Warping Strategies for Intersubject Registration**. *Handbook of Medical Imaging*, pages 569–601, 2000. 77

[232] Peter Pieperhoff, Lars Hömke, Frank Schneider, Ute Habel, et al. **Deformation Field Morphometry Reveals Age-Related Structural Differences between the Brains of Adults up to 51 Years**. *Journal of Neuroscience*, **28**(4):828–842, 2008. 77, 88

[233] Katja Frankea, Robert Dahnke, Geoffrey Clarke, Anderson Kuo, et al. **MRI based biomarker for brain aging in rodents and non-human primates**. *PRNI 2016 - 6th International Workshop on Pattern Recognition in Neuroimaging*, 2016. 78

[234] Babak A. Ardekani and Alvin H. Bachman. **Model-based automatic detection of the anterior and posterior commissures on MRI scans**. *NeuroImage*, **46**(3):677–682, 2009. 79, 109

[235] Frank E. Harrell, Kerry L. Lee, Robert M. Califf, David B. Pryor, and Robert A. Rosati. **Regression modelling strategies for improved prognostic prediction**. *Statistics in Medicine*, **3**(2):143–152, 1984. 79

[236] Martin Lotze, Martin Domin, Florian H. Gerlach, Christian Gaser, et al. **Novel findings from 2,838 Adult Brains on Sex Differences in Gray Matter Brain Volume**. *Scientific Reports 2019 9:1*, **9**(1):1–7, 2019. 79

[237] Akihiko Shiino, Yen-wei Chen, Kenji Tanigaki, Atsushi Yamada, et al. **Sex-related difference in human white matter volumes studied: Inspection of the corpus callosum and other white matter by VBM**. *Scientific Reports 2017 7:1*, **7**(1):1–7, 2017. 79

[238] F Pedregosa, G Varoquaux, A Gramfort, V Michel, et al. **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research*, **12**:2825–2830, 2011. 81

[239] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. 81

[240] Hualou Liang, Fengqing Zhang, and Xin Niu. **Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders**. *Human brain mapping*, **40**(11):3143–3152, 2019. 85

[241] Stephen M. Smith, D. Vidaurre, F. Alfaro-Almagro, Thomas E. Nichols, and Karla L. Miller. **Estimation of brain age delta from brain imaging**. *NeuroImage*, **200**:528–539, 2019. 85

[242] Ann Marie G. de Lange and James H. Cole. **Commentary: Correction procedures in brain-age prediction**. *NeuroImage : Clinical*, **26**, 2020. 85

[243] Uicheul Yoon, Jong Min Lee, Jun Soo Kwon, Hyun Pil Kim, et al. **An MRI study of structural variations in schizophrenia using deformation field morphometry**. *Psychiatry Research: Neuroimaging*, **146**(2):171–177, 2006. 88

[244] Ana L. Manera, Mahsa Dadar, D. Louis Collins, and Simon Ducharme. **Deformation based morphometry study of longitudinal MRI changes in behavioral variant frontotemporal dementia**. *NeuroImage: Clinical*, **24**:102079, 2019. 88

[245] Koutsouleris N, Davatzikos C, Borgwardt S, Gaser C, et al. **Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders**. *Schizophrenia bulletin*, **40**(5):1140–1153, 2014. 88

[246] Naftali Raz, Ulman Lindenberger, Karen M. Rodrigue, Kristen M. Kennedy, et al. **Regional brain changes in aging healthy adults: General trends, individual differences and modifiers**. *Cerebral Cortex*, **15**(11):1676–1689, 2005. 88

[247] Sarah K. Madsen, Boris A. Gutman, Shantanu H. Joshi, Arthur W. Toga, et al. **Mapping ventricular expansion onto cortical gray matter in older adults**. *Neurobiology of Aging*, **36**(S1):S32–S41, 2015. 88

[248] Rafael Garcia-Dias, Cristina Scarpazza, Lea Baecker, Sandra Vieira, et al. **Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners**. *NeuroImage*, **220**, 2020. 89

[249] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, et al. **Imagenet: A large-scale hierarchical image database**. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 92

[250] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, et al. **Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization**. *Neurology*, **74**:201, 2010. 92, 93, 108, 114, 118

[251] Bharat B. Biswal, Maarten Mennes, Xi Nian Zuo, Suril Gohel, et al. **Toward discovery science of human brain function**. *Proceedings of the National Academy of Sciences of the United States of America*, **107**:4734–4739, 2010. 92, 109, 114, 118

[252] Avram J. Holmes, Marisa O. Hollinshead, Timothy M. O'Keefe, Victor I. Petrov, et al. **Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures**. *Scientific Data*, 2015. 92, 109, 114, 118

[253] Adriana Di Martino, David O'Connor, Bosi Chen, Kaat Alaerts, et al. **Enhancing studies of the connectome in autism using the autism brain imaging data exchange II**. *Scientific data*, **4**, 2017. 92, 93, 108, 114, 118

[254] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, et al. **Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults**. *Journal of Cognitive Neuroscience*, 2007. 92, 93, 109, 114, 118

[255] Daniel S. Marcus, Anthony F. Fotenos, John G. Csernansky, John C. Morris, and Randy L. Buckner. **Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults**. *Journal of Cognitive Neuroscience*, 2010. 92, 93, 109, 114, 118

[256] Pamela LaMontagne, Tammie Benzinger, John Morris, Sarah Keefe, et al. **OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease**. *medRxiv*, page 2019.12.13.19014902, 2019. 92, 93, 109, 114, 118

[257] Farheen Ramzan, Muhammad Usman Ghani Khan, Asim Rehmat, Sajid Iqbal, et al. **A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks**. *Journal of Medical Systems*, **44**(2):1–16, 2020. 92, 93

[258] Atif Mehmood, Shuyuan Yang, Zhixi Feng, Min Wang, et al. **A Transfer Learning Approach for Early Diagnosis of Alzheimer's Disease on MRI Images**. *Neuroscience*, **460**:43–52, 2021. 92, 93

[259] Muazzam Maqsood, Faria Nazir, Umair Khan, Farhan Aadil, et al. **Transfer Learning Assisted Classification and Detection of**

Alzheimer's Disease Stages Using 3D MRI Scans. *Sensors (Basel, Switzerland)*, **19**(11), 2019. 92, 93

[260] Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, et al. **Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks**. *NeuroImage: Clinical*, **21**:101645, 2019. 92, 93

[261] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U. Brandt, et al. **Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation**. *NeuroImage: Clinical*, **24**:102003, 2019. 92, 93

[262] Kanghan Oh, Young Chul Chung, Ko Woon Kim, Woo Sung Kim, and Il Seok Oh. **Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning**. *Scientific Reports 2019 9:1*, **9**(1):1–16, 2019. 92, 93

[263] Pál Vakli, Regina J. Deák-Meszlényi, Petra Hermann, and Zoltán Vidnyánszky. **Transfer learning improves resting-state functional connectivity pattern analysis using convolutional neural networks**. *GigaScience*, **7**(12), 2018. 92

[264] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, et al. **Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **10435 LNCS**:516–524, 2017. 92, 107

[265] Zaniar Ardalan and Vignesh Subbian. **Transfer Learning Approaches for Neuroimaging Analysis: A Scoping Review**. *Frontiers in Artificial Intelligence*, **5**:15, 2022. 93

[266] Wenjie Kang, Lan Lin, Baiwen Zhang, Xiaoqi Shen, and Shuicai Wu. **Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis**. *Computers in Biology and Medicine*, **136**:104678, 2021. 93

[267] Karen Simonyan and Andrew Zisserman. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2014. 93

[268] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. **ImageNet classification with deep convolutional neural networks**. In *Advances in Neural Information Processing Systems*, 2012. 93

[269] Adrien Payan and Giovanni Montana. **Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks**. *ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings*, **2**:355–362, 2015. 93

[270] Nenadić I, Dietzek M, Langbein K, Sauer H, and Gaser C. **BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder**. *Psychiatry research. Neuroimaging*, **266**:86–89, 2017. 93, 110

[271] Stijn Denissen, Denis Alexander Engemann, Alexander De Cock, Lars Costers, et al. **Brain age as a surrogate marker for cognitive performance in multiple sclerosis**. *European journal of neurology*, **29**(10):3039–3049, 2022. 94, 110

[272] Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, David L. Thomas, et al. **Multimodal population brain imaging in the UK Biobank prospective epidemiological study**. *Nature Neuroscience 2016 19:11*, **19**(11):1523–1536, 2016. 94

[273] Bennett A. Shaywitz, Sally E. Shaywltz, Kenneth R. Pugh, R. Todd Constable, et al. **Sex differences in the functional organization of the brain for language**. *Nature 1995 373:6515*, **373**(6515):607–609, 1995. 94

[274] Anca Larisa Sandu, Gordon D. Waiter, Roger T. Staff, Nafeesa Nazlee, et al. **Sexual dimorphism in the relationship between brain complexity, volume and general intelligence (g): a cross-cohort study**. *Scientific Reports 2022 12:1*, **12**(1):1–12, 2022. 94

[275] Antonia N. Kaczkurkin, Armin Raznahan, and Theodore D. Satterthwaite. **Sex differences in the developing brain: insights from multimodal neuroimaging**. *Neuropsychopharmacology 2018 44:1*, **44**(1):71–85, 2018. 94

[276] Simon Baron-Cohen, Rebecca C. Knickmeyer, and Matthew K. Belmonte. **Sex differences in the brain: implications for explaining autism**. *Science (New York, N.Y.)*, **310**(5749):819–823, 2005. 94

[277] Sofia Santos, Helena Ferreira, João Martins, Joana Gonçalves, and Miguel Castelo-Branco. **Male sex bias in early and late onset**

neurodevelopmental disorders: Shared aspects and differences in Autism Spectrum Disorder, Attention Deficit/hyperactivity Disorder, and Schizophrenia. *Neuroscience and biobehavioral reviews*, **135**, 2022. 94

[278] Lília Jorge, Nádia Canário, Hugo Quental, Rui Bernardes, and Miguel Castelo-Branco. **Is the Retina a Mirror of the Aging Brain? Aging of Neural Retina Layers and Primary Visual Cortex Across the Lifespan**. *Frontiers in aging neuroscience*, **11**, 2020. 94

[279] Susanne Weis, Kaustubh R. Patil, Felix Hoffstaedter, Alessandra Nostro, et al. **Sex Classification by Resting State Brain Connectivity**. *Cerebral Cortex*, **30**(2):824–835, 2020. 94

[280] Matthew Leming and John Suckling. **Deep learning for sex classification in resting-state and task functional brain networks from the UK Biobank**. *Neuroimage*, **241**, 2021. 94

[281] Marc Andre Schulz, B. T.Thomas Yeo, Joshua T. Vogelstein, Janaina Mourao-Miranada, et al. **Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets**. *Nature communications*, **11**, 2020. 106

[282] Tong He, Ru Kong, Avram J. Holmes, Minh Nguyen, et al. **Deep Neural Networks and Kernel Regression Achieve Comparable Accuracies for Functional Connectivity Prediction of Behavior and Demographics**. *NeuroImage*, **206**:116276, 2020. 106

[283] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. **Models Genesis**. *Medical image analysis*, **67**, 2021. 107, 137

[284] Maria de Fátima Machado Dias, Paulo Carvalho, Miguel Castelo-Branco, and João Valente Duarte. **Cortical thickness in brain imaging studies using FreeSurfer and CAT12: A matter of reproducibility**. *Neuroimage: Reports*, **2**:100137, 2022. 109, 119

[285] Marko Wilke, Mekibib Altaye, and Scott K. Holland. **CerebroMatic: A Versatile Toolbox for Spline-Based MRI Template Creation**. *Frontiers in computational neuroscience*, **11**, 2017. 109, 119

[286] **Measurement and clinical effect of grey matter pathology in multiple sclerosis**. *The Lancet Neurology*, **11**(12):1082–1092, 2012. 110

[287] Lília Jorge, Ricardo Martins, Nádia Canário, Carolina Xavier, et al. **Investigating the Spatial Associations Between Amyloid-β Deposition, Grey Matter Volume, and Neuroinflammation in Alzheimer's Disease**. *Journal of Alzheimer's disease : JAD*, **80**:113–132, 2021. 110, 119

[288] Diederik P. Kingma and Jimmy Lei Ba. **Adam: A Method for Stochastic Optimization**. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2014. 111

[289] Maria De Fátima Mac Hado Dias, Paulo Carvalho, João Valente Duarte, and Miguel Castelo-Branco. **Deformation fields: a new source of information to predict brain age**. *Journal of neural engineering*, **19**, 2022. 112

[290] Jacob Cohen. *Statistical power for the behaviour sciences*. 1977. 113

[291] Shlomo S. Sawilowsky. **New Effect Size Rules of Thumb**. *Journal of Modern Applied Statistical Methods*, **8**(2):26, 2009. 113

[292] Yujun Hou, Xiuli Dan, Mansi Babbar, Yong Wei, et al. **Ageing as a risk factor for neurodegenerative disease**. *Nature Reviews Neurology 2019 15:10*, **15**:565–581, 2019. 116

[293] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. **The hallmarks of aging**. *Cell*, **153**:1194, 2013. 116

[294] George T. Baker and Richard L. Sprott. **Biomarkers of aging**. *Experimental gerontology*, **23**:223–239, 1988. 117

[295] Joana Crisóstomo, João V. Duarte, Nádia Canário, Carolina Moreno, et al. **The longitudinal impact of type 2 diabetes on brain gyrification**. *The European journal of neuroscience*, **58**, 2023. 117

[296] Meredith A. Shafto, Lorraine K. Tyler, Marie Dixon, Jason R. Taylor, et al. **The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing**. *BMC neurology*, **14**, 2014. 119

[297] Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, et al. **The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample**. *NeuroImage*, **144**:262–269, 2017. 119

[298] João V. Duarte, João M.S. Pereira, Bruno Quendera, Miguel Raimundo, et al. **Early disrupted neurovascular coupling and changed event level hemodynamic response function in type 2 diabetes: an fMRI study**. *Journal of Cerebral Blood Flow & Metabolism*, **35**:1671, 2015. 119

[299] **SPM**. 119

[300] Maria de Fátima Machado Dias, Tiago FT Cerqueira, João Valente Duarte, Miguel Castelo-Branco, and Paulo Carvalho. **3DCAE-MRI: Overcoming Data Availability in Small Sample Size MRI Studies**. 2023. 120

[301] Yoav Benjamini, Dan Drai, Greg Elmer, Neri Kafkafi, and Ilan Golani. **Controlling the false discovery rate in behavior genetics research**. *Behavioural brain research*, **125**:279–284, 2001. 121

[302] Fengling Hu, Andrew A. Chen, Hannah Horng, Vishnu Bashyam, et al. **Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization**. *NeuroImage*, **274**:120125, 2023. 136

[303] Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, et al. **Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan**. *NeuroImage*, **208**:116450, 2020. 136

Appendix A

Cortical thickness in brain imaging studies using FreeSurfer and CAT12: A matter of reproducibility

Table A.1: Reproducibility in paediatric versus early adults' groups: Demographics of the participants.

Site	*Total of participants*	*Number males*	Mean and standard deviation [years]	Min Age [years]	Max Age [years]	Voxel Size [mm]
California Institute of Technology	19	15	28.87+/- 11.21	17.00	56.20	1.0×1.0×1.0
Carnegie Mellon University	13	10	26.85+/- 5.74	20.00	40.00	1.0×1.0×1.0
Kennedy Krieger Institute	33	24	10.16+/- 1.26	8.07	12.77	1.0×1.0×1.0
Ludwig Maximilians University Munich	32	28	26.19+/- 9.96	7.00	48.00	1.0×1.0×1.0
NYU Langone Medical Center	104	78	15.83+/- 6.28	6.47	31.78	1.3×1.0×1.3
Olin Institute of Living at Hartford Hospital	16	14	16.94+/- 3.68	10.00	23.00	1.0×1.0×1.0
Oregon Health and Science University	15	15	10.06+/- 1.08	8.20	11.99	1.0×1.0×1.0
San Diego State University	22	16	14.22+/- 1.90	8.67	16.88	1.0×1.0×1.0
Social Brain Lab	15	15	33.73+/- 6.61	20.00	42.00	1.0×1.0×1.0
Stanford University	19	15	9.97+/- 1.64	7.75	12.43	0.859x1.500x0.859
Trinity Centre for Health Sciences	25	25	17.08+/- 3.77	12.04	25.66	1.0×1.0×1.0
University of California Los Angeles	45	39	12.96+/- 1.92	9.21	17.79	1.0×1.0×1.2
University of Leuven	35	30	18.17+/- 4.99	12.20	29.00	0.977x0.97x1.20
University of Michigan	75	58	14.82+/- 3.62	8.20	28.80	1.0×1.0×1.2
University of Pittsburgh School of Medicine	27	23	18.88+/- 6.64	9.44	33.24	1.1×1.1×1.1
University of Utah School of Medicine	43	43	21.36+/- 7.64	8.77	39.39	1.0×1.0×1.2
Yale Child Study Center	28	20	12.68+/- 2.75	7.66	17.83	1.0×1.0×1.0

Table A.2: Reproducibility in different acquisition settings: Demographics of the participants.

Repository		*Total of participants*	Number males	Mean and standard deviation [years]	Min Age [years]	Max Age [years]	Voxel Size [mm]
IXI	GH	312	139	50.73+/- 15.98	20.07	86.20	0.9375x0.9375x1.2000
	HH	179	85	47.63+/- 16.61	20.17	81.94	0.9375x0.9375x1.2000
	IOP	67	24	42.13+/- 16.60	19.98	86.32	0.9375x0.9375x1.2000
OASIS3		491	199	68.33+/- 8.86	42.66	95.20	1x1x1

Table A.3: Test-retest reliability: Demographics of the participants.

Repository	*Total of participants*	*Number males*	Mean and standard deviation [years]	Min Age [years]	Max Age [years]	Voxel Size [mm]
OASIS3	296	111	68.94+/- 8.86	45.78	88.86	1x1x1

Table A.4: Test-retest reliability: Mean R^2 per lobe and framework.

	R^2	
Framework	CAT12	FREESURFER
Area		
Frontal Lobe	0.76	0.63
Insula	0.77	0.72
Limbic lobe	0.77	0.72
Parietal lobe	0.83	0.73
Temporal and occipital lobes	0.80	0.77

Table A.5: Test-retest reliability: R^2 for each framework and brain hemisphere.

	R^2			
software	CAT12		FREESURFER	
Hemisphere	Left	Right	Left	Right
name				
Angular gyrus	8.14e-01	7.63e-01	7.51e-01	6.67e-01
Anterior occipital sulcus and preoccipital notch	8.63e-01	8.73e-01	7.04e-01	7.45e-01
Anterior part of the cingulate gyrus and sulcus	8.37e-01	8.18e-01	6.87e-01	6.61e-01
Anterior segment of the circular sulcus of the ...	7.79e-01	6.57e-01	6.75e-01	7.32e-01
Anterior transverse collateral sulcus	7.04e-01	6.89e-01	8.15e-01	8.01e-01
Anterior transverse temporal gyrus	7.19e-01	7.10e-01	7.39e-01	6.93e-01
Calcarine sulcus	8.58e-01	7.80e-01	7.56e-01	7.41e-01
Central sulcus	5.31e-01	5.95e-01	5.73e-01	5.91e-01
Cuneus	7.91e-01	8.17e-01	7.28e-01	6.89e-01
Fronto-marginal gyrus and sulcus	7.76e-01	7.24e-01	4.96e-01	4.95e-01
Horizontal ramus of the anterior segment of the...	8.57e-01	8.51e-01	6.73e-01	7.75e-01
Inferior frontal sulcus	7.72e-01	8.07e-01	6.67e-01	6.64e-01
Inferior occipital gyrus and sulcus	8.72e-01	8.15e-01	8.17e-01	7.98e-01
Inferior part of the precentral sulcus	7.78e-01	7.03e-01	7.24e-01	5.84e-01
Inferior segment of the circular sulcus of the ...	9.05e-01	8.48e-01	8.89e-01	8.38e-01
Inferior temporal gyrus	7.94e-01	7.73e-01	7.84e-01	7.93e-01
Inferior temporal sulcus	8.61e-01	8.94e-01	7.52e-01	7.49e-01
Intraparietal sulcus and transverse parietal s...	8.81e-01	8.85e-01	7.43e-01	8.11e-01
Lateral aspect of the superior temporal gyrus	8.21e-01	8.23e-01	8.10e-01	8.12e-01
Lateral occipito-temporal gyrus	7.22e-01	8.10e-01	8.22e-01	8.05e-01
Lateral occipito-temporal sulcus	8.28e-01	8.25e-01	6.99e-01	8.09e-01
Lateral orbital sulcus	7.71e-01	7.31e-01	4.04e-01	3.74e-01
Lingual gyrus	8.25e-01	7.76e-01	7.63e-01	7.32e-01
Long insular gyrus and central sulcus of the in...	5.93e-01	6.62e-01	6.45e-01	6.18e-01
Marginal branch of the cingulate sulcus	9.14e-01	8.60e-01	8.11e-01	7.65e-01
Medial occipito-temporal sulcus and lingual su...	8.45e-01	8.46e-01	8.54e-01	8.71e-01
Medial orbital sulcus	7.67e-01	7.00e-01	4.64e-01	6.17e-01
Middle frontal gyrus	8.18e-01	8.37e-01	6.78e-01	7.02e-01
Middle frontal sulcus	8.12e-01	7.62e-01	5.29e-01	5.08e-01
Middle occipital gyrus	8.83e-01	7.85e-01	7.61e-01	6.98e-01
Middle occipital sulcus and lunatus sulcus	8.93e-01	9.12e-01	7.71e-01	8.08e-01
Middle temporal gyrus	8.61e-01	8.43e-01	8.04e-01	8.27e-01
Middle-anterior part of the cingulate gyrus and...	9.20e-01	9.39e-01	8.21e-01	6.20e-01
Middle-posterior part of the cingulate gyrus an...	8.94e-01	8.57e-01	6.59e-01	6.59e-01
Occipital pole	8.20e-01	8.00e-01	6.93e-01	7.56e-01
Opercular part of the inferior frontal gyrus	8.40e-01	7.79e-01	7.68e-01	6.32e-01
Orbital gyri	7.42e-01	7.26e-01	5.36e-01	6.93e-01
Orbital part of the inferior frontal gyrus	8.16e-01	7.52e-01	7.17e-01	6.09e-01
Orbital sulci	7.12e-01	7.55e-01	5.73e-01	5.43e-01
Paracentral lobule and sulcus	8.20e-01	7.96e-01	6.84e-01	6.41e-01
Parahippocampal gyrus	5.28e-01	4.57e-01	8.22e-01	7.39e-01
Parieto-occipital sulcus	9.04e-01	8.43e-01	8.28e-01	7.89e-01
Pericallosal sulcus	7.52e-01	5.94e-01	7.16e-01	7.50e-01
Planum polare of the superior temporal gyrus	4.86e-01	7.36e-01	6.90e-01	6.61e-01
Planum temporale or temporal plane of the super...	8.21e-01	7.53e-01	7.62e-01	7.60e-01
Postcentral gyrus	8.29e-01	8.57e-01	7.91e-01	7.37e-01
Postcentral sulcus	8.10e-01	8.52e-01	7.48e-01	7.50e-01
Posterior ramus	8.50e-01	9.14e-01	8.39e-01	8.66e-01
Posterior transverse collateral sulcus	8.43e-01	8.48e-01	7.60e-01	7.34e-01
Posterior-dorsal part of the cingulate gyrus	6.41e-01	6.78e-01	6.77e-01	6.64e-01
Posterior-ventral part of the cingulate gyrus	6.98e-01	7.67e-01	7.62e-01	7.97e-01
Precentral gyrus	7.24e-01	7.34e-01	6.31e-01	6.05e-01
Precuneus	8.35e-01	8.22e-01	7.76e-01	7.67e-01
Short insular gyri	6.39e-01	6.30e-01	6.94e-01	5.70e-01
Straight gyrus	6.33e-01	7.22e-01	5.18e-01	5.75e-01
Subcallosal area	7.20e-01	6.10e-01	5.84e-01	4.25e-01
Subcentral gyrus and sulci	8.19e-01	8.40e-01	7.26e-01	7.86e-01
Suborbital sulcus	6.75e-01	7.00e-01	6.18e-01	4.70e-01
Subparietal sulcus	8.28e-01	7.99e-01	6.98e-01	7.58e-01
Sulcus intermedius primus	7.84e-01	7.92e-01	6.15e-01	6.23e-01
Superior frontal gyrus	8.47e-01	8.62e-01	8.41e-01	7.75e-01
Superior frontal sulcus	8.13e-01	8.39e-01	7.45e-01	6.26e-01
Superior occipital gyrus	8.79e-01	8.91e-01	8.07e-01	7.90e-01
Superior occipital sulcus and transverse occipi...	8.99e-01	8.91e-01	7.78e-01	8.15e-01
Superior parietal lobule	8.47e-01	8.52e-01	7.68e-01	7.54e-01
Superior part of the precentral sulcus	7.05e-01	7.51e-01	6.74e-01	5.35e-01
Superior segment of the circular sulcus of the ...	8.62e-01	8.30e-01	7.12e-01	6.62e-01
Superior temporal sulcus	9.20e-01	8.79e-01	8.01e-01	8.10e-01
Supramarginal gyrus	8.41e-01	7.53e-01	7.60e-01	7.03e-01
Temporal pole	6.07e-01	6.40e-01	6.24e-01	6.35e-01
Transverse frontopolar gyri and sulci	6.74e-01	6.54e-01	5.05e-01	5.35e-01
Transverse temporal sulcus	7.83e-01	7.70e-01	7.12e-01	7.08e-01
Triangular part of the inferior frontal gyrus	8.35e-01	6.95e-01	6.85e-01	6.57e-01
Vertical ramus of the anterior segment of the l...	7.84e-01	7.03e-01	6.92e-01	6.18e-01

Table A.6: Test-retest reliability: ICC and Confidence interval (CI) for each framework and brain hemisphere.

Framework	CAT12				FREESURFER			
Hemisphere	Left		Right		Left		Right	
	ICC	CI95%	ICC	CI95%	ICC	CI95%	ICC	CI95%
ROI Name								
Angular gyrus	0.90	[0.88, 0.92]	0.87	[0.84, 0.9]	0.87	[0.83, 0.89]	0.82	[0.77, 0.85]
Anterior occipital sulcus and preoccipital notch	0.93	[0.91, 0.94]	0.93	[0.92, 0.95]	0.84	[0.8, 0.87]	0.86	[0.83, 0.89]
Anterior part of the cingulate gyrus and sulcus	0.91	[0.89, 0.93]	0.90	[0.88, 0.92]	0.83	[0.79, 0.86]	0.81	[0.77, 0.85]
Anterior segment of the circular sulcus of the ...	0.88	[0.85, 0.91]	0.81	[0.77, 0.85]	0.82	[0.78, 0.86]	0.85	[0.82, 0.88]
Anterior transverse collateral sulcus	0.84	[0.8, 0.87]	0.83	[0.79, 0.86]	0.90	[0.88, 0.92]	0.89	[0.87, 0.92]
Anterior transverse temporal gyrus	0.85	[0.81, 0.88]	0.84	[0.81, 0.87]	0.86	[0.83, 0.89]	0.83	[0.79, 0.86]
Calcarine sulcus	0.93	[0.91, 0.94]	0.88	[0.86, 0.91]	0.87	[0.84, 0.89]	0.86	[0.83, 0.89]
Central sulcus	0.73	[0.67, 0.78]	0.77	[0.72, 0.81]	0.76	[0.7, 0.8]	0.77	[0.72, 0.81]
Cuneus	0.89	[0.86, 0.91]	0.90	[0.88, 0.92]	0.85	[0.82, 0.88]	0.83	[0.79, 0.86]
Fronto-marginal gyrus and sulcus	0.88	[0.85, 0.9]	0.85	[0.82, 0.88]	0.70	[0.64, 0.75]	0.70	[0.64, 0.75]
Horizontal ramus of the anterior segment of the...	0.92	[0.9, 0.94]	0.92	[0.9, 0.94]	0.82	[0.78, 0.85]	0.88	[0.85, 0.9]
Inferior frontal sulcus	0.88	[0.85, 0.9]	0.90	[0.87, 0.92]	0.82	[0.77, 0.85]	0.81	[0.77, 0.85]
Inferior occipital gyrus and sulcus	0.93	[0.92, 0.95]	0.90	[0.88, 0.92]	0.90	[0.88, 0.92]	0.89	[0.87, 0.91]
Inferior part of the precentral sulcus	0.88	[0.85, 0.9]	0.84	[0.8, 0.87]	0.85	[0.82, 0.88]	0.76	[0.71, 0.81]
Inferior segment of the circular sulcus of the ...	0.95	[0.94, 0.96]	0.92	[0.9, 0.94]	0.94	[0.93, 0.95]	0.91	[0.89, 0.93]
Inferior temporal gyrus	0.89	[0.86, 0.91]	0.88	[0.85, 0.9]	0.89	[0.86, 0.91]	0.89	[0.86, 0.91]
Inferior temporal sulcus	0.93	[0.91, 0.94]	0.95	[0.93, 0.96]	0.87	[0.83, 0.89]	0.86	[0.83, 0.89]
Intraparietal sulcus and transverse parietal s...	0.94	[0.92, 0.95]	0.94	[0.93, 0.95]	0.86	[0.83, 0.89]	0.90	[0.87, 0.92]
Lateral aspect of the superior temporal gyrus	0.91	[0.88, 0.92]	0.91	[0.88, 0.93]	0.90	[0.88, 0.92]	0.90	[0.88, 0.92]
Lateral occipito-temporal gyrus	0.85	[0.81, 0.88]	0.90	[0.87, 0.92]	0.91	[0.88, 0.92]	0.90	[0.87, 0.92]
Lateral occipito-temporal sulcus	0.91	[0.89, 0.93]	0.91	[0.89, 0.93]	0.84	[0.8, 0.87]	0.90	[0.87, 0.92]
Lateral orbital sulcus	0.88	[0.85, 0.9]	0.85	[0.82, 0.88]	0.62	[0.55, 0.69]	0.61	[0.53, 0.67]
Lingual gyrus	0.91	[0.89, 0.93]	0.88	[0.85, 0.9]	0.87	[0.84, 0.9]	0.86	[0.82, 0.88]
Long insular gyrus and central sulcus of the in...	0.77	[0.72, 0.81]	0.81	[0.77, 0.85]	0.80	[0.76, 0.84]	0.79	[0.74, 0.83]
Marginal branch of the cingulate sulcus	0.96	[0.94, 0.96]	0.93	[0.91, 0.94]	0.90	[0.88, 0.92]	0.87	[0.84, 0.9]
Medial occipito-temporal sulcus and lingual su...	0.92	[0.9, 0.93]	0.92	[0.9, 0.94]	0.92	[0.91, 0.94]	0.93	[0.92, 0.95]
Medial orbital sulcus	0.88	[0.85, 0.9]	0.84	[0.8, 0.87]	0.68	[0.61, 0.74]	0.78	[0.74, 0.82]
Middle frontal gyrus	0.90	[0.88, 0.92]	0.91	[0.89, 0.93]	0.82	[0.78, 0.86]	0.84	[0.8, 0.87]
Middle frontal sulcus	0.90	[0.88, 0.92]	0.87	[0.84, 0.9]	0.73	[0.67, 0.78]	0.71	[0.65, 0.76]
Middle occipital gyrus	0.94	[0.92, 0.95]	0.89	[0.86, 0.91]	0.87	[0.84, 0.9]	0.84	[0.8, 0.87]
Middle occipital sulcus and lunatus sulcus	0.94	[0.93, 0.96]	0.95	[0.94, 0.96]	0.88	[0.85, 0.9]	0.90	[0.87, 0.92]
Middle temporal gyrus	0.93	[0.91, 0.94]	0.92	[0.9, 0.93]	0.90	[0.87, 0.92]	0.91	[0.89, 0.93]
Middle-anterior part of the cingulate gyrus and...	0.96	[0.95, 0.97]	0.97	[0.96, 0.97]	0.91	[0.88, 0.92]	0.79	[0.74, 0.83]
Middle-posterior part of the cingulate gyrus an...	0.95	[0.93, 0.96]	0.93	[0.91, 0.94]	0.81	[0.77, 0.85]	0.81	[0.77, 0.85]
Occipital pole	0.91	[0.88, 0.92]	0.89	[0.87, 0.92]	0.83	[0.79, 0.86]	0.87	[0.84, 0.89]
Opercular part of the inferior frontal gyrus	0.92	[0.9, 0.93]	0.88	[0.85, 0.91]	0.88	[0.85, 0.9]	0.80	[0.75, 0.83]
Orbital gyri	0.86	[0.83, 0.89]	0.85	[0.82, 0.88]	0.72	[0.66, 0.77]	0.83	[0.79, 0.86]
Orbital part of the inferior frontal gyrus	0.90	[0.88, 0.92]	0.87	[0.83, 0.89]	0.85	[0.81, 0.88]	0.78	[0.73, 0.82]
Orbital sulci	0.84	[0.81, 0.87]	0.87	[0.84, 0.89]	0.76	[0.7, 0.8]	0.73	[0.68, 0.78]
Paracentral lobule and sulcus	0.90	[0.88, 0.92]	0.89	[0.87, 0.91]	0.83	[0.79, 0.86]	0.80	[0.75, 0.84]
Parahippocampal gyrus	0.73	[0.67, 0.78]	0.68	[0.61, 0.73]	0.91	[0.88, 0.92]	0.86	[0.83, 0.89]
Parieto-occipital sulcus	0.95	[0.94, 0.96]	0.92	[0.9, 0.93]	0.91	[0.89, 0.93]	0.89	[0.86, 0.91]
Pericallosal sulcus	0.87	[0.84, 0.89]	0.77	[0.72, 0.81]	0.85	[0.81, 0.88]	0.87	[0.83, 0.89]
Planum polare of the superior temporal gyrus	0.70	[0.63, 0.75]	0.86	[0.82, 0.89]	0.83	[0.79, 0.86]	0.81	[0.77, 0.85]
Planum temporale or temporal plane of the super...	0.91	[0.88, 0.92]	0.87	[0.84, 0.89]	0.87	[0.84, 0.9]	0.87	[0.84, 0.9]
Postcentral gyrus	0.91	[0.89, 0.93]	0.93	[0.91, 0.94]	0.89	[0.86, 0.91]	0.86	[0.83, 0.89]
Postcentral sulcus	0.90	[0.88, 0.92]	0.92	[0.9, 0.94]	0.86	[0.83, 0.89]	0.87	[0.83, 0.89]
Posterior ramus	0.92	[0.9, 0.94]	0.96	[0.94, 0.96]	0.91	[0.89, 0.93]	0.93	[0.91, 0.94]
Posterior transverse collateral sulcus	0.92	[0.9, 0.93]	0.92	[0.9, 0.94]	0.87	[0.84, 0.9]	0.86	[0.82, 0.88]
Posterior-dorsal part of the cingulate gyrus	0.80	[0.76, 0.84]	0.82	[0.78, 0.86]	0.82	[0.78, 0.86]	0.81	[0.77, 0.85]
Posterior-ventral part of the cingulate gyrus	0.84	[0.8, 0.87]	0.88	[0.85, 0.9]	0.87	[0.84, 0.9]	0.89	[0.87, 0.91]
Precentral gyrus	0.85	[0.82, 0.88]	0.86	[0.82, 0.88]	0.79	[0.75, 0.83]	0.78	[0.73, 0.82]
Precuneus	0.91	[0.89, 0.93]	0.91	[0.88, 0.92]	0.88	[0.85, 0.9]	0.88	[0.85, 0.9]
Short insular gyri	0.80	[0.75, 0.84]	0.79	[0.75, 0.83]	0.83	[0.79, 0.86]	0.75	[0.7, 0.8]
Straight gyrus	0.80	[0.75, 0.83]	0.85	[0.81, 0.88]	0.72	[0.66, 0.77]	0.76	[0.71, 0.8]
Subcallosal area	0.85	[0.81, 0.88]	0.78	[0.73, 0.82]	0.76	[0.71, 0.81]	0.65	[0.58, 0.71]
Subcentral gyrus and sulci	0.91	[0.88, 0.92]	0.92	[0.9, 0.93]	0.85	[0.82, 0.88]	0.89	[0.86, 0.91]
Suborbital sulcus	0.82	[0.78, 0.85]	0.84	[0.8, 0.87]	0.79	[0.74, 0.83]	0.68	[0.62, 0.74]
Subparietal sulcus	0.91	[0.89, 0.93]	0.89	[0.87, 0.91]	0.83	[0.8, 0.87]	0.87	[0.84, 0.9]
Sulcus intermedius primus	0.89	[0.86, 0.91]	0.89	[0.86, 0.91]	0.78	[0.74, 0.82]	0.79	[0.74, 0.83]
Superior frontal gyrus	0.92	[0.9, 0.94]	0.93	[0.91, 0.94]	0.92	[0.9, 0.93]	0.88	[0.85, 0.9]
Superior frontal sulcus	0.90	[0.88, 0.92]	0.92	[0.9, 0.93]	0.86	[0.83, 0.89]	0.79	[0.74, 0.83]
Superior occipital gyrus	0.94	[0.92, 0.95]	0.94	[0.93, 0.96]	0.90	[0.87, 0.92]	0.89	[0.86, 0.91]
Superior occipital sulcus and transverse occipi...	0.95	[0.94, 0.96]	0.94	[0.93, 0.96]	0.88	[0.85, 0.9]	0.90	[0.88, 0.92]
Superior parietal lobule	0.92	[0.9, 0.94]	0.92	[0.9, 0.94]	0.88	[0.85, 0.9]	0.87	[0.84, 0.89]
Superior part of the precentral sulcus	0.84	[0.8, 0.87]	0.87	[0.84, 0.89]	0.82	[0.78, 0.85]	0.73	[0.67, 0.78]
Superior segment of the circular sulcus of the ...	0.93	[0.91, 0.94]	0.91	[0.89, 0.93]	0.84	[0.81, 0.87]	0.81	[0.77, 0.85]
Superior temporal sulcus	0.96	[0.95, 0.97]	0.94	[0.92, 0.95]	0.89	[0.87, 0.92]	0.90	[0.88, 0.92]
Supramarginal gyrus	0.92	[0.9, 0.93]	0.87	[0.84, 0.89]	0.87	[0.84, 0.9]	0.84	[0.8, 0.87]
Temporal pole	0.78	[0.73, 0.82]	0.80	[0.75, 0.84]	0.79	[0.74, 0.83]	0.80	[0.75, 0.83]
Transverse frontopolar gyri and sulci	0.82	[0.78, 0.85]	0.81	[0.77, 0.84]	0.71	[0.65, 0.76]	0.73	[0.67, 0.78]
Transverse temporal sulcus	0.88	[0.86, 0.91]	0.88	[0.85, 0.9]	0.84	[0.81, 0.87]	0.84	[0.8, 0.87]
Triangular part of the inferior frontal gyrus	0.91	[0.89, 0.93]	0.83	[0.79, 0.86]	0.83	[0.79, 0.86]	0.81	[0.77, 0.85]
Vertical ramus of the anterior segment of the l...	0.89	[0.86, 0.91]	0.84	[0.8, 0.87]	0.83	[0.79, 0.86]	0.79	[0.74, 0.83]

Table A.7: Test-retest reliability: Cohen'd values for the paired t-test per ROI and framework. The * represents the ROIs in which the null hypothesis was rejected

Framework	CAT12		FREESURFER	
Hemisphere	Left	Right	Left	Right
ROI Name				
Angular gyrus	-0.01	0.0008	-0.008	-0.02
Anterior occipital sulcus and preoccipital notch	-0.007	-0.02	0.02	-0.02
Anterior part of the cingulate gyrus and sulcus	-0.02	-0.05	0.007	-0.01
Anterior segment of the circular sulcus of the ...	-0.02	-0.05	0.04	0.03
Anterior transverse collateral sulcus	0.06	0.05	0.07	0.08
Anterior transverse temporal gyrus	0.007	0.01	0.03	-0.02
Calcarine sulcus	-0.001	-0.02	0.04	-0.02
Central sulcus	-0.06	0.01	0.02	-0.07
Cuneus	-0.002	0.04	-0.02	0e+00
Fronto-marginal gyrus and sulcus	-0.01	0.01	-0.02	0.03
Horizontal ramus of the anterior segment of the...	-0.03	-0.03	0.06	-0.05
Inferior frontal sulcus	-0.04	-0.03	0.06	0.007
Inferior occipital gyrus and sulcus	-0.02	-0.04	0.05	0.04
Inferior part of the precentral sulcus	0.02	-0.08	-0.0008	0.003
Inferior segment of the circular sulcus of the ...	-0.009	0.04	0.03	0.04
Inferior temporal gyrus	0.01	-0.003	0.005	0.05
Inferior temporal sulcus	0.01	-0.001	0.04	0.01
Intraparietal sulcus and transverse parietal s...	0.0005	0.02	0.03	-0.02
Lateral aspect of the superior temporal gyrus	0.03	0.01	0.07	-0.01
Lateral occipito-temporal gyrus	0.01	0.01	0.08	0.05
Lateral occipito-temporal sulcus	0.01	-0.008	0.05	0.06
Lateral orbital sulcus	-0.02	-0.02	-0.004	0.02
Lingual gyrus	-0.02	-0.02	0.05	0.04
Long insular gyrus and central sulcus of the in...	-0.02	0.02	0.1	0.07
Marginal branch of the cingulate sulcus	0.04	0.03	0.03	0.009
Medial occipito-temporal sulcus and lingual su...	0.003	0.03	0.03	0.02
Medial orbital sulcus	0.004	0.01	0.01	0.05
Middle frontal gyrus	0.003	-0.02	0.03	-0.04
Middle frontal sulcus	-0.03	0.007	0.05	0.03
Middle occipital gyrus	-0.004	0.02	0.03	0.05
Middle occipital sulcus and lunatus sulcus	-0.03	-0.005	-0.02	-0.02
Middle temporal gyrus	0.03	0.01	0.04	0.03
Middle-anterior part of the cingulate gyrus and...	-0.02	0.01	0.03	0.1
Middle-posterior part of the cingulate gyrus an...	0.003	0.02	0.03	0.07
Occipital pole	0.003	-0.05	0.01	0.05
Opercular part of the inferior frontal gyrus	-0.01	-0.04	0.04	-0.05
Orbital gyri	0.001	0.03	0.04	0.03
Orbital part of the inferior frontal gyrus	-0.002	-0.009	0.04	-0.07
Orbital sulci	-0.01	0.01	0.1	0.06
Paracentral lobule and sulcus	-0.007	0.01	0.04	0.06
Parahippocampal gyrus	0.05	0.001	0.04	0.08
Parieto-occipital sulcus	0.03	-0.005	0.03	0.002
Pericallosal sulcus	-0.02	0.03	0.04	0.01
Planum polare of the superior temporal gyrus	0.1	0.05	0.1	0.02
Planum temporale or temporal plane of the super...	0.003	-0.04	-0.008	-0.02
Postcentral gyrus	-0.02	-0.01	-0.002	-0.06
Postcentral sulcus	-0.007	-0.04	-0.004	-0.02
Posterior ramus	0.02	0.0002	0.04	-0.02
Posterior transverse collateral sulcus	0.03	-0.04	0.01	0.006
Posterior-dorsal part of the cingulate gyrus	-0.05	-0.06	0.03	0.09
Posterior-ventral part of the cingulate gyrus	-0.03	0.02	0.005	0.06
Precentral gyrus	-0.007	-0.04	0.06	0.02
Precuneus	0.06	0.04	0.05	0.04
Short insular gyri	-0.05	-0.02	0.1*	0.09
Straight gyrus	-0.02	0.01	0.02	0.02
Subcallosal area	0.04	0.02	0.009	0.04
Subcentral gyrus and sulci	-0.02	-0.03	0.02	-0.03
Suborbital sulcus	-0.01	-0.05	-0.03	0.03
Subparietal sulcus	0.01	-0.03	0.01	-0.005
Sulcus intermedius primus	-0.01	-0.05	-0.03	-0.04
Superior frontal gyrus	0.008	0.02	0.04	0.07
Superior frontal sulcus	-0.02	0.002	0.02	-0.009
Superior occipital gyrus	-0.03	0.0004	0.03	0.01
Superior occipital sulcus and transverse occipi...	0.02	0.04	0.04	0.008
Superior parietal lobule	0.03	0.03	0.03	-0.009
Superior part of the precentral sulcus	0.04	-0.05	0.04	0.05
Superior segment of the circular sulcus of the ...	0.04	0.04	0.1	0.06
Superior temporal sulcus	0.003	-0.006	0.06	-0.02
Supramarginal gyrus	-0.02	-0.05	-0.006	-0.03
Temporal pole	0.06	0.04	0.1	0.1
Transverse frontopolar gyri and sulci	-0.02	0.004	-0.004	0.02
Transverse temporal sulcus	-0.01	-0.01	0.01	0.05
Triangular part of the inferior frontal gyrus	-0.04	-0.03	0.01	0.02
Vertical ramus of the anterior segment of the l...	0.003	0.02	0.1	0.01

Table A.8: Reproducibility in paediatric versus early adults' groups: ANCOVA for the participant R^2 values.

	df	sumsq	statistic	p-value	Cohen'd
Acquisition setting	16.00	3.05	16.30	0.00	0.69
Age	1.00	0.98	84.09	0.00	0.39
Age group	1.00	0.13	10.75	0.00	0.14
SNR	1.00	0.47	39.97	0.00	0.27
Residuals	546.00	6.38			

Table A.9: Reproducibility in paediatric versus early adults' groups: ANCOVA for the participant ICC values.

	df	sumsq	statistic	p-value	Cohen'd
Acquisition setting	16.00	1.64	13.59	0.00	0.63
Age	1.00	0.46	61.32	0.00	0.34
Age group	1.00	0.07	8.59	0.00	0.12
SNR	1.00	0.29	37.90	0.00	0.26
Residuals	546.00	4.12			

Table A.10: Reproducibility in paediatric versus early adults' groups: Mean R^2 per lobe and age group.

	R^2	
Age group	Adult	Paediatric
Area		
Frontal Lobe	0.35	0.28
Insula	0.24	0.16
Limbic lobe	0.37	0.27
Parietal lobe	0.49	0.41
Temporal and occipital lobes	0.41	0.31

Table A.11: Reproducibility in paediatric versus early adults' groups: R^2 of each ROI.

Age Group	Adult		Paediatric	
Hemisphere	Left	Right	Left	Right
ROI Name				
Angular gyrus	0.36	0.43	0.39	0.41
Anterior occipital sulcus and preoccipital notch	0.43	0.44	0.3	0.26
Anterior part of the cingulate gyrus and sulcus	0.26	0.15	0.17	0.16
Anterior segment of the circular sulcus of the ...	0.091	0.16	0.19	0.13
Anterior transverse collateral sulcus	0.067	0.06	0.017	0.073
Anterior transverse temporal gyrus	0.52	0.43	0.26	0.25
Calcarine sulcus	0.55	0.56	0.24	0.26
Central sulcus	0.3	0.23	0.36	0.23
Cuneus	0.59	0.56	0.31	0.36
Fronto-marginal gyrus and sulcus	0.34	0.34	0.26	0.19
Horizontal ramus of the anterior segment of the...	0.34	0.26	0.13	0.15
Inferior frontal sulcus	0.56	0.43	0.45	0.31
Inferior occipital gyrus and sulcus	0.37	0.43	0.21	0.26
Inferior part of the precentral sulcus	0.6	0.39	0.42	0.34
Inferior segment of the circular sulcus of the ...	0.25	0.32	0.14	0.16
Inferior temporal gyrus	0.38	0.34	0.29	0.35
Inferior temporal sulcus	0.13	0.13	0.19	0.15
Intraparietal sulcus and transverse parietal s...	0.51	0.55	0.45	0.4
Lateral aspect of the superior temporal gyrus	0.26	0.39	0.44	0.5
Lateral occipito-temporal gyrus	0.42	0.36	0.37	0.41
Lateral occipito-temporal sulcus	0.24	0.23	0.13	0.13
Lateral orbital sulcus	0.24	0.23	0.14	0.13
Lingual gyrus	0.59	0.47	0.24	0.35
Long insular gyrus and central sulcus of the in...	0.26	0.18	0.18	0.14
Marginal branch of the cingulate sulcus	0.63	0.63	0.37	0.39
Medial occipito-temporal sulcus and lingual su...	0.39	0.38	0.34	0.35
Medial orbital sulcus	0.045	0.06	0.023	0.067
Middle frontal gyrus	0.55	0.42	0.51	0.35
Middle frontal sulcus	0.36	0.21	0.23	0.22
Middle occipital gyrus	0.53	0.52	0.31	0.35
Middle occipital sulcus and lunatus sulcus	0.39	0.4	0.21	0.32
Middle temporal gyrus	0.29	0.41	0.45	0.44
Middle-anterior part of the cingulate gyrus and...	0.44	0.35	0.36	0.38
Middle-posterior part of the cingulate gyrus an...	0.6	0.47	0.48	0.41
Occipital pole	0.74	0.7	0.15	0.16
Opercular part of the inferior frontal gyrus	0.47	0.45	0.46	0.42
Orbital gyri	0.24	0.43	0.29	0.26
Orbital part of the inferior frontal gyrus	0.36	0.33	0.29	0.24
Orbital sulci	0.13	0.15	0.058	0.16
Paracentral lobule and sulcus	0.59	0.55	0.47	0.45
Parahippocampal gyrus	0.33	0.32	0.48	0.43
Parieto-occipital sulcus	0.52	0.59	0.39	0.25
Pericallosal sulcus	0.26	0.28	0.33	0.35
Planum polare of the superior temporal gyrus	0.22	0.17	0.38	0.42
Planum temporale or temporal plane of the super...	0.46	0.49	0.48	0.45
Postcentral gyrus	0.48	0.47	0.44	0.39
Postcentral sulcus	0.53	0.58	0.58	0.53
Posterior ramus	0.51	0.46	0.24	0.17
Posterior transverse collateral sulcus	0.37	0.41	0.16	0.18
Posterior-dorsal part of the cingulate gyrus	0.24	0.24	0.11	0.18
Posterior-ventral part of the cingulate gyrus	0.51	0.39	0.27	0.25
Precentral gyrus	0.54	0.41	0.54	0.37
Precuneus	0.57	0.6	0.42	0.39
Short insular gyri	0.19	0.2	0.29	0.18
Straight gyrus	0.4	0.14	0.092	0.087
Subcallosal area	0.0044	0.015	0.0062	0.035
Subcentral gyrus and sulci	0.53	0.5	0.4	0.39
Suborbital sulcus	0.11	0.099	0.12	0.095
Subparietal sulcus	0.49	0.51	0.22	0.18
Sulcus intermedius primus	0.3	0.26	0.27	0.17
Superior frontal gyrus	0.54	0.38	0.52	0.44
Superior frontal sulcus	0.51	0.34	0.53	0.4
Superior occipital gyrus	0.64	0.6	0.38	0.28
Superior occipital sulcus and transverse occipi...	0.46	0.42	0.42	0.29
Superior parietal lobule	0.45	0.48	0.49	0.46
Superior part of the precentral sulcus	0.42	0.41	0.44	0.32
Superior segment of the circular sulcus of the ...	0.16	0.21	0.089	0.053
Superior temporal sulcus	0.57	0.62	0.41	0.43
Supramarginal gyrus	0.5	0.48	0.51	0.52
Temporal pole	0.3	0.28	0.44	0.51
Transverse frontopolar gyri and sulci	0.38	0.43	0.31	0.3
Transverse temporal sulcus	0.48	0.41	0.33	0.094
Triangular part of the inferior frontal gyrus	0.38	0.39	0.38	0.28
Vertical ramus of the anterior segment of the l...	0.16	0.12	0.15	0.13

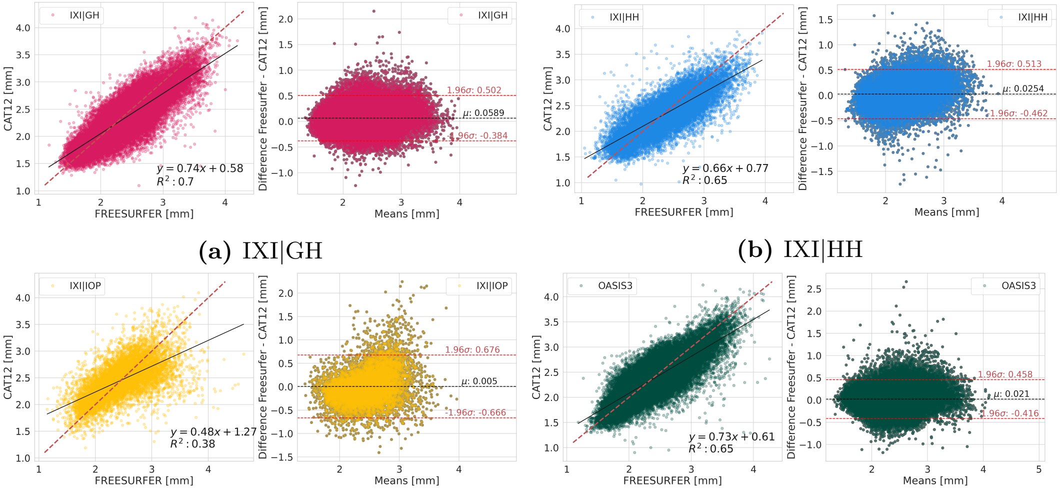
Table A.12: Reproducibility in paediatric versus early adults' groups: ICC and Confidence interval (CI) for each group and brain hemisphere.

Age group	Adult				Paediatric			
Hemisphere	Left		Right		Left		Right	
	ICC	CI95%	ICC	CI95%	ICC	CI95%	ICC	CI95%
ROI Name								
Angular gyrus	0.59	[0.48, 0.67]	0.65	[0.56, 0.73]	0.63	[0.56, 0.68]	0.63	[0.57, 0.69]
Anterior occipital sulcus and preoccipital notch	0.65	[0.56, 0.73]	0.66	[0.57, 0.73]	0.54	[0.47, 0.61]	0.51	[0.43, 0.58]
Anterior part of the cingulate gyrus and sulcus	0.51	[0.39, 0.61]	0.38	[0.25, 0.5]	0.42	[0.33, 0.5]	0.4	[0.31, 0.48]
Anterior segment of the circular sulcus of the ...	0.3	[0.16, 0.42]	0.4	[0.27, 0.52]	0.39	[0.31, 0.48]	0.34	[0.25, 0.43]
Anterior transverse collateral sulcus	0.26	[0.12, 0.39]	0.25	[0.1, 0.38]	0.13	[0.03, 0.23]	0.27	[0.17, 0.36]
Anterior transverse temporal gyrus	0.71	[0.63, 0.77]	0.65	[0.56, 0.73]	0.5	[0.42, 0.57]	0.48	[0.4, 0.55]
Calcarine sulcus	0.74	[0.67, 0.8]	0.75	[0.67, 0.8]	0.49	[0.41, 0.56]	0.51	[0.43, 0.58]
Central sulcus	0.5	[0.39, 0.6]	0.44	[0.32, 0.55]	0.6	[0.54, 0.66]	0.48	[0.4, 0.55]
Cuneus	0.77	[0.7, 0.82]	0.74	[0.67, 0.8]	0.52	[0.44, 0.59]	0.56	[0.49, 0.63]
Fronto-marginal gyrus and sulcus	0.57	[0.46, 0.66]	0.58	[0.47, 0.67]	0.51	[0.43, 0.58]	0.44	[0.35, 0.52]
Horizontal ramus of the anterior segment of the...	0.58	[0.47, 0.67]	0.51	[0.4, 0.61]	0.37	[0.28, 0.45]	0.38	[0.3, 0.47]
Inferior frontal sulcus	0.75	[0.67, 0.8]	0.66	[0.56, 0.73]	0.67	[0.61, 0.72]	0.55	[0.48, 0.62]
Inferior occipital gyrus and sulcus	0.6	[0.5, 0.69]	0.63	[0.53, 0.71]	0.46	[0.38, 0.53]	0.5	[0.42, 0.57]
Inferior part of the precentral sulcus	0.77	[0.71, 0.83]	0.63	[0.53, 0.71]	0.64	[0.58, 0.7]	0.59	[0.52, 0.65]
Inferior segment of the circular sulcus of the ...	0.48	[0.36, 0.58]	0.53	[0.42, 0.63]	0.31	[0.22, 0.4]	0.36	[0.27, 0.45]
Inferior temporal gyrus	0.59	[0.49, 0.68]	0.56	[0.45, 0.65]	0.53	[0.45, 0.6]	0.57	[0.49, 0.63]
Inferior temporal sulcus	0.36	[0.23, 0.48]	0.35	[0.22, 0.47]	0.41	[0.33, 0.49]	0.37	[0.28, 0.45]
Intraparietal sulcus and transverse parietal s...	0.7	[0.62, 0.77]	0.74	[0.66, 0.8]	0.66	[0.6, 0.71]	0.63	[0.57, 0.69]
Lateral aspect of the superior temporal gyrus	0.51	[0.39, 0.61]	0.62	[0.53, 0.7]	0.66	[0.6, 0.72]	0.7	[0.65, 0.75]
Lateral occipito-temporal gyrus	0.61	[0.51, 0.7]	0.57	[0.46, 0.66]	0.61	[0.54, 0.67]	0.64	[0.58, 0.7]
Lateral occipito-temporal sulcus	0.49	[0.38, 0.6]	0.48	[0.36, 0.59]	0.36	[0.27, 0.44]	0.36	[0.27, 0.44]
Lateral orbital sulcus	0.48	[0.36, 0.58]	0.48	[0.36, 0.59]	0.37	[0.28, 0.45]	0.36	[0.27, 0.45]
Lingual gyrus	0.74	[0.67, 0.8]	0.64	[0.54, 0.72]	0.47	[0.39, 0.54]	0.57	[0.5, 0.64]
Long insular gyrus and central sulcus of the in...	0.51	[0.39, 0.61]	0.43	[0.3, 0.54]	0.38	[0.29, 0.46]	0.37	[0.28, 0.46]
Marginal branch of the cingulate sulcus	0.79	[0.73, 0.84]	0.79	[0.73, 0.84]	0.61	[0.54, 0.67]	0.63	[0.56, 0.68]
Medial occipito-temporal sulcus and lingual su...	0.63	[0.53, 0.71]	0.61	[0.51, 0.7]	0.58	[0.51, 0.64]	0.58	[0.51, 0.64]
Medial orbital sulcus	0.21	[0.07, 0.35]	0.24	[0.1, 0.38]	0.15	[0.05, 0.25]	0.25	[0.16, 0.35]
Middle frontal gyrus	0.74	[0.66, 0.8]	0.65	[0.55, 0.72]	0.69	[0.63, 0.74]	0.57	[0.5, 0.64]
Middle frontal sulcus	0.6	[0.5, 0.69]	0.46	[0.34, 0.57]	0.47	[0.39, 0.55]	0.46	[0.38, 0.54]
Middle occipital gyrus	0.64	[0.55, 0.72]	0.69	[0.6, 0.76]	0.55	[0.48, 0.62]	0.59	[0.52, 0.65]
Middle occipital sulcus and lunatus sulcus	0.61	[0.52, 0.7]	0.62	[0.52, 0.7]	0.46	[0.37, 0.53]	0.56	[0.49, 0.63]
Middle temporal gyrus	0.53	[0.42, 0.63]	0.64	[0.55, 0.72]	0.67	[0.61, 0.72]	0.65	[0.59, 0.71]
Middle-anterior part of the cingulate gyrus and...	0.66	[0.57, 0.73]	0.58	[0.47, 0.67]	0.59	[0.52, 0.65]	0.6	[0.54, 0.66]
Middle-posterior part of the cingulate gyrus an...	0.77	[0.71, 0.83]	0.69	[0.6, 0.76]	0.69	[0.64, 0.74]	0.64	[0.57, 0.69]
Occipital pole	0.81	[0.76, 0.86]	0.78	[0.71, 0.83]	0.36	[0.27, 0.44]	0.39	[0.3, 0.47]
Opercular part of the inferior frontal gyrus	0.68	[0.59, 0.75]	0.67	[0.58, 0.74]	0.66	[0.6, 0.71]	0.63	[0.57, 0.69]
Orbital gyri	0.48	[0.36, 0.59]	0.64	[0.55, 0.72]	0.54	[0.47, 0.61]	0.51	[0.44, 0.58]
Orbital part of the inferior frontal gyrus	0.6	[0.5, 0.69]	0.57	[0.47, 0.66]	0.53	[0.46, 0.6]	0.47	[0.39, 0.54]
Orbital sulci	0.35	[0.22, 0.47]	0.38	[0.25, 0.5]	0.23	[0.13, 0.32]	0.4	[0.31, 0.48]
Paracentral lobule and sulcus	0.76	[0.69, 0.82]	0.74	[0.66, 0.8]	0.68	[0.63, 0.73]	0.67	[0.61, 0.72]
Parahippocampal gyrus	0.58	[0.47, 0.67]	0.56	[0.46, 0.66]	0.69	[0.63, 0.74]	0.65	[0.59, 0.71]
Parieto-occipital sulcus	0.72	[0.64, 0.78]	0.77	[0.7, 0.82]	0.62	[0.55, 0.68]	0.5	[0.42, 0.57]
Pericallosal sulcus	0.47	[0.35, 0.57]	0.48	[0.36, 0.58]	0.55	[0.47, 0.61]	0.55	[0.47, 0.61]
Planum polare of the superior temporal gyrus	0.46	[0.34, 0.57]	0.4	[0.27, 0.52]	0.62	[0.55, 0.68]	0.65	[0.59, 0.7]
Planum temporale or temporal plane of the super...	0.67	[0.58, 0.74]	0.69	[0.6, 0.76]	0.69	[0.64, 0.74]	0.67	[0.61, 0.72]
Postcentral gyrus	0.67	[0.58, 0.74]	0.67	[0.58, 0.74]	0.67	[0.61, 0.72]	0.62	[0.56, 0.68]
Postcentral sulcus	0.73	[0.65, 0.79]	0.76	[0.69, 0.82]	0.76	[0.72, 0.8]	0.73	[0.68, 0.77]
Posterior ramus	0.7	[0.62, 0.77]	0.62	[0.52, 0.7]	0.48	[0.4, 0.55]	0.35	[0.25, 0.43]
Posterior transverse collateral sulcus	0.59	[0.48, 0.67]	0.61	[0.52, 0.7]	0.4	[0.31, 0.48]	0.43	[0.34, 0.51]
Posterior-dorsal part of the cingulate gyrus	0.49	[0.37, 0.59]	0.49	[0.37, 0.59]	0.34	[0.24, 0.42]	0.42	[0.34, 0.5]
Posterior-ventral part of the cingulate gyrus	0.71	[0.63, 0.78]	0.62	[0.52, 0.7]	0.52	[0.44, 0.59]	0.5	[0.42, 0.57]
Precentral gyrus	0.73	[0.65, 0.79]	0.63	[0.53, 0.71]	0.73	[0.68, 0.78]	0.61	[0.54, 0.67]
Precuneus	0.72	[0.64, 0.78]	0.76	[0.69, 0.81]	0.64	[0.58, 0.7]	0.63	[0.56, 0.68]
Short insular gyri	0.42	[0.29, 0.53]	0.43	[0.3, 0.54]	0.5	[0.43, 0.58]	0.42	[0.33, 0.5]
Straight gyrus	0.62	[0.52, 0.7]	0.36	[0.22, 0.48]	0.3	[0.21, 0.39]	0.29	[0.2, 0.38]
Subcallosal area	-0.066	[-0.21, 0.08]	0.12	[-0.02, 0.26]	0.079	[-0.02, 0.18]	0.19	[0.09, 0.28]
Subcentral gyrus and sulci	0.71	[0.63, 0.78]	0.7	[0.62, 0.77]	0.63	[0.57, 0.69]	0.62	[0.56, 0.68]
Suborbital sulcus	0.32	[0.18, 0.45]	0.31	[0.17, 0.43]	0.34	[0.25, 0.43]	0.29	[0.2, 0.38]
Subparietal sulcus	0.7	[0.61, 0.77]	0.72	[0.64, 0.78]	0.46	[0.38, 0.54]	0.42	[0.34, 0.5]
Sulcus intermedius primus	0.53	[0.42, 0.63]	0.51	[0.39, 0.61]	0.5	[0.42, 0.57]	0.41	[0.32, 0.49]
Superior frontal gyrus	0.73	[0.66, 0.79]	0.62	[0.52, 0.7]	0.7	[0.65, 0.75]	0.65	[0.59, 0.7]
Superior frontal sulcus	0.72	[0.64, 0.78]	0.58	[0.48, 0.67]	0.73	[0.67, 0.77]	0.63	[0.56, 0.68]
Superior occipital gyrus	0.74	[0.67, 0.8]	0.73	[0.66, 0.79]	0.6	[0.53, 0.66]	0.52	[0.45, 0.59]
Superior occipital sulcus and transverse occipi...	0.67	[0.58, 0.74]	0.63	[0.54, 0.71]	0.64	[0.58, 0.7]	0.54	[0.46, 0.6]
Superior parietal lobule	0.66	[0.56, 0.73]	0.69	[0.6, 0.76]	0.7	[0.64, 0.75]	0.67	[0.61, 0.72]
Superior part of the precentral sulcus	0.65	[0.56, 0.73]	0.64	[0.54, 0.71]	0.66	[0.6, 0.72]	0.57	[0.5, 0.63]
Superior segment of the circular sulcus of the ...	0.36	[0.23, 0.48]	0.44	[0.31, 0.55]	0.26	[0.17, 0.35]	0.21	[0.11, 0.3]
Superior temporal sulcus	0.75	[0.68, 0.81]	0.79	[0.72, 0.84]	0.63	[0.57, 0.69]	0.65	[0.59, 0.71]
Supramarginal gyrus	0.7	[0.62, 0.77]	0.69	[0.61, 0.76]	0.71	[0.66, 0.76]	0.71	[0.66, 0.76]
Temporal pole	0.55	[0.44, 0.64]	0.53	[0.41, 0.62]	0.66	[0.6, 0.71]	0.71	[0.66, 0.76]
Transverse frontopolar gyri and sulci	0.62	[0.52, 0.7]	0.65	[0.56, 0.73]	0.55	[0.48, 0.62]	0.53	[0.46, 0.6]
Transverse temporal sulcus	0.69	[0.6, 0.76]	0.61	[0.51, 0.69]	0.56	[0.49, 0.63]	0.31	[0.21, 0.39]
Triangular part of the inferior frontal gyrus	0.62	[0.52, 0.7]	0.62	[0.53, 0.71]	0.61	[0.54, 0.67]	0.52	[0.44, 0.59]
Vertical ramus of the anterior segment of the l...	0.4	[0.27, 0.52]	0.35	[0.21, 0.47]	0.38	[0.29, 0.46]	0.36	[0.27, 0.44]

Table A.13: Reproducibility in paediatric versus early adults' groups: Cohen'd value for the paired *t*-test between the cortical thickness estimations of the FreeSurfer and CAT12 per ROI. The * represents the ROIs in which the null hypothesis was rejected.

Age group	Adult		Paediatric	
Hemisphere	Left	Right	Left	Right
ROI Name				
Angular gyrus	0.45*	0.73*	0.85*	0.9*
Anterior occipital sulcus and preoccipital notch	0.08	0.02	0.04	0.06
Anterior part of the cingulate gyrus and sulcus	0.16	0.2	0.34*	0.14
Anterior segment of the circular sulcus of the ...	0.08	0.17	0.09	0.45*
Anterior transverse collateral sulcus	0.28	0.38*	0.45*	0.48*
Anterior transverse temporal gyrus	0.12	0.0	0.39*	0.36*
Calcarine sulcus	1.24*	0.93*	0.92*	0.58*
Central sulcus	0.04	0.17	0.97*	1.09*
Cuneus	0.72*	0.85*	0.68*	0.69*
Fronto-marginal gyrus and sulcus	0.07	0.18	0.29*	0.31*
Horizontal ramus of the anterior segment of the...	0.68*	1.08*	0.4*	0.52*
Inferior frontal sulcus	0.84*	1.13*	0.78*	0.87*
Inferior occipital gyrus and sulcus	0.38*	0.66*	0.79*	1.04*
Inferior part of the precentral sulcus	0.54*	0.62*	0.56*	0.67*
Inferior segment of the circular sulcus of the ...	0.3*	0.34*	0.27*	0.17
Inferior temporal gyrus	1.26*	1.4*	1.32*	1.21*
Inferior temporal sulcus	0.09	0.09	0.01	0.01
Intraparietal sulcus and transverse parietal s...	0.58*	0.52*	0.51*	0.35*
Lateral aspect of the superior temporal gyrus	1.76*	1.22*	1.58*	1.22*
Lateral occipito-temporal gyrus	1.71*	1.9*	1.97*	2.05*
Lateral occipito-temporal sulcus	0.14	0.32*	0.3*	0.4*
Lateral orbital sulcus	0.78*	0.95*	0.56*	0.54*
Lingual gyrus	0.15	0.25*	0.38*	0.34*
Long insular gyrus and central sulcus of the in...	0.39*	0.87*	0.51*	0.69*
Marginal branch of the cingulate sulcus	0.77*	0.73*	0.65*	0.56*
Medial occipito-temporal sulcus and lingual su...	0.3*	0.6*	0.35*	0.53*
Medial orbital sulcus	0.18	0.0	0.16	0.18
Middle frontal gyrus	0.1	0.22	0.28*	0.53*
Middle frontal sulcus	1.17*	1.29*	0.7*	0.75*
Middle occipital gyrus	0.26*	0.74*	0.9*	1.22*
Middle occipital sulcus and lunatus sulcus	0.5*	0.55*	0.44*	0.38*
Middle temporal gyrus	1.04*	0.93*	0.92*	0.77*
Middle-anterior part of the cingulate gyrus and...	0.74*	0.51*	0.34*	0.22*
Middle-posterior part of the cingulate gyrus an...	0.16	0.3*	0.06	0.09
Occipital pole	0.09	0.13	0.47*	0.73*
Opercular part of the inferior frontal gyrus	0.39*	0.21	0.75*	0.76*
Orbital gyri	0.44*	0.68*	1.03*	1.11*
Orbital part of the inferior frontal gyrus	0.85*	1.07*	1.32*	1.31*
Orbital sulci	0.59*	0.14	0.99*	0.4*
Paracentral lobule and sulcus	0.58*	0.76*	0.4*	0.58*
Parahippocampal gyrus	1.01*	1.1*	0.9*	1.07*
Parieto-occipital sulcus	0.33*	0.51*	0.36*	0.45*
Pericallosal sulcus	0.32*	0.17	0.41*	0.23*
Planum polare of the superior temporal gyrus	1.39*	0.64*	0.86*	0.46*
Planum temporale or temporal plane of the super...	0.17	0.3*	0.35*	0.46*
Postcentral gyrus	0.59*	0.45*	0.36*	0.26*
Postcentral sulcus	0.49*	0.44*	0.49*	0.42*
Posterior ramus	0.22*	0.47*	0.03	0.08
Posterior transverse collateral sulcus	0.19	0.24	0.08	0.4*
Posterior-dorsal part of the cingulate gyrus	1.75*	1.5*	2.29*	1.86*
Posterior-ventral part of the cingulate gyrus	0.9*	1.12*	1.03*	1.15*
Precentral gyrus	0.81*	0.84*	0.77*	0.73*
Precuneus	0.4*	0.32*	0.89*	0.75*
Short insular gyri	0.73*	0.78*	0.51*	0.54*
Straight gyrus	1.04*	0.62*	1.14*	0.89*
Subcallosal area	0.28	0.36	0.03	0.33*
Subcentral gyrus and sulci	0.5*	0.51*	0.76*	0.7*
Suborbital sulcus	0.18	0.26	0.29*	0.44*
Subparietal sulcus	0.37*	0.21	0.1	0.23*
Sulcus intermedius primus	0.33*	0.37*	0.07	0.3*
Superior frontal gyrus	0.15	0.19	0.36*	0.38*
Superior frontal sulcus	0.85*	1.1*	0.71*	0.82*
Superior occipital gyrus	0.01	0.0	0.39*	0.34*
Superior occipital sulcus and transverse occipi...	0.4*	0.24*	0.42*	0.32*
Superior parietal lobule	0.24*	0.43*	0.39*	0.43*
Superior part of the precentral sulcus	0.43*	0.56*	0.65*	0.7*
Superior segment of the circular sulcus of the ...	1.21*	1.26*	0.63*	0.62*
Superior temporal sulcus	0.32*	0.52*	0.29*	0.31*
Supramarginal gyrus	0.42*	0.68*	0.64*	0.81*
Temporal pole	0.9*	0.93*	0.75*	0.74*
Transverse frontopolar gyri and sulci	0.68*	0.49*	1.0*	0.74*
Transverse temporal sulcus	0.26*	0.52*	0.56*	0.82*
Triangular part of the inferior frontal gyrus	0.56*	0.38*	1.02*	0.76*
Vertical ramus of the anterior segment of the l...	0.9*	0.45*	0.55*	0.16

Figure A.1: Regression and Bland-Altman for the adult dataset with different acquisition settings. On the left the distribution of cortical thickness extracted with FreeSurfer and CAT12. The dashed red line on the left plot represents the equation y=x. On the right is the plot with the mean and difference (FreeSurfer – CAT12) between the cortical thickness estimates of the frameworks. In this plot each point corresponds to a participant's ROI.



(a) IXI|GH

(b) IXI|HH

(c) IXI|IOP

(d) IXI|OASIS3

Table A.14: Reproducibility in different acquisition settings: ANCOVA for the participant' R^2 values.

	df	sumsq	statistic	*p*-value	Cohen'd
Acquisition settting	3.00	5.39	329.03	0.00	0.97
Age	1.00	0.07	13.71	0.00	0.12
SNR	1.00	0.23	42.20	0.00	0.20
Residuals	1043.00	5.70			

Table A.15: Reproducibility in different acquisition settings: ANCOVA pos-hoc analysis comparing the participant' R^2s of different sites.

	Estimate	Std. Error	*t-value*	*p*-value
IXI|HH-IXI|GH	-0.076	0.009	-8.86	<0,0001
IXI|IOP-IXI|GH	-0.347	0.011	-31.37	<0,0001
OASIS3-IXI|GH	0.062	0.016	3.98	0.000274
IXI|IOP-IXI|HH	-0.271	0.011	-25.49	<0,0001
OASIS3-IXI|HH	0.138	0.020	6.67	<0,0001
OASIS3-IXI|IOP	0.409	0.022	18.85	<0,0001

Table A.16: Reproducibility in different acquisition settings: ANCOVA for the participant' ICC values.

	df	sumsq	statistic	*p*-value	Cohen'd
Acquisition setting	3.00	2.95	311.52	0.00	0.95
Age	1.00	0.01	3.14	0.08	0.06
SNR	1.00	0.16	49.78	0.00	0.22
Residuals	1043.00	3.29			

Table A.17: Reproducibility in different acquisition settings: ANCOVA pos-hoc analysis comparing the participant' ICC of different acquisition settings.

	Estimate	Std. Error	*t value*	*p*-value
IXI|HH-IXI|GH	-0.062	0.006	-9.61	<0,00001
IXI|IOP-IXI|GH	-0.256	0.008	-30.42	<0,00001
OASIS3-IXI|GH	0.056	0.012	4.77	0,000011
IXI|IOP-IXI|HH	-0.193	0.008	-23.90	<0,00001
OASIS3-IXI|HH	0.119	0.016	7.57	<0,00001
OASIS3-IXI|IOP	0.312	0.016	18.93	<0,00001

Table A.18: Reproducibility in different acquisition settings: Comparison of the slope^2_R for different acquisition settings.

	t-value	p-value	Cohen'd
IXI|GH-IXI|HH	-18.62	1.52e-61	-1.75
IXI|GH-IXI|IOP	-9.61	3.56e-20	-1.29
IXI|GH-OASIS3	-72.64	1.02e-305	-5.26
IXI|HH-IXI|IOP	0.70	9.63e-01	0.10

Table A.19: Reproducibility in different acquisition settings: Comparison of the slope$_{ICC}$ for different acquisition settings.

	t-value	p-value	*Cohen'd*
IXI|GH-IXI|HH	-19.56	1.84e-66	-1.83
IXI|GH-IXI|IOP	-2.71	1.40e-02	-0.36
IXI|GH-OASIS3	-66.99	2.38e-286	-4.85
IXI|HH-IXI|IOP	5.44	2.02e-07	0.78

Table A.20: Reproducibility in different acquisition settings: Mean R^2 per lobe and site.

	R^2			
Acquisition setting	IXI|GH	IXI|HH	IXI|IOP	OASIS3
Area				
Frontal Lobe	0.40	0.42	0.26	0.33
Insula	0.34	0.23	0.20	0.37
Limbic lobe	0.32	0.32	0.19	0.31
Parietal lobe	0.51	0.49	0.30	0.57
Temporal and occipital lobes	0.41	0.43	0.25	0.42

Table A.21: Reproducibility in different acquisition settings: R^2 for each for and acquisition setting.

Acquisition setting	IXI|GH		IXI|HH		IXI|IOP		OASIS3	
Hemisphere	Left	Right	Left	Right	Left	Right	Left	Right
ROI Name								
Angular gyrus	0.63	0.63	0.69	0.61	0.43	0.36	0.58	0.69
Anterior occipital sulcus and preoccipital notch	0.49	0.54	0.37	0.52	0.5	0.39	0.35	0.53
Anterior part of the cingulate gyrus and sulcus	0.35	0.33	0.32	0.43	0.11	0.14	0.19	0.2
Anterior segment of the circular sulcus of the ...	0.11	0.25	0.089	0.17	0.057	0.012	0.23	0.22
Anterior transverse collateral sulcus	0.0037	0.013	0.015	0.02	0.0024	0.00022	5.2e-05	0.0056
Anterior transverse temporal gyrus	0.24	0.17	0.28	0.16	0.21	0.41	0.51	0.38
Calcarine sulcus	0.35	0.42	0.42	0.35	0.23	0.46	0.39	0.43
Central sulcus	0.2	0.18	0.27	0.26	0.57	0.41	0.33	0.31
Cuneus	0.19	0.3	0.27	0.25	0.16	0.2	0.34	0.39
Fronto-marginal gyrus and sulcus	0.29	0.26	0.22	0.3	0.095	0.095	0.15	0.061
Horizontal ramus of the anterior segment of the...	0.26	0.3	0.32	0.39	0.15	0.31	0.29	0.14
Inferior frontal sulcus	0.56	0.55	0.49	0.7	0.4	0.3	0.42	0.31
Inferior occipital gyrus and sulcus	0.43	0.5	0.56	0.55	0.16	0.22	0.35	0.39
Inferior part of the precentral sulcus	0.56	0.6	0.56	0.59	0.27	0.39	0.65	0.44
Inferior segment of the circular sulcus of the ...	0.47	0.36	0.31	0.11	0.25	0.33	0.67	0.53
Inferior temporal gyrus	0.44	0.44	0.43	0.4	0.15	0.093	0.31	0.38
Inferior temporal sulcus	0.3	0.39	0.23	0.25	0.24	0.093	0.17	0.2
Intraparietal sulcus and transverse parietal s...	0.58	0.57	0.41	0.55	0.11	0.15	0.59	0.7
Lateral aspect of the superior temporal gyrus	0.47	0.5	0.41	0.49	0.28	0.23	0.57	0.63
Lateral occipito-temporal gyrus	0.45	0.32	0.34	0.37	0.035	0.079	0.28	0.28
Lateral occipito-temporal sulcus	0.27	0.22	0.32	0.18	0.21	0.23	0.079	0.081
Lateral orbital sulcus	0.23	0.19	0.25	0.26	0.038	0.09	0.19	0.009
Lingual gyrus	0.28	0.38	0.3	0.36	0.069	0.091	0.42	0.39
Long insular gyrus and central sulcus of the in...	0.24	0.17	0.072	0.091	0.088	0.073	0.28	0.14
Marginal branch of the cingulate sulcus	0.54	0.43	0.5	0.52	0.19	0.11	0.6	0.52
Medial occipito-temporal sulcus and lingual su...	0.47	0.5	0.6	0.54	0.22	0.44	0.33	0.45
Medial orbital sulcus	0.0081	0.082	0.028	0.05	0.049	0.21	0.021	0.0012
Middle frontal gyrus	0.7	0.73	0.74	0.72	0.43	0.12	0.57	0.38
Middle frontal sulcus	0.39	0.36	0.23	0.48	0.15	0.12	0.29	0.16
Middle occipital gyrus	0.59	0.58	0.59	0.62	0.53	0.29	0.6	0.61
Middle occipital sulcus and lunatus sulcus	0.52	0.47	0.53	0.42	0.45	0.34	0.39	0.53
Middle temporal gyrus	0.62	0.64	0.58	0.52	0.47	0.14	0.61	0.66
Middle-anterior part of the cingulate gyrus and...	0.35	0.39	0.41	0.38	0.15	0.23	0.51	0.45
Middle-posterior part of the cingulate gyrus an...	0.53	0.4	0.51	0.43	0.24	0.43	0.49	0.43
Occipital pole	0.38	0.38	0.35	0.4	0.01	0.15	0.45	0.45
Opercular part of the inferior frontal gyrus	0.68	0.72	0.69	0.68	0.56	0.58	0.72	0.61
Orbital gyri	0.35	0.44	0.32	0.41	0.31	0.27	0.32	0.13
Orbital part of the inferior frontal gyrus	0.4	0.4	0.38	0.44	0.31	0.34	0.42	0.18
Orbital sulci	0.12	0.14	0.33	0.27	0.0081	0.19	0.098	0.089
Paracentral lobule and sulcus	0.21	0.29	0.26	0.35	0.22	0.3	0.4	0.35
Parahippocampal gyrus	0.44	0.34	0.37	0.46	0.3	0.15	0.49	0.4
Parieto-occipital sulcus	0.52	0.54	0.6	0.58	0.39	0.09	0.6	0.53
Pericallosal sulcus	0.3	0.18	0.099	0.06	0.19	0.12	0.23	0.15
Planum polare of the superior temporal gyrus	0.04	0.1	0.12	0.14	0.04	0.074	0.27	0.34
Planum temporale or temporal plane of the super...	0.6	0.63	0.56	0.61	0.5	0.43	0.67	0.72
Postcentral gyrus	0.44	0.45	0.47	0.51	0.39	0.42	0.65	0.61
Postcentral sulcus	0.64	0.65	0.63	0.7	0.29	0.39	0.7	0.69
Posterior ramus	0.7	0.62	0.54	0.26	0.3	0.57	0.68	0.67
Posterior transverse collateral sulcus	0.3	0.36	0.36	0.39	0.16	0.32	0.3	0.28
Posterior-dorsal part of the cingulate gyrus	0.14	0.14	0.16	0.17	2.9e-05	0.13	0.11	0.27
Posterior-ventral part of the cingulate gyrus	0.35	0.4	0.24	0.3	0.24	0.011	0.36	0.41
Precentral gyrus	0.29	0.3	0.44	0.3	0.2	0.22	0.52	0.42
Precuneus	0.66	0.65	0.52	0.58	0.23	0.36	0.62	0.65
Short insular gyri	0.33	0.21	0.067	0.11	0.11	0.046	0.38	0.17
Straight gyrus	0.2	0.21	0.23	0.23	0.008	0.015	0.11	0.18
Subcallosal area	0.00071	0.0061	0.0057	0.005	0.0031	0.00073	7.7e-09	0.0086
Subcentral gyrus and sulci	0.58	0.61	0.58	0.42	0.49	0.48	0.67	0.63
Suborbital sulcus	0.1	0.053	0.21	0.083	0.079	0.0039	0.037	0.047
Subparietal sulcus	0.25	0.31	0.3	0.33	0.16	0.24	0.25	0.24
Sulcus intermedius primus	0.15	0.24	0.31	0.15	0.18	0.079	0.25	0.28
Superior frontal gyrus	0.69	0.69	0.69	0.67	0.58	0.6	0.68	0.64
Superior frontal sulcus	0.64	0.63	0.54	0.56	0.4	0.21	0.62	0.57
Superior occipital gyrus	0.51	0.46	0.57	0.56	0.34	0.26	0.63	0.66
Superior occipital sulcus and transverse occipi...	0.6	0.58	0.54	0.55	0.37	0.23	0.54	0.56
Superior parietal lobule	0.68	0.65	0.59	0.56	0.21	0.39	0.72	0.67
Superior part of the precentral sulcus	0.48	0.62	0.61	0.54	0.14	0.24	0.56	0.32
Superior segment of the circular sulcus of the ...	0.51	0.44	0.43	0.29	0.29	0.27	0.53	0.38
Superior temporal sulcus	0.63	0.64	0.64	0.61	0.45	0.43	0.6	0.68
Supramarginal gyrus	0.65	0.65	0.67	0.66	0.57	0.48	0.69	0.74
Temporal pole	0.18	0.32	0.5	0.37	0.19	0.095	0.21	0.27
Transverse frontopolar gyri and sulci	0.33	0.47	0.44	0.52	0.095	0.28	0.19	0.12
Transverse temporal sulcus	0.51	0.25	0.42	0.2	0.5	0.13	0.37	0.32
Triangular part of the inferior frontal gyrus	0.63	0.67	0.59	0.58	0.56	0.48	0.63	0.28
Vertical ramus of the anterior segment of the l...	0.28	0.24	0.26	0.24	0.2	0.19	0.35	0.18

Table A.22: Reproducibility in different acquisition settings: ICC and Confidence interval (CI) for the IXI-GH and IXI-HH acquisition settings and brain hemisphere.

Acquisition setting	IXI|GH				IXI|HH			
Hemisphere	Left		Right		Left		Right	
	ICC	CI95%	ICC	CI95%	ICC	CI95%	ICC	CI95%
ROI Name								
Angular gyrus	0.78	[0.73, 0.82]	0.79	[0.74, 0.82]	0.82	[0.76, 0.86]	0.77	[0.71, 0.83]
Anterior occipital sulcus and preoccipital notch	0.7	[0.64, 0.75]	0.73	[0.68, 0.78]	0.58	[0.47, 0.67]	0.71	[0.63, 0.78]
Anterior part of the cingulate gyrus and sulcus	0.59	[0.51, 0.66]	0.57	[0.5, 0.64]	0.55	[0.44, 0.65]	0.65	[0.55, 0.72]
Anterior segment of the circular sulcus of the ...	0.33	[0.22, 0.42]	0.5	[0.41, 0.58]	0.3	[0.16, 0.43]	0.39	[0.26, 0.51]
Anterior transverse collateral sulcus	0.06	[-0.05, 0.17]	0.11	[0.0, 0.22]	0.12	[-0.02, 0.26]	0.14	[-0.01, 0.28]
Anterior transverse temporal gyrus	0.49	[0.4, 0.57]	0.42	[0.32, 0.5]	0.51	[0.39, 0.61]	0.4	[0.27, 0.51]
Calcarine sulcus	0.59	[0.51, 0.65]	0.65	[0.58, 0.71]	0.65	[0.55, 0.72]	0.59	[0.49, 0.68]
Central sulcus	0.44	[0.35, 0.53]	0.41	[0.32, 0.5]	0.52	[0.4, 0.62]	0.51	[0.39, 0.61]
Cuneus	0.41	[0.31, 0.5]	0.54	[0.46, 0.62]	0.5	[0.38, 0.6]	0.49	[0.37, 0.6]
Fronto-marginal gyrus and sulcus	0.53	[0.44, 0.6]	0.51	[0.42, 0.59]	0.47	[0.35, 0.58]	0.54	[0.43, 0.64]
Horizontal ramus of the anterior segment of the...	0.51	[0.42, 0.58]	0.55	[0.46, 0.62]	0.56	[0.45, 0.65]	0.62	[0.52, 0.7]
Inferior frontal sulcus	0.75	[0.69, 0.79]	0.74	[0.68, 0.78]	0.7	[0.62, 0.77]	0.82	[0.77, 0.87]
Inferior occipital gyrus and sulcus	0.65	[0.59, 0.71]	0.68	[0.62, 0.74]	0.74	[0.67, 0.80]	0.72	[0.64, 0.78]
Inferior part of the precentral sulcus	0.74	[0.68, 0.79]	0.76	[0.71, 0.8]	0.74	[0.66, 0.8]	0.74	[0.67, 0.8]
Inferior segment of the circular sulcus of the ...	0.68	[0.61, 0.73]	0.6	[0.53, 0.67]	0.56	[0.45, 0.65]	0.33	[0.19, 0.45]
Inferior temporal gyrus	0.67	[0.6, 0.72]	0.66	[0.59, 0.72]	0.64	[0.55, 0.72]	0.63	[0.53, 0.71]
Inferior temporal sulcus	0.54	[0.46, 0.61]	0.62	[0.55, 0.69]	0.48	[0.36, 0.58]	0.5	[0.38, 0.6]
Intraparietal sulcus and transverse parietal s...	0.76	[0.71, 0.8]	0.74	[0.69, 0.79]	0.6	[0.5, 0.69]	0.74	[0.67, 0.8]
Lateral aspect of the superior temporal gyrus	0.69	[0.62, 0.74]	0.71	[0.65, 0.76]	0.63	[0.53, 0.71]	0.67	[0.58, 0.75]
Lateral occipito-temporal gyrus	0.66	[0.6, 0.72]	0.56	[0.48, 0.64]	0.5	[0.39, 0.61]	0.57	[0.46, 0.66]
Lateral occipito-temporal sulcus	0.52	[0.43, 0.59]	0.47	[0.38, 0.55]	0.55	[0.44, 0.65]	0.42	[0.3, 0.54]
Lateral orbital sulcus	0.47	[0.38, 0.55]	0.43	[0.34, 0.52]	0.5	[0.38, 0.6]	0.5	[0.38, 0.6]
Lingual gyrus	0.5	[0.42, 0.58]	0.61	[0.53, 0.67]	0.51	[0.4, 0.61]	0.58	[0.47, 0.67]
Long insular gyrus and central sulcus of the in...	0.49	[0.4, 0.57]	0.4	[0.3, 0.49]	0.27	[0.13, 0.4]	0.3	[0.16, 0.43]
Marginal branch of the cingulate sulcus	0.73	[0.67, 0.78]	0.65	[0.58, 0.71]	0.7	[0.62, 0.77]	0.7	[0.62, 0.77]
Medial occipito-temporal sulcus and lingual su...	0.68	[0.62, 0.74]	0.71	[0.65, 0.76]	0.74	[0.66, 0.8]	0.73	[0.65, 0.79]
Medial orbital sulcus	0.085	[-0.03, 0.19]	0.28	[0.17, 0.38]	0.17	[0.02, 0.31]	0.22	[0.08, 0.36]
Middle frontal gyrus	0.83	[0.8, 0.86]	0.85	[0.82, 0.88]	0.86	[0.81, 0.89]	0.85	[0.8, 0.88]
Middle frontal sulcus	0.62	[0.54, 0.68]	0.58	[0.51, 0.65]	0.46	[0.34, 0.57]	0.68	[0.6, 0.75]
Middle occipital gyrus	0.77	[0.72, 0.81]	0.76	[0.71, 0.8]	0.75	[0.68, 0.81]	0.77	[0.7, 0.82]
Middle occipital sulcus and lunatus sulcus	0.71	[0.65, 0.76]	0.69	[0.62, 0.74]	0.72	[0.64, 0.78]	0.64	[0.55, 0.72]
Middle temporal gyrus	0.79	[0.74, 0.83]	0.8	[0.75, 0.83]	0.75	[0.68, 0.81]	0.71	[0.63, 0.78]
Middle-anterior part of the cingulate gyrus and...	0.59	[0.51, 0.66]	0.62	[0.55, 0.69]	0.63	[0.54, 0.71]	0.61	[0.5, 0.69]
Middle-posterior part of the cingulate gyrus an...	0.73	[0.67, 0.78]	0.63	[0.56, 0.69]	0.69	[0.6, 0.76]	0.63	[0.53, 0.71]
Occipital pole	0.6	[0.52, 0.66]	0.61	[0.53, 0.67]	0.59	[0.49, 0.68]	0.63	[0.53, 0.71]
Opercular part of the inferior frontal gyrus	0.81	[0.77, 0.85]	0.84	[0.8, 0.87]	0.78	[0.72, 0.83]	0.79	[0.73, 0.84]
Orbital gyri	0.59	[0.52, 0.66]	0.66	[0.59, 0.72]	0.56	[0.45, 0.65]	0.63	[0.53, 0.71]
Orbital part of the inferior frontal gyrus	0.63	[0.56, 0.69]	0.63	[0.56, 0.69]	0.61	[0.51, 0.69]	0.65	[0.55, 0.72]
Orbital sulci	0.33	[0.23, 0.43]	0.37	[0.27, 0.46]	0.54	[0.43, 0.64]	0.47	[0.34, 0.57]
Paracentral lobule and sulcus	0.45	[0.35, 0.53]	0.52	[0.44, 0.6]	0.5	[0.39, 0.61]	0.58	[0.47, 0.67]
Parahippocampal gyrus	0.66	[0.59, 0.72]	0.58	[0.5, 0.65]	0.6	[0.5, 0.69]	0.68	[0.59, 0.75]
Parieto-occipital sulcus	0.71	[0.65, 0.76]	0.73	[0.68, 0.78]	0.75	[0.68, 0.81]	0.75	[0.68, 0.81]
Pericallosal sulcus	0.53	[0.44, 0.6]	0.38	[0.29, 0.47]	0.28	[0.14, 0.41]	0.22	[0.08, 0.35]
Planum polare of the superior temporal gyrus	0.2	[0.09, 0.3]	0.32	[0.22, 0.42]	0.35	[0.21, 0.47]	0.36	[0.23, 0.48]
Planum temporale or temporal plane of the super...	0.75	[0.7, 0.8]	0.78	[0.73, 0.82]	0.73	[0.65, 0.79]	0.75	[0.67, 0.81]
Postcentral gyrus	0.67	[0.6, 0.72]	0.67	[0.61, 0.73]	0.68	[0.6, 0.75]	0.71	[0.63, 0.78]
Postcentral sulcus	0.8	[0.76, 0.84]	0.8	[0.76, 0.84]	0.8	[0.73, 0.84]	0.84	[0.79, 0.87]
Posterior ramus	0.82	[0.78, 0.86]	0.78	[0.73, 0.82]	0.71	[0.62, 0.77]	0.51	[0.39, 0.61]
Posterior transverse collateral sulcus	0.54	[0.46, 0.61]	0.59	[0.51, 0.65]	0.59	[0.48, 0.68]	0.59	[0.49, 0.68]
Posterior-dorsal part of the cingulate gyrus	0.38	[0.28, 0.47]	0.37	[0.27, 0.46]	0.33	[0.2, 0.46]	0.38	[0.24, 0.5]
Posterior-ventral part of the cingulate gyrus	0.59	[0.51, 0.66]	0.64	[0.56, 0.7]	0.49	[0.37, 0.59]	0.55	[0.44, 0.64]
Precentral gyrus	0.53	[0.45, 0.61]	0.54	[0.46, 0.62]	0.66	[0.57, 0.74]	0.54	[0.43, 0.64]
Precuneus	0.78	[0.73, 0.82]	0.78	[0.74, 0.82]	0.66	[0.57, 0.73]	0.7	[0.62, 0.77]
Short insular gyri	0.49	[0.4, 0.57]	0.42	[0.33, 0.51]	0.26	[0.12, 0.39]	0.32	[0.19, 0.45]
Straight gyrus	0.44	[0.34, 0.52]	0.45	[0.36, 0.54]	0.45	[0.32, 0.56]	0.46	[0.33, 0.57]
Subcallosal area	-0.025	[-0.14, 0.09]	0.073	[-0.04, 0.18]	-0.074	[-0.22, 0.07]	-0.07	[-0.21, 0.08]
Subcentral gyrus and sulci	0.75	[0.69, 0.79]	0.77	[0.72, 0.81]	0.72	[0.64, 0.79]	0.62	[0.52, 0.7]
Suborbital sulcus	0.32	[0.21, 0.41]	0.21	[0.1, 0.32]	0.46	[0.34, 0.57]	0.28	[0.14, 0.41]
Subparietal sulcus	0.5	[0.41, 0.58]	0.54	[0.46, 0.62]	0.53	[0.41, 0.63]	0.55	[0.44, 0.64]
Sulcus intermedius primus	0.33	[0.23, 0.43]	0.45	[0.36, 0.53]	0.51	[0.4, 0.61]	0.37	[0.23, 0.49]
Superior frontal gyrus	0.83	[0.79, 0.86]	0.83	[0.79, 0.86]	0.83	[0.78, 0.87]	0.82	[0.77, 0.86]
Superior frontal sulcus	0.8	[0.75, 0.84]	0.79	[0.74, 0.83]	0.73	[0.66, 0.79]	0.75	[0.67, 0.8]
Superior occipital gyrus	0.71	[0.65, 0.76]	0.68	[0.61, 0.73]	0.74	[0.66, 0.8]	0.74	[0.66, 0.8]
Superior occipital sulcus and transverse occipi...	0.77	[0.72, 0.81]	0.76	[0.71, 0.8]	0.72	[0.65, 0.79]	0.73	[0.66, 0.79]
Superior parietal lobule	0.83	[0.79, 0.86]	0.8	[0.76, 0.84]	0.77	[0.7, 0.82]	0.75	[0.67, 0.8]
Superior part of the precentral sulcus	0.68	[0.62, 0.74]	0.78	[0.74, 0.82]	0.77	[0.71, 0.83]	0.73	[0.65, 0.79]
Superior segment of the circular sulcus of the ...	0.71	[0.65, 0.76]	0.67	[0.6, 0.72]	0.65	[0.55, 0.72]	0.52	[0.41, 0.62]
Superior temporal sulcus	0.78	[0.74, 0.82]	0.8	[0.76, 0.84]	0.78	[0.72, 0.84]	0.78	[0.71, 0.83]
Supramarginal gyrus	0.78	[0.74, 0.82]	0.79	[0.74, 0.83]	0.8	[0.74, 0.84]	0.8	[0.75, 0.85]
Temporal pole	0.42	[0.32, 0.5]	0.55	[0.47, 0.63]	0.7	[0.62, 0.77]	0.61	[0.51, 0.7]
Transverse frontopolar gyri and sulci	0.56	[0.48, 0.63]	0.66	[0.59, 0.71]	0.65	[0.56, 0.73]	0.7	[0.61, 0.77]
Transverse temporal sulcus	0.66	[0.59, 0.72]	0.47	[0.38, 0.55]	0.59	[0.49, 0.68]	0.4	[0.27, 0.52]
Triangular part of the inferior frontal gyrus	0.77	[0.72, 0.81]	0.81	[0.77, 0.84]	0.75	[0.67, 0.81]	0.74	[0.67, 0.8]
Vertical ramus of the anterior segment of the l...	0.49	[0.4, 0.57]	0.47	[0.37, 0.55]	0.5	[0.38, 0.6]	0.48	[0.36, 0.58]

Table A.23: Reproducibility in different acquisition settings: ICC and Confidence interval (CI) for the IXI-IOP and OASIS3 acquisition settings and brain hemisphere.

Acquisition setting	IXI|IOP				OASIS3			
Hemisphere	Left		Right		Left		Right	
	ICC	CI95%	ICC	CI95%	ICC	CI95%	ICC	CI95%
ROI Name								
Angular gyrus	0.58	[0.39, 0.72]	0.53	[0.33, 0.68]	0.76	[0.72, 0.79]	0.82	[0.78, 0.84]
Anterior occipital sulcus and preoccipital notch	0.68	[0.53, 0.79]	0.61	[0.44, 0.74]	0.59	[0.53, 0.64]	0.72	[0.68, 0.76]
Anterior part of the cingulate gyrus and sulcus	0.33	[0.1, 0.53]	0.36	[0.13, 0.55]	0.43	[0.36, 0.5]	0.45	[0.37, 0.51]
Anterior segment of the circular sulcus of the ...	0.23	[-0.01, 0.44]	0.11	[-0.13, 0.34]	0.48	[0.41, 0.55]	0.47	[0.4, 0.54]
Anterior transverse collateral sulcus	-0.048	[-0.28, 0.19]	0.015	[-0.22, 0.25]	-0.0072	[-0.1, 0.08]	0.074	[-0.01, 0.16]
Anterior transverse temporal gyrus	0.46	[0.25, 0.63]	0.64	[0.47, 0.76]	0.71	[0.66, 0.75]	0.62	[0.56, 0.67]
Calcarine sulcus	0.48	[0.27, 0.65]	0.68	[0.52, 0.79]	0.62	[0.57, 0.68]	0.66	[0.6, 0.71]
Central sulcus	0.73	[0.6, 0.83]	0.64	[0.47, 0.76]	0.57	[0.51, 0.63]	0.55	[0.49, 0.61]
Cuneus	0.37	[0.14, 0.56]	0.44	[0.23, 0.61]	0.58	[0.52, 0.64]	0.63	[0.57, 0.68]
Fronto-marginal gyrus and sulcus	0.29	[0.06, 0.5]	0.3	[0.07, 0.51]	0.37	[0.3, 0.45]	0.23	[0.14, 0.31]
Horizontal ramus of the anterior segment of the...	0.38	[0.16, 0.57]	0.55	[0.36, 0.7]	0.54	[0.47, 0.6]	0.38	[0.3, 0.45]
Inferior frontal sulcus	0.63	[0.46, 0.76]	0.55	[0.35, 0.69]	0.65	[0.59, 0.69]	0.55	[0.49, 0.61]
Inferior occipital gyrus and sulcus	0.4	[0.18, 0.58]	0.43	[0.21, 0.61]	0.59	[0.53, 0.65]	0.61	[0.56, 0.67]
Inferior part of the precentral sulcus	0.51	[0.31, 0.67]	0.62	[0.45, 0.75]	0.8	[0.77, 0.83]	0.66	[0.61, 0.71]
Inferior segment of the circular sulcus of the ...	0.49	[0.29, 0.66]	0.56	[0.37, 0.7]	0.81	[0.78, 0.84]	0.73	[0.68, 0.77]
Inferior temporal gyrus	0.37	[0.14, 0.56]	0.3	[0.07, 0.51]	0.55	[0.49, 0.61]	0.62	[0.56, 0.67]
Inferior temporal sulcus	0.48	[0.28, 0.65]	0.3	[0.07, 0.51]	0.41	[0.33, 0.48]	0.45	[0.37, 0.52]
Intraparietal sulcus and transverse parietal s...	0.32	[0.09, 0.52]	0.38	[0.16, 0.57]	0.76	[0.72, 0.8]	0.83	[0.8, 0.85]
Lateral aspect of the superior temporal gyrus	0.53	[0.33, 0.68]	0.45	[0.24, 0.63]	0.75	[0.71, 0.79]	0.79	[0.76, 0.82]
Lateral occipito-temporal gyrus	0.18	[-0.06, 0.4]	0.25	[0.01, 0.46]	0.53	[0.47, 0.59]	0.53	[0.46, 0.59]
Lateral occipito-temporal sulcus	0.44	[0.23, 0.62]	0.47	[0.26, 0.63]	0.28	[0.2, 0.36]	0.28	[0.2, 0.36]
Lateral orbital sulcus	0.18	[-0.06, 0.4]	0.3	[0.06, 0.5]	0.44	[0.36, 0.51]	0.092	[0.0, 0.18]
Lingual gyrus	0.26	[0.02, 0.47]	0.29	[0.06, 0.5]	0.63	[0.58, 0.68]	0.61	[0.55, 0.66]
Long insular gyrus and central sulcus of the in...	0.29	[0.05, 0.49]	0.27	[0.03, 0.48]	0.53	[0.46, 0.59]	0.38	[0.3, 0.45]
Marginal branch of the cingulate sulcus	0.43	[0.21, 0.61]	0.32	[0.08, 0.52]	0.77	[0.74, 0.81]	0.72	[0.67, 0.76]
Medial occipito-temporal sulcus and lingual su...	0.46	[0.24, 0.63]	0.64	[0.48, 0.76]	0.57	[0.51, 0.63]	0.67	[0.62, 0.72]
Medial orbital sulcus	0.21	[-0.03, 0.43]	0.43	[0.21, 0.61]	0.14	[0.05, 0.22]	0.034	[-0.05, 0.12]
Middle frontal gyrus	0.65	[0.49, 0.77]	0.35	[0.12, 0.54]	0.75	[0.71, 0.79]	0.62	[0.56, 0.67]
Middle frontal sulcus	0.38	[0.16, 0.57]	0.33	[0.1, 0.53]	0.54	[0.47, 0.6]	0.39	[0.32, 0.46]
Middle occipital gyrus	0.69	[0.54, 0.8]	0.53	[0.34, 0.69]	0.77	[0.73, 0.8]	0.76	[0.72, 0.8]
Middle occipital sulcus and lunatus sulcus	0.65	[0.49, 0.77]	0.58	[0.39, 0.72]	0.62	[0.56, 0.67]	0.72	[0.68, 0.76]
Middle temporal gyrus	0.65	[0.49, 0.77]	0.37	[0.15, 0.56]	0.78	[0.74, 0.81]	0.81	[0.78, 0.84]
Middle-anterior part of the cingulate gyrus and...	0.38	[0.16, 0.57]	0.48	[0.28, 0.65]	0.71	[0.67, 0.75]	0.66	[0.61, 0.71]
Middle-posterior part of the cingulate gyrus an...	0.49	[0.28, 0.65]	0.63	[0.46, 0.75]	0.7	[0.65, 0.74]	0.65	[0.6, 0.7]
Occipital pole	0.087	[-0.15, 0.32]	0.36	[0.13, 0.55]	0.65	[0.59, 0.7]	0.65	[0.6, 0.7]
Opercular part of the inferior frontal gyrus	0.74	[0.61, 0.83]	0.75	[0.63, 0.84]	0.85	[0.82, 0.87]	0.78	[0.74, 0.81]
Orbital gyri	0.55	[0.36, 0.7]	0.51	[0.31, 0.67]	0.57	[0.5, 0.62]	0.36	[0.28, 0.43]
Orbital part of the inferior frontal gyrus	0.55	[0.36, 0.7]	0.56	[0.37, 0.7]	0.65	[0.59, 0.69]	0.42	[0.35, 0.49]
Orbital sulci	0.079	[-0.16, 0.31]	0.41	[0.19, 0.59]	0.3	[0.22, 0.38]	0.29	[0.21, 0.37]
Paracentral lobule and sulcus	0.44	[0.23, 0.62]	0.53	[0.33, 0.68]	0.63	[0.57, 0.68]	0.59	[0.53, 0.65]
Parahippocampal gyrus	0.54	[0.34, 0.69]	0.38	[0.15, 0.56]	0.69	[0.65, 0.74]	0.63	[0.58, 0.68]
Parieto-occipital sulcus	0.58	[0.4, 0.72]	0.3	[0.06, 0.5]	0.77	[0.74, 0.81]	0.72	[0.68, 0.76]
Pericallosal sulcus	0.42	[0.21, 0.6]	0.29	[0.06, 0.49]	0.47	[0.4, 0.54]	0.35	[0.27, 0.43]
Planum polare of the superior temporal gyrus	0.2	[-0.04, 0.42]	0.27	[0.03, 0.48]	0.51	[0.45, 0.58]	0.58	[0.52, 0.63]
Planum temporale or temporal plane of the super...	0.65	[0.49, 0.77]	0.65	[0.49, 0.77]	0.81	[0.78, 0.84]	0.83	[0.8, 0.86]
Postcentral gyrus	0.57	[0.38, 0.71]	0.64	[0.48, 0.76]	0.81	[0.77, 0.84]	0.78	[0.74, 0.81]
Postcentral sulcus	0.54	[0.35, 0.69]	0.63	[0.46, 0.75]	0.84	[0.81, 0.86]	0.83	[0.8, 0.85]
Posterior ramus	0.54	[0.35, 0.69]	0.72	[0.58, 0.82]	0.82	[0.79, 0.85]	0.82	[0.79, 0.85]
Posterior transverse collateral sulcus	0.4	[0.18, 0.58]	0.56	[0.37, 0.7]	0.54	[0.48, 0.6]	0.52	[0.45, 0.58]
Posterior-dorsal part of the cingulate gyrus	-0.0053	[-0.24, 0.23]	0.31	[0.08, 0.51]	0.32	[0.24, 0.4]	0.52	[0.45, 0.58]
Posterior-ventral part of the cingulate gyrus	0.49	[0.29, 0.66]	0.1	[-0.14, 0.34]	0.6	[0.54, 0.66]	0.64	[0.58, 0.69]
Precentral gyrus	0.4	[0.18, 0.58]	0.45	[0.24, 0.63]	0.72	[0.68, 0.76]	0.64	[0.58, 0.69]
Precuneus	0.43	[0.22, 0.61]	0.57	[0.38, 0.71]	0.77	[0.74, 0.81]	0.8	[0.77, 0.83]
Short insular gyri	0.32	[0.09, 0.52]	0.21	[-0.03, 0.43]	0.59	[0.52, 0.64]	0.37	[0.29, 0.44]
Straight gyrus	0.085	[-0.16, 0.32]	0.12	[-0.12, 0.35]	0.32	[0.24, 0.4]	0.41	[0.33, 0.48]
Subcallosal area	0.05	[-0.19, 0.29]	-0.024	[-0.26, 0.22]	8.5e-05	[-0.09, 0.09]	0.087	[-0.0, 0.17]
Subcentral gyrus and sulci	0.64	[0.47, 0.76]	0.68	[0.53, 0.79]	0.81	[0.78, 0.84]	0.79	[0.75, 0.82]
Suborbital sulcus	0.28	[0.04, 0.49]	0.062	[-0.18, 0.3]	0.18	[0.1, 0.27]	0.19	[0.1, 0.27]
Subparietal sulcus	0.38	[0.15, 0.57]	0.48	[0.28, 0.65]	0.5	[0.43, 0.56]	0.49	[0.41, 0.55]
Sulcus intermedius primus	0.4	[0.18, 0.59]	0.24	[-0.0, 0.45]	0.48	[0.4, 0.54]	0.49	[0.42, 0.56]
Superior frontal gyrus	0.75	[0.62, 0.84]	0.77	[0.64, 0.85]	0.82	[0.79, 0.85]	0.79	[0.75, 0.82]
Superior frontal sulcus	0.63	[0.47, 0.76]	0.46	[0.25, 0.63]	0.79	[0.75, 0.82]	0.75	[0.71, 0.79]
Superior occipital gyrus	0.55	[0.35, 0.69]	0.49	[0.29, 0.66]	0.78	[0.74, 0.81]	0.8	[0.76, 0.83]
Superior occipital sulcus and transverse occipi...	0.61	[0.43, 0.74]	0.46	[0.25, 0.63]	0.73	[0.69, 0.77]	0.74	[0.7, 0.78]
Superior parietal lobule	0.41	[0.19, 0.59]	0.61	[0.43, 0.74]	0.85	[0.82, 0.87]	0.81	[0.78, 0.84]
Superior part of the precentral sulcus	0.34	[0.11, 0.53]	0.43	[0.21, 0.61]	0.75	[0.71, 0.79]	0.56	[0.5, 0.62]
Superior segment of the circular sulcus of the ...	0.52	[0.33, 0.68]	0.5	[0.3, 0.66]	0.72	[0.68, 0.76]	0.61	[0.55, 0.67]
Superior temporal sulcus	0.66	[0.51, 0.78]	0.65	[0.49, 0.77]	0.78	[0.74, 0.81]	0.82	[0.79, 0.85]
Supramarginal gyrus	0.7	[0.55, 0.8]	0.69	[0.54, 0.8]	0.83	[0.8, 0.85]	0.84	[0.81, 0.86]
Temporal pole	0.43	[0.21, 0.61]	0.3	[0.06, 0.5]	0.45	[0.38, 0.52]	0.5	[0.43, 0.56]
Transverse frontopolar gyri and sulci	0.31	[0.08, 0.51]	0.53	[0.33, 0.68]	0.42	[0.35, 0.49]	0.35	[0.27, 0.42]
Transverse temporal sulcus	0.69	[0.54, 0.8]	0.32	[0.09, 0.52]	0.59	[0.53, 0.65]	0.53	[0.47, 0.59]
Triangular part of the inferior frontal gyrus	0.75	[0.62, 0.84]	0.69	[0.54, 0.8]	0.79	[0.76, 0.82]	0.53	[0.46, 0.59]
Vertical ramus of the anterior segment of the l...	0.43	[0.21, 0.6]	0.43	[0.21, 0.61]	0.58	[0.52, 0.64]	0.41	[0.34, 0.49]

Table A.24: Reproducibility in different acquisition settings: Cohen'd values for the paired t-test of each ROI, comparing all the cortical thickness estimations of individuals within acquisition setting. The * represents the ROIs in which the null hypothesis was rejected.

Acquisition setting	IXI|GH		IXI|HH		IXI|IOP		OASIS3	
Hemisphere	Left	Right	Left	Right	Left	Right	Left	Right
ROI Name								
Angular gyrus	0.73*	0.88*	0.5*	0.55*	0.22	0.34	0.25*	0.69*
Anterior occipital sulcus and preoccipital notch	0.02	0.02	0.09	0.31*	0.01	0.13	0.09	0.28*
Anterior part of the cingulate gyrus and sulcus	0.94*	0.73*	0.81*	0.48*	0.64*	0.56	0.21*	0.44*
Anterior segment of the circular sulcus of the ...	0.82*	0.11	0.17	0.42*	0.48	0.42	1.14*	0.89*
Anterior transverse collateral sulcus	0.67*	0.48*	0.88*	0.85*	2.38*	1.85*	0.19	0.44*
Anterior transverse temporal gyrus	0.47*	0.49*	0.26	0.12	1.07*	0.49*	0.74*	0.86*
Calcarine sulcus	2.71*	2.03*	2.25*	1.47*	2.18*	1.72*	1.79*	1.26*
Central sulcus	0.85*	0.8*	0.32*	0.26	0.61*	0.84*	1.1*	0.87*
Cuneus	0.66*	0.88*	0.93*	1.16*	1.33*	1.63*	0.36*	0.23*
Fronto-marginal gyrus and sulcus	0.15	0.19	0.39*	0.06	0.03	0.25	0.02	0.03
Horizontal ramus of the anterior segment of the...	0.84*	1.25*	0.59*	0.93*	1.4*	1.64*	1.04*	1.5*
Inferior frontal sulcus	0.71*	0.72*	0.76*	0.67*	1.4*	1.95*	1.17*	1.46*
Inferior occipital gyrus and sulcus	0.76*	1.13*	0.06	0.25*	0.78*	0.9*	0.43*	1.29*
Inferior part of the precentral sulcus	0.18*	0.2*	0.33*	0.18	0.98*	1.55*	0.82*	0.93*
Inferior segment of the circular sulcus of the ...	1.46*	1.5*	0.49*	0.22	1.13*	1.38*	1.1*	1.56*
Inferior temporal gyrus	2.35*	2.05*	1.82*	1.51*	2.28*	2.46*	1.36*	1.71*
Inferior temporal sulcus	0.8*	0.97*	0.7*	0.7*	0.82*	0.62*	0.21*	0.43*
Intraparietal sulcus and transverse parietal s...	0.79*	0.86*	0.86*	1.19*	1.62*	1.56*	0.71*	0.49*
Lateral aspect of the superior temporal gyrus	2.21*	1.76*	1.82*	1.25*	2.42*	1.77*	1.31*	0.91*
Lateral occipito-temporal gyrus	1.78*	1.61*	0.53*	0.5*	0.59	0.17	2.07*	2.55*
Lateral occipito-temporal sulcus	0.61*	0.51*	0.54*	0.33*	0.46	0.79*	0.18	0.61*
Lateral orbital sulcus	0.47*	0.49*	0.67*	0.59*	1.35*	1.01*	1.08*	1.2*
Lingual gyrus	0.56*	0.82*	1.32*	1.2*	1.44*	2.02*	0.28*	0.3*
Long insular gyrus and central sulcus of the in...	0.04	0.5*	0.06	0.64*	1.15*	0.68*	0.14	0.59*
Marginal branch of the cingulate sulcus	1.18*	1.1*	1.41*	1.03*	1.37*	1.27*	1.23*	1.14*
Medial occipito-temporal sulcus and lingual su...	0.16	0.51*	0.05	0.24*	0.12	0.15	0.17*	0.75*
Medial orbital sulcus	0.36*	0.53*	0.48*	0.2	3.27*	2.26*	0.64*	0.48*
Middle frontal gyrus	0.15*	0.34*	0.2*	0.06	0.33	0.25	0.1	0.22*
Middle frontal sulcus	1.36*	1.3*	1.83*	1.4*	1.68*	1.72*	1.48*	1.77*
Middle occipital gyrus	0.79*	1.39*	0.24*	0.57*	0.12	0.69*	0.43*	1.14*
Middle occipital sulcus and lunatus sulcus	0.48*	0.53*	0.9*	0.93*	0.68*	0.58*	0.43*	0.12
Middle temporal gyrus	1.65*	1.63*	1.19*	1.16*	1.25*	0.91*	0.81*	0.75*
Middle-anterior part of the cingulate gyrus and...	0.4*	0.11	0.41*	0.17	0.6*	0.19	0.98*	0.89*
Middle-posterior part of the cingulate gyrus an...	0.02	0.14	0.33*	0.14	0.07	0.23	0.39*	0.29*
Occipital pole	0.67*	0.62*	0.12	0.33*	0.6	0.46	0.5*	0.65*
Opercular part of the inferior frontal gyrus	0.25*	0.15*	0.21*	0.22*	0.22	0.4*	0.04	0.18*
Orbital gyri	0.29*	0.89*	0.11	0.39*	1.64*	2.17*	0.11	0.82*
Orbital part of the inferior frontal gyrus	1.04*	1.34*	0.84*	1.13*	0.45	1.21*	0.66*	0.9*
Orbital sulci	0.93*	0.75*	0.52*	0.67*	0.37	0.68*	0.43*	0.84*
Paracentral lobule and sulcus	2.47*	2.11*	1.57*	1.56*	0.28	0.21	1.44*	1.52*
Parahippocampal gyrus	1.0*	1.31*	0.85*	0.94*	1.22*	0.97*	1.0*	1.14*
Parieto-occipital sulcus	0.86*	0.68*	0.88*	0.77*	0.95*	1.12*	0.6*	0.51*
Pericallosal sulcus	0.44*	0.55*	0.83*	0.83*	1.05*	1.04*	0.56*	0.69*
Planum polare of the superior temporal gyrus	1.35*	0.7*	1.25*	0.8*	1.02*	0.74*	0.73*	0.18*
Planum temporale or temporal plane of the super...	0.21*	0.32*	0.11	0.08	0.06	0.09	0.04	0.37*
Postcentral gyrus	1.4*	1.24*	1.16*	1.02*	0.45	0.32	0.82*	0.88*
Postcentral sulcus	0.72*	0.7*	0.65*	0.73*	1.65*	1.8*	0.73*	0.68*
Posterior ramus	0.47*	1.17*	0.38*	0.8*	1.2*	1.86*	0.44*	1.23*
Posterior transverse collateral sulcus	0.45*	0.08	0.69*	0.3*	0.48	0.18	0.03	0.74*
Posterior-dorsal part of the cingulate gyrus	1.69*	1.93*	1.23*	1.32*	1.62*	1.82*	1.38*	1.34*
Posterior-ventral part of the cingulate gyrus	0.86*	1.17*	0.73*	1.02*	0.24	0.31	0.76*	0.8*
Precentral gyrus	2.54*	2.82*	1.89*	1.95*	0.87*	0.75*	1.72*	1.55*
Precuneus	0.29*	0.34*	0.15	0.13	0.02	0.11	0.13*	0.12*
Short insular gyri	0.59*	0.66*	0.62*	0.46*	0.01	0.05	0.34*	0.62*
Straight gyrus	1.15*	0.85*	0.4*	0.52*	2.08*	2.07*	1.65*	0.75*
Subcallosal area	0.67*	0.86*	0.57*	0.67*	0.11	0.64	0.49*	0.12
Subcentral gyrus and sulci	1.06*	1.09*	0.61*	0.8*	0.61*	0.42*	0.29*	0.25*
Suborbital sulcus	0.09	0.69*	0.0	0.28	0.0	0.3	0.08	0.3*
Subparietal sulcus	0.33*	0.02	0.66*	0.08	0.1	0.23	0.67*	0.32*
Sulcus intermedius primus	0.09	0.25*	0.15	0.53*	0.31	1.33*	0.67*	0.52*
Superior frontal gyrus	1.16*	1.19*	0.38*	0.41*	0.45*	0.43*	0.1	0.1
Superior frontal sulcus	0.48*	0.46*	1.08*	0.95*	1.44*	1.89*	0.95*	1.55*
Superior occipital gyrus	0.33*	0.31*	0.06	0.2	0.4	0.72*	0.48*	0.33*
Superior occipital sulcus and transverse occipi...	0.56*	0.32*	0.76*	0.91*	1.13*	0.87*	0.27*	0.04
Superior parietal lobule	0.34*	0.42*	0.04	0.11	0.7*	0.36	0.16*	0.22*
Superior part of the precentral sulcus	0.02	0.05	0.36*	0.47*	1.03*	1.31*	0.31*	0.62*
Superior segment of the circular sulcus of the ...	1.96*	1.74*	1.47*	1.19*	2.44*	2.19*	2.6*	2.82*
Superior temporal sulcus	0.33*	0.37*	0.2*	0.35*	0.29	0.8*	0.71*	0.71*
Supramarginal gyrus	0.75*	1.06*	0.59*	0.86*	0.01	0.46*	0.17*	0.52*
Temporal pole	1.19*	1.35*	1.06*	1.1*	1.91*	2.04*	0.83*	0.95*
Transverse frontopolar gyri and sulci	0.76*	0.35*	0.12	0.0	0.45	0.25	0.65*	0.51*
Transverse temporal sulcus	0.25*	0.57*	0.04	0.33*	0.17	0.26	0.45*	0.75*
Triangular part of the inferior frontal gyrus	0.82*	0.82*	0.54*	0.62*	0.42*	0.65*	0.28*	0.11
Vertical ramus of the anterior segment of the l...	1.09*	0.42*	1.07*	0.31*	1.53*	1.4*	1.32*	1.12*

Table A.25: Brain age model results using cortical thickness estimations, extracted by FreeSurfer, to predict the participant' age (coefficient and p-value) and the mean reproducibility R^2 for the corresponding ROI. The significant ROIs are shown in bold.

ROI Name	coefficient	p-value	R^2
Right Inferior segment of the circular sulcus of the insula	**-10.80**	**<0.001**	**0.334**
Right Central sulcus	**15.96**	**<0.001**	**0.289**
Left Subcallosal area	**-7.09**	**<0.001**	**0.002**
Right Lateral orbital sulcus	**-5.67**	**0.001**	**0.136**
Left Superior frontal gyrus	**-13.69**	**0.004**	**0.66**
Right Anterior segment of the circular sulcus of the insula	**-4.89**	**0.004**	**0.163**
Right Occipital pole	**10.14**	**0.004**	**0.343**
Right Lingual gyrus	**9.76**	**0.005**	**0.308**
Right Angular gyrus	**-9.89**	**0.005**	**0.571**
Right Superior occipital sulcus and transverse occipital sulcus	**8.00**	**0.007**	**0.481**
Left Lingual gyrus	**10.19**	**0.009**	**0.266**
Right Medial occipito-temporal sulcus and lingual sulcus	**-7.06**	**0.012**	**0.484**
Right Transverse temporal sulcus	**-2.84**	**0.016**	**0.224**
Right Paracentral lobule and sulcus	**-5.78**	**0.016**	**0.32**
Right Inferior occipital gyrus and sulcus	**4.42**	**0.016**	**0.415**
Left Orbital sulci	**4.94**	**0.017**	**0.138**
Left Medial orbital sulcus	**-4.23**	**0.018**	**0.027**
Right Superior part of the precentral sulcus	**-5.69**	**0.019**	**0.431**
Left Triangular part of the inferior frontal gyrus	**-6.10**	**0.019**	**0.603**
Right Anterior occipital sulcus and preoccipital notch	**-4.82**	**0.024**	**0.496**
Right Long insular gyrus and central sulcus of the insula	**2.96**	**0.028**	**0.119**
Right Pericallosal sulcus	**3.65**	**0.028**	**0.128**
Right Vertical ramus of the anterior segment of the lateral sulcus	**-2.87**	**0.03**	**0.21**
Right Lateral aspect of the superior temporal gyrus	**-5.56**	**0.033**	**0.463**
Right Temporal pole	**4.25**	**0.04**	**0.264**
Left Superior segment of the circular sulcus of the insula	**-7.24**	**0.041**	**0.44**
Left Medial occipito-temporal sulcus and lingual sulcus	**-5.83**	**0.042**	**0.403**
Right Superior segment of the circular sulcus of the insula	**-6.55**	**0.044**	**0.345**
Right Anterior transverse temporal gyrus	**3.48**	**0.048**	**0.281**
Left Postcentral gyrus	6.12	0.05	0.487
Right Parahippocampal gyrus	3.58	0.058	0.338
Right Calcarine sulcus	-6.88	0.059	0.416
Right Fronto-marginal gyrus and sulcus	3.75	0.066	0.18
Right Suborbital sulcus	-1.79	0.067	0.047
Right Middle-anterior part of the cingulate gyrus and sulcus	-5.33	0.069	0.362
Right Superior occipital gyrus	-5.05	0.071	0.486
Left Middle-posterior part of the cingulate gyrus and sulcus	5.80	0.072	0.443
Right Horizontal ramus of the anterior segment of the lateral sulcus	-2.89	0.072	0.284
Right Superior frontal sulcus	-6.72	0.079	0.493
Left Posterior ramus	-4.61	0.081	0.553
Left Temporal pole	3.40	0.087	0.267
Left Calcarine sulcus	-6.70	0.092	0.348
Left Intraparietal sulcus and transverse parietal sulci	5.54	0.093	0.422
Right Middle frontal gyrus	6.30	0.099	0.491
Right Middle occipital sulcus and lunatus sulcus	-3.92	0.101	0.439

Left Posterior transverse collateral sulcus	-3.30	0.112	0.281
Left Cuneus	5.57	0.112	0.241
Left Orbital gyri	-4.81	0.116	0.326
Left Occipital pole	4.52	0.128	0.297
Left Anterior part of the cingulate gyrus and sulcus	4.59	0.128	0.242
Right Orbital sulci	3.24	0.129	0.174
Left Suborbital sulcus	-2.35	0.133	0.108
Right Straight gyrus	-2.97	0.133	0.158
Right Precuneus	4.79	0.133	0.56
Right Postcentral gyrus	-4.25	0.139	0.497
Right Planum polare of the superior temporal gyrus	-2.27	0.141	0.163
Left Horizontal ramus of the anterior segment of the lateral sulcus	2.02	0.147	0.254
Right Superior temporal sulcus	6.76	0.149	0.59
Left Superior part of the precentral sulcus	-3.46	0.15	0.448
Right Middle temporal gyrus	-4.65	0.152	0.49
Left Supramarginal gyrus	4.64	0.165	0.648
Left Superior temporal sulcus	-6.43	0.169	0.58
Left Transverse frontopolar gyri and sulci	-2.87	0.175	0.261
Right Anterior part of the cingulate gyrus and sulcus	4.15	0.184	0.274
Right Posterior ramus	-3.89	0.185	0.531
Right Supramarginal gyrus	4.49	0.187	0.634
Left Marginal branch of the cingulate sulcus	-3.82	0.191	0.457
Left Lateral occipito-temporal sulcus	-2.82	0.191	0.222
Left Anterior transverse temporal gyrus	2.31	0.192	0.311
Left Short insular gyri	2.05	0.195	0.221
Right Postcentral sulcus	4.46	0.195	0.609
Left Postcentral sulcus	-4.83	0.203	0.568
Left Inferior occipital gyrus and sulcus	2.78	0.209	0.376
Right Posterior-dorsal part of the cingulate gyrus	-2.93	0.217	0.177
Left Posterior-dorsal part of the cingulate gyrus	-3.16	0.226	0.102
Left Anterior transverse collateral sulcus	1.91	0.229	0.005
Left Planum polare of the superior temporal gyrus	-1.88	0.234	0.118
Left Fronto-marginal gyrus and sulcus	2.86	0.236	0.19
Left Orbital part of the inferior frontal gyrus	-2.19	0.239	0.377
Left Long insular gyrus and central sulcus of the insula	1.88	0.252	0.171
Right Intraparietal sulcus and transverse parietal sulci	4.64	0.256	0.492
Right Inferior temporal gyrus	2.84	0.258	0.328
Left Paracentral lobule and sulcus	2.64	0.29	0.272
Right Subcallosal area	1.69	0.294	0.005
Right Orbital gyri	-3.03	0.298	0.309
Left Precuneus	3.53	0.307	0.506
Left Lateral occipito-temporal gyrus	2.47	0.308	0.277
Right Short insular gyri	1.40	0.317	0.133
Right Lateral occipito-temporal gyrus	2.34	0.342	0.262
Left Precentral gyrus	-2.48	0.356	0.365
Right Subcentral gyrus and sulci	-2.28	0.36	0.534
Left Middle occipital gyrus	-2.91	0.373	0.576
Right Cuneus	3.26	0.375	0.287
Left Vertical ramus of the anterior segment of the lateral sulcus	-1.31	0.376	0.27

Left Subcentral gyrus and sulci	-2.31	0.385	0.579
Right Subparietal sulcus	2.28	0.393	0.279
Left Anterior segment of the circular sulcus of the insula	-1.53	0.393	0.123
Left Anterior occipital sulcus and preoccipital notch	1.87	0.398	0.427
Right Sulcus intermedius primus	-1.34	0.404	0.189
Left Middle frontal sulcus	-2.43	0.421	0.263
Right Middle-posterior part of the cingulate gyrus and sulcus	2.65	0.422	0.422
Left Parahippocampal gyrus	1.42	0.438	0.401
Left Pericallosal sulcus	1.31	0.465	0.203
Left Middle-anterior part of the cingulate gyrus and sulcus	-1.79	0.498	0.352
Right Orbital part of the inferior frontal gyrus	1.18	0.506	0.337
Right Parieto-occipital sulcus	2.11	0.516	0.435
Right Lateral occipito-temporal sulcus	1.25	0.52	0.18
Left Transverse temporal sulcus	-0.78	0.534	0.451
Left Central sulcus	2.24	0.536	0.341
Right Inferior frontal sulcus	2.13	0.54	0.464
Left Superior occipital gyrus	-1.51	0.553	0.513
Left Lateral aspect of the superior temporal gyrus	1.53	0.554	0.432
Right Medial orbital sulcus	-1.03	0.556	0.086
Left Middle occipital sulcus and lunatus sulcus	1.35	0.574	0.474
Right Anterior transverse collateral sulcus	-0.88	0.586	0.01
Left Inferior segment of the circular sulcus of the insula	-1.13	0.608	0.424
Left Posterior-ventral part of the cingulate gyrus	-0.71	0.636	0.298
Right Triangular part of the inferior frontal gyrus	-1.10	0.648	0.506
Left Inferior part of the precentral sulcus	1.38	0.65	0.507
Left Lateral orbital sulcus	-0.71	0.652	0.18
Left Superior parietal lobule	-1.63	0.672	0.55
Left Inferior temporal sulcus	-1.01	0.694	0.233
Left Superior occipital sulcus and transverse occipital sulcus	1.21	0.706	0.512
Left Opercular part of the inferior frontal gyrus	1.08	0.71	0.664
Right Middle occipital gyrus	-1.22	0.723	0.527
Right Superior parietal lobule	-1.18	0.724	0.57
Right Opercular part of the inferior frontal gyrus	-0.99	0.727	0.651
Right Transverse frontopolar gyri and sulci	0.75	0.753	0.35
Left Subparietal sulcus	-0.86	0.753	0.241
Left Parieto-occipital sulcus	-0.89	0.766	0.527
Left Planum temporale or temporal plane of the superior temporal gyrus	-0.64	0.769	0.582
Right Planum temporale or temporal plane of the superior temporal gyrus	0.64	0.774	0.597
Right Middle frontal sulcus	0.71	0.828	0.276
Left Superior frontal sulcus	0.77	0.84	0.551
Left Inferior temporal gyrus	-0.47	0.864	0.331
Right Superior frontal gyrus	-0.78	0.871	0.649
Right Precentral gyrus	-0.39	0.879	0.308
Right Posterior-ventral part of the cingulate gyrus	0.25	0.883	0.282
Right Inferior temporal sulcus	-0.28	0.913	0.233
Right Inferior part of the precentral sulcus	0.30	0.918	0.503
Left Angular gyrus	-0.31	0.921	0.58
Left Straight gyrus	-0.13	0.953	0.135
Left Middle temporal gyrus	0.16	0.956	0.569

Left Inferior frontal sulcus	-0.20	0.957	0.468
Left Sulcus intermedius primus	0.04	0.968	0.222
Right Marginal branch of the cingulate sulcus	0.10	0.976	0.395
Right Posterior transverse collateral sulcus	0.01	0.995	0.339
Left Middle frontal gyrus	0.02	0.996	0.611

Table A.26: Brain age model results using cortical thickness estimations, extracted by CAT12, to predict the participant' age (coefficient and p-value) and the mean reproducibility R^2 for the corresponding ROI. The significant ROIs are shown in bold.

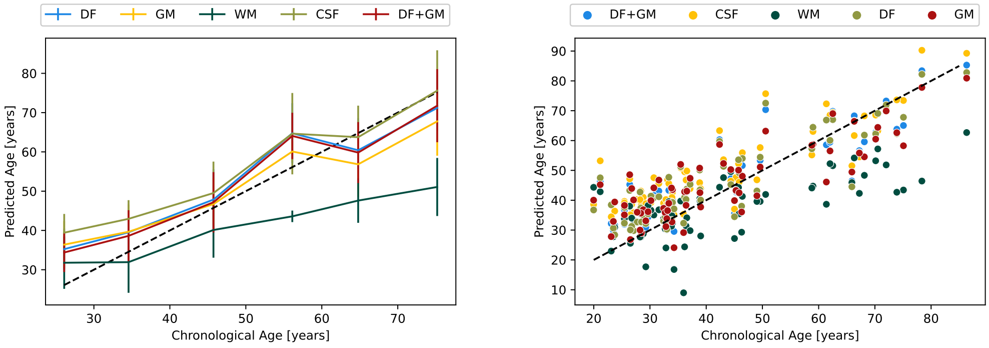
ROI Name	coefficient	p-value	R^2
Left Central sulcus	**-28.07**	**<0.001**	**0.341**
Left Anterior transverse temporal gyrus	**-13.59**	**<0.001**	**0.311**
Right Cuneus	**20.30**	**<0.001**	**0.287**
Left Lateral aspect of the superior temporal gyrus	**15.51**	**<0.001**	**0.432**
Right Anterior transverse temporal gyrus	**-12.42**	**<0.001**	**0.281**
Right Planum temporale or temporal plane of the superior temporal gyrus	**-9.20**	**0.001**	**0.597**
Right Angular gyrus	**-13.42**	**0.001**	**0.571**
Right Posterior ramus	**9.44**	**0.001**	**0.531**
Right Superior frontal gyrus	**-17.16**	**0.001**	**0.649**
Right Superior part of the precentral sulcus	**10.59**	**0.001**	**0.431**
Left Anterior occipital sulcus and preoccipital notch	**6.71**	**0.002**	**0.427**
Left Paracentral lobule and sulcus	**9.87**	**0.002**	**0.272**
Left Superior occipital gyrus	**-11.22**	**0.003**	**0.513**
Left Precuneus	**12.90**	**0.003**	**0.506**
Left Inferior temporal sulcus	**-7.03**	**0.004**	**0.233**
Left Inferior part of the precentral sulcus	**9.64**	**0.006**	**0.507**
Left Superior temporal sulcus	**-11.21**	**0.006**	**0.58**
Left Superior frontal gyrus	**-13.90**	**0.009**	**0.66**
Right Occipital pole	**11.25**	**0.01**	**0.343**
Left Triangular part of the inferior frontal gyrus	**-8.61**	**0.011**	**0.603**
Right Superior occipital gyrus	**-9.49**	**0.011**	**0.486**
Right Middle occipital gyrus	**9.72**	**0.013**	**0.527**
Right Subcallosal area	**2.60**	**0.013**	**0.005**
Left Straight gyrus	**-6.64**	**0.014**	**0.135**
Right Calcarine sulcus	**-8.16**	**0.015**	**0.416**
Left Medial orbital sulcus	**-3.79**	**0.016**	**0.027**
Left Planum temporale or temporal plane of the superior temporal gyrus	**-7.19**	**0.016**	**0.582**
Right Medial orbital sulcus	**4.24**	**0.02**	**0.086**
Left Lingual gyrus	**9.94**	**0.023**	**0.266**
Left Occipital pole	**9.44**	**0.026**	**0.297**
Right Inferior part of the precentral sulcus	**8.09**	**0.026**	**0.503**
Right Superior occipital sulcus and transverse occipital sulcus	**6.96**	**0.03**	**0.481**
Left Transverse temporal sulcus	**3.82**	**0.03**	**0.451**
Left Calcarine sulcus	**-7.47**	**0.032**	**0.348**
Left Superior occipital sulcus and transverse occipital sulcus	**6.90**	**0.034**	**0.512**
Left Cuneus	**-10.63**	**0.035**	**0.241**
Left Superior parietal lobule	**-9.52**	**0.04**	**0.55**
Right Precuneus	**8.03**	**0.048**	**0.56**
Right Orbital gyri	-5.97	0.052	0.309
Left Long insular gyrus and central sulcus of the insula	2.68	0.052	0.171
Right Postcentral gyrus	-6.90	0.054	0.497
Left Opercular part of the inferior frontal gyrus	-6.87	0.054	0.664
Left Subcentral gyrus and sulci	6.91	0.055	0.579
Left Lateral occipito-temporal gyrus	4.83	0.061	0.277

Right Sulcus intermedius primus	4.69	0.063	0.189
Right Horizontal ramus of the anterior segment of the lateral sulcus	-3.10	0.066	0.284
Left Precentral gyrus	6.83	0.066	0.365
Left Inferior temporal gyrus	5.10	0.07	0.331
Right Suborbital sulcus	-2.48	0.073	0.047
Left Middle occipital sulcus and lunatus sulcus	4.70	0.092	0.474
Left Marginal branch of the cingulate sulcus	-5.21	0.094	0.457
Right Middle frontal sulcus	6.49	0.095	0.276
Right Postcentral sulcus	5.98	0.104	0.609
Left Orbital sulci	4.30	0.104	0.138
Left Anterior transverse collateral sulcus	2.53	0.108	0.005
Right Orbital sulci	4.47	0.114	0.174
Right Superior segment of the circular sulcus of the insula	-5.07	0.119	0.345
Left Fronto-marginal gyrus and sulcus	-4.38	0.129	0.19
Left Posterior ramus	-4.75	0.131	0.553
Left Short insular gyri	1.53	0.141	0.221
Left Posterior-dorsal part of the cingulate gyrus	2.79	0.156	0.102
Right Straight gyrus	3.47	0.163	0.158
Right Lateral occipito-temporal gyrus	-3.45	0.171	0.262
Left Middle frontal gyrus	-7.03	0.177	0.611
Left Inferior occipital gyrus and sulcus	3.09	0.178	0.376
Left Orbital gyri	4.07	0.179	0.326
Left Postcentral sulcus	5.27	0.186	0.568
Left Supramarginal gyrus	6.24	0.188	0.648
Right Planum polare of the superior temporal gyrus	2.12	0.189	0.163
Right Temporal pole	-2.13	0.19	0.264
Right Posterior-ventral part of the cingulate gyrus	1.88	0.193	0.282
Right Lateral aspect of the superior temporal gyrus	3.44	0.196	0.463
Left Anterior segment of the circular sulcus of the insula	-1.99	0.2	0.123
Right Central sulcus	-6.57	0.214	0.289
Right Parahippocampal gyrus	2.13	0.222	0.338
Right Superior temporal sulcus	-5.42	0.23	0.59
Right Superior frontal sulcus	-5.21	0.233	0.493
Left Middle occipital gyrus	-4.37	0.239	0.576
Right Inferior frontal sulcus	-4.67	0.249	0.464
Right Marginal branch of the cingulate sulcus	4.19	0.258	0.395
Left Planum polare of the superior temporal gyrus	-1.72	0.259	0.118
Left Lateral orbital sulcus	2.31	0.266	0.18
Left Middle temporal gyrus	3.08	0.305	0.569
Right Anterior part of the cingulate gyrus and sulcus	-2.67	0.307	0.274
Right Inferior temporal sulcus	-2.48	0.31	0.233
Right Middle frontal gyrus	-4.77	0.321	0.491
Left Transverse frontopolar gyri and sulci	2.43	0.323	0.261
Right Inferior segment of the circular sulcus of the insula	-1.86	0.343	0.334
Right Subparietal sulcus	2.85	0.345	0.279
Right Long insular gyrus and central sulcus of the insula	1.11	0.348	0.119
Right Anterior transverse collateral sulcus	-1.48	0.353	0.01
Left Intraparietal sulcus and transverse parietal sulci	4.06	0.363	0.422
Right Intraparietal sulcus and transverse parietal sulci	-4.06	0.367	0.492
Left Medial occipito-temporal sulcus and lingual sulcus	-2.31	0.37	0.403
Left Inferior segment of the circular sulcus of the insula	-1.96	0.373	0.424
Right Lingual gyrus	3.17	0.401	0.308
Right Middle-posterior part of the cingulate gyrus and sulcus	-2.61	0.404	0.422
Right Vertical ramus of the anterior segment of the lateral sulcus	-1.30	0.407	0.21
Right Supramarginal gyrus	4.16	0.416	0.634
Left Orbital part of the inferior frontal gyrus	1.11	0.473	0.377
Right Precentral gyrus	-2.49	0.484	0.308

Right Medial occipito-temporal sulcus and lingual sulcus	-1.73	0.495	0.484
Left Parahippocampal gyrus	-1.18	0.51	0.401
Left Superior segment of the circular sulcus of the insula	-2.21	0.546	0.44
Left Vertical ramus of the anterior segment of the lateral sulcus	1.25	0.553	0.27
Left Subparietal sulcus	1.75	0.567	0.241
Left Posterior transverse collateral sulcus	-1.25	0.58	0.281
Right Middle occipital sulcus and lunatus sulcus	-1.37	0.581	0.439
Left Pericallosal sulcus	1.21	0.593	0.203
Right Orbital part of the inferior frontal gyrus	-0.80	0.598	0.337
Right Superior parietal lobule	-2.18	0.616	0.57
Left Anterior part of the cingulate gyrus and sulcus	-1.14	0.64	0.242
Left Suborbital sulcus	-0.77	0.643	0.108
Left Middle-anterior part of the cingulate gyrus and sulcus	1.10	0.653	0.352
Left Angular gyrus	-1.84	0.659	0.58
Right Posterior-dorsal part of the cingulate gyrus	0.91	0.663	0.177
Left Middle-posterior part of the cingulate gyrus and sulcus	-1.30	0.664	0.443
Right Transverse temporal sulcus	0.82	0.664	0.224
Right Short insular gyri	0.39	0.699	0.133
Left Posterior-ventral part of the cingulate gyrus	0.50	0.707	0.298
Right Paracentral lobule and sulcus	1.10	0.707	0.32
Left Lateral occipito-temporal sulcus	-0.73	0.709	0.222
Right Middle-anterior part of the cingulate gyrus and sulcus	0.86	0.718	0.362
Right Anterior occipital sulcus and preoccipital notch	-0.71	0.749	0.496
Right Pericallosal sulcus	-0.82	0.755	0.128
Left Superior frontal sulcus	1.22	0.77	0.551
Right Parieto-occipital sulcus	1.01	0.777	0.435
Right Anterior segment of the circular sulcus of the insula	-0.52	0.778	0.163
Right Inferior temporal gyrus	0.71	0.788	0.328
Left Sulcus intermedius primus	-0.40	0.812	0.222
Right Fronto-marginal gyrus and sulcus	0.56	0.834	0.18
Left Inferior frontal sulcus	-0.84	0.835	0.468
Right Transverse frontopolar gyri and sulci	0.55	0.841	0.35
Right Lateral orbital sulcus	0.33	0.869	0.136
Left Parieto-occipital sulcus	-0.57	0.872	0.527
Left Superior part of the precentral sulcus	0.49	0.873	0.448
Right Opercular part of the inferior frontal gyrus	-0.45	0.89	0.651
Right Lateral occipito-temporal sulcus	-0.23	0.894	0.18
Right Triangular part of the inferior frontal gyrus	-0.38	0.906	0.506
Right Middle temporal gyrus	-0.36	0.917	0.49
Right Inferior occipital gyrus and sulcus	-0.21	0.923	0.415
Right Posterior transverse collateral sulcus	0.18	0.937	0.339
Left Middle frontal sulcus	0.24	0.939	0.263
Left Temporal pole	-0.12	0.948	0.267
Left Horizontal ramus of the anterior segment of the lateral sulcus	0.06	0.963	0.254
Left Subcallosal area	0.01	0.992	0.002
Left Postcentral gyrus	0.04	0.992	0.487
Right Subcentral gyrus and sulci	-0.02	0.995	0.534

Appendix B

Deformation Fields: A new source of information to predict brain age



(a)

(b) Min processed Diamarker

Figure B.1: Chronological Age versus predicted age of the model of Deformation Fields (DF), GM, WM, CSF and fusion. (a) Each point represents the corresponding result for an individual of the test set (b) Mean age and mean age prediction for the test set grouped by modality and age bins of 10 years, the error bars represent the standard deviation of the predictions for the corresponding bin.

Table B.1: Difference between the mean predicted age and the mean chronological for each age bin for the deformation fields model.

Age bin	**20-29**		**40-49**		**60-69**	**>=70.0**
Mean age [years]	26.09	34.56	45.74	56.12	64.79	75.19
Mean predictions [years]	35.23	39.52	47.86	64.74	60.42	71.2
Difference [years]	9.14	4.96	2.12	8.62	-4.37	-3.99

Table B.2: Difference between the mean predicted age and the mean chronological for each age bin for the grey matter model.

Age bin	**20-29**		**40-49**		**60-69**	**>=70.0**
Mean age [years]	26.09	34.56	45.74	56.12	64.79	75.19
Mean predictions [years]	36.38	39.66	46.7	60.09	56.84	67.77
Difference [years]	10.29	5.1	0.96	3.97	-7.95	-7.42

Table B.3: Difference between the mean predicted age and the mean chronological for each age bin for the white matter model.

Age bin	**20-29**		**40-49**		**60-69**	**>=70.0**
Mean age [years]	26.09	34.56	45.74	56.12	64.79	75.19
Mean predictions [years]	31.78	31.93	40.12	43.59	47.6	51.07
Difference [years]	5.69	-2.63	-5.62	-12.53	-17.19	-24.12

Table B.4: Difference between the mean predicted age and the mean chronological for each age bin for the CSF model.

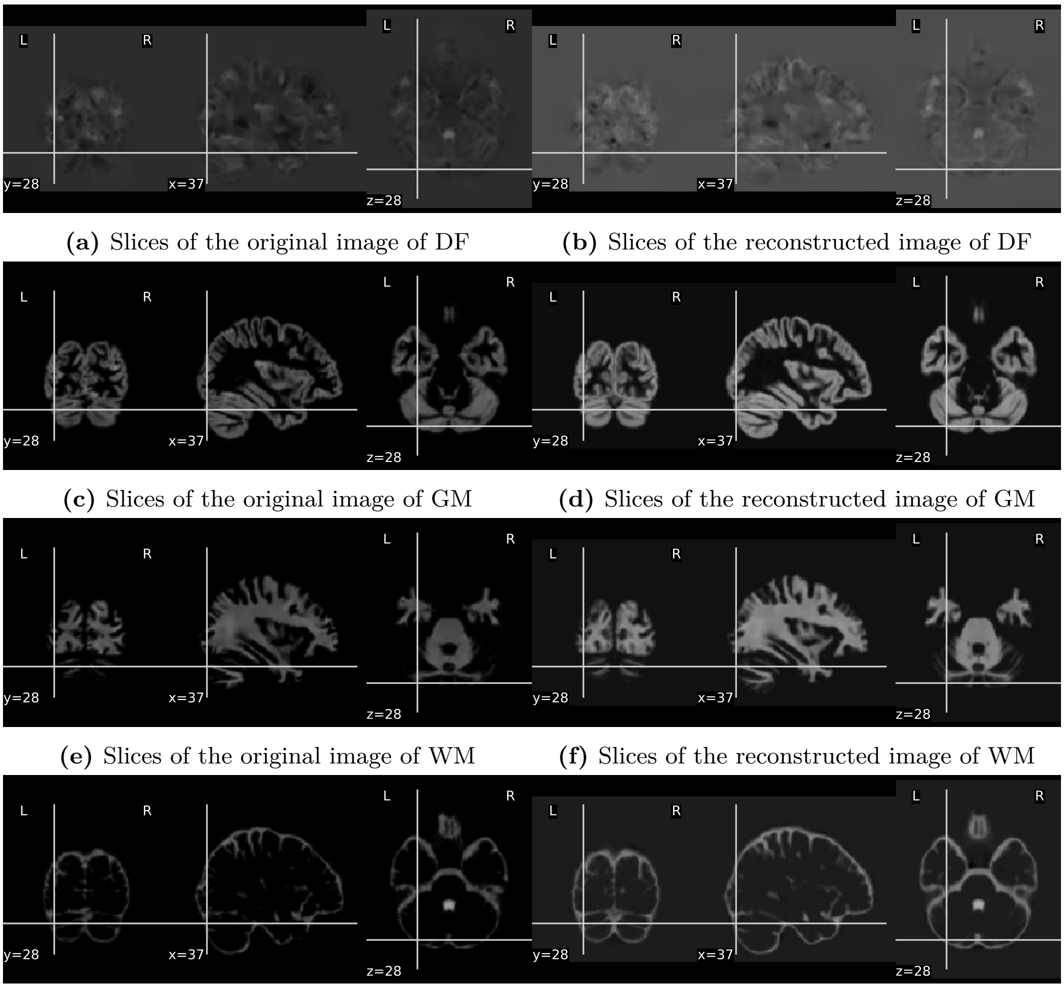
Age bin	**20-29**		**40-49**		**60-69**	**>=70.0**
Mean age [years]	26.09	34.56	45.74	56.12	64.79	75.19
Mean predictions [years]	39.43	42.96	49.52	64.64	63.75	75.59
Difference [years]	13.34	8.4	3.78	8.52	-1.04	0.4

Table B.5: Difference between the mean predicted age and the mean chronological for each age bin for the fusion model (deformation fields and grey matter).

Age bin	**20-29**		**40-49**		**60-69**	**>=70.0**
Mean age [years]	26.09	34.56	45.74	56.12	64.79	75.19
Mean predictions [years]	34.37	38.61	47.17	64.08	59.84	71.79
Difference [years]	8.28	4.05	1.43	7.96	-4.95	-3.4

Table B.6: Mean and standard deviations of the MAE (in years) and R^2 for the validation set, 30-fold cross-validation, and MAE and R^2 for the test for two fusion models: one with deformation fields (DF) and grey matter (GM) and another with GM and white matter (WM).

	MAE [years]		R2	
	Validation	Test	Validation	Test
DF+GM	5.55 ± 1.14	6.90	0.79	0.76
GM+WM	5.45 ± 1.29	7.92	0.79	0.64



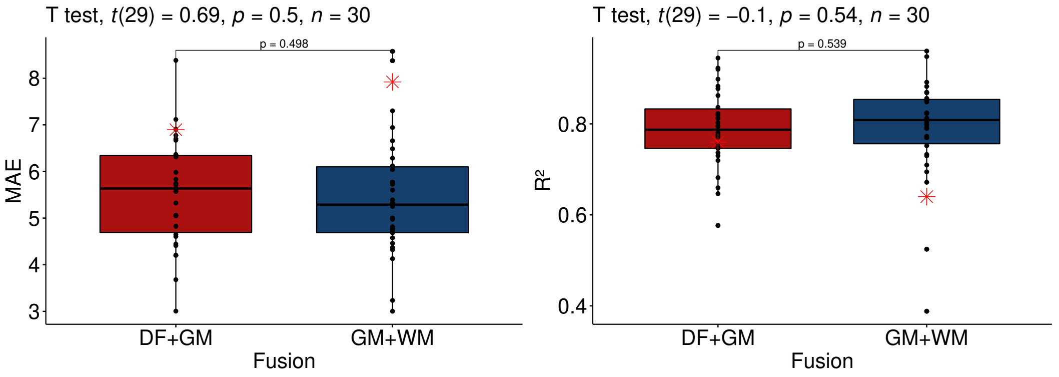
(a) Slices of the original image of DF **(b)** Slices of the reconstructed image of DF

(c) Slices of the original image of GM **(d)** Slices of the reconstructed image of GM

(e) Slices of the original image of WM **(f)** Slices of the reconstructed image of WM

(g) Slices of the original image of CSF **(h)** Slices of the reconstructed image of CSF

Figure B.2: Original image and reconstructed image from 40-components of the deformation fields (DF), GM, WM and CSF. The image is from the subject ID 510 of the IOP site.



(a) **(b)** Min processed Diamarker

Figure B.3: Boxplot of the MAE per fusion model: The statistical results for a paired t-test. Each point in the boxplot is the (a) MAE (b) R^2 for a cross-validation fold. The horizontal black line represents the median value, and the limits of the boxes represent the 1^{st} and 3^{rd} quartiles. The red star identifies the result on the test set.

Appendix C

Overcoming Data Availability in MRI Studies: Leveraging Off-the-Shelf Deep Learning Models

C.1 Age Prediction and Sex classification

C.1.1 Methods

Table C.1: Demographics of the participants used to train and validate the autoencoder.

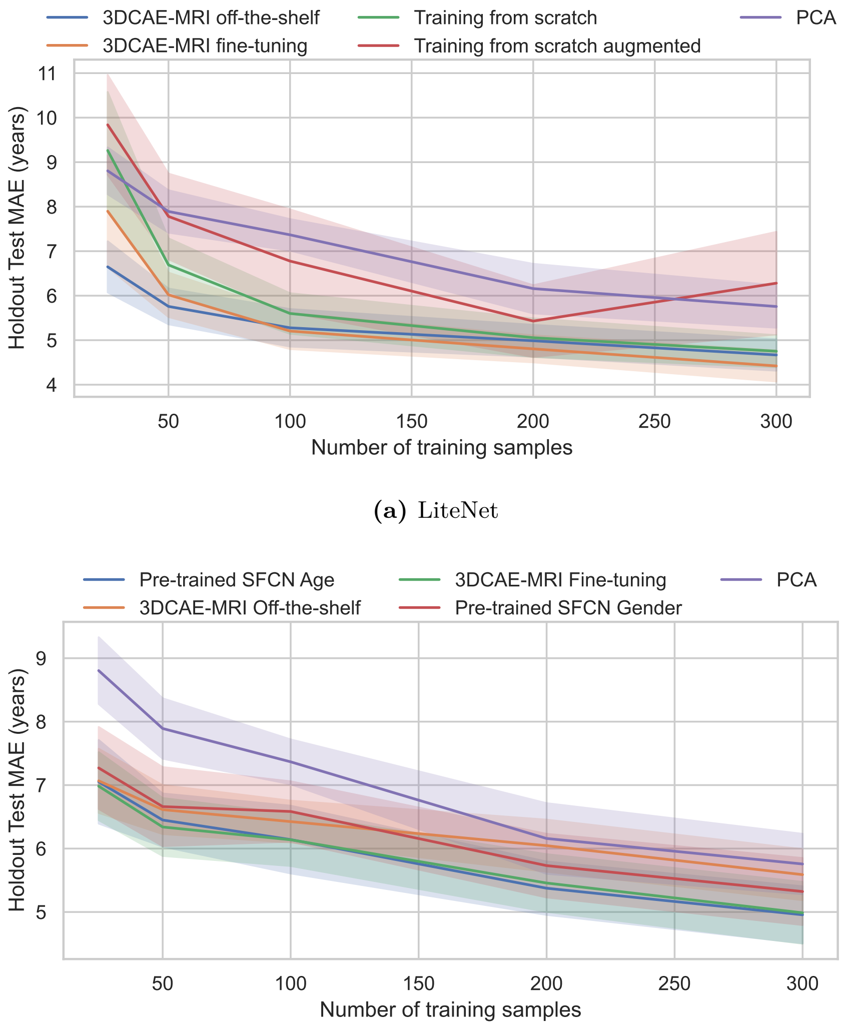
Repository	**Site**	*Total of participants*	Number males	Mean and standard deviation [years]	Min Age [years]	Max Age [years]
ABIDE I	California Institute of Technology	37	30	28.4 ± 10.7	17.0	56.2
ABIDE I	Carnegie Mellon University	26	20	26.8 ± 5.7	19.0	40.0
ABIDE I	Kennedy Krieger Institute	54	41	10.1 ± 1.3	8.1	12.8
ABIDE I	Ludwig Maximilians University Munich	56	49	25.7 ± 11.7	7.0	58.0
ABIDE I	NYU Langone Medical Center	182	145	15.3 ± 6.6	6.5	39.1
ABIDE I	Olin Institute of Livingat Hartford Hospital	16	14	16.9 ± 3.7	10.0	23.0
ABIDE I	Oregon Health and Science University	28	28	10.8 ± 1.9	8.0	15.2
ABIDE I	San Diego State University	36	29	14.4 ± 1.8	8.7	17.2
ABIDE I	Social Brain Lab	28	28	33.4 ± 6.8	20.0	49.0
ABIDE I	Stanford University	37	30	9.9 ± 1.6	7.5	12.9
ABIDE I	Trinity Centre for Health Sciences	49	49	17.2 ± 3.6	12.0	25.9
ABIDE I	University of California Los Angeles	97	86	13.0 ± 2.2	8.4	17.9
ABIDE I	University of Leuven	64	56	18.0 ± 5.0	12.1	32.0
ABIDE I	University of Michigan	143	116	14.0 ± 3.2	8.2	28.8
ABIDE I	University of Pittsburgh School of Medicine	55	48	18.9 ± 6.9	9.3	35.2
ABIDE I	University of Utah School of Medicine	100	100	22.1 ± 7.7	8.8	50.2
ABIDE I	Yale Child Study Center	56	40	12.7 ± 2.9	7.0	17.8
ABIDE II	Barrow Neurological Institute	58	58	38.5 ± 15.5	18.0	64.0
ABIDE II	ETH Zurich	37	37	22.7 ± 4.4	13.8	30.7
ABIDE II	Erasmus University Medical Center Rotterdam	54	44	8.1 ± 1.1	6.2	10.7
ABIDE II	Georgetown University	103	68	10.7 ± 1.6	8.1	13.9
ABIDE II	Indiana University	26	20	24.8 ± 8.5	17.0	54.0
ABIDE II	Institut Pasteur and Robert Debré Hospital	55	25	20.1 ± 10.5	6.1	46.6
ABIDE II	Katholieke Universiteit Leuven	28	28	23.6 ± 4.8	18.0	35.0
ABIDE II	Kennedy Krieger Institute	207	137	10.3 ± 1.3	8.0	13.0
ABIDE II	NYU Langone Medical Center Sample 1	74	67	9.9 ± 5.0	5.2	34.8
ABIDE II	NYU Langone Medical Center Sample 2	27	24	6.8 ± 1.1	5.1	8.8
ABIDE II	Oregon Health and Science University	93	57	10.9 ± 2.0	7.0	15.0
ABIDE II	SanDiego State University	56	47	12.9 ± 3.1	7.4	18.0
ABIDE II	Stanford University	41	37	11.1 ± 1.2	8.4	13.2
ABIDE II	Trinity Centre for Health Sciences	42	42	15.2 ± 3.2	10.0	20.0
ABIDE II	University of California Davis	32	24	14.8 ± 1.8	12.0	17.8
ABIDE II	University of California Los Angeles	31	25	10.8 ± 2.4	7.8	15.0
ABIDE II	University of California Los Angeles Longitudinal Sample	37	35	13.5 ± 1.9	10.0	17.2
ABIDE II	University of Miami	26	20	9.8 ± 2.1	7.1	14.3
ABIDE II	University of Pittsburgh	34	26	14.9 ± 2.4	9.3	19.5
ABIDE II	University of Utah School of Medicine	32	27	20.9 ± 7.9	9.1	38.9
ADNI	–	18705	10233	75.3 ± 7.4	51.0	97.0
GSP	–	1558	661	21.5 ± 2.9	19.0	35.0
OASIS1	–	1683	638	51.5 ± 25.3	18.0	96.0
OASIS2	–	1345	576	76.9 ± 7.6	60.0	98.0
OASIS3	–	2768	1185	70.7 ± 9.3	42.7	97.0
FCP1000	AnnArbor a	25	20	20.4 ± 7.7	13.0	40.0
FCP1000	AnnArbor b	36	17	348.0 ± 1732.7	19.0	9999.0
FCP1000	Atlanta	28	11	30.6 ± 9.2	23.0	54.0
FCP1000	Baltimore	23	8	29.3 ± 5.5	20.0	40.0
FCP1000	Bangor	20	16	22.6 ± 4.6	19.0	38.0
FCP1000	Beijing Zang	197	68	21.1 ± 1.8	18.0	26.0
FCP1000	Berlin Margulies	26	12	29.9 ± 5.2	24.0	44.0
FCP1000	Cambridge Buckner	198	68	20.9 ± 2.1	18.0	29.0
FCP1000	Dallas	24	10	42.9 ± 20.4	20.0	71.0
FCP1000	ICBM	86	0	–	–	–
FCP1000	Leiden 2180	12	9	23.6 ± 2.6	20.0	27.0
FCP1000	Leiden 2200	19	11	21.8 ± 2.7	18.0	28.0
FCP1000	Leipzig	37	13	25.8 ± 5.1	20.0	42.0
FCP1000	Milwaukee a	18	0	–	–	–
FCP1000	Milwaukee b	46	14	53.7 ± 5.9	44.0	65.0
FCP1000	Munchen	16	9	68.3 ± 4.1	63.0	74.0
FCP1000	NYU TRT session1b	12	0	–		
FCP1000	NewHaven a	18	10	31.6 ± 10.3	18.0	48.0
FCP1000	NewHaven b	15	7	27.6 ± 6.4	18.0	42.0
FCP1000	NewYork a	84	40	24.3 ± 10.1	7.0	49.0
FCP1000	NewYork a ADHD	25	18	34.9 ± 9.6	20.0	50.0
FCP1000	NewYork b	20	1	$40.0 \pm nan$	40.0	40.0
FCP1000	Newark	19	9	24.1 ± 3.9	21.0	39.0
FCP1000	Ontario	9	0	–	–	–
FCP1000	Orangeburg	20	12	41.6 ± 11.2	20.0	55.0
FCP1000	Oulu	102	33	21.5 ± 0.6	20.0	23.0
FCP1000	Oxford	22	11	29.3 ± 3.3	21.0	35.0
FCP1000	PaloAlto	17	2	31.6 ± 7.6	22.0	46.0
FCP1000	Pittsburgh	16	9	37.6 ± 8.7	25.0	54.0
FCP1000	Queensland	19	10	25.9 ± 4.1	20.0	34.0
FCP1000	SaintLouis	31	13	25.3 ± 2.3	21.0	29.0
FCP1000	Taipei a	14	0	–	–	–
FCP1000	Taipei b	8	0	–	–	–

Table C.2: Demographics of the participants used to train, validate and test the age prediction and sex classification models.

Site	*Total of participants*	Number males	Mean and standard deviation [years]	Min Age [years]	Max Age [years]
GH	312	139	50.73 ± 15.98	20.07	86.20
HH	179	85	47.63 ± 16.61	20.17	81.94
IOP	67	24	42.13 ± 16.60	19.98	86.32

C.2 Results

C.2.1 Age Prediction



(a) LiteNet

(b) SFCN

Figure C.1: Holdout test set MAE results for age prediction problem. Training evolution for the age prediction problem with different dataset training sizes for the four training strategies (3DCAE-MRI off-the-shelf, 3DCAE-MRI fine-tuning, training from scratch and PCA-RVM). 3DCAE-MRI off-the-shelf and 3DCAE-MRI fine-tuning represents the transfer learning from the 3DCAE-MRI to the CNN. The shaded band represents to the standard deviation.

Table C.3: LiteNet: ANOVA results for the holdout test set comparing the different training strategies for the age prediction problem and for the different training dataset sizes. The significant $p-$values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	63.84	2.01×10^{-28}
50	4	116	95.79	7.91×10^{-36}
100	4	116	104.66	1.51×10^{-37}
200	4	116	55.10	5.9×10^{-26}
300	4	116	66.32	4.41×10^{-29}

Table C.4: LiteNet Age prediction: Post-hoc results for the statistically significant ANOVA comparing the MAE on holdout test set across different training strategies for different training dataset sizes. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	5.98	29.0	1.28	**1.67×10^{-6}**
25	3DCAE-MRI fine-tuning	PCA	-4.33	29.0	-0.94	**0.00016**
25	3DCAE-MRI fine-tuning	Training from scratch	-4.68	29.0	-1.06	**6.23×10^{-5}**
25	3DCAE-MRI fine-tuning	Training from scratch augmented	-7.10	29.0	-1.63	**8.11×10^{-8}**
25	3DCAE-MRI off-the-shelf	PCA	-24.36	29.0	-3.91	**7.38×10^{-21}**
25	3DCAE-MRI off-the-shelf	Training from scratch	-11.26	29.0	-2.58	**4.18×10^{-12}**
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	-17.01	29.0	-3.58	**1.26×10^{-16}**
25	PCA	Training from scratch	-1.96	29.0	-0.46	0.06
25	PCA	Training from scratch augmented	-4.80	29.0	-1.18	**4.42×10^{-5}**
25	Training from scratch	Training from scratch augmented	-2.72	29.0	-0.47	**0.011**
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	4.13	29.0	0.58	**0.00028**
50	3DCAE-MRI fine-tuning	PCA	-15.83	29.0	-3.84	**8.33×10^{-16}**
50	3DCAE-MRI fine-tuning	Training from scratch	-5.64	29.0	-1.22	**4.24×10^{-6}**
50	3DCAE-MRI fine-tuning	Training from scratch augmented	-9.41	29.0	-2.30	**2.58×10^{-10}**
50	3DCAE-MRI off-the-shelf	PCA	-26.62	29.0	-4.84	**6.23×10^{-22}**
50	3DCAE-MRI off-the-shelf	Training from scratch	-9.56	29.0	-1.83	**1.82×10^{-10}**
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	-11.33	29.0	-2.74	**3.59×10^{-12}**
50	PCA	Training from scratch	10.26	29.0	2.22	**3.69×10^{-11}**
50	PCA	Training from scratch augmented	0.65	29.0	0.15	0.52
50	Training from scratch	Training from scratch augmented	-5.26	29.0	-1.36	**1.23×10^{-5}**
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-1.13	29.0	-0.18	0.27
100	3DCAE-MRI fine-tuning	PCA	-34.15	29.0	-5.68	**5.69×10^{-25}**
100	3DCAE-MRI fine-tuning	Training from scratch	-6.93	29.0	-0.92	**1.28×10^{-7}**
100	3DCAE-MRI fine-tuning	Training from scratch augmented	-8.04	29.0	-1.81	**7.33×10^{-9}**
100	3DCAE-MRI off-the-shelf	PCA	-28.12	29.0	-5.37	**1.36×10^{-22}**
100	3DCAE-MRI off-the-shelf	Training from scratch	-4.13	29.0	-0.73	**0.00028**
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	-7.62	29.0	-1.72	**2.12×10^{-8}**
100	PCA	Training from scratch	23.48	29.0	4.33	**2.03×10^{-20}**
100	PCA	Training from scratch augmented	2.86	29.0	0.69	**0.0078**
100	Training from scratch	Training from scratch augmented	-6.57	29.0	-1.33	**3.35×10^{-7}**
200	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-3.66	29.0	-0.55	**0.001**
200	3DCAE-MRI fine-tuning	PCA	-13.41	29.0	-3.02	**5.83×10^{-14}**
200	3DCAE-MRI fine-tuning	Training from scratch	-4.09	29.0	-0.67	**0.00031**
200	3DCAE-MRI fine-tuning	Training from scratch augmented	-5.20	29.0	-1.03	**1.45×10^{-5}**
200	3DCAE-MRI off-the-shelf	PCA	-11.66	29.0	-2.50	**1.8×10^{-12}**
200	3DCAE-MRI off-the-shelf	Training from scratch	-1.17	29.0	-0.17	0.25
200	3DCAE-MRI off-the-shelf	Training from scratch augmented	-3.53	29.0	-0.71	**0.0014**
200	PCA	Training from scratch	10.84	29.0	2.21	**1.02×10^{-11}**
200	PCA	Training from scratch augmented	5.28	29.0	1.05	**1.18×10^{-5}**
200	Training from scratch	Training from scratch augmented	-3.01	29.0	-0.58	**0.0053**
300	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-5.91	29.0	-0.70	**2.03×10^{-6}**
300	3DCAE-MRI fine-tuning	PCA	-18.20	29.0	-3.20	**2.07×10^{-17}**
300	3DCAE-MRI fine-tuning	Training from scratch	-6.28	29.0	-0.93	**7.33×10^{-7}**
300	3DCAE-MRI fine-tuning	Training from scratch augmented	-9.13	29.0	-2.18	**4.99×10^{-10}**
300	3DCAE-MRI off-the-shelf	PCA	-17.07	29.0	-2.60	**1.15×10^{-16}**
300	3DCAE-MRI off-the-shelf	Training from scratch	-1.43	29.0	-0.23	0.16
300	3DCAE-MRI off-the-shelf	Training from scratch augmented	-7.79	29.0	-1.89	**1.35×10^{-8}**
300	PCA	Training from scratch	12.37	29.0	2.38	**4.33×10^{-13}**
300	PCA	Training from scratch augmented	-2.33	29.0	-0.59	**0.027**
300	Training from scratch	Training from scratch augmented	-8.20	29.0	-1.79	**4.82×10^{-9}**

Table C.5: SFCN Age prediction: ANOVA results for the holdout test set when comparing different training strategies for the age prediction problem and for the different training dataset sizes. The significant $p-$values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	123.54	7.3×10^{-41}
50	4	116	96.11	6.82×10^{-36}
100	4	116	54.00	1.26×10^{-25}
200	4	116	21.85	1.85×10^{-13}
300	4	116	19.69	2.16×10^{-12}

Table C.6: SFCN Age prediction: Post-hoc results for the statistically significant ANOVA comparing the MAE of the holdout test set across different training strategies for different training dataset sizes. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-1.28	29.0	-0.16	0.21
25	3DCAE-MRI Fine-tuning	PCA	-15.98	29.0	-3.43	**6.45×10^{-16}**
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-0.82	29.0	-0.11	0.42
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-3.01	29.0	-0.48	**0.0054**
25	3DCAE-MRI Off-the-shelf	PCA	-17.34	29.0	-3.37	**7.51×10^{-17}**
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	0.18	29.0	0.03	0.86
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-2.64	29.0	-0.35	**0.013**
25	PCA	Pre-trained SFCN Age	15.08	29.0	2.94	**2.91×10^{-15}**
25	PCA	Pre-trained SFCN Gender	13.65	29.0	2.60	**3.75×10^{-14}**
25	Pre-trained SFCN Age	Pre-trained SFCN Gender	-1.84	29.0	-0.33	0.076
50	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-2.93	29.0	-0.65	**0.0066**
50	3DCAE-MRI Fine-tuning	PCA	-15.63	29.0	-3.33	**1.15×10^{-15}**
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-1.22	29.0	-0.25	0.23
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-3.25	29.0	-0.59	**0.003**
50	3DCAE-MRI Off-the-shelf	PCA	-14.38	29.0	-2.95	**9.9×10^{-15}**
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	2.11	29.0	0.41	**0.044**
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-0.54	29.0	-0.09	0.6
50	PCA	Pre-trained SFCN Age	17.23	29.0	3.22	**8.97×10^{-17}**
50	PCA	Pre-trained SFCN Gender	13.37	29.0	2.22	**6.22×10^{-14}**
50	Pre-trained SFCN Age	Pre-trained SFCN Gender	-2.31	29.0	-0.40	**0.028**
100	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-5.30	29.0	-0.77	**1.11×10^{-5}**
100	3DCAE-MRI Fine-tuning	PCA	-13.14	29.0	-3.20	**9.74×10^{-14}**
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-0.03	29.0	-0.01	0.98
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-3.87	29.0	-1.00	**0.00057**
100	3DCAE-MRI Off-the-shelf	PCA	-11.48	29.0	-2.75	**2.62×10^{-12}**
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	3.25	29.0	0.64	**0.0029**
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-1.65	29.0	-0.39	0.11
100	PCA	Pre-trained SFCN Age	10.91	29.0	2.71	**8.76×10^{-12}**
100	PCA	Pre-trained SFCN Gender	7.62	29.0	1.87	**2.13×10^{-8}**
100	Pre-trained SFCN Age	Pre-trained SFCN Gender	-4.20	29.0	-0.88	**0.00023**
200	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-5.89	29.0	-1.36	**2.18×10^{-6}**
200	3DCAE-MRI Fine-tuning	PCA	-5.34	29.0	-1.38	**9.91×10^{-6}**
200	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	0.87	29.0	0.19	0.39
200	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-2.23	29.0	-0.57	**0.034**
200	3DCAE-MRI Off-the-shelf	PCA	-1.46	29.0	-0.23	0.15
200	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	7.05	29.0	1.61	**9.33×10^{-8}**
200	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	3.37	29.0	0.69	**0.0021**
200	PCA	Pre-trained SFCN Age	6.73	29.0	1.59	**2.2×10^{-7}**
200	PCA	Pre-trained SFCN Gender	4.41	29.0	0.81	**0.00013**
200	Pre-trained SFCN Age	Pre-trained SFCN Gender	-3.24	29.0	-0.77	**0.003**
300	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-6.96	29.0	-1.35	**1.19×10^{-7}**
300	3DCAE-MRI Fine-tuning	PCA	-5.95	29.0	-1.60	**1.81×10^{-6}**
300	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	0.24	29.0	0.07	0.82
300	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-2.92	29.0	-0.66	**0.0068**
300	3DCAE-MRI Off-the-shelf	PCA	-1.59	29.0	-0.38	0.12
300	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	6.11	29.0	1.48	**1.17×10^{-6}**
300	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	2.36	29.0	0.57	**0.025**
300	PCA	Pre-trained SFCN Age	7.02	29.0	1.73	**1.0×10^{-7}**
300	PCA	Pre-trained SFCN Gender	4.07	29.0	0.86	**0.00033**
300	Pre-trained SFCN Age	Pre-trained SFCN Gender	-3.02	29.0	-0.75	**0.0052**

Table C.7: LiteNet Age prediction: ANOVA comparing the means of MAE of the external test set across different training strategies for the different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	87.00	5.35×10^{-34}
50	4	116	85.27	1.27×10^{-33}
100	4	116	84.51	1.87×10^{-33}
200	4	116	57.27	1.37×10^{-26}
300	4	116	78.89	3.49×10^{-32}

Table C.8: LiteNet Age prediction: Post-hoc results for the statistically significant ANOVA comparing the mean MAE of the external test set across different training strategies for different number of training instances. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	8.97	29.0	1.47	**7.26×10^{-10}**
25	3DCAE-MRI fine-tuning	PCA	-7.51	29.0	-1.83	**2.81×10^{-8}**
25	3DCAE-MRI fine-tuning	Training from scratch	-9.22	29.0	-1.65	**3.99×10^{-10}**
25	3DCAE-MRI fine-tuning	Training from scratch augmented	-9.74	29.0	-1.71	**1.19×10^{-10}**
25	3DCAE-MRI off-the-shelf	PCA	-19.21	29.0	-4.09	**4.88×10^{-18}**
25	3DCAE-MRI off-the-shelf	Training from scratch	-13.46	29.0	-2.74	**5.29×10^{-14}**
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	-13.83	29.0	-2.80	**2.69×10^{-14}**
25	PCA	Training from scratch	-2.51	29.0	-0.58	**0.018**
25	PCA	Training from scratch augmented	-2.82	29.0	-0.65	**0.0085**
25	Training from scratch	Training from scratch augmented	-0.39	29.0	-0.05	0.7
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	6.62	29.0	1.30	**2.95×10^{-7}**
50	3DCAE-MRI fine-tuning	PCA	-13.24	29.0	-2.61	**8.04×10^{-14}**
50	3DCAE-MRI fine-tuning	Training from scratch	-7.20	29.0	-1.48	**6.26×10^{-8}**
50	3DCAE-MRI fine-tuning	Training from scratch augmented	-11.53	29.0	-2.26	**2.35×10^{-12}**
50	3DCAE-MRI off-the-shelf	PCA	-26.23	29.0	-5.15	**9.45×10^{-22}**
50	3DCAE-MRI off-the-shelf	Training from scratch	-9.73	29.0	-2.30	**1.24×10^{-10}**
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	-14.69	29.0	-3.28	**5.8×10^{-15}**
50	PCA	Training from scratch	-0.16	29.0	-0.03	0.88
50	PCA	Training from scratch augmented	-3.26	29.0	-0.69	**0.0028**
50	Training from scratch	Training from scratch augmented	-2.58	29.0	-0.47	**0.015**
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	4.45	29.0	1.10	**0.00012**
100	3DCAE-MRI fine-tuning	PCA	-14.57	29.0	-3.04	**7.14×10^{-15}**
100	3DCAE-MRI fine-tuning	Training from scratch	-5.27	29.0	-0.85	**1.19×10^{-5}**
100	3DCAE-MRI fine-tuning	Training from scratch augmented	-11.63	29.0	-2.18	**1.92×10^{-12}**
100	3DCAE-MRI off-the-shelf	PCA	-29.09	29.0	-6.35	**5.2×10^{-23}**
100	3DCAE-MRI off-the-shelf	Training from scratch	-6.26	29.0	-1.55	**7.83×10^{-7}**
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	-12.01	29.0	-2.90	**8.93×10^{-13}**
100	PCA	Training from scratch	3.83	29.0	0.85	**0.00063**
100	PCA	Training from scratch augmented	-3.45	29.0	-0.75	**0.0017**
100	Training from scratch	Training from scratch augmented	-8.58	29.0	-1.17	**1.9×10^{-9}**
200	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	3.51	29.0	0.91	**0.0015**
200	3DCAE-MRI fine-tuning	PCA	-7.90	29.0	-1.91	**1.03×10^{-8}**
200	3DCAE-MRI fine-tuning	Training from scratch	-8.76	29.0	-1.86	**1.21×10^{-9}**
200	3DCAE-MRI fine-tuning	Training from scratch augmented	-9.55	29.0	-1.98	**1.84×10^{-10}**
200	3DCAE-MRI off-the-shelf	PCA	-16.00	29.0	-3.74	**6.26×10^{-16}**
200	3DCAE-MRI off-the-shelf	Training from scratch	-8.98	29.0	-2.44	**7.2×10^{-10}**
200	3DCAE-MRI off-the-shelf	Training from scratch augmented	-9.52	29.0	-2.46	**2.0×10^{-10}**
200	PCA	Training from scratch	-3.74	29.0	-0.97	**0.00082**
200	PCA	Training from scratch augmented	-5.05	29.0	-1.22	**2.18×10^{-5}**
200	Training from scratch	Training from scratch augmented	-1.85	29.0	-0.31	0.074
300	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	4.88	29.0	1.22	**3.57×10^{-5}**
300	3DCAE-MRI fine-tuning	PCA	-2.41	29.0	-0.49	**0.023**
300	3DCAE-MRI fine-tuning	Training from scratch	-7.64	29.0	-1.80	**2.03×10^{-8}**
300	3DCAE-MRI fine-tuning	Training from scratch augmented	-10.15	29.0	-2.58	**4.75×10^{-11}**
300	3DCAE-MRI off-the-shelf	PCA	-12.40	29.0	-3.08	**4.06×10^{-13}**
300	3DCAE-MRI off-the-shelf	Training from scratch	-10.05	29.0	-2.63	**5.84×10^{-11}**
300	3DCAE-MRI off-the-shelf	Training from scratch augmented	-12.85	29.0	-3.25	**1.68×10^{-13}**
300	PCA	Training from scratch	-7.07	29.0	-1.71	**8.98×10^{-8}**
300	PCA	Training from scratch augmented	-9.70	29.0	-2.53	**1.3×10^{-10}**
300	Training from scratch	Training from scratch augmented	-3.80	29.0	-0.97	**0.00068**

Table C.9: SFCN Age prediction: ANOVA comparing the means of MAE of the holdout test set across different training strategies for the different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	306.94	1.07×10^{-60}
50	4	116	240.50	3.71×10^{-55}
100	4	116	232.30	2.22×10^{-54}
200	4	116	63.12	3.13×10^{-28}
300	4	116	28.59	1.56×10^{-16}

Table C.10: SFCN Age prediction: Post-hoc results for the statistically significant ANOVA comparing the mean MAE of the external test set across different training strategies for different number of training instances. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	1.43	29.0	0.18	0.16
25	3DCAE-MRI Fine-tuning	PCA	-30.54	29.0	-4.87	**1.33×10^{-23}**
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-0.68	29.0	-0.09	0.5
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-3.42	29.0	-0.55	**0.0019**
25	3DCAE-MRI Off-the-shelf	PCA	-31.94	29.0	-5.60	**3.77×10^{-24}**
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	-1.95	29.0	-0.30	0.06
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-4.57	29.0	-0.74	**8.46×10^{-5}**
25	PCA	Pre-trained SFCN Age	30.20	29.0	5.37	**1.82×10^{-23}**
25	PCA	Pre-trained SFCN Gender	18.42	29.0	3.75	**1.51×10^{-17}**
25	Pre-trained SFCN Age	Pre-trained SFCN Gender	-2.77	29.0	-0.51	**0.0096**
50	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-1.31	29.0	-0.30	0.2
50	3DCAE-MRI Fine-tuning	PCA	-30.20	29.0	-6.14	**1.83×10^{-23}**
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-5.28	29.0	-1.00	**1.17×10^{-5}**
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-2.90	29.0	-0.70	**0.007**
50	3DCAE-MRI Off-the-shelf	PCA	-26.89	29.0	-6.47	**4.72×10^{-22}**
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	-4.44	29.0	-0.82	**0.00012**
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-2.12	29.0	-0.51	**0.043**
50	PCA	Pre-trained SFCN Age	23.60	29.0	4.97	**1.76×10^{-20}**
50	PCA	Pre-trained SFCN Gender	18.71	29.0	4.57	**9.97×10^{-18}**
50	Pre-trained SFCN Age	Pre-trained SFCN Gender	0.67	29.0	0.17	0.51
100	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	2.00	29.0	0.49	0.055
100	3DCAE-MRI Fine-tuning	PCA	-29.97	29.0	-6.17	**2.25×10^{-23}**
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-1.48	29.0	-0.32	0.15
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-3.33	29.0	-0.78	**0.0024**
100	3DCAE-MRI Off-the-shelf	PCA	-28.57	29.0	-7.89	**8.7×10^{-23}**
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	-4.05	29.0	-0.84	**0.00035**
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-5.22	29.0	-1.24	**1.39×10^{-5}**
100	PCA	Pre-trained SFCN Age	22.93	29.0	5.59	**3.93×10^{-20}**
100	PCA	Pre-trained SFCN Gender	16.21	29.0	4.20	**4.45×10^{-16}**
100	Pre-trained SFCN Age	Pre-trained SFCN Gender	-2.33	29.0	-0.49	**0.027**
200	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	2.81	29.0	0.73	**0.0087**
200	3DCAE-MRI Fine-tuning	PCA	-16.75	29.0	-3.75	**1.9×10^{-16}**
200	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-2.17	29.0	-0.59	**0.038**
200	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-2.62	29.0	-0.60	**0.014**
200	3DCAE-MRI Off-the-shelf	PCA	-22.16	29.0	-5.28	**9.95×10^{-20}**
200	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	-5.19	29.0	-1.22	**1.51×10^{-5}**
200	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-5.16	29.0	-1.24	**1.61×10^{-5}**
200	PCA	Pre-trained SFCN Age	8.76	29.0	2.32	**1.22×10^{-9}**
200	PCA	Pre-trained SFCN Gender	11.32	29.0	2.36	**3.69×10^{-12}**
200	Pre-trained SFCN Age	Pre-trained SFCN Gender	-0.01	29.0	-0.00	1.0
300	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	6.98	29.0	2.05	**1.12×10^{-7}**
300	3DCAE-MRI Fine-tuning	PCA	-3.67	29.0	-1.05	**0.00096**
300	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-2.59	29.0	-0.64	**0.015**
300	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	1.33	29.0	0.36	0.2
300	3DCAE-MRI Off-the-shelf	PCA	-21.15	29.0	-5.05	**3.59×10^{-19}**
300	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	-8.01	29.0	-2.09	**7.75×10^{-9}**
300	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-5.34	29.0	-1.36	**9.98×10^{-6}**
300	PCA	Pre-trained SFCN Age	-0.22	29.0	-0.06	0.83
300	PCA	Pre-trained SFCN Gender	5.42	29.0	1.35	**7.93×10^{-6}**
300	Pre-trained SFCN Age	Pre-trained SFCN Gender	3.33	29.0	0.89	**0.0024**

Table C.11: LiteNet Age prediction: ANOVA results comparing the stability metric across different training strategies for the different dataset sizes tested. The significant p-values are shown in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	3	87	244.51	2.87×10^{-42}
50	3	87	71.63	2.0×10^{-23}
100	3	87	89.77	1.52×10^{-26}
200	3	87	120.87	6.32×10^{-31}
300	3	87	102.65	1.75×10^{-28}

Table C.12: LiteNet Age prediction: Post-hoc results for the statistically significant ANOVA comparing the stability metric across different training strategies for the different dataset sizes tested. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	27.22	29.0	5.22	$\mathbf{3.36 \times 10^{-22}}$
25	3DCAE-MRI fine-tuning	Training from scratch	-7.14	29.0	-1.49	$\mathbf{7.49 \times 10^{-8}}$
25	3DCAE-MRI fine-tuning	Training from scratch augmented	-2.28	29.0	-0.43	**0.03**
25	3DCAE-MRI off-the-shelf	Training from scratch	-20.72	29.0	-5.71	$\mathbf{6.25 \times 10^{-19}}$
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	-26.19	29.0	-4.88	$\mathbf{9.89 \times 10^{-22}}$
25	Training from scratch	Training from scratch augmented	3.76	29.0	0.98	**0.00077**
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	13.34	29.0	3.29	$\mathbf{6.68 \times 10^{-14}}$
50	3DCAE-MRI fine-tuning	Training from scratch	-3.68	29.0	-0.72	**0.00094**
50	3DCAE-MRI fine-tuning	Training from scratch augmented	-5.87	29.0	-1.59	$\mathbf{2.31 \times 10^{-6}}$
50	3DCAE-MRI off-the-shelf	Training from scratch	-11.02	29.0	-2.75	$\mathbf{6.96 \times 10^{-12}}$
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	-14.27	29.0	-3.36	$\mathbf{1.21 \times 10^{-14}}$
50	Training from scratch	Training from scratch augmented	-3.28	29.0	-0.81	**0.0027**
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	11.50	29.0	2.57	$\mathbf{2.52 \times 10^{-12}}$
100	3DCAE-MRI fine-tuning	Training from scratch	-4.54	29.0	-0.79	$\mathbf{9.15 \times 10^{-5}}$
100	3DCAE-MRI fine-tuning	Training from scratch augmented	-8.21	29.0	-1.98	$\mathbf{4.77 \times 10^{-9}}$
100	3DCAE-MRI off-the-shelf	Training from scratch	-10.77	29.0	-2.51	$\mathbf{1.18 \times 10^{-11}}$
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	-14.97	29.0	-3.90	$\mathbf{3.54 \times 10^{-15}}$
100	Training from scratch	Training from scratch augmented	-4.26	29.0	-0.95	**0.0002**
200	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	10.42	29.0	2.06	$\mathbf{2.6 \times 10^{-11}}$
200	3DCAE-MRI fine-tuning	Training from scratch	-5.41	29.0	-1.03	$\mathbf{8.02 \times 10^{-6}}$
200	3DCAE-MRI fine-tuning	Training from scratch augmented	-12.05	29.0	-2.68	$\mathbf{8.16 \times 10^{-13}}$
200	3DCAE-MRI off-the-shelf	Training from scratch	-9.64	29.0	-2.19	$\mathbf{1.49 \times 10^{-10}}$
200	3DCAE-MRI off-the-shelf	Training from scratch augmented	-16.39	29.0	-3.92	$\mathbf{3.33 \times 10^{-16}}$
200	Training from scratch	Training from scratch augmented	-6.86	29.0	-1.22	$\mathbf{1.54 \times 10^{-7}}$
300	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	10.64	29.0	2.36	$\mathbf{1.6 \times 10^{-11}}$
300	3DCAE-MRI fine-tuning	Training from scratch	-6.72	29.0	-1.56	$\mathbf{2.26 \times 10^{-7}}$
300	3DCAE-MRI fine-tuning	Training from scratch augmented	-9.79	29.0	-2.61	$\mathbf{1.07 \times 10^{-10}}$
300	3DCAE-MRI off-the-shelf	Training from scratch	-11.81	29.0	-2.79	$\mathbf{1.33 \times 10^{-12}}$
300	3DCAE-MRI off-the-shelf	Training from scratch augmented	-15.22	29.0	-4.03	$\mathbf{2.3 \times 10^{-15}}$
300	Training from scratch	Training from scratch augmented	-3.49	29.0	-0.69	**0.0016**

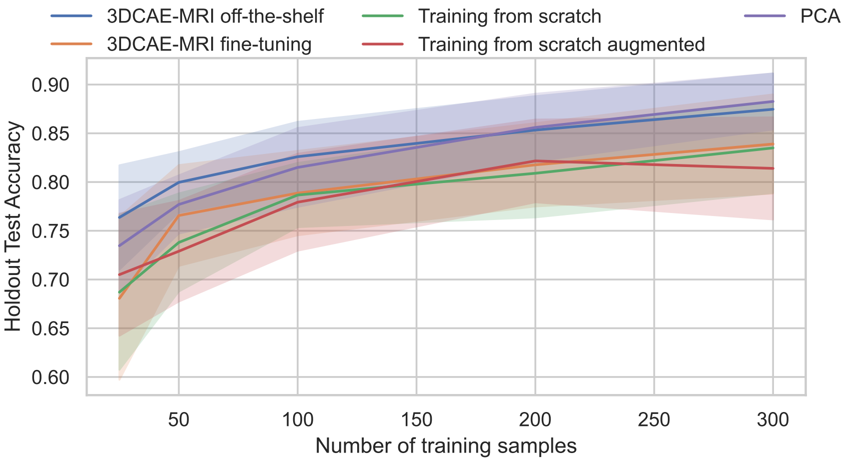
Table C.13: SFCN Age prediction: ANOVA results comparing the stability metric across different training strategies for the different dataset sizes tested. The significant p-values are shown in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	3	87	14.66	8.25×10^{-8}
50	3	87	39.38	3.58×10^{-16}
100	3	87	32.52	3.42×10^{-14}
200	3	87	67.84	1.05×10^{-22}
300	3	87	108.71	2.48×10^{-29}

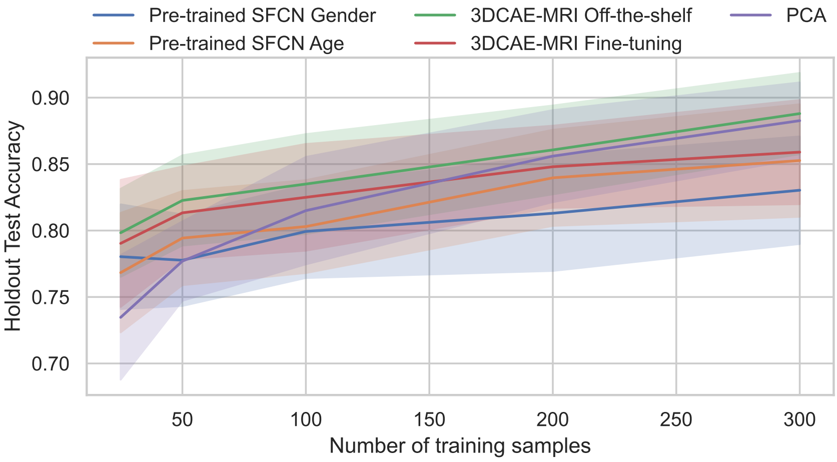
Table C.14: SFCN Age prediction: Post-hoc results for the statistically significant ANOVA comparing the stability metric across different training strategies for the different dataset sizes tested. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	1.83	29.0	0.47	0.077
25	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	7.68	29.0	1.79	**1.79×10^{-8}**
25	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	0.49	29.0	0.13	0.63
25	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	5.38	29.0	1.24	**8.84×10^{-6}**
25	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	-1.29	29.0	-0.29	0.21
25	Pre-trained SFCN Age	Pre-trained SFCN Gender	-5.33	29.0	-1.38	**1.0×10^{-5}**
50	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	0.33	29.0	0.09	0.74
50	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	10.19	29.0	2.62	**4.26×10^{-11}**
50	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	3.68	29.0	0.91	**0.00094**
50	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	9.75	29.0	2.34	**1.17×10^{-10}**
50	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	2.86	29.0	0.75	**0.0077**
50	Pre-trained SFCN Age	Pre-trained SFCN Gender	-6.84	29.0	-1.86	**1.63×10^{-7}**
100	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-3.21	29.0	-0.75	**0.0033**
100	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	7.01	29.0	2.04	**1.05×10^{-7}**
100	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	1.03	29.0	0.27	0.31
100	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	8.23	29.0	2.00	**4.55×10^{-9}**
100	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	3.37	29.0	0.92	**0.0021**
100	Pre-trained SFCN Age	Pre-trained SFCN Gender	-8.17	29.0	-1.91	**5.17×10^{-9}**
200	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-5.19	29.0	-1.32	**1.51×10^{-5}**
200	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	11.78	29.0	2.85	**1.42×10^{-12}**
200	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	0.62	29.0	0.18	0.54
200	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	11.84	29.0	2.93	**1.25×10^{-12}**
200	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	5.71	29.0	1.44	**3.54×10^{-6}**
200	Pre-trained SFCN Age	Pre-trained SFCN Gender	-13.29	29.0	-2.78	**7.26×10^{-14}**
300	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-9.84	29.0	-2.37	**9.48×10^{-11}**
300	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	7.20	29.0	2.06	**6.33×10^{-8}**
300	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	-1.15	29.0	-0.26	0.26
300	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	17.88	29.0	4.19	**3.32×10^{-17}**
300	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	8.03	29.0	2.24	**7.47×10^{-9}**
300	Pre-trained SFCN Age	Pre-trained SFCN Gender	-10.10	29.0	-2.54	**5.27×10^{-11}**

C.2.2 Sex Classification



(a) LiteNet



(b) SFCN

Figure C.2: Holdout test set accuracy results for sex classification problem. Training evolution for the sex classification problem with different dataset training sizes for the four training strategies. 3DCAE-MRI off-the-shelf and 3DCAE-MRI fine-tuning represents the transfer learning from the 3DCAE-MRI to the CNN. The shaded band represents to the standard deviation.

Table C.15: LiteNet Sex prediction: ANOVA comparing the accuracy of the holdout test set across training strategies for different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	10.47	2.82×10^{-7}
50	4	116	13.66	3.72×10^{-9}
100	4	116	10.98	1.38×10^{-7}
200	4	116	13.39	5.3×10^{-9}
300	4	116	17.54	2.79×10^{-11}

Table C.16: LiteNet Sex prediction: Post-hoc results for the statistically significant ANOVA comparing the accuracy of the holdout test set across training strategies for different training dataset sizes. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-5.69	29.0	-1.18	**3.73×10^{-6}**
25	3DCAE-MRI fine-tuning	PCA	-3.54	29.0	-0.80	**0.0014**
25	3DCAE-MRI fine-tuning	Training from scratch	-0.32	29.0	-0.08	0.75
25	3DCAE-MRI fine-tuning	Training from scratch augmented	-1.37	29.0	-0.33	0.18
25	3DCAE-MRI off-the-shelf	PCA	3.82	29.0	0.58	**0.00065**
25	3DCAE-MRI off-the-shelf	Training from scratch	5.05	29.0	1.13	**2.2×10^{-5}**
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	4.39	29.0	1.01	**0.00014**
25	PCA	Training from scratch	3.25	29.0	0.73	**0.0029**
25	PCA	Training from scratch augmented	2.28	29.0	0.54	**0.03**
25	Training from scratch	Training from scratch augmented	-1.04	29.0	-0.25	0.31
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-3.68	29.0	-0.80	**0.00095**
50	3DCAE-MRI fine-tuning	PCA	-1.07	29.0	-0.27	0.29
50	3DCAE-MRI fine-tuning	Training from scratch	2.35	29.0	0.54	**0.026**
50	3DCAE-MRI fine-tuning	Training from scratch augmented	2.73	29.0	0.71	**0.011**
50	3DCAE-MRI off-the-shelf	PCA	3.69	29.0	0.74	**0.00092**
50	3DCAE-MRI off-the-shelf	Training from scratch	5.42	29.0	1.48	**8.0×10^{-6}**
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	6.13	29.0	1.66	**1.11×10^{-6}**
50	PCA	Training from scratch	3.61	29.0	0.94	**0.0011**
50	PCA	Training from scratch augmented	3.89	29.0	1.14	**0.00054**
50	Training from scratch	Training from scratch augmented	0.79	29.0	0.18	0.44
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-4.95	29.0	-0.94	**2.95×10^{-5}**
100	3DCAE-MRI fine-tuning	PCA	-3.36	29.0	-0.63	**0.0022**
100	3DCAE-MRI fine-tuning	Training from scratch	0.31	29.0	0.05	0.76
100	3DCAE-MRI fine-tuning	Training from scratch augmented	0.83	29.0	0.20	0.41
100	3DCAE-MRI off-the-shelf	PCA	1.56	29.0	0.29	0.13
100	3DCAE-MRI off-the-shelf	Training from scratch	5.17	29.0	1.14	**1.59×10^{-5}**
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	4.46	29.0	1.08	**0.00011**
100	PCA	Training from scratch	3.81	29.0	0.77	**0.00066**
100	PCA	Training from scratch augmented	3.61	29.0	0.79	**0.0011**
100	Training from scratch	Training from scratch augmented	0.78	29.0	0.17	0.44
200	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-5.14	29.0	-0.92	**1.73×10^{-5}**
200	3DCAE-MRI fine-tuning	PCA	-4.58	29.0	-0.99	**8.2×10^{-5}**
200	3DCAE-MRI fine-tuning	Training from scratch	0.99	29.0	0.20	0.33
200	3DCAE-MRI fine-tuning	Training from scratch augmented	-0.50	29.0	-0.09	0.62
200	3DCAE-MRI off-the-shelf	PCA	-0.42	29.0	-0.08	0.68
200	3DCAE-MRI off-the-shelf	Training from scratch	5.43	29.0	1.10	**7.63×10^{-6}**
200	3DCAE-MRI off-the-shelf	Training from scratch augmented	3.69	29.0	0.82	**0.00093**
200	PCA	Training from scratch	5.82	29.0	1.17	**2.65×10^{-6}**
200	PCA	Training from scratch augmented	4.01	29.0	0.89	**0.00039**
200	Training from scratch	Training from scratch augmented	-1.15	29.0	-0.29	0.26
300	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-3.71	29.0	-0.80	**0.00087**
300	3DCAE-MRI fine-tuning	PCA	-5.09	29.0	-1.06	**1.99×10^{-5}**
300	3DCAE-MRI fine-tuning	Training from scratch	0.39	29.0	0.08	0.7
300	3DCAE-MRI fine-tuning	Training from scratch augmented	2.04	29.0	0.48	0.05
300	3DCAE-MRI off-the-shelf	PCA	-1.53	29.0	-0.24	0.14
300	3DCAE-MRI off-the-shelf	Training from scratch	3.94	29.0	0.95	**0.00047**
300	3DCAE-MRI off-the-shelf	Training from scratch augmented	6.10	29.0	1.34	**1.2×10^{-6}**
300	PCA	Training from scratch	5.02	29.0	1.24	**2.38×10^{-5}**
300	PCA	Training from scratch augmented	7.55	29.0	1.63	**2.53×10^{-8}**
300	Training from scratch	Training from scratch augmented	1.89	29.0	0.43	0.069

Table C.17: SFCN Sex prediction: ANOVA comparing the accuracy of the holdout test set across training strategies for different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	16.55	9.35×10^{-11}
50	4	116	12.98	9.16×10^{-9}
100	4	116	6.56	8.52×10^{-5}
200	4	116	11.16	1.07×10^{-7}
300	4	116	14.92	7.24×10^{-10}

Table C.18: SFCN Sex prediction: Post-hoc results for the statistically significant ANOVA comparing the accuracy of the holdout test set across training strategies for different training dataset sizes. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-1.01	29.0	-0.19	0.32
25	3DCAE-MRI Fine-tuning	PCA	4.61	29.0	1.18	**7.44×10^{-5}**
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	2.43	29.0	0.47	**0.021**
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	1.22	29.0	0.23	0.23
25	3DCAE-MRI Off-the-shelf	PCA	6.45	29.0	1.57	**4.64×10^{-7}**
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	4.44	29.0	0.76	**0.00012**
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	2.70	29.0	0.49	**0.011**
25	PCA	Pre-trained SFCN Age	-3.90	29.0	-0.73	**0.00053**
25	PCA	Pre-trained SFCN Gender	-5.42	29.0	-1.06	**7.87×10^{-6}**
25	Pre-trained SFCN Age	Pre-trained SFCN Gender	-1.58	29.0	-0.28	0.12
50	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-0.99	29.0	-0.27	0.33
50	3DCAE-MRI Fine-tuning	PCA	3.97	29.0	1.12	**0.00043**
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	1.94	29.0	0.54	0.062
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	4.13	29.0	1.03	**0.00028**
50	3DCAE-MRI Off-the-shelf	PCA	6.38	29.0	1.43	**5.61×10^{-7}**
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	3.80	29.0	0.82	**0.00069**
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	5.64	29.0	1.32	**4.27×10^{-6}**
50	PCA	Pre-trained SFCN Age	-2.62	29.0	-0.53	**0.014**
50	PCA	Pre-trained SFCN Gender	-0.09	29.0	-0.02	0.93
50	Pre-trained SFCN Age	Pre-trained SFCN Gender	2.37	29.0	0.48	**0.024**
100	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-1.31	29.0	-0.26	0.2
100	3DCAE-MRI Fine-tuning	PCA	1.21	29.0	0.25	0.24
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	2.39	29.0	0.58	**0.024**
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	3.50	29.0	0.68	**0.0015**
100	3DCAE-MRI Off-the-shelf	PCA	2.45	29.0	0.51	**0.021**
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	3.88	29.0	0.88	**0.00055**
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	3.78	29.0	0.98	**0.00073**
100	PCA	Pre-trained SFCN Age	1.44	29.0	0.32	0.16
100	PCA	Pre-trained SFCN Gender	2.19	29.0	0.41	**0.037**
100	Pre-trained SFCN Age	Pre-trained SFCN Gender	0.45	29.0	0.10	0.66
200	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-1.57	29.0	-0.39	0.13
200	3DCAE-MRI Fine-tuning	PCA	-1.04	29.0	-0.24	0.31
200	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	1.07	29.0	0.25	0.3
200	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	4.06	29.0	0.93	**0.00034**
200	3DCAE-MRI Off-the-shelf	PCA	0.84	29.0	0.14	0.41
200	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	2.34	29.0	0.60	**0.027**
200	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	5.26	29.0	1.23	**1.24×10^{-5}**
200	PCA	Pre-trained SFCN Age	2.01	29.0	0.46	0.054
200	PCA	Pre-trained SFCN Gender	5.17	29.0	1.10	**1.59×10^{-5}**
200	Pre-trained SFCN Age	Pre-trained SFCN Gender	3.96	29.0	0.67	**0.00045**
300	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-3.36	29.0	-0.83	**0.0022**
300	3DCAE-MRI Fine-tuning	PCA	-3.37	29.0	-0.69	**0.0021**
300	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	0.78	29.0	0.16	0.44
300	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	3.33	29.0	0.72	**0.0024**
300	3DCAE-MRI Off-the-shelf	PCA	0.65	29.0	0.18	0.52
300	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	3.62	29.0	0.96	**0.0011**
300	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	6.25	29.0	1.61	**8.11×10^{-7}**
300	PCA	Pre-trained SFCN Age	3.64	29.0	0.83	**0.0011**
300	PCA	Pre-trained SFCN Gender	5.65	29.0	1.49	**4.22×10^{-6}**
300	Pre-trained SFCN Age	Pre-trained SFCN Gender	2.63	29.0	0.54	**0.014**

Table C.19: LiteNet Sex prediction: ANOVA comparing the accuracy of the external test set across training strategies for different training dataset sizes. The significant *p*-values are in bold.

	ddof1	ddof2	F	*p*-value
Number of training samples				
25	4	116	22.05	1.49×10^{-13}
50	4	116	12.14	2.8×10^{-8}
100	4	116	10.30	3.6×10^{-7}
200	4	116	17.91	1.78×10^{-11}
300	4	116	20.94	5.17×10^{-13}

Table C.20: LiteNet Sex prediction: Post-hoc results for the statistically significant ANOVA comparing the accuracy of the external test set across training strategies for different training dataset sizes. The significant *p*-values are shown in bold.

	A	B	T	dof	cohen	*p*-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-5.77	29.0	-1.37	**2.99×10^{-6}**
25	3DCAE-MRI fine-tuning	PCA	-6.75	29.0	-1.70	**2.08×10^{-7}**
25	3DCAE-MRI fine-tuning	Training from scratch	-0.39	29.0	-0.11	0.7
25	3DCAE-MRI fine-tuning	Training from scratch augmented	-2.42	29.0	-0.47	**0.022**
25	3DCAE-MRI off-the-shelf	PCA	-1.77	29.0	-0.33	0.088
25	3DCAE-MRI off-the-shelf	Training from scratch	5.45	29.0	1.37	**7.21×10^{-6}**
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	4.86	29.0	0.94	**3.75×10^{-5}**
25	PCA	Training from scratch	7.21	29.0	1.73	**6.23×10^{-8}**
25	PCA	Training from scratch augmented	5.32	29.0	1.27	**1.03×10^{-5}**
25	Training from scratch	Training from scratch augmented	-1.62	29.0	-0.40	0.12
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-2.98	29.0	-0.57	**0.0058**
50	3DCAE-MRI fine-tuning	PCA	-5.13	29.0	-1.13	**1.78×10^{-5}**
50	3DCAE-MRI fine-tuning	Training from scratch	2.23	29.0	0.54	**0.034**
50	3DCAE-MRI fine-tuning	Training from scratch augmented	0.42	29.0	0.11	0.68
50	3DCAE-MRI off-the-shelf	PCA	-2.01	29.0	-0.46	0.054
50	3DCAE-MRI off-the-shelf	Training from scratch	4.56	29.0	1.04	**8.58×10^{-5}**
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	2.47	29.0	0.59	**0.02**
50	PCA	Training from scratch	6.28	29.0	1.57	**7.41×10^{-7}**
50	PCA	Training from scratch augmented	4.37	29.0	1.02	**0.00015**
50	Training from scratch	Training from scratch augmented	-1.91	29.0	-0.37	0.067
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-2.10	29.0	-0.47	**0.044**
100	3DCAE-MRI fine-tuning	PCA	-4.79	29.0	-0.96	**4.53×10^{-5}**
100	3DCAE-MRI fine-tuning	Training from scratch	1.41	29.0	0.31	0.17
100	3DCAE-MRI fine-tuning	Training from scratch augmented	1.45	29.0	0.39	0.16
100	3DCAE-MRI off-the-shelf	PCA	-4.19	29.0	-0.69	**0.00024**
100	3DCAE-MRI off-the-shelf	Training from scratch	2.95	29.0	0.79	**0.0063**
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	3.94	29.0	0.96	**0.00047**
100	PCA	Training from scratch	4.90	29.0	1.22	**3.34×10^{-5}**
100	PCA	Training from scratch augmented	5.45	29.0	1.45	**7.32×10^{-6}**
100	Training from scratch	Training from scratch augmented	0.17	29.0	0.05	0.87
200	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-3.12	29.0	-0.68	**0.0041**
200	3DCAE-MRI fine-tuning	PCA	-5.37	29.0	-1.26	**9.02×10^{-6}**
200	3DCAE-MRI fine-tuning	Training from scratch	0.24	29.0	0.06	0.81
200	3DCAE-MRI fine-tuning	Training from scratch augmented	1.62	29.0	0.38	0.12
200	3DCAE-MRI off-the-shelf	PCA	-3.52	29.0	-0.72	**0.0014**
200	3DCAE-MRI off-the-shelf	Training from scratch	3.70	29.0	0.80	**0.00091**
200	3DCAE-MRI off-the-shelf	Training from scratch augmented	5.13	29.0	1.12	**1.74×10^{-5}**
200	PCA	Training from scratch	6.96	29.0	1.44	**1.18×10^{-7}**
200	PCA	Training from scratch augmented	7.88	29.0	1.72	**1.07×10^{-8}**
200	Training from scratch	Training from scratch augmented	1.72	29.0	0.35	0.097
300	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-2.00	29.0	-0.45	0.055
300	3DCAE-MRI fine-tuning	PCA	-2.97	29.0	-0.59	**0.0059**
300	3DCAE-MRI fine-tuning	Training from scratch	3.66	29.0	0.74	**0.001**
300	3DCAE-MRI fine-tuning	Training from scratch augmented	4.39	29.0	1.09	**0.00014**
300	3DCAE-MRI off-the-shelf	PCA	-0.26	29.0	-0.06	0.8
300	3DCAE-MRI off-the-shelf	Training from scratch	5.05	29.0	1.11	**2.21×10^{-5}**
300	3DCAE-MRI off-the-shelf	Training from scratch augmented	7.01	29.0	1.45	**1.05×10^{-7}**
300	PCA	Training from scratch	5.70	29.0	1.28	**3.66×10^{-6}**
300	PCA	Training from scratch augmented	6.64	29.0	1.64	**2.79×10^{-7}**
300	Training from scratch	Training from scratch augmented	1.35	29.0	0.33	0.19

Table C.21: SFCN Sex prediction: ANOVA comparing the accuracy of the external test set across training strategies for different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	4.87	0.0011
50	4	116	12.92	9.85×10^{-9}
100	4	116	5.61	0.00037
200	4	116	15.64	2.93×10^{-10}
300	4	116	6.11	0.00017

Table C.22: SFCN Sex prediction: Post-hoc results for the statistically significant ANOVA comparing the accuracy of the external test set across training strategies for different training dataset sizes. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-0.50	29.0	-0.12	0.62
25	3DCAE-MRI Fine-tuning	PCA	0.78	29.0	0.20	0.44
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	3.62	29.0	0.71	**0.0011**
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	3.06	29.0	0.75	**0.0047**
25	3DCAE-MRI Off-the-shelf	PCA	1.21	29.0	0.29	0.24
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	2.91	29.0	0.77	**0.0069**
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	3.27	29.0	0.80	**0.0028**
25	PCA	Pre-trained SFCN Age	1.68	29.0	0.45	0.1
25	PCA	Pre-trained SFCN Gender	2.03	29.0	0.49	0.052
25	Pre-trained SFCN Age	Pre-trained SFCN Gender	0.26	29.0	0.06	0.79
50	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	1.47	29.0	0.39	0.15
50	3DCAE-MRI Fine-tuning	PCA	1.96	29.0	0.54	0.059
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	4.66	29.0	1.19	$\mathbf{6.48 \times 10^{-5}}$
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	5.83	29.0	1.42	$\mathbf{2.54 \times 10^{-6}}$
50	3DCAE-MRI Off-the-shelf	PCA	0.85	29.0	0.22	0.4
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	3.96	29.0	0.91	**0.00045**
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	5.38	29.0	1.15	$\mathbf{8.84 \times 10^{-6}}$
50	PCA	Pre-trained SFCN Age	2.71	29.0	0.65	**0.011**
50	PCA	Pre-trained SFCN Gender	3.34	29.0	0.87	**0.0023**
50	Pre-trained SFCN Age	Pre-trained SFCN Gender	1.02	29.0	0.21	0.32
100	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-1.18	29.0	-0.22	0.25
100	3DCAE-MRI Fine-tuning	PCA	-1.50	29.0	-0.31	0.15
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	2.66	29.0	0.69	**0.013**
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	1.62	29.0	0.37	0.12
100	3DCAE-MRI Off-the-shelf	PCA	-0.69	29.0	-0.12	0.5
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	3.67	29.0	0.97	**0.00098**
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	2.16	29.0	0.55	**0.04**
100	PCA	Pre-trained SFCN Age	3.78	29.0	1.00	**0.00072**
100	PCA	Pre-trained SFCN Gender	2.57	29.0	0.60	**0.015**
100	Pre-trained SFCN Age	Pre-trained SFCN Gender	-0.70	29.0	-0.18	0.49
200	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-0.45	29.0	-0.12	0.65
200	3DCAE-MRI Fine-tuning	PCA	-1.86	29.0	-0.49	0.074
200	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	3.17	29.0	0.89	**0.0036**
200	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	3.92	29.0	1.10	**0.0005**
200	3DCAE-MRI Off-the-shelf	PCA	-1.89	29.0	-0.48	0.069
200	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	4.18	29.0	1.13	**0.00024**
200	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	5.12	29.0	1.40	$\mathbf{1.8 \times 10^{-5}}$
200	PCA	Pre-trained SFCN Age	6.04	29.0	1.34	$\mathbf{1.43 \times 10^{-6}}$
200	PCA	Pre-trained SFCN Gender	6.58	29.0	1.57	$\mathbf{3.29 \times 10^{-7}}$
200	Pre-trained SFCN Age	Pre-trained SFCN Gender	0.66	29.0	0.16	0.51
300	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-1.07	29.0	-0.25	0.29
300	3DCAE-MRI Fine-tuning	PCA	-2.91	29.0	-0.57	**0.0068**
300	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	2.00	29.0	0.42	0.055
300	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	2.22	29.0	0.46	**0.034**
300	3DCAE-MRI Off-the-shelf	PCA	-1.54	29.0	-0.39	0.13
300	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	3.13	29.0	0.66	**0.004**
300	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	2.41	29.0	0.64	**0.022**
300	PCA	Pre-trained SFCN Age	4.12	29.0	0.90	**0.00029**
300	PCA	Pre-trained SFCN Gender	3.29	29.0	0.83	**0.0026**
300	Pre-trained SFCN Age	Pre-trained SFCN Gender	0.47	29.0	0.11	0.65

Table C.23: LiteNet Sex prediction: ANOVA results comparing the stability metric across different training strategies for the different dataset sizes tested. The significant p-values are shown in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	3	87	92.86	4.99×10^{-27}
50	3	87	34.96	6.39×10^{-15}
100	3	87	23.18	4.05×10^{-11}
200	3	87	1.95	0.13
300	3	87	8.03	8.75×10^{-5}

Table C.24: LiteNet Sex prediction: Post-hoc results for the statistically significant ANOVA comparing the stability metric across different training strategies (training the network from scratch, off-the-shelf and fine-tuning learning using the 3DCAE-MRI) for the different dataset sizes tested. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	21.55	29.0	6.17	**2.14×10^{-19}**
25	3DCAE-MRI fine-tuning	Training from scratch	5.83	29.0	1.41	**2.54×10^{-6}**
25	3DCAE-MRI fine-tuning	Training from scratch augmented	6.43	29.0	1.68	**4.98×10^{-7}**
25	3DCAE-MRI off-the-shelf	Training from scratch	-8.80	29.0	-2.24	**1.1×10^{-9}**
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	-16.03	29.0	-3.73	**6.03×10^{-16}**
25	Training from scratch	Training from scratch augmented	-0.81	29.0	-0.22	0.43
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	7.72	29.0	1.72	**1.65×10^{-8}**
50	3DCAE-MRI fine-tuning	Training from scratch	11.73	29.0	2.94	**1.56×10^{-12}**
50	3DCAE-MRI fine-tuning	Training from scratch augmented	12.62	29.0	3.47	**2.63×10^{-13}**
50	3DCAE-MRI off-the-shelf	Training from scratch	-0.70	29.0	-0.17	0.49
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	0.36	29.0	0.10	0.72
50	Training from scratch	Training from scratch augmented	1.79	29.0	0.47	0.084
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	7.26	29.0	1.41	**5.4×10^{-8}**
100	3DCAE-MRI fine-tuning	Training from scratch	6.05	29.0	1.39	**1.39×10^{-6}**
100	3DCAE-MRI fine-tuning	Training from scratch augmented	6.85	29.0	1.38	**1.59×10^{-7}**
100	3DCAE-MRI off-the-shelf	Training from scratch	-2.10	29.0	-0.50	**0.044**
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	-1.45	29.0	-0.33	0.16
100	Training from scratch	Training from scratch augmented	0.85	29.0	0.19	0.4
300	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	3.25	29.0	0.81	**0.0029**
300	3DCAE-MRI fine-tuning	Training from scratch	3.54	29.0	0.84	**0.0014**
300	3DCAE-MRI fine-tuning	Training from scratch augmented	7.08	29.0	1.46	**8.72×10^{-8}**
300	3DCAE-MRI off-the-shelf	Training from scratch	-0.35	29.0	-0.09	0.73
300	3DCAE-MRI off-the-shelf	Training from scratch augmented	0.92	29.0	0.23	0.36
300	Training from scratch	Training from scratch augmented	1.54	29.0	0.39	0.14

Table C.25: SFCN Sex classification: ANOVA results comparing the stability metric across different training strategies for the different dataset sizes tested. The significant p-values are shown in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	3	87	17.23	7.26×10^{-9}
50	3	87	14.78	7.4×10^{-8}
100	3	87	11.97	1.23×10^{-6}
200	3	87	26.82	2.24×10^{-12}
300	3	87	33.85	1.35×10^{-14}

Table C.26: SFCN Sex classification: Post-hoc results for the statistically significant ANOVA comparing the stability metric across different training strategies for the different dataset sizes tested. The significant *p*-values are shown in bold.

	A	B	T	dof	cohen	*p*-value
Number of training samples						
25	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-2.42	29.0	-0.62	**0.022**
25	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	3.23	29.0	0.87	**0.0031**
25	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	2.74	29.0	0.74	**0.01**
25	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	7.65	29.0	1.79	$\mathbf{1.98 \times 10^{-8}}$
25	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	6.39	29.0	1.58	$\mathbf{5.46 \times 10^{-7}}$
25	Pre-trained SFCN Age	Pre-trained SFCN Gender	-0.53	29.0	-0.14	0.6
50	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-3.39	29.0	-0.83	**0.0021**
50	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	2.85	29.0	0.69	**0.0079**
50	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	2.01	29.0	0.45	0.054
50	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	7.03	29.0	1.77	$\mathbf{9.8 \times 10^{-8}}$
50	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	4.70	29.0	1.33	$\mathbf{5.77 \times 10^{-5}}$
50	Pre-trained SFCN Age	Pre-trained SFCN Gender	-0.65	29.0	-0.17	0.52
100	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-4.48	29.0	-1.08	**0.00011**
100	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	-0.87	29.0	-0.24	0.39
100	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	-0.03	29.0	-0.01	0.98
100	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	5.38	29.0	1.30	$\mathbf{8.72 \times 10^{-6}}$
100	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	5.93	29.0	1.54	$\mathbf{1.91 \times 10^{-6}}$
100	Pre-trained SFCN Age	Pre-trained SFCN Gender	1.26	29.0	0.36	0.22
200	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-3.26	29.0	-0.78	**0.0028**
200	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	1.50	29.0	0.41	0.14
200	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	4.75	29.0	1.18	$\mathbf{5.02 \times 10^{-5}}$
200	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	6.88	29.0	1.38	$\mathbf{1.46 \times 10^{-7}}$
200	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	8.76	29.0	2.13	$\mathbf{1.21 \times 10^{-9}}$
200	Pre-trained SFCN Age	Pre-trained SFCN Gender	4.20	29.0	0.94	**0.00023**
300	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-4.47	29.0	-1.10	**0.00011**
300	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	3.28	29.0	0.59	**0.0027**
300	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	4.58	29.0	1.18	$\mathbf{8.04 \times 10^{-5}}$
300	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	7.13	29.0	1.82	$\mathbf{7.6 \times 10^{-8}}$
300	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	9.71	29.0	2.34	$\mathbf{1.27 \times 10^{-10}}$
300	Pre-trained SFCN Age	Pre-trained SFCN Gender	2.59	29.0	0.66	**0.015**

C.3 AD Classification

C.3.1 Methods

C.3.1.1 Data

The data considered for the Alzheimer's disease (AD) classification problem was from the OASIS3 and OASIS4 repositories. The number of healthy controls and AD subjects is unbalanced in these repositories. Furthermore, the age range of both groups also differs. Therefore, a subset of subjects matched for age and sex were selected. In total, 220 and 68 individuals were selected from OASIS3 and OASIS4; the demographics are in Table C.27. The data considered to train and validate the models was from the OASIS3 dataset. The holdout dataset comprised a subset of 50 samples from the OASIS3 subsample. The OASIS4 subset was considered to assess the performance of different models on an external test set.

Table C.27: Demographics of the participants used to train, validate and test the age prediction and sex classification models.

Dataset	Total	Number of males	Mean and standard deviation [years]	Min age [years]	Max age [years]	Number of Control
OASIS3	220	104	75.57 ± 7.59	49	95	110
OASIS4	68	32	76.09 ± 5.32	65	88	34

C.3.2 Results

C.3.3 Holdout test set

The results for the holdout test set for the different shallow models and different training strategies are in Table C.28. For LiteNet architecture, the results show that the training strategy that attains the highest mean absolute error (MAE) is transfer learning from the 3D-convolutional autoencoder magnetic resonance imaging (3DCAE-MRI) either by fine-tuning or off-the-shelf. Concerning the simple fully convolutional network (SFCN), the results show that off-the-shelf transfer learning from the 3DCAE-MRI attains better or equivalent performance than pre-trained models. The comparison between shallow and deep learning models outlined that SFCN yields statistically similar performance to shallow learning for all training instances. Fort the LiteNet architecture, off-the-shelf and fine-tuning yielded significantly higher performance for 25 training samples, yet for 50 and 100 training samples, the performance of both models was similar.

C.3.4 External test set

Table C.29 shows the accuracy of LiteNet and SFCN on an external test set. The results for LiteNet evidence that the deep learning training strategy that attains the highest performance is off-the-shelf transfer learning from the 3DCAE-MRI. The difference between off-the-shelf and training from scratch (with and without) is

Table C.28: Holdout test set mean and standard deviation of AD classification accuracy for PCA-RVM and two CNN architectures, different training strategies and training set sizes. The lowest value for each CNN architecture and training size is represented in bold; the symbols * and † represent that PCA-RVM obtained better performance than the LiteNet and SFCN architectures, respectively.

Training Strategy	PCA-RVM	LiteNet				SFCN			
		Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
		No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	0.73 ± 0.07	0.70 ± 0.09	0.65 ± 0.11	**0.75 ± 0.07**	**0.75 ± 0.06**	0.72 ± 0.07	0.74 ± 0.06	**0.75 ± 0.07**	**0.75 ± 0.07**
50	$0.77 \pm 0.05 * \dagger$	0.74 ± 0.08	0.73 ± 0.07	**0.77 ± 0.05**	**0.77 ± 0.04**	0.75 ± 0.05	**0.77 ± 0.06**	0.76 ± 0.04	**0.77 ± 0.06**
100	$0.79 \pm 0.05 * \dagger$	0.75 ± 0.06	0.74 ± 0.06	0.78 ± 0.06	**0.79 ± 0.06**	0.77 ± 0.06	0.77 ± 0.06	**0.78 ± 0.07**	**0.78 ± 0.06**

significant in both cases. Concerning the SFCN, the results evidence that, in general, transfer learning from age or sex pre-trained models yields equivalent performance to transfer learning from 3DCAE-MRI either by off-the-shelf or fine-tuning. The comparison of deep learning architectures and shallow learning outlines that shallow learning,principal component analyis (PCA)-relevant vector machine (RVM), yields higher or equivalent performance than SFCN and LiteNet.

Table C.29: External test set mean and standard deviation of AD classification accuracy for PCA-RVM and two CNN architectures, different training strategies and training set sizes. The lowest value for each CNN architecture and training size is represented in bold; the symbols * and † represent that PCA-RVM performed better than the LiteNet and SFCN architectures, respectively.

Training Strategy	PCA-RVM	LiteNet				SFCN			
		Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
		No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	$0.82 \pm 0.03 * \dagger$	0.74 ± 0.06	0.69 ± 0.11	0.80 ± 0.05	**0.82 ± 0.03**	**0.78 ± 0.06**	0.77 ± 0.04	0.76 ± 0.05	**0.78 ± 0.03**
50	$0.84 \pm 0.03 * \dagger$	0.76 ± 0.04	0.76 ± 0.04	0.80 ± 0.04	**0.81 ± 0.04**	**0.79 ± 0.04**	0.78 ± 0.04	0.78 ± 0.04	0.78 ± 0.04
100	$0.84 \pm 0.02 * \dagger$	0.78 ± 0.04	0.76 ± 0.04	0.80 ± 0.03	**0.82 ± 0.03**	0.80 ± 0.03	**0.81 ± 0.04**	0.80 ± 0.03	0.80 ± 0.02

C.3.5 Stability

The variation variability results are summarized in Table C.30. The results highlight that for LitNet architecture, off-the-shelf transfer learning from 3DCAE-MRI is significantly more stable than the other three deep learning training strategies. Regarding SFCN, the results outline similar stability across the different transfer learning strategies.

Table C.30: Mean and standard deviation of validation variability across different training sizes and training strategies for the AD classification problem. The lowest value for each training size is represented in bold.

Training Strategy	LiteNet				SFCN			
	Training from scratch		3DCAE-MRI		Pretrained		3DCAE-MRI	
	No augmentation	Augmentation	Fine-tuning	Off-the-shelf	Age	Sex	Fine-tuning	Off-the-shelf
25	0.11 ± 0.01	0.11 ± 0.01	0.10 ± 0.01	**0.07 ± 0.01**	**0.11 ± 0.01**	**0.11 ± 0.01**	**0.11 ± 0.01**	**0.11 ± 0.01**
50	0.11 ± 0.01	0.11 ± 0.01	0.08 ± 0.01	**0.06 ± 0.01**	**0.11 ± 0.01**	0.12 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
100	0.10 ± 0.01	0.10 ± 0.01	0.11 ± 0.01	**0.07 ± 0.01**	**0.12 ± 0.01**	**0.12 ± 0.01**	**0.12 ± 0.01**	**0.12 ± 0.01**

Table C.31: LiteNet AD prediction: ANOVA comparing the accuracy of the holdout test set across training strategies for different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	16.39	1.14×10^{-10}
50	4	116	4.78	0.0013
100	4	116	8.78	3.16×10^{-6}

Table C.32: LiteNet AD prediction: Post-hoc results for the statistically significant ANOVA comparing the accuracy of the holdout test set across training strategies for different training dataset sizes. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-0.67	29.0	-0.09	0.51
25	3DCAE-MRI fine-tuning	PCA	1.22	29.0	0.20	0.23
25	3DCAE-MRI fine-tuning	Training from scratch	3.75	29.0	0.64	**0.00078**
25	3DCAE-MRI fine-tuning	Training from scratch augmented	5.61	29.0	1.05	$\mathbf{4.62 \times 10^{-6}}$
25	3DCAE-MRI off-the-shelf	PCA	1.82	29.0	0.31	0.08
25	3DCAE-MRI off-the-shelf	Training from scratch	4.32	29.0	0.74	**0.00017**
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	5.24	29.0	1.15	$\mathbf{1.29 \times 10^{-5}}$
25	PCA	Training from scratch	2.78	29.0	0.48	**0.0095**
25	PCA	Training from scratch augmented	4.71	29.0	0.93	$\mathbf{5.67 \times 10^{-5}}$
25	Training from scratch	Training from scratch augmented	2.31	29.0	0.47	**0.029**
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	0.07	29.0	0.01	0.95
50	3DCAE-MRI fine-tuning	PCA	0.21	29.0	0.05	0.83
50	3DCAE-MRI fine-tuning	Training from scratch	1.98	29.0	0.50	0.058
50	3DCAE-MRI fine-tuning	Training from scratch augmented	3.03	29.0	0.69	**0.0051**
50	3DCAE-MRI off-the-shelf	PCA	0.25	29.0	0.04	0.81
50	3DCAE-MRI off-the-shelf	Training from scratch	2.68	29.0	0.52	**0.012**
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	3.46	29.0	0.72	**0.0017**
50	PCA	Training from scratch	2.13	29.0	0.45	**0.042**
50	PCA	Training from scratch augmented	2.55	29.0	0.63	**0.016**
50	Training from scratch	Training from scratch augmented	0.71	29.0	0.13	0.48
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-0.59	29.0	-0.08	0.56
100	3DCAE-MRI fine-tuning	PCA	-0.53	29.0	-0.10	0.6
100	3DCAE-MRI fine-tuning	Training from scratch	3.17	29.0	0.54	**0.0036**
100	3DCAE-MRI fine-tuning	Training from scratch augmented	3.53	29.0	0.75	**0.0014**
100	3DCAE-MRI off-the-shelf	PCA	-0.07	29.0	-0.01	0.95
100	3DCAE-MRI off-the-shelf	Training from scratch	3.70	29.0	0.59	**0.0009**
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	3.64	29.0	0.80	**0.001**
100	PCA	Training from scratch	3.39	29.0	0.64	**0.002**
100	PCA	Training from scratch augmented	4.65	29.0	0.88	$\mathbf{6.63 \times 10^{-5}}$
100	Training from scratch	Training from scratch augmented	0.74	29.0	0.15	0.46

Table C.33: SFCN AD prediction: ANOVA comparing the accuracy of the holdout test set across training strategies for different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	1.70	0.15
50	4	116	1.04	0.39
100	4	116	0.31	0.87

Table C.34: LiteNet AD prediction: ANOVA comparing the accuracy of the external test set across training strategies for different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	29.74	5.04×10^{-17}
50	4	116	28.32	2.04×10^{-16}
100	4	116	39.72	6.43×10^{-21}

Table C.35: LiteNet AD prediction: Post-hoc results for the statistically significant ANOVA comparing the accuracy of the external test set across training strategies for different training dataset sizes. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-2.04	29.0	-0.47	0.051
25	3DCAE-MRI fine-tuning	PCA	-2.04	29.0	-0.47	0.051
25	3DCAE-MRI fine-tuning	Training from scratch	4.85	29.0	0.96	**3.89×10^{-5}**
25	3DCAE-MRI fine-tuning	Training from scratch augmented	5.91	29.0	1.23	**2.02×10^{-6}**
25	3DCAE-MRI off-the-shelf	PCA	0.08	29.0	0.02	0.94
25	3DCAE-MRI off-the-shelf	Training from scratch	8.78	29.0	1.52	**1.15×10^{-9}**
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	5.96	29.0	1.54	**1.79×10^{-6}**
25	PCA	Training from scratch	7.84	29.0	1.54	**1.2×10^{-8}**
25	PCA	Training from scratch augmented	6.70	29.0	1.54	**2.4×10^{-7}**
25	Training from scratch	Training from scratch augmented	2.66	29.0	0.59	**0.013**
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-0.90	29.0	-0.23	0.38
50	3DCAE-MRI fine-tuning	PCA	-4.46	29.0	-1.10	**0.00011**
50	3DCAE-MRI fine-tuning	Training from scratch	3.70	29.0	0.98	**0.0009**
50	3DCAE-MRI fine-tuning	Training from scratch augmented	5.30	29.0	1.22	**1.09×10^{-5}**
50	3DCAE-MRI off-the-shelf	PCA	-4.15	29.0	-0.82	**0.00027**
50	3DCAE-MRI off-the-shelf	Training from scratch	4.68	29.0	1.17	**6.17×10^{-5}**
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	5.03	29.0	1.41	**2.36×10^{-5}**
50	PCA	Training from scratch	8.80	29.0	2.06	**1.11×10^{-9}**
50	PCA	Training from scratch augmented	8.85	29.0	2.36	**9.83×10^{-10}**
50	Training from scratch	Training from scratch augmented	0.99	29.0	0.20	0.33
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	-3.16	29.0	-0.71	**0.0037**
100	3DCAE-MRI fine-tuning	PCA	-5.99	29.0	-1.58	**1.62×10^{-6}**
100	3DCAE-MRI fine-tuning	Training from scratch	2.80	29.0	0.55	**0.009**
100	3DCAE-MRI fine-tuning	Training from scratch augmented	6.77	29.0	1.23	**1.97×10^{-7}**
100	3DCAE-MRI off-the-shelf	PCA	-3.28	29.0	-0.68	**0.0027**
100	3DCAE-MRI off-the-shelf	Training from scratch	5.27	29.0	1.11	**1.19×10^{-5}**
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	8.21	29.0	1.76	**4.74×10^{-9}**
100	PCA	Training from scratch	6.79	29.0	1.79	**1.89×10^{-7}**
100	PCA	Training from scratch augmented	9.67	29.0	2.53	**1.41×10^{-10}**
100	Training from scratch	Training from scratch augmented	3.14	29.0	0.57	**0.0039**

Table C.36: SFCN AD prediction: ANOVA comparing the accuracy of the external test set across training strategies for different training dataset sizes. The significant p-values are in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	4	116	11.97	3.57×10^{-8}
50	4	116	15.79	2.4×10^{-10}
100	4	116	9.39	1.31×10^{-6}

Table C.37: SFCN AD prediction: Post-hoc results for the statistically significant ANOVA comparing the accuracy of the external test set across training strategies for different training dataset sizes. The significant p-values are shown in bold.

	A	B	T	dof	cohen	p-value
Number of training samples						
25	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	-2.64	29.0	-0.51	**0.013**
25	3DCAE-MRI Fine-tuning	PCA	-7.05	29.0	-1.54	**9.39×10^{-8}**
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-2.13	29.0	-0.50	**0.041**
25	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-2.14	29.0	-0.37	**0.041**
25	3DCAE-MRI Off-the-shelf	PCA	-6.03	29.0	-1.27	**1.48×10^{-6}**
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	-0.56	29.0	-0.13	0.58
25	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	0.60	29.0	0.12	0.55
25	PCA	Pre-trained SFCN Age	3.66	29.0	0.72	**0.001**
25	PCA	Pre-trained SFCN Gender	6.90	29.0	1.25	**1.41×10^{-7}**
25	Pre-trained SFCN Age	Pre-trained SFCN Gender	0.89	29.0	0.21	0.38
50	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	0.36	29.0	0.11	0.72
50	3DCAE-MRI Fine-tuning	PCA	-7.38	29.0	-1.63	**3.97×10^{-8}**
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-0.90	29.0	-0.19	0.37
50	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-0.27	29.0	-0.06	0.79
50	3DCAE-MRI Off-the-shelf	PCA	-6.36	29.0	-1.81	**5.93×10^{-7}**
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	-1.15	29.0	-0.30	0.26
50	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-0.66	29.0	-0.18	0.51
50	PCA	Pre-trained SFCN Age	6.97	29.0	1.42	**1.16×10^{-7}**
50	PCA	Pre-trained SFCN Gender	6.86	29.0	1.68	**1.55×10^{-7}**
50	Pre-trained SFCN Age	Pre-trained SFCN Gender	0.58	29.0	0.14	0.57
100	3DCAE-MRI Fine-tuning	3DCAE-MRI Off-the-shelf	0.23	29.0	0.05	0.82
100	3DCAE-MRI Fine-tuning	PCA	-6.21	29.0	-1.46	**8.97×10^{-7}**
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Age	-0.76	29.0	-0.17	0.45
100	3DCAE-MRI Fine-tuning	Pre-trained SFCN Gender	-1.48	29.0	-0.38	0.15
100	3DCAE-MRI Off-the-shelf	PCA	-8.65	29.0	-2.01	**1.61×10^{-9}**
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Age	-1.19	29.0	-0.27	0.24
100	3DCAE-MRI Off-the-shelf	Pre-trained SFCN Gender	-1.60	29.0	-0.48	0.12
100	PCA	Pre-trained SFCN Age	5.42	29.0	1.38	**7.9×10^{-6}**
100	PCA	Pre-trained SFCN Gender	2.83	29.0	0.74	**0.0083**
100	Pre-trained SFCN Age	Pre-trained SFCN Gender	-0.94	29.0	-0.25	0.36

Table C.38: LiteNet AD prediction: ANOVA results comparing the stability metric across different training strategies for the different dataset sizes tested. The significant p-values are shown in bold.

	ddof1	ddof2	F	p-value
Number of training samples				
25	3	87	137.38	6.78×10^{-33}
50	3	87	154.64	9.35×10^{-35}
100	3	87	67.37	1.29×10^{-22}

Table C.39: LiteNet AD classification: Post-hoc results for the statistically significant ANOVA comparing the stability metric across different training strategies for the different dataset sizes tested. The significant *p*-values are shown in bold.

	A	B	T	dof	cohen	*p*-value
Number of training samples						
25	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	9.93	29.0	2.68	**7.66×10^{-11}**
25	3DCAE-MRI fine-tuning	Training from scratch	-3.39	29.0	-0.99	**0.002**
25	3DCAE-MRI fine-tuning	Training from scratch augmented	-5.55	29.0	-1.37	**5.51×10^{-6}**
25	3DCAE-MRI off-the-shelf	Training from scratch	-25.51	29.0	-4.87	**2.05×10^{-21}**
25	3DCAE-MRI off-the-shelf	Training from scratch augmented	-22.61	29.0	-5.15	**5.75×10^{-20}**
25	Training from scratch	Training from scratch augmented	-2.44	29.0	-0.56	**0.021**
50	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	7.82	29.0	1.77	**1.26×10^{-8}**
50	3DCAE-MRI fine-tuning	Training from scratch	-12.11	29.0	-2.97	**7.3×10^{-13}**
50	3DCAE-MRI fine-tuning	Training from scratch augmented	-7.90	29.0	-2.42	**1.04×10^{-8}**
50	3DCAE-MRI off-the-shelf	Training from scratch	-18.70	29.0	-4.61	**1.0×10^{-17}**
50	3DCAE-MRI off-the-shelf	Training from scratch augmented	-15.65	29.0	-4.13	**1.12×10^{-15}**
50	Training from scratch	Training from scratch augmented	2.42	29.0	0.69	**0.022**
100	3DCAE-MRI fine-tuning	3DCAE-MRI off-the-shelf	11.12	29.0	2.50	**5.66×10^{-12}**
100	3DCAE-MRI fine-tuning	Training from scratch	0.95	29.0	0.22	0.35
100	3DCAE-MRI fine-tuning	Training from scratch augmented	1.65	29.0	0.29	0.11
100	3DCAE-MRI off-the-shelf	Training from scratch	-10.47	29.0	-2.23	**2.28×10^{-11}**
100	3DCAE-MRI off-the-shelf	Training from scratch augmented	-11.75	29.0	-2.66	**1.5×10^{-12}**
100	Training from scratch	Training from scratch augmented	0.10	29.0	0.02	0.92

Table C.40: SFCN AD classification: ANOVA results comparing the stability metric across different training strategies for the different dataset sizes tested. The significant *p*-values are shown in bold.

	ddof1	ddof2	F	*p*-value
Number of training samples				
25	3	87	1.90	0.14
50	3	87	6.35	0.00061
100	3	87	3.79	0.013

Table C.41: SFCN AD classification: Post-hoc results for the statistically significant ANOVA comparing the stability metric across different training strategies for the different dataset sizes tested. The significant *p*-values are shown in bold.

	A	B	T	dof	cohen	*p*-value
Number of training samples						
50	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	0.95	29.0	0.22	0.35
50	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	1.65	29.0	0.44	0.11
50	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	-2.30	29.0	-0.58	**0.029**
50	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	1.02	29.0	0.22	0.32
50	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	-3.06	29.0	-0.83	**0.0047**
50	Pre-trained SFCN Age	Pre-trained SFCN Gender	-3.91	29.0	-1.13	**0.00051**
100	3D-CAEMRI Fine-tuning	3D-CAEMRI Off-the-shelf	-0.94	29.0	-0.26	0.35
100	3D-CAEMRI Fine-tuning	Pre-trained SFCN Age	-0.96	29.0	-0.25	0.34
100	3D-CAEMRI Fine-tuning	Pre-trained SFCN Gender	-3.07	29.0	-0.76	**0.0047**
100	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Age	0.05	29.0	0.01	0.96
100	3D-CAEMRI Off-the-shelf	Pre-trained SFCN Gender	-2.68	29.0	-0.62	**0.012**
100	Pre-trained SFCN Age	Pre-trained SFCN Gender	-2.35	29.0	-0.63	**0.026**

Appendix D

Unravelling pathological ageing with BrainAGE on Alzheimer's Disease, Diabetes, and Schizophrenia

D.1 Methods

D.1.1 Data

Table D.1: Demographics of the participants used to train and validate the autoencoder.

Repository	**Site**	*Total of participants*	Number males	Mean and standard deviation [years]	Min Age [years]	Max Age [years]
ABIDE I	California Institute of Technology	37	30	28.4 ± 10.7	17.0	56.2
ABIDE I	Carnegie Mellon University	26	20	26.8 ± 5.7	19.0	40.0
ABIDE I	Kennedy Krieger Institute	54	41	10.1 ± 1.3	8.1	12.8
ABIDE I	Ludwig Maximilians University Munich	56	49	25.7 ± 11.7	7.0	58.0
ABIDE I	NYU Langone Medical Center	182	145	15.3 ± 6.6	6.5	39.1
ABIDE I	Olin Institute of Livingat Hartford Hospital	16	14	16.9 ± 3.7	10.0	23.0
ABIDE I	Oregon Health and Science University	28	28	10.8 ± 1.9	8.0	15.2
ABIDE I	San Diego State University	36	29	14.4 ± 1.8	8.7	17.2
ABIDE I	Social Brain Lab	28	28	33.4 ± 6.8	20.0	49.0
ABIDE I	Stanford University	37	30	9.9 ± 1.6	7.5	12.9
ABIDE I	Trinity Centre for Health Sciences	49	49	17.2 ± 3.6	12.0	25.9
ABIDE I	University of California Los Angeles	97	86	13.0 ± 2.2	8.4	17.9
ABIDE I	University of Leuven	64	56	18.0 ± 5.0	12.1	32.0
ABIDE I	University of Michigan	143	116	14.0 ± 3.2	8.2	28.8
ABIDE I	University of Pittsburgh School of Medicine	55	48	18.9 ± 6.9	9.3	35.2
ABIDE I	University of Utah School of Medicine	100	100	22.1 ± 7.7	8.8	50.2
ABIDE I	Yale Child Study Center	56	40	12.7 ± 2.9	7.0	17.8
ABIDE II	Barrow Neurological Institute	58	58	38.5 ± 15.5	18.0	64.0
ABIDE II	ETH Zurich	37	37	22.7 ± 4.4	13.8	30.7
ABIDE II	Erasmus University Medical Center Rotterdam	54	44	8.1 ± 1.1	6.2	10.7
ABIDE II	Georgetown University	103	68	10.7 ± 1.6	8.1	13.9
ABIDE II	Indiana University	26	20	24.8 ± 8.5	17.0	54.0
ABIDE II	Institut Pasteur and Robert Debré Hospital	55	25	20.1 ± 10.5	6.1	46.6
ABIDE II	Katholieke Universiteit Leuven	28	28	23.6 ± 4.8	18.0	35.0
ABIDE II	Kennedy Krieger Institute	207	137	10.3 ± 1.3	8.0	13.0
ABIDE II	NYU Langone Medical Center Sample 1	74	67	9.9 ± 5.0	5.2	34.8
ABIDE II	NYU Langone Medical Center Sample 2	27	24	6.8 ± 1.1	5.1	8.8
ABIDE II	Oregon Health and Science University	93	57	10.9 ± 2.0	7.0	15.0
ABIDE II	SanDiego State University	56	47	12.9 ± 3.1	7.4	18.0
ABIDE II	Stanford University	41	37	11.1 ± 1.2	8.4	13.2
ABIDE II	Trinity Centre for Health Sciences	42	42	15.2 ± 3.2	10.0	20.0
ABIDE II	University of California Davis	32	24	14.8 ± 1.8	12.0	17.8
ABIDE II	University of California Los Angeles	31	25	10.8 ± 2.4	7.8	15.0
ABIDE II	University of California Los Angeles Longitudinal Sample	37	35	13.5 ± 1.9	10.0	17.2
ABIDE II	University of Miami	26	20	9.8 ± 2.1	7.1	14.3
ABIDE II	University of Pittsburgh	34	26	14.9 ± 2.4	9.3	19.5
ABIDE II	University of Utah School of Medicine	32	27	20.9 ± 7.9	9.1	38.9
ADNI	–	18705	10233	75.3 ± 7.4	51.0	97.0
GSP	–	1558	661	21.5 ± 2.9	19.0	35.0
OASIS1	–	1683	638	51.5 ± 25.3	18.0	96.0
OASIS2	–	1345	576	76.9 ± 7.6	60.0	98.0
OASIS3	–	2768	1185	70.7 ± 9.3	42.7	97.0
FCP1000	AnnArbor a	25	20	20.4 ± 7.7	13.0	40.0
FCP1000	AnnArbor b	36	17	348.0 ± 1732.7	19.0	9999.0
FCP1000	Atlanta	28	11	30.6 ± 9.2	23.0	54.0
FCP1000	Baltimore	23	8	29.3 ± 5.5	20.0	40.0
FCP1000	Bangor	20	16	22.6 ± 4.6	19.0	38.0
FCP1000	Beijing Zang	197	68	21.1 ± 1.8	18.0	26.0
FCP1000	Berlin Margulies	26	12	29.9 ± 5.2	24.0	44.0
FCP1000	Cambridge Buckner	198	68	20.9 ± 2.1	18.0	29.0
FCP1000	Dallas	24	10	42.9 ± 20.4	20.0	71.0
FCP1000	ICBM	86	0	–	–	–
FCP1000	Leiden 2180	12	9	23.6 ± 2.6	20.0	27.0
FCP1000	Leiden 2200	19	11	21.8 ± 2.7	18.0	28.0
FCP1000	Leipzig	37	13	25.8 ± 5.1	20.0	42.0
FCP1000	Milwaukee a	18	0	–	–	–
FCP1000	Milwaukee b	46	14	53.7 ± 5.9	44.0	65.0
FCP1000	Munchen	16	9	68.3 ± 4.1	63.0	74.0
FCP1000	NYU TRT session1b	12	0	–		
FCP1000	NewHaven a	18	10	31.6 ± 10.3	18.0	48.0
FCP1000	NewHaven b	15	7	27.6 ± 6.4	18.0	42.0
FCP1000	NewYork a	84	40	24.3 ± 10.1	7.0	49.0
FCP1000	NewYork a ADHD	25	18	34.9 ± 9.6	20.0	50.0
FCP1000	NewYork b	20	1	$40.0 \pm nan$	40.0	40.0
FCP1000	Newark	19	9	24.1 ± 3.9	21.0	39.0
FCP1000	Ontario	9	0	–	–	–
FCP1000	Orangeburg	20	12	41.6 ± 11.2	20.0	55.0
FCP1000	Oulu	102	33	21.5 ± 0.6	20.0	23.0
FCP1000	Oxford	22	11	29.3 ± 3.3	21.0	35.0
FCP1000	PaloAlto	17	2	31.6 ± 7.6	22.0	46.0
FCP1000	Pittsburgh	16	9	37.6 ± 8.7	25.0	54.0
FCP1000	Queensland	19	10	25.9 ± 4.1	20.0	34.0
FCP1000	SaintLouis	31	13	25.3 ± 2.3	21.0	29.0
FCP1000	Taipei a	14	0	–	–	–
FCP1000	Taipei b	8	0	–	–	–

D.2 Results

D.2.1 MAE and BAG

Table D.2: MAE and the pearson correlation between BrainAGE and age, with and without the bias correction.

		MAE		r	
tissue	method	original	corrected	original	corrected
min proc	HH	4.46	4.08	**-0.29 ($p = 7.46 \times 10^{-5}$)**	-9.16×10^{-6} ($p = 1.00$)
	IOP	4.66	4.45	**-0.43 ($p = 0.00026$)**	-0.17 ($p = 0.17$)
GM	HH	5.75	4.68	**-0.37 ($p = 2.71 \times 10^{-7}$)**	-3.29×10^{-6} ($p = 1.00$)
	IOP	5.36	4.51	**-0.69 ($p = 8.93 \times 10^{-11}$)**	**-0.43 ($p = 0.00023$)**
WM	HH	5.26	4.01	**-0.46 ($p = 1.37 \times 10^{-10}$)**	-5.20×10^{-6} ($p = 1.00$)
	IOP	6.10	6.48	**-0.78 ($p = 2.60 \times 10^{-15}$)**	**-0.59 (1.51×10^{-7})**
CSF	HH	5.44	4.21	**-0.42 ($p = 4.11 \times 10^{-9}$)**	-5.72×10^{-6} ($p = 1.00$)
	IOP	5.61	5.47	**-0.51 ($p = 8.85 \times 10^{-6}$)**	-0.17 ($p = 0.18$)
DF	HH	6.53	4.14	**-0.57 ($p = 7.45 \times 10^{-17}$)**	1.74×10^{-5} ($p = 1.00$)
	IOP	5.48	6.07	**-0.49 (p=2.56×10^{-5})**	0.044 ($p = 0.72$)

D.2.1.1 Schizophrenia

Table D.3: MAE and mean of BrainAGE, in years, results of the brain age models for the schizophrenia analysis.

Metric [in years]	Clinical condition	Minimally processed	GM	WM	CSF	Deformation fields
MAE	Control	3.56	4.10	4.09	4.45	4.16
	Schizophrenia	**4.79**	**5.30**	**6.09**	**5.64**	**6.22**
BAG	Control	-0.01	1.12	2.09	1.98	1.13
	Schizophrenia	2.39	2.59	4.91	3.90	3.50

D.2.1.2 Diabetes

Table D.4: MAE and mean of BrainAGE, in years, results of the brain age models for the T2D analysis.

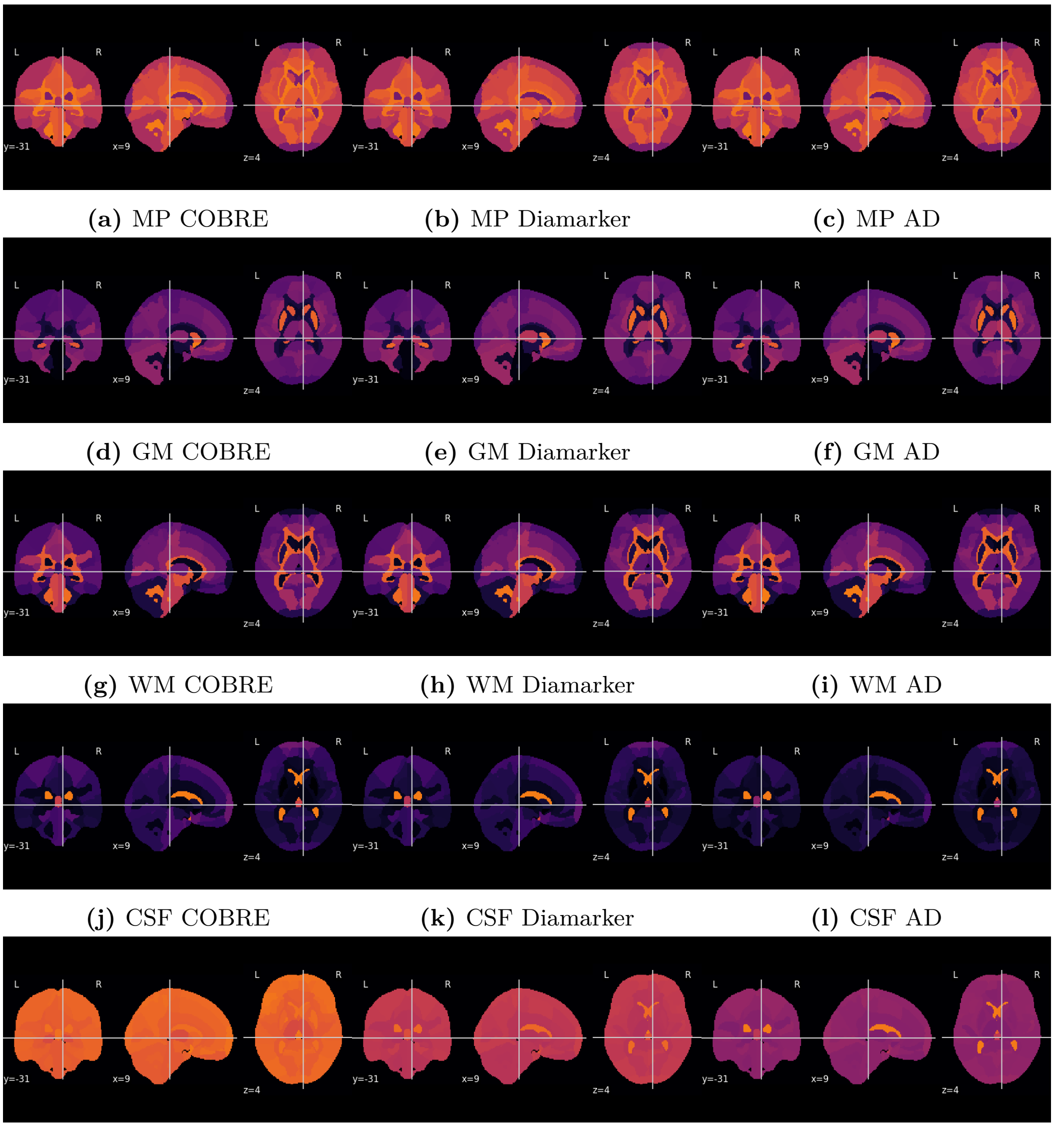
Metric [in years]	Clinical group	Minimally processed	GM	WM	CSF	Deformation fields
MAE	Control	4.44	4.24	4.71	4.79	4.56
	T2D	**6.43**	**8.59**	**7.86**	**7.71**	**11.13**
BAG mean	Control	-3.17	0.57	0.88	-1.21	0.90
	T2D	3.58	8.24	6.47	5.99	9.37

D.2.1.3 Alzheimer

Table D.5: MAE and mean of BrainAGE, in years, results of the brain age models for the AD dataset.

Metric [in years]	Clinical condition	Minimally processed	GM	WM	CSF	Deformation fields
MAE	Control	5.44	5.87	4.12	5.85	6.95
	AD	**6.36**	**11.43**	**8.41**	**10.31**	**12.70**
BAG	Control	-3.81	1.25	0.98	-0.68	3.37
	AD	5.23	11.21	8.41	9.01	12.41

D.2.2 Morphometry and sensitivity maps analysis



(a) MP COBRE (b) MP Diamarker (c) MP AD

(d) GM COBRE (e) GM Diamarker (f) GM AD

(g) WM COBRE (h) WM Diamarker (i) WM AD

(j) CSF COBRE (k) CSF Diamarker (l) CSF AD

(m) DF COBRE (n) DF Diamarker (o) DF AD

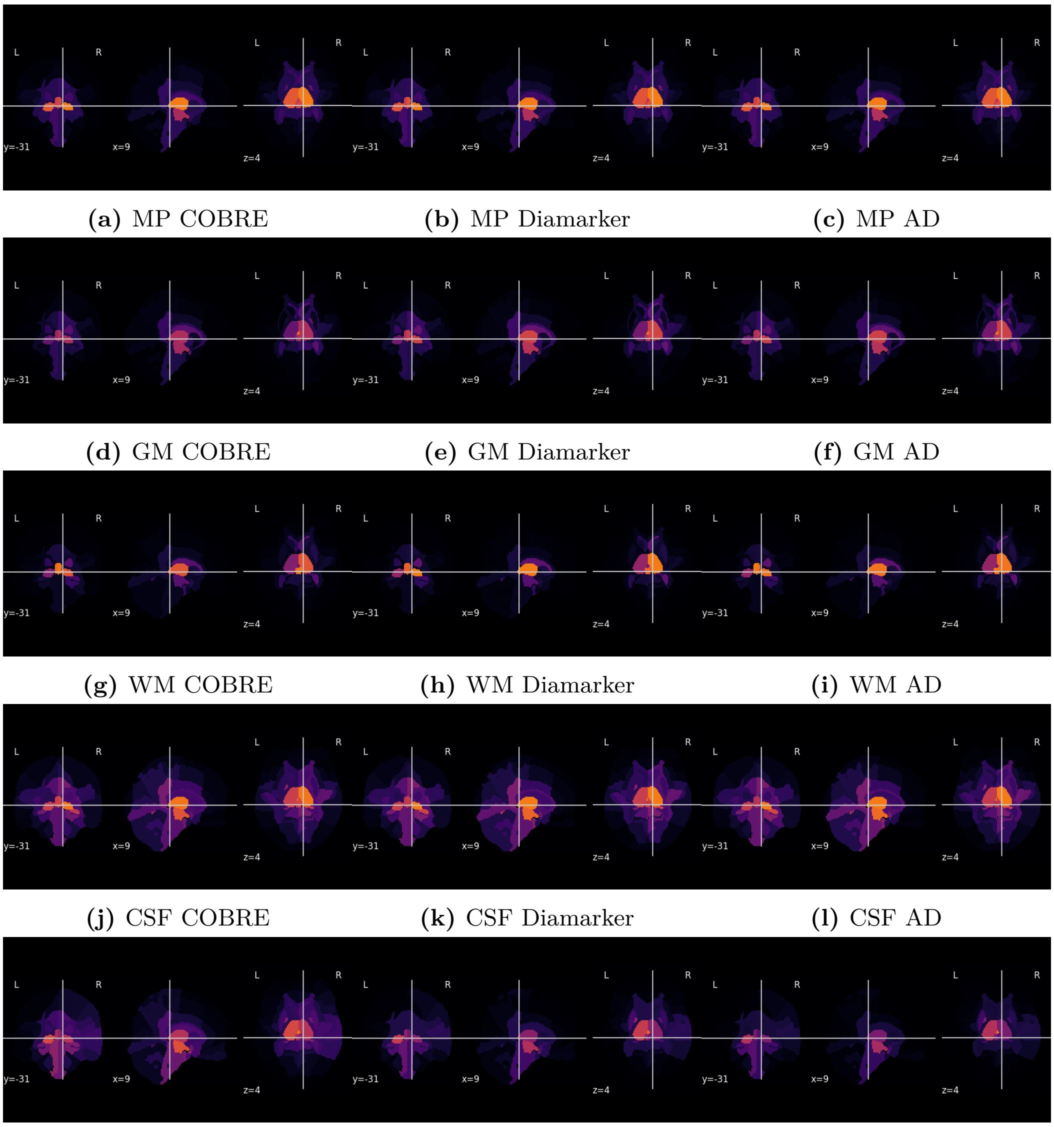
Figure D.1: Mean of the morphometric maps per ROI for minimally processed (MP), grey matter (GM), white matter (WM), Cerebrospinal fluid (CSF) and deformation fields (DF).

Table D.6: Percentage ROIs which were considered significant for the age factor on the ANCOVA test on the morphometrics analysis.

dataset	IXI–HH	COBRE	DIAMARKER	CIBIT AD
tissue				
Minimally processed	86.43	73.57	57.86	8.57
GM	95.71	94.29	84.29	0.00
WM	83.57	4.29	63.57	0.00
CSF	97.14	95.00	91.43	0.00
DF	75.00	66.43	41.43	0.00

Table D.7: Percentage of the ROIs which were considered significant for the age factor on the ANCOVA test on the sensitivity map analysis.

dataset	IXI–HH	COBRE	DIAMARKER	CIBIT AD
tissue				
Minimally processed	100.00	100.00	100.00	99.29
GM	100.00	100.00	72.14	0.00
WM	98.57	97.86	98.57	0.00
CSF	100.00	100.00	10.00	0.00
DF	65.00	54.29	60.71	30.00



(a) MP COBRE **(b)** MP Diamarker **(c)** MP AD

(d) GM COBRE **(e)** GM Diamarker **(f)** GM AD

(g) WM COBRE **(h)** WM Diamarker **(i)** WM AD

(j) CSF COBRE **(k)** CSF Diamarker **(l)** CSF AD

(m) DF COBRE **(n)** DF Diamarker **(o)** DF AD

Figure D.2: Mean of the sensitivity maps per ROI for minimally processed (MP), grey matter (GM), white matter (WM), Cerebrospinal fluid (CSF) and deformation fields (DF).

Table D.8: Percentage ROIs which were considered significant for the clinical condition factor on the ANCOVA test on the morphometrics analysis.

dataset tissue	IXI–HH	COBRE	DIAMARKER	CIBIT AD
Minimally processed	0.00	20.00	34.29	30.00
GM	0.00	1.43	61.43	0.00
WM	0.00	0.00	74.29	0.00
CSF	0.00	56.43	33.57	85.00
DF	0.00	12.14	11.43	30.00

Table D.9: Percentage of ROIs which were considered significant for the clinical condition factor on the ANCOVA test on the sensitivity map analysis.

dataset tissue	IXI–HH	COBRE	DIAMARKER	CIBIT AD
Minimally processed	0.00	80.00	97.86	69.29
GM	0.00	0.00	18.57	0.00
WM	0.00	0.00	94.29	0.00
CSF	0.00	0.00	0.00	100.00
DF	0.00	0.00	26.43	0.00

Table D.10: Jaccard index comparing the significant ROIs on age factor of morphometric with the sensitivity analysis.

dataset	COBRE	DIAMARKER	CIBIT AD
Minimally processed	0.74	0.58	0.09
GM	0.94	0.72	0.00
WM	0.04	0.62	0.00
CSF	0.95	0.07	0.00
DF	0.41	0.31	0.00

D.2.3 Noise impact on sensitivity maps per tissue

The relation of correlation between sensitivity and age with noise is depicted in Figure D.3 for the different modalities. The results reveal that each tissue achieves the maximum correlation at a different noise level. Moreover, the correlation evolution is specific to each tissue. white matter (WM) peaks at 2% and decreases drastically afterwards, the grey matter (GM) and deformation fields exhibit a smoother relationship with noise, reaching their maximum at a noise level of 2% and 4%, respectively. Minimally processed images plateau between 10% and 20%, reaching their maximum at 28% and decreasing more abruptly subsequently. Cerebrospinal fluid (CSF) experiences an abrupt transition to the maximum at 6%, with a smoother decrease afterwards. The maximum correlation values attained are 0.87 for minimally processed images, followed by GM, WM, and CSF with 0.82, 0.80, 0.63, and deformation fields with 0.30, respectively.

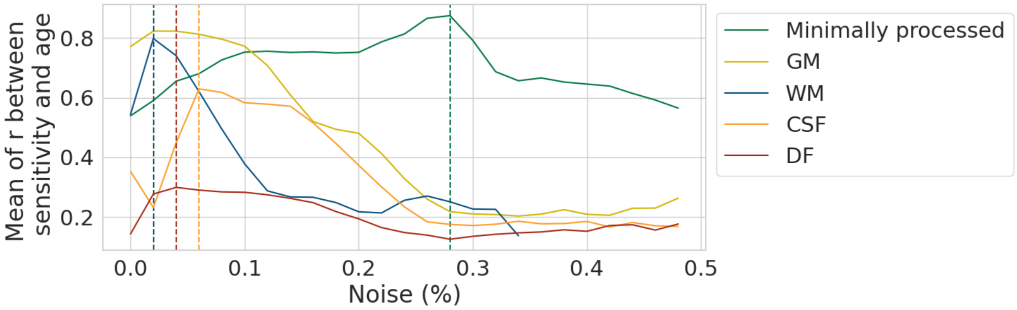


Figure D.3: Evolution of the Pearson correlation means between saliency and age for different noise levels using the IXI–HH data.

The correlation between sensitivity and age varies, drastically, with different noise degrees. Furthermore, the correlation of evolution with age depends upon the modality. This finding is in agreement with the results reported by Smilkov *et al*, the most suitable noise level depends upon the input type. For natural images, it is advantageous to apply noise within the 10-25% range, while the MNIST dataset (black and white images) exhibited optimal results at approximately 50%. Similarly, in these cases, GM, WM, CSF and deformation field images require reduced noise values, while minimally processed images require higher levels of noise. Moreover, in our results, minimally processed seems to be the modality more robust to noise.

Sensitivity maps yield reproducible results across datasets concerning the regions correlated with age. A perfect agreement is reported between the baseline (IXI-HH set) and the Cobre dataset. Regarding the diamarker, the agreement is perfect on minimally processed images and WM and very high on GM and deformation fields. Finally, for the Alzheimer's disease (AD) dataset, the agreement with baseline results is also perfect for the minimally processed image, but not for the other modalities. The morphometric analysis also yields reproducible results between the baseline dataset and Diamarker and Cobre. The AD dataset yields poor agreement with the baseline regarding the significant regions. As previously discussed this result might be related to the age range and the reduced number of samples of the AD dataset.