Exploring Image Super-Resolution to Optimise Studies of Atrophy in Frontotemporal Dementia

Master's Thesis

Thesis submitted as partial fulfilment of the MSc in Clinical Neuroscience

University College London

MSc Clinical Neuroscience

Word count:

10,844

Primary Supervisor: Dr. James Cole

Secondary Supervisor: Dr. Martina Bocchetta

UCL Queen Square Institute of Neurology – MSc/MRes Programmes

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. James Cole, for his assistance and direction at every stage of the research project. Dr. Cole supported my intuitions throughout this investigation and provided invaluable insight into the world of Deep Learning and Data Science. I would also like to offer my special thanks to Dr. Martina Bocchetta for lending me some of her infinite knowledge on Frontotemporal Dementia and the local UCL database.

I am deeply grateful to Yannick Hollenweger for his unwavering support. Finally, I would like to thank my parents, Nancy and Matt Holland, for their insightful comments, suggestions, and consistent encouragement.

Abstract

Background:

Frontotemporal Dementia (FTD) is a heterogenous group of progressive neurological disorders primarily impacting linguistic abilities and social behaviour. To enable disorder management and treatment, new biomarkers are required that accurately identify FTD. FTD biomarkers can be developed from measuring changes associated with pathology in MRI scans which are statistically different from healthy subject physiology. As MRI technology has improved, inconsistencies have emerged between the MRI data acquired on different machines which impact the quality of the collected data and biases the accuracy of brain atrophy measurements. This hinders research that combines data from across scanners, either across institutions or within the same centres over time as machinery has been upgraded over time. This reduces the quality of potential findings by limiting study analysis to either smaller data sets with identical acquisition protocols and technology or larger datasets with inherent biases.

Aim:

In this study, different combinations of preprocessing techniques were explored to reduce the heterogeneity of MRI data collection. Additionally, differences between healthy subjects and patients with FTD were explored using brain measures estimated from MRI data.

Methods:

MRI scans from healthy people and patients with FTD were cross-sectionally and longitudinally compared using the difference between chronological and predicted ages of each participant. As the quality of MRI scans changes with acquisition parameters and

scanner resolution, preprocessing methods including SynthSR deep learning 'super-resolution' were explored to reduce noise and biases in the data and improve lower-quality scans. This algorithm will be developed and tested using an initial sample size of N=129 Control participants and N=446 FTD participants, with the potential to scale.

Results:

The difference between each participant's chronological and predicted age was used to statistically differentiate between healthy and control datasets. BrainageR most accurately predicted healthy ages without the inclusion of additional preprocessing steps. When SynthSR preprocessing was applied before brainageR, inaccurate age predictions were produced which improved with data-specific thresholding.

Conclusions & Implications:

This study demonstrated that brainageR accurately predicts healthy ages across the various scanner magnetic fields of 1.5T and 3.0T scanners, enabling future investigations to combine datasets, leading to more robust findings. This investigation also highlights that brainageR can identify significant differences between FTD patients and controls, and it has the potential to be further investigated as a potential biomarker for FTD.

Table of Contents

Terms

bvFTD – Behavioural Variant Frontotemporal Dementia

Cam-CAN – Cambridge Centre for Ageing and Neuroscience

C9orf72 – Chromosome 9 Open Reading Frame 72

FTD – Frontotemporal Dementia

FTLD – Frontotemporal Lobar Degeneration

UCL – University College London

GRN – Progranulin

LMEM – Linear Mixed Effects Model

MAPT – Microtubule Associated Protein Tau

MR – Magnetic Resonance

MRI – Magnetic Resonance Imaging

MWU – Mann Whitney U

PAD – Predicted Age Difference

PNFA – Progressive Non-fluent Aphasia

PPA – Primary Progressive Aphasia

SD – Semantic Dementia

1.0. Introduction

Frontotemporal Dementia (FTD) is a group of common progressive neurodegenerative clinical diseases that primarily impact complex behaviours including social interactions and language capabilities (Sivasathiaseelan *et al.*, 2019). There are currently no definitive diagnostic tools for FTD, so diagnosis depends almost entirely on clinical assessment. FTD currently has no disease-modifying treatments and is characterized by a variable rate of disease progression.

Biomarkers are a useful tool for better understanding and treating FTD, and FTD biomarker development should provide a prognostic tool to aid clinicians in the early detection of FTD.

FTD mainly affects the brain, foremost in the frontal and temporal lobes, and can be measured, visualized, and identified using structural brain MRIs.

FTD MRI biomarkers, however, are particularly limited when applied to combined datasets collected using different MRI scanners due to the biases resulting from variable quality or acquisition parameters.

The first step in quantifying biomarkers for FTD is to identify group differences between control and FTD participant data and identify changes in FTD participants over time. Therefore, eliminating the inconsistencies in the processing of larger datasets is critical.

This study determines and then applies an optimal biomarker algorithm to a large and variable dataset to move towards an FTD prognostic tool.

1.1. Frontotemporal Dementia and Early Diagnosis

FTD is the second most common young onset dementia behind Alzheimer's Disease, and can clinically present as a variety of symptoms and diagnoses (Ratnavalli *et al.*, 2002). The age of

onset for FTD is usually between 45 and 65 years, and it has been estimated that 10% of cases are less than 45 year old and 30% of cases are above 65 (Knopman and Roberts, 2011). FTD incidence depends on the country assessed, ranging from 1 to 17 cases per 100,000 people (Onyike and Diehl-Schmid, 2013). FTD subtypes can be described in terms of clinical, pathologic, or genetic diagnoses, which can vary for a given patient. FTD subtypes have varying pathology, epidemiology, and survival factors which are important in diagnoses to manage patient and carer expectations (Erkkinen, Kim and Geschwind, 2018).

The most common FTD diagnoses are Behavioural Variant FTD (bvFTD) and Primary Progressive Aphasias (PPA), which respectively produce behavioural changes and language decline(Rascovsky *et al.*, 2011). PPA is further broken into two diagnoses relevant in this study: Progressive Non-Fluent Aphasia (PNFA, also referred to as Non-Fluent Variant PPA (nfvPPA)) and Semantic Dementia (SD, also referred to as Semantic Variant PPA (svPPA)), which are both also primarily associated with language deficiencies (Gorno-Tempini *et al.*, 2011). PPA patients presenting with core features PNFA, SD, or logopenic PPA (LPA or lvPPA) but not meeting diagnostic criteria, are classified as PPA not otherwise specified (PPA-NOS) (Harris *et al.*, 2013a). The different subtypes have been organized by clinical presentation, as shown below in Figure 1.

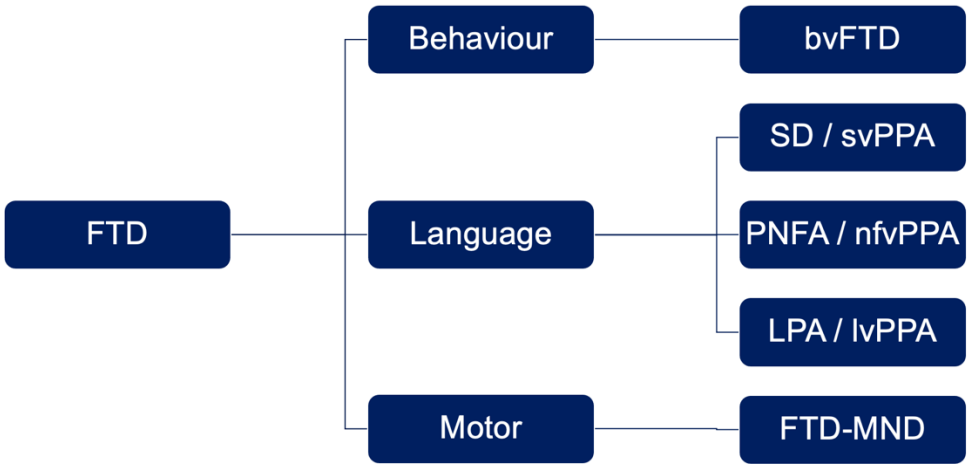

Figure 1: Hierarchy of FTD clinical diagnostics, organized by symptom presentation. The pathologies which produce FTD are defined as Frontotemporal Lobar Degeneration (FTLD) and have unique categorizations.

The average age at onset and disease duration for the FTD subtypes relevant in this study with known population statistics are shown in Table 1 below.

Table 1: Population statistics including age at onset and disease duration for bvFTD, SD, and PNFA FTD subtypes (Hodges et al.*, 2003).*

FTD Subtype	Age at Onset (years)	Disease Duration (years)
bvFTD	56.5 ± 7.6	8.2 ± 0.8
SD	59.4 ± 8.6	6.9 ± 1.2
PNFA	63.3 ± 5.1	10.6 ± 1.8

Misdiagnosis at early disorder stages is an important issue. As FTD is extremely heterogeneous between both patients and subtypes, a precise early diagnosis is extremely challenging to achieve.

Due to significant symptomatic overlap, early clinical diagnosis of FTD is often misdiagnosed as psychiatric illness, stroke, Alzheimer's Disease, Atypical Parkinson's Disorders, and Motor Neuron Disease (Espay and Litvan, 2011; Harris *et al.*, 2013b; Emma Devenney *et al.*, 2015; Sivasathiaseelan *et al.*, 2019). This early-stage misdiagnosis complicates the management process, impacting treatment options which could improve the disorder prognosis.

To improve diagnostic accuracy early in the disorder, clinical assessments have expanded to include genetic testing and neuroimaging analysis (Gossye, Van Broeckhoven and Engelborghs, 2019). These assessments which identify the underlying pathology are particularly useful as pathology begins before symptom presentation (Rohrer *et al.*, 2015). Genetic testing is used to approximate symptom cause or disorder risk, while neuroimaging analysis is used to identify changes in the brain which may be caused by disorder pathology.

FTD is largely heritable, with 30-50% of patients having a family history of dementia. Heritability results from various genetic mutations which produce pathogenic proteins that aggregate in neural tissue, interfering with normal cellular interactions and pathways (Ferrari *et al.*, 2014). The genetic mutations have complex relationships with diagnostics, as mutations are linked to a range of diseases and impact various neural mechanisms.

Mutations commonly associated with FTD include *MAPT*, *GRN*, *C9orf72*, and *TBK1* (Gossye, Van Broeckhoven and Engelborghs, 2019; Moore *et al.*, 2020). The average symptom onset, disease duration, and prevalence in familial FTD cases of each genetic mutation patient group can be seen in Table 2 below.

Table 2: The average and standard deviation of the symptom onset and disease duration are shown for each genetic mutation patient group, produced from two reviews (Gijselinck et al.*, 2015; Moore* et al.*, 2020). The MAPT, GRN, and TBK1 mutation studies were conducted in Belgian while the C9orf72 study was conducted in Western Europe (Cruts* et al.*, 2006; van der Zee* et al.*, 2013).*

Genetic Mutation	Symptom Onset (years)	Disease Duration (years)	Familial FTD Cases (Percentage)
MAPT	49.5 ± 11	9.3 ± 6.4	7
C9orf72	58.2 ± 9.8	6.4 ± 4.9	18.52
GRN	61.3 ± 8.8	7.1 ± 3.9	25.6
TBK1	69.1 ± 7.7	6.4 ± 3.9	3.2

The genetic mutations are most often associated with bvFTD, and correlate with apathy, social withdrawal, and memory loss (Gossye, Van Broeckhoven and Engelborghs, 2019;

Moore *et al.*, 2020). Each gene mutation produces abnormal proteins, which result in the accumulation of toxic proteins in the brain. *MAPT* mutations are associated with the tau protein and lead to pathological aggregation, while *GRN* and *C9orf72* both involve TDP-43 aggregation. *MAPT* mutations are often associated with dysfunctional neural plasticity and microtubule-based transportation. *GRN* is located on the same chromosome as *MAPT* and produces the progranulin growth factor involved in healing and inflammation (Che *et al.*, 2018). The *TBK1* mutation influences autophagy, neuroinflammation, and phosphorylation, upstream processes each involved in maintaining healthy neural equilibrium. *TBK1* and *C9ORF72* mutations are often associated with ALS as well as FTD, suggesting an FTD-ALS continuum (Van Mossevelde *et al.*, 2016).

The severity and regions impacted by the atrophy vary across FTD subtypes, genetic mutation and participants (Rohrer *et al.*, 2015). Macroscopically, this pathology produces primarily frontal and anterior temporal atrophy, which can be quantified and tracked over time using MRI scans (Rosen *et al.*, 2002; Peelle *et al.*, 2008; Hornberger *et al.*, 2010).

Longitudinal studies can quantify how the pathology is expanding over time and might enable predictions of future disease developments (Gossye, Van Broeckhoven and Engelborghs, 2019).

Patients with different genetic mutations can be tracked over time, and mutation-specific atrophy can be quantified to develop biomarkers which can be used in the early diagnosis of FTD.

1.2. Biomarkers

Biomarkers are quantifiable and trackable changes in molecules, genes, or characteristics which can be used to identify physiological processes, informing healthy or diseased patient

states. Biomarkers enable earlier identification and intervention, as they identify underlying pathology and inform changes occurring in the patients. This can lead to early qualification for clinical trials unavailable at later disease stages, potentially drastically impacting prognosis as the impact of intervention will have changed over time (Morovic *et al.*, 2019). Proven biomarkers can also be used as outcome measures in clinical trials to quantify the impact of an intervention.

Biomarkers enable researchers to better define and understand pathologies and improve clinical diagnostics. They extend the researcher's understanding of the disease by highlighting pathological mechanisms which can be leveraged in treatment development. Extending biomarker accuracy and clinical applications would provide more diagnostic information in hospitals, empowering clinicians to determine and communicate with improved diagnostic certainty and precision, and potentially use the new information to improve disorder management.

Biomarkers can be quantified using various technologies and physiological changes. Fluid biomarkers for FTD track the presence of certain molecules, such as progranulin, p-tau, and dipeptide repeat proteins, in cerebrospinal fluid, serum or plasma (Swift *et al.*, 2021). Fluid biomarkers are limited in the earlier stages of the disease, as they often correlate with symptom severity (Younes and Miller, 2020).

Non-invasive neuroimaging also tracks various neural biomarkers, with each technique optimized for certain physiological phenomena, including molecular, structural or functional changes. FDG-PET, for example, is a molecular imaging method that measures the uptake of the radiotracer ^{18}F-Fluorodeoxyglucose to quantify brain glucose metabolism and identify hypometabolism in mild cognitive impairment (Veldsman and Egorova, 2017). Structural

Magnetic Resonance Imaging (MRI) is another commonly used tool, which leverages hydrogen proton behaviour to quantify grey matter, white matter, vasculature, and pathological tissue. Over time, these neural structures can be monitored for macroscale indicators of neurodegeneration such as atrophy, potentially informing pathology or prognosis.

1.3. Brain-Age Paradigm

Biomarkers can be used to develop algorithms which better inform the diagnosis and prognosis associated with the data. Neuroimaging has been used, for example, to estimate brain age from an MRI scan (Franke *et al.*, 2010; Basodi *et al.*, 2021). This predicted brain age, compared with the chronological patient age, is the brain age gap or brain-predicted age difference (brain-PAD). This metric has been used in early diagnosis of various neurological disorders, including Alzheimer's Disease and Parkinson's Disease (Beheshti *et al.*, 2019).

An algorithm based out of UCL, brainageR, quantifies volumetric atrophy from T1-weighted MRI scans to produce predicted brain ages (Cole, 2022). Developed and tested using MRI data from healthy subjects, it quantifies volumetric measurements of neural tissues to predict a subject's biological brain age. When used on longitudinal study data, this algorithm could anticipate pathologic changes over time and allow clinicians, researchers, and participants to better understand and predict patterns of FTD. The brain-PAD produced with brainageR can be calculated for each scan or as a function of time over multiple scans. Larger brain-PADs imply more pathology and can inform clinical and diagnostic decisions.

Applying brainageR to MRI data could identify group-level differences between healthy and diseased brains, potentially producing a new biomarker for this disease.

1.4. MRI Technology and Data Heterogeneity

Anatomical MRI is a useful tool for clinical and research FTD applications and can be used to develop numerous biomarkers for prognosis, diagnosis, and trial outcome measurement. Using MRI to visualise tissues in the living brain enables non-invasive identification of abnormalities, pathologies, and structures. MRI technology has high spatial resolution which enables increased scientific understanding of neural structure and function (Yousaf, Dervenoulas and Politis, 2018).

MRI technology is consistently improved, with different institutions leveraging each technological iteration for different studies. Each application requires varying technologies and protocols for recording MRI data, with unique data quality and features produced with each MRI scanner and acquisition parameter.

Scanner differences resulting in variable MRI measurements can falsify the results of studies, particularly with small sample sizes where measurement variability could be mistaken as a disease-based biomarker. As changes in atrophy are minimal to begin with, small artifacts from different measurements could drastically alter the findings in dementia studies. Each scanner has specific hardware such as field strength and the number of channels in the head coil, which influence the resulting data. As each MRI company produces distinct scanners, with a unique selection of individual models available, cross-scanner variability can drastically impact study conclusions. Even studies comparing identical scanners have found different measurements to produce potentially confounding factors. One study evaluated the impact of collecting control and Alzheimer's Disease data at one centre, using 6 identical scanners over 10 years (Stonnington *et al.*, 2008). All scans were performed using General Electric Signa 1.5 T scanners with identical major hardware

elements, slice thickness, and matrix dimensions, with varying recording parameters (TR, TE, and flip angle). Although differences across scanners were negligible for AD investigations, inaccurate segmentation of the thalamus was observed which could drastically impact investigations for other pathologies. Combining data across scanners without image enhancement complicates conclusions or applications of findings from this data due to inherent biases and noise.

MRI technology improvements have yielded scans appropriate for different applications which leverage different hardware, field strengths, and resolutions to produce data for the analysis (Balchandani and Naidich, 2015; Obusez *et al.*, 2018). As some MRI scanners have been in use longer than others, the older technology is more widely implemented, and the data is better investigated in publications. Additionally, MRI scanners are costly to replace, limiting the number of institutions with access to cutting-edge technology. As newer machinery, such as some higher resolution scanners, often produces greater detail of intracranial pathology and anatomy, well-funded groups will purchase upgraded technology to enable new research findings (Balchandani and Naidich, 2015). These studies are challenging to compare, however, exemplified by different resolutions of 1.5T and 3.0T producing wide-ranging differences in images recorded from the same participants under identical conditions (Buchanan *et al.*, 2021).

Limited resources, such as the time, money, and patient comfort required to record MRI data, often dictate the acquisition parameters for an MRI scan. These acquisition parameters determine the amount of data collected, determining the level of detail available for analysis. Some research groups have the resources to record the entire brain with limited space between slices. Minimising the space between each recorded slice takes

longer but maximises the amount of detailed data available for analysis. In contrast, hospitals tend to minimize recording times for each patient, with scanning times allotted for each patient constraining the number of slices which can be recorded (van Beek *et al.*, 2019). This maximizes the number of subjects imaged daily while minimizing the cost required for each scan. Additionally, MRI recordings are often uncomfortable for the patient, particularly if they are claustrophobic. Limiting recording times limits unnecessary patient discomfort. These limited recording practices produce less precise patient data which only include 20-30 slices that are spaced at distances of 5-7 mm (Iglesias *et al.*, 2021). These discrepancies limit the ability to combine datasets or leverage research-proven algorithms in clinical environments.

The variability across MRI scans across institutions and clinics results in small sets of standardized data. The amalgamation of multiple such smaller sets into a sufficiently large dataset, usable for the evaluation of numerical biomarkers and algorithms, gives rise to unwanted biases stemming from the differences in recordings.

1.5. MRI-Based Algorithms

Current algorithms are developed and validated using distinct data types, with specified resolutions and acquisition parameters. As research datasets are often recorded at higher resolution MRI and highly detailed acquisition parameters, algorithms are often developed using these data specifications (Laird, 2021). Combining or mismatching datasets and algorithms leads to a bias in the data analysis, impacting the information which can be experimentally extracted.

This variability in resolutions and acquisition parameters presents obstacles for all users of the data. The contrast between high and low data quality reduces the applicability of

validated biomarkers, limiting the translation of these tools to clinical applications, treatment development and improved understanding of pathology (Alexander *et al.*, 2017). For example, biomarkers validated using research-grade resolution and acquisition parameters are challenging to effectively translate to clinics which tend to have lower-grade technology.

Additionally, the inability to combine datasets limits investigations to small datasets of identically collected research-based data. Conclusions from smaller datasets have reduced statistical accuracy and clinical relevancy as they are more susceptible to biases from a subset of pathology or sample factors and are less accurate when extrapolated to a larger population. Due to the variations, phenotypes, and pathologies of FTD, small datasets often do not provide enough samples of each subtype to garner subtype-specific statistically meaningful conclusions (Ferrari *et al.*, 2014). Enabling combined datasets would allow for subtype-specific biomarker investigation of large datasets featuring various FTD subtypes which may yield novel information on pathology and progression.

1.6. MRI Signal and Noise

MRIs can be visualized using methods that each highlight certain image features. Identified patterns can be leveraged to improve the resulting data through the removal of corrupted data and noise reduction.

MRI data often includes noise, which interferes with signal quality and can result from imperfect technical measurements during data acquisition (Bhonsle, Chandra and Sinha, 2012). Noise can also include signal components that negatively impact algorithm efficacy. Noise can be added to an MRI signal in myriad ways including the technology used for the

recording, any artefacts during the measurement, or any brightness and colour variations that may result from variations in the scanner or analysis conducted (Sujitha *et al.*, 2017).

All variations of noise sources similarly impact the data. For example, patient discomfort from prolonged recording sessions increases the likelihood of shifting during the session, producing artefacts in the image including ringing and blurring which lead to reduced quality. Similarly, artefacts resulting from scanner malfunction would impact the image in the same way and thus can be collectively identified and removed.

Noise is often depicted using histograms, which plot the intensity of each frequency included in the 3D nifty image (*FSLeyes — FSLeyes 1.4.6 documentation*, no date). Histograms depict the power and frequency of different signal components of the image and often display technological noise as spikes at low and high frequencies (Sujitha *et al.*, 2017). The spikes often bookend useful data for analysis and confound statistical tools if applied to the entire dataset. Noise reduction is a common feature of preprocessing techniques and improves signal quality to harvest more useful information from the data. For example, thresholding is a popular image preprocessing step which removes noise by reducing the sample to only include data within a defined intensity range (Sujitha *et al.*, 2017).

Once identified, image enhancement can be applied to data which includes artefacts to improve the poor image quality. Data with any type of artefact source can benefit from image enhancement. Image Quality Transfer (IQT), for example, learns the mapping of matched clinical and research-grade images to augment the lower quality clinical images by improving image resolution or associated information content (Alexander *et al.*, 2017). Similarly, consumer software such as FDA-approved iRAD uses image enhancement for MRI,

CT, and X-Ray to reduce noise and improve both image contrast and entropy (*iRAD 510(k) Premarket Notification*, 2021). Each image enhancement software improves certain aspects of the data, solving issues individual to each technique or data type considered.

1.7. Preprocessing Methods

Preprocessing methods have been developed which preprocess the MRI data to algorithmically reduce the differences from resolution or acquisition parameters. These methods extract features and biomarkers from MRI scans, remove MRI artefacts like magnetic field inhomogeneities, and spatially register images to enable voxel-wise statistical testing across participants. Collectively, this allows for the combination of datasets from data recorded across research and clinical groups. Currently, MRI preprocessing methods require specific input parameters determined by resolution and acquisition parameters including slice spacing, voxel size, and anisotropy. When the data used in these methods differ from these input parameters, the methods are not validated to accurately preprocess the data (Park and Han, 2018).

To combat data inconsistency and algorithm-compliancy limitations, researchers collaborated to develop SynthSR which standardizes data across modalities, resolution, and acquisition parameters (Iglesias *et al.*, 2021). SynthSR aims to create a new standardised model for processing data, allowing data collected from different MRI scanners to be combined and included in analysis, to enable investigations limited by dataset sample size. SynthSR algorithmically modulates the image dimensions by artificially adding and manipulating data slices to enable research algorithms to use data collected in routine clinical brain MRI protocols. Additionally, SynthSR 'equalises' MRI data quality from different sources, enabling the combination of data from various recording technologies. SynthSR

algorithms leveraging super-resolution were shown to improve data quality for healthy brain data and Alzheimer's Disease data (Iglesias *et al.*, 2021).

1.8. Proposed Investigation

This study aims to test whether image enhancement methods can be used to improve the sensitivity of brain-age prediction to group differences between FTD patients and healthy controls. An additional aim of the investigation is to assess changes in FTD patients over time using the best-performing image enhancement method. This study also aims to develop a longitudinal model using factors such as genetic mutations and FTD clinical subtypes which influence image enhancement performance.

Finally, this study aims to validate an image enhancement-based pipeline that accepts data of varying quality and acquisition parameters by testing three separate preprocessing pipelines in conjunction with brainageR. If a pipeline is found that can be successfully applied to data and identify group differences, this would enable FTD biomarker discovery using larger, previously unexplored, datasets.

The three preprocessing pipelines were compared as follows: no preprocessing; super-resolution SynthSR preprocessing; and a novel pipeline featuring both SynthSR and additional preprocessing steps. The novel pipeline will be developed to maximize age prediction accuracy, before differentiation between FTD and control data and developing the longitudinal model. If successful, this will provide initial validation of brainageR age predictions as biomarkers for FTD and enable the use of larger FTD datasets with higher statistical power for biomarker development.

In general, this investigation will evaluate four hypotheses:

1. An image enhancement pipeline with additional preprocessing will outperform SynthSR image enhancement when applied before brain age prediction algorithms.

2. Applying SynthSR image enhancement methods to the data will produce the best performing cross-sectional brain age predictions.

3. The cross-sectional group differences will show that the produced brain-PAD for FTD and control data will be statistically dissimilar.

4. The longitudinal within-subject differences will show that brain-PAD is statistically impacted by time, genetic mutation, and FTD subtype.

2.0. Methods

BrainageR, a brain-PAD biomarker algorithm, was investigated as a potential tool for clinical FTD identification and monitoring. The exploration sought to statistically differentiate between labelled healthy and FTD participant MRI data from a local UCL database. Additionally, a longitudinal study considered patient brain-PAD changes over time and aimed to aid clinical management by identifying patterns of pathology progression. Three computational preprocessing methods were applied and tested to reduce noise from heterogeneous scanners and acquisition parameters by minimizing the brainageR brain-PAD for healthy participants. The primary preprocessing method investigated was the SynthSR algorithm, which produces outputs of a common resolution and dimension, irrespective of the input resolution or acquisition parameters (Iglesias et al., 2021). The best-performing pipeline was leveraged to compare the healthy and FTD data, assess the factors contributing to brain-PAD in the FTD data, and to conduct a longitudinal analysis.

2.1. Overview of Data

Investigations were conducted using two datasets including labelled data from healthy participants and participants diagnosed with Frontotemporal Dementia (FTD).

The IXI dataset is an open-source dataset of 581 MRI images collected from healthy subjects ('IXI Dataset – Brain Development', no date). This dataset was collected by institutions in London including the Hammersmith Hospital using a Philips 3.0T system, the Institute of Psychiatry using a GE 1.5T system, and Guy's Hospital using a Philips 1.5T system. The MRI images used in this dataset are T1-weighted images with 1mm spacing between slices, as described in Table 4. The participants included in this dataset have a mean age of 48.6 years at the time of recording, with a standard deviation of 16.5 years.

Secondly, a local UCL dataset collected at the Dementia Research Centre (London, UK) was used which includes control subjects and participants with various clinical subtypes of FTD. The data collection was approved by a local ethics committee, and each participant provided signed informed consent. Most participants are British residents and have the financial means to travel to central London and dedicate time to research. 317 participants have multiple scans recorded over variable years, producing a total of 382 control and 885 FTD scans. The FTD subtypes present in this data include bvFTD, PPA, PNFA, and SD. Most patients had FTD of sporadic origin; only 88 patients were included with known genetic factors including Progranulin (*GRN*) mutations, Microtubule Associated Protein Tau (*MAPT*) mutations, and non-coding repeat expansions in the *C9orf72* gene. The collected data is from different scanners using different protocols including 523 on 1.5T and 744 on 3.0T scanner. The participants included in this dataset have a mean age of 64.2 years old at the time of recording, with a standard deviation of 9.4 years. The number of participants with each associated label is demonstrated below in Table 3.

Table 3: UCL Participant Data Specifications.

UCL Dataset	Controls	FTD	bvFTD	PNFA	SD	PPA-NOS	MAPT	GRN	C9orf72	TBK1	GRN + C9orf72	C9orf72 + SQSTM1
Number of Participants	129	446	211	119	97	19	30	23	31	2	1	1
Average Age at First Scan	60.7	64.4	62.4	67.9	64.4	63.9	55.5	64.0	62.3	67.3	63.0	68.1
Longitudinal Participants	101	216	100	52	53	11	18	11	16	1	0	0
Longitudinal Scans	3.5	3.0	3.2	2.6	3.0	3.2	4.2	2.9	3.9	6.0	-	-
Average Disease Onset	-	59.7	57.2	63.7	59.7	60.7	50.0	60.3	55.9	63.0	58.0	63.0
Std. Disease Onset	-	8.3	7.7	8.6	7.7	6.3	6.6	6.6	7.7	2.8	-	-
Average Disease Duration	-	4.7	5.1	4.2	4.6	3.2	5.5	3.7	6.4	4.3	5.0	5.1
Std. Disease Duration	-	2.8	3.2	2.2	2.4	1.6	3.0	2.5	4.6	1.7	-	-

The spatial resolution of the data varies, depending on the recording protocols used. Data were recorded across three systems, each with unique protocols: 1.5T GE Signa, 3.0T Siemens Trio, and 3.0T Siemens Prisma, as presented in Table 4 below.

Table 4: Scanner protocol and acquisition parameters for the UCL and IXI databases. Some information for the IXI scanner parameters is not available. The scanner information for the GE 1.5T system used at Institute of Psychiatry in the IXI dataset is unavailable.

Scanner	Manufacturer	Thickness (mm)	Acquisition Matrix	TR (ms)	TI (ms)	TE (ms)
1.5T Signa	GE Medical systems, Milwaukee, Wisconsin, USA	1	256 x 256	1	650	5
3.0T Trio	Siemens, Erlangen, Germany	1.1	256 x 256	220	900	2.9
3.0T Prisma	Siemens, Erlangen, Germany	1.1	256 x 256	2000	850	2.93
3.0T Intera	Philips Medical Systems	1.2	208 x 208	9.6	N/A	4.6
1.5T Intera	Philips Medical Systems Gyroscan	N/A	N/A	9813	N/A	4.6

Data were removed from both datasets due to database errors, including corruption or inconsistencies between the file names in the data and csv files. The IXI and UCL databases respectively had 563 and 1256 files used in the final analysis, with 18 and 10 files removed from each database.

2.2. Overview of Algorithms

BrainageR. BrainageR is an algorithm that accepts T1-weighted MRI data with 1mm resolution for volumetric atrophy analysis to approximate the size of specific brain regions and predict the brain age of a subject. BrainageR preprocesses the data, segmenting and normalizing the input data using voxel-based morphology implemented through Statistical Parametric Mapping (SPM12; Wellcome Department of Cognitive Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm) (Friston, 2007). After preprocessing, the data is saved as images for manual qualitative inspection and quality control. The data is then vectorized, and brain regions containing white matter, grey matter, and cerebrospinal fluid have their vectors first masked and then combined. The data then undergoes a principal component

analysis with 80% variance to reduce the dimension of the data to 435 principal components. These 435 variables are the inputs into a trained algorithm that outputs a predicted brain age of the participant at the time of recording. The data involved in the development of brainageR includes 7 different projects with public datasets, including the IXI dataset used as validation in this investigation. The model was developed using only healthy subjects and randomly split the combined dataset into 3377 scans for training and 857 scans for testing.

SynthSR. SynthSR is a preprocessing software that can be applied to MRI scans for image enhancement, computing registrations and segmentations for analysis which improve image quality and resolution (Iglesias *et al.*, 2021). SynthSR is distinct from other image enhancement algorithms as it provides an 'all-purpose' model which can recover high-resolution features from data recorded with any scanner, magnetic field strength, or acquisition parameters. This model was uniquely developed from synthetic data according to contrast, resolution, and orientation, and accounts for movement and other artefacts. SynthSR outputs images which are equalised across scans with improved image quality and normalized spacing between slices. In this investigation, SynthSR was explored as a preprocessing technique to prepare the data with varying scan resolutions and acquisition parameters for brainageR analysis.

2.3. Preprocessing Pipelines

SynthSR preprocessing was explored to improve brain-PAD accuracy, 'equalising' neuroimaging data with varying quality and structures. Three preprocessing techniques were combined with brainageR and applied to healthy MRI data to identify which technique

produces minimal brain-PAD across heterogeneous recording methods. The best-performing pipeline was used for cross-sectional and longitudinal analysis of the data.

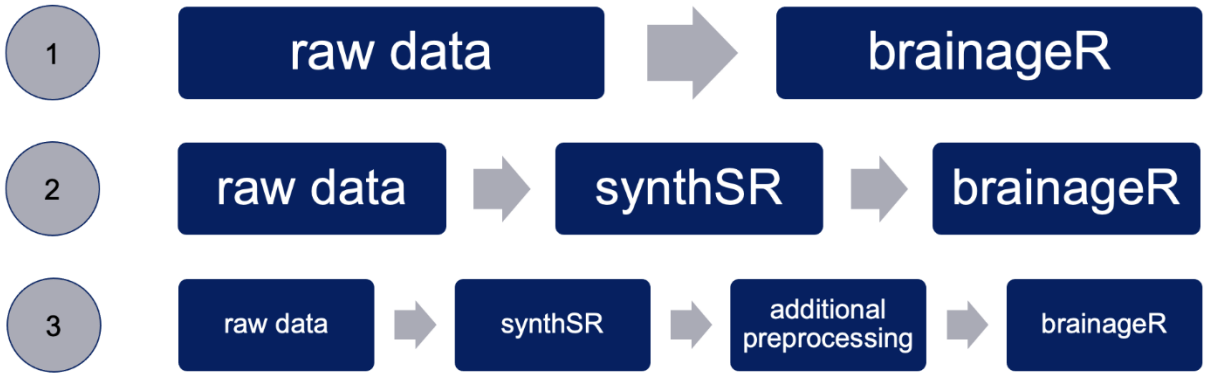

Figure 2: The three preprocessing pipelines considered are labelled 1, 2, and 3. The first pipeline acts as a control pipeline and is conducted solely with the brainageR software, without additional preprocessing. The second pipeline first processes the data with SynthSR to improve the data quality and then passes the data through the brainageR software. The third pipeline first applied SynthSR, then additional preprocessing steps, then brainageR.

In each pipeline, the raw data were reoriented using the FSL toolbox (fslreorient2Std) to ensure standard orientations were used for all pipelines. The impact of reorienting the data was assessed using brain-PAD values before and after reorientation on a subset of data.

2.4. Pipeline Analysis Methods

Each pipeline was assessed using the brain-PAD, calculated by the difference between the chronological participant ages and the brainageR age estimations. In pipeline development, brain-PAD from healthy data was minimized to optimize algorithm accuracy. In cross-sectional and longitudinal analysis, brain-PAD was considered to quantify patient pathology. The absolute brain-PAD was used to calculate the mean average error (MAE) and associated standard deviation (Std.) to ensure that under-predicted and over-predicted ages did not cancel each other out in the analysis.

The Pearson's correlation r-value of the estimated and chronological ages was also used to assess brainageR accuracy. The r^2 value was also used to reduce influence by test set age distribution.

The MAE, Std., r-value, and r^2 value are used to compare accuracy across different preprocessing pipelines, with more accurate results showing minimized MAE and Std and maximized r-value.

The Mann-Whitney U (MWU) statistical test was used to compare all independent and non-parametric datasets. The MWU test compares the distribution of two datasets to assess if they are likely to result from the same population, with p-values < 0.05 determined to be dissimilar. Similarly, the closer the p-value is to 1, the more similar the datasets are. As required by the MWU, this analysis assumes the datasets compared include samples which are randomly selected from their populations.

All statistical analyses were conducted using Python and various toolboxes, most predominantly SciPy (Virtanen *et al.*, 2020).

2.5. Image Enhancement Pipeline 3

Pipeline 3 was developed to maximise prediction accuracy by applying additional preprocessing to Pipeline 2. This leveraged SynthSR preprocessing benefits while replicating raw data features to improve compatibility with brainageR.

Healthy patient data were used throughout pipeline development as they are assumed to have similar chronological and biological brain ages, allowing brain-PAD to be minimized as a measure of best-performing pipeline accuracy.

Pipeline 3 was first developed using the IXI dataset and recreated using healthy UCL subjects. Pipeline 3 was developed using the IXI dataset as it was involved in brainageR development and includes features expected by brainageR, ensuring changes in the data appearance or brainageR output result from SynthSR. The specifications of applying Pipeline 3 to the UCL dataset instead of the IXI dataset can be seen in Appendix 1: Pipeline 3 UCL Dataset Specifications. Although both scanner and spatial resolution potentially contribute to subset differences in the UCL dataset, for simplification these categories are referred to solely by their magnetic field strength of 1.5T and 3.0T throughout this study.

Differences in the data before and after SynthSR preprocessing were first identified then minimized to remove features which may not be compatible with brainageR. The preprocessing method which produced the most accurate age estimation was included as Pipeline 3 in the final comparison. The development steps discussed below are summarized in Figure 3 below.

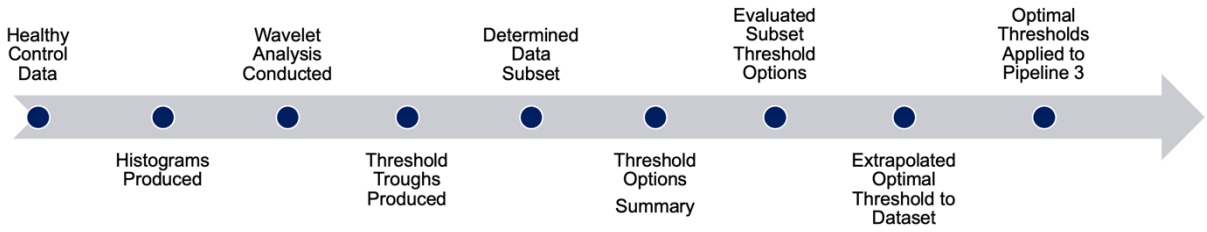

Figure 3: A summary of the steps used to develop Pipeline 3, described below in depth.

Visual inspections were conducted (using the FSLEyes software tool) on randomly selected segmented MRI scans, and showed that images after SynthSR preprocessing appeared brighter, implying higher levels of noise (Sujitha *et al.*, 2017; *FSLeyes — FSLeyes 1.4.6 documentation*, no date).

Histograms identified the largest difference in noise resulting from SynthSR preprocessing to be a large low-frequency spike demonstrated in Figure 4, identified as a potential confounding factor impacting brainageR.

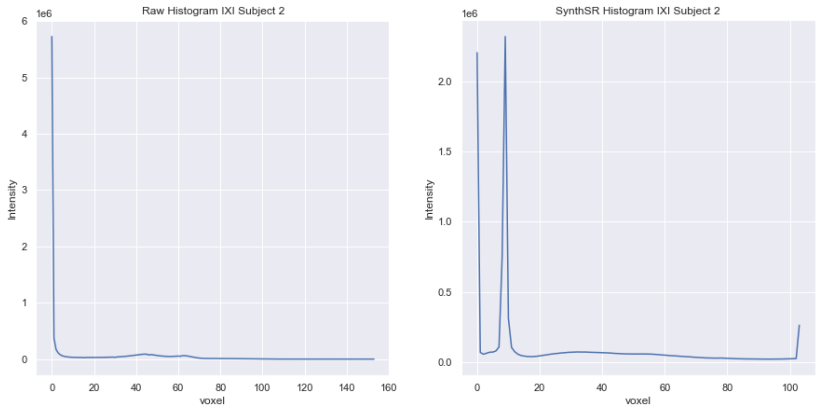

Figure 4: Example histograms for IXI subject 2 before and after the application of SynthSR preprocessing.

Thresholding was applied to reduce the impact noise may have on brain-PAD accuracy by removing the largest low- and high-frequency peaks. The useful signal was maximized by selecting thresholds which were far enough from the spikes to remove noise without compromising signal components. The thresholds for each sample were calculated as the troughs immediately following the low-frequency peak and immediately preceding the high-frequency peak, as shown in Figure 5.

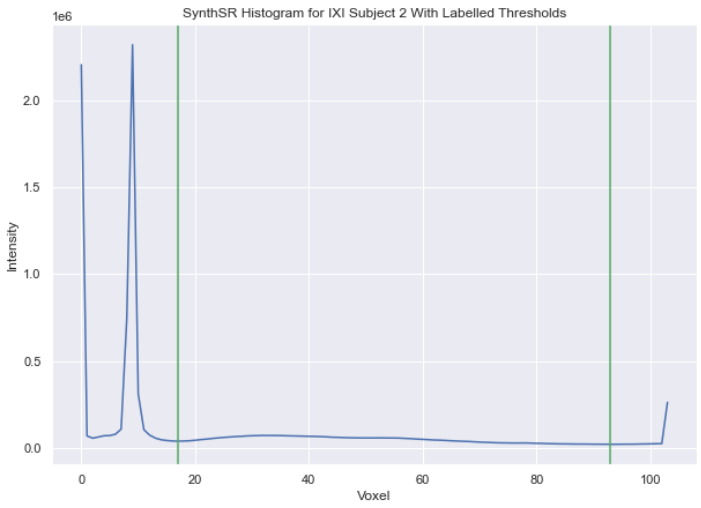

Figure 5: The histogram is depicted after SynthSR preprocessing for IXI subject 2, including labelled threshold values.

The threshold values for each sample in a dataset were used to determine the minimum, maximum, median, and mean threshold options for both the lower and upper thresholds. Each threshold option was individually evaluated using a subset of the healthy data, described in more detail in Appendix 2: Sample Size.

Different threshold combinations were assessed using the subset, with the lowest brain-PADs considered optimal. First, each lower threshold option was combined with the maximum upper threshold option, to limit the influence of high-frequency noise. Then, each upper threshold option was combined with the best-performing lower threshold.

The best-performing thresholds were then extrapolated to the entire healthy dataset, using both the exact threshold values shown to be successful and the equivalent threshold options for the large dataset. To determine the threshold values appropriate for all healthy data, a wavelet analysis for the entire dataset produced lower and upper threshold options. Both thresholds were applied to the entire dataset and compared, with the best-performing threshold combination used as Pipeline 3.

2.6. Cross Sectional Group Difference for Healthy and FTD UCL Data

Statistical tests were applied to the output of the best-performing pipeline, directly comparing the healthy and FTD-diagnosed participant data to identify group-level differences resultant of the pathology.

Only the first MRI recording for each patient was leveraged for the statistical comparison. The data were also separated for analysis by resolution, respectively comparing 1.5T and 3.0T control data to 1.5T and 3.0T FTD data. The data included two numerical and independent sample groups. The null hypothesis stated there would be no difference in the predicted and actual ages of the participants across the two sample groups. The data were determined to be non-parametric through Shapiro Wilk normality tests (scipy.stats.shapiro) and histograms of the relevant datasets. The statistical comparison of these groups was conducted using the Mann-Whitney-U test, as outlined in Section 2.4. Pipeline Analysis Methods.

2.7. Confounding Diagnostic, Genetic Mutation, and Chronological Age Factors

Factors thought to influence the brain-PAD include the patient-specific genetic mutation and FTD clinical diagnosis. These confounding factors were investigated by statistically comparing the brain-PAD of each subset of labelled data by mutation or diagnosis. Each labelled subset was compared to the rest of the FTD group using MWU tests. Assessments were conducted using MAE, Std., r-value, and r^2-value as outlined in Section 2.4. Pipeline Analysis Method.

Previous investigations involving the Cambridge Centre for Ageing and Neuroscience (Cam-CAN) dataset showed a chronological age dependency, with a significant relationship between the brain-PAD and chronological age (r-value = -0.379) (Cole, 2022). To investigate

the current dataset and the impact of processing techniques on age dependence, the brain-PAD and chronological age relationship was investigated for each pipeline using the UCL dataset. Linear regressions were produced depicting age-based errors for analysis and applied to 'correct' the age predictions in post-processing. This post-processing subtracted the age-based error from each predicted age for each sample using the following equation:

Equation 1: Predicted Age Correction Calculation, used in chronological age dependency post-processing. Here, the error is according to the chronological age and was calculated with a linear regression.

$$Predicted\ Age_{Corrected} = Predicted\ Age_{Uncorrected} - Error(Chronological\ Age)$$

Corrected and uncorrected predictions were compared and assessed using the MAE, Std, and r-values produced from the respective brain-PADs.

2.8. Longitudinal Model Development

The change in patient data over time was assessed to identify the impact of increased disease pathology on the produced brain-PAD value. The Pipeline 1 age predictions were investigated as a longitudinal study, with the confounding factors included in a linear mixed-effects model (LMEM). This model should accurately predict brain-PAD for a given patient and time.

Participant-specific variables were developed for each sample including the brain-PAD and time since the primary scan was recorded. This time-based dataset was analysed for patterns across FTD and control participants. Linear regressions were developed for each participant describing the relationship between time since the primary scan and brain-PAD. The slopes and intercepts of the regressions were used to assess similarity across participants and groups.

LMEMs expand regression analysis to include random effects, alongside fixed effects. The samples collected for each patient were assessed using an LMEM mapping the relationship

between brain-PAD and the time since the first patient scan was recorded. This model considered the strength of the brain-PAD and time since primary scan relationship across subtypes, including diagnosis, sex, and chronological age. The impact of genetic mutation was analysed using a separate LMEM as most cases were sporadic. The FTD subtypes were considered relative to the control group. The binary groups were directly compared.

3.0. Results

3.1. Reorientation Impact

Reorienting the MRI images to correct mismatched coordinate systems had little impact on the brainageR outputs, as shown by the similar MAE, Std., r-value, and r^2-value shown below in Table 5. The impact of reorientation was assessed using 60 UCL dataset scans which were separately entered into the brainageR algorithm before and after reorientation.

Table 5: BrainageR age prediction accuracy with and without preprocessing reorientation applied to a 60-sample subset of the UCL dataset.

Processing Method	MAE	Std.	R value	R^2 Value
BrainageR	5.45	3.74	0.8983	0.8069
Reoriented Raw and BrainageR	5.40	3.71	0.8979	0.8061

3.2. Developing the Image Enhancement Pipeline with IXI Data

The significantly lowered MAE, lowered Std, and higher r-value and r^2 value shown in Table 6 demonstrate that Pipeline 3 significantly outperformed Pipeline 2. Therefore, the IXI development process validated the image enhancement pipeline (Pipeline 3). The pipeline development interim results can be seen in Appendix 3: Pipeline Development with IXI.

Table 6: BrainageR output results for Pipelines 2 and 3, applied to the IXI data, including MAE, Std., r value, and r^2 value.

Processing Method	Algorithms Applied	MAE	Std.	R Value	R^2 Value
Pipeline 2	Reoriented Raw, SynthSR and BrainAgeR	16.57	9.19	0.8414	0.6945
Pipeline 3	Reoriented Raw, SynthSR, Additional Preprocessing and BrainAgeR	6.88	4.84	0.8950	0.8009

3.3. Applying and Evaluating the Image Enhancement Pipeline

The control data in the UCL dataset was used to produce a data-specific Pipeline 3 using the steps outlined in Section 2.5 Image Enhancement Pipeline 3 which could be compared to Pipeline 1.

The threshold options for the UCL dataset, which can be seen in Appendix 5: Threshold Options, were found to vary with diagnosis and resolution, implying a dependence which may impact brainageR interpretation and results. All data show similar spike locations as in the complete IXI dataset, confirming the continued spike location consistency of SynthSR. As such, 60 control scans were randomly selected: 30 on 1.5T scans and 30 on 3.0T scans, described in detail in Appendix 2 Sample Size. The details of the wavelet analysis results can be seen in Appendix 4: Pipeline 3 Application to the UCL Dataset.

Pipelines 1, 2, and 3 were each applied to the entire UCL dataset, with the age prediction results shown in Table 7. Pipeline 2 is shown to perform significantly worse than Pipeline 1 or 3. Pipeline 1 outperforms Pipeline 3 with a slightly lower MAE and Std. Additionally, Pipeline 1 demonstrates a significantly better r-value and r^2-value than Pipeline 2 or 3.

Table 7: BrainageR output results for the UCL dataset processed with Pipelines 1, 2, and 3, including MAE, Std., and r values.

Pipeline Assessed	MAE	Std.	R Value	R^2 Value
Pipeline 1	6.18	4.34	0.8594	0.7386
Pipeline 2	22.76	8.14	0.7250	0.5256
Pipeline 3	6.44	5.06	0.7461	0.5566

Pipelines 1 and 3 were further assessed by subgroup to identify trends based on diagnostic and recording resolutions. These results show that Pipeline 1 outperformed Pipeline 3 in predicting the ages of control participant data recorded with both 1.5T and 3.0T. The Pipeline 1 data showed better prediction accuracy represented by the MAE, Std., r-value, and r^2 value. Pipeline 1 best predicts control data recorded at 1.5T, which likely results from this data sample being most similar to the T1-weighted healthy data used to train the algorithm. Control data recorded with 3.0T resolution produced very similar MAE in both pipelines, however Pipeline 1 resulted in a lower Std. and a higher r-value and r^2-value,

depicting a closer relationship between the actual participant ages and the brainageR predicted ages, shown in Table 8. Therefore, these results demonstrate that Pipeline 1 best predicts age in control data, minimizing the influence from different data collection methods. Pipeline 1, with the thresholds optimized using only control data, was therefore applied to both control and FTD data to best compare the brain-PADs produced.

Table 8: BrainageR output results for the UCL dataset processed with Pipelines 1, and 3, including MAE, Std., and r values. The brainageR age prediction errors were assessed for varying subsets of the data processed with Pipelines 1 and 3, comparing the impact of resolution and diagnosis on pipeline accuracy. Pipeline 3 includes resolution-specific thresholding of 28.9 to 122.4 for 1.5T and 25.2 to 123.3 for 3.0T. The bolded values highlight the pipeline performance on all control value, with Pipeline 1 outperforming Pipeline 3 in MAE, Std., and R value.

Pipeline Assessed	Test Sample Considered	MAE	Std.	R Value	R^2 Value
Pipeline 1	**1.5T**	6.64	5.18	0.6877	0.4723
	3.0T	7.84	5.47	0.6334	0.4012
	All Controls	**6.18**	**4.34**	**0.8594**	**0.7386**
	All FTD	7.85	5.71	0.4843	0.2345
	Controls with 1.5T	5.03	4.00	0.9117	0.8312
	Controls with 3.0T	6.96	4.40	0.8288	0.6868
	FTDs with 1.5T	7.31	5.47	0.5506	0.3031
	FTDs with 3.0T	8.23	5.85	0.44101	0.1945
Pipeline 3	**1.5T**	7.26	5.76	0.4952	0.2452
	3.0T	7.05	5.05	0.5685	0.3232
	All Controls	**6.44**	**5.06**	**0.7461**	**0.5566**
	All FTD	7.44	5.45	0.3817	0.1457
	Controls with 1.5T	5.63	4.79	0.7789	0.6067
	Controls with 3.0T	6.98	5.17	0.7402	0.5479
	FTDs with 1.5T	7.94	5.99	0.3593	0.1291
	FTDs with 3.0T	7.08	5.00	0.4157	0.1728

3.4. Cross Sectional Group Difference

The 1.5T and 3.0T control and FTD data were separately compared using the best pipeline output results to assess impact of recording scanner and protocol.

To determine the appropriate statistical approach for comparison, the data were determined to be non-parametric, through qualitative analysis of histograms for the relevant subgroups of the data. The histograms show that all data processed with Pipeline 1

has similar distribution shape even through the data have varying predicted age and actual age differences for each group.

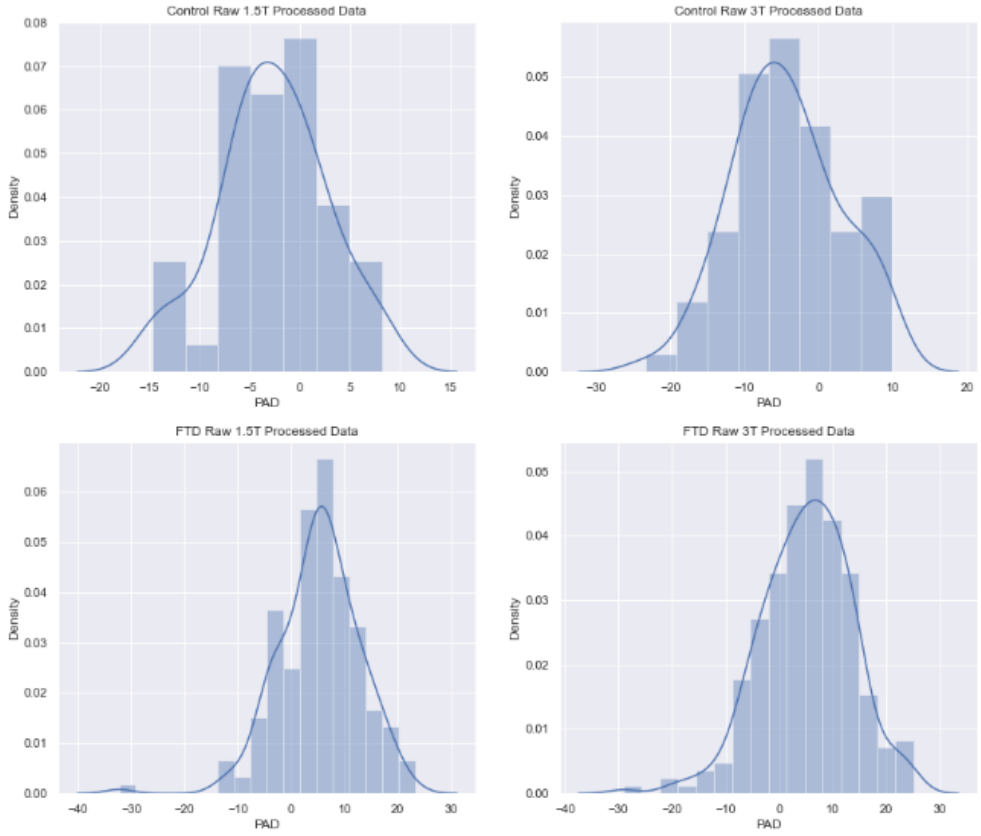

Figure 6: Histograms for resolution-specific control UCL data subsets were assessed for normality.

These results were further confirmed through quantitative assessment using the Shapiro Wilk normality test. The Pipeline 1 MAE for the resolution and diagnosis specific subtypes were each found to be non-parametric.

As both subsets to be compared are not non-parametric and independent, the MWU test was applied to resolution-specific subgroups to compare the diagnostic groups. These results compared Dataset 1 and Dataset 2, as labelled in Table 9 below to produce a statistical value and a p-value. The resulting statistic and p-values differed by large factors, implying that resolution has a large impact on the final statistical analysis. Additionally, the p-value for the FTD and Control data comparison was below 0.05 at both 1.5T and 3.0T,

defining the datasets as statistically dissimilar. Additionally, the two datasets were assessed and found to be statistically dissimilar without scanner resolution dependence. This aligns with the results shown in Table 8, which demonstrate a higher brain-PAD of 1.67 ± 1.37 years in FTD patients relative to the control participants.

Table 9: The UCL data processed with Pipeline 1 was compared for statistical similarity using the Mann-Whitney-U tests.

Dataset 1	Dataset 2	p Value
FTD at 1.5T	Control at 1.5T	1.2667e-11
FTD at 3.0T	Control at 3.0T	1.1050e-16
All FTD	All Control	8.8756e-27

3.5. Confounding Diagnostic, Genetic Mutation, and Chronological Age Factors

The FTD data were further investigated across specific diagnoses and the genetic mutations of the participants.

To quantify the diagnostic capability of Pipeline 1, the subset of scans for each FTD subtype were compared to see if the scan could be statistically differentiated by subtype. If successful, this could provide clinicians with more information of the disease and could further support diagnosis. The Pipeline 1 brain-PAD was assessed individually for each clinical subtype of FTD present in the dataset: bvFTD, PNFA, SD, and PPA-NOS. These results showed the largest brain-PAD by MAE for bvFTD, as seen in Table 10 below. The brain-PAD is slightly lower for PNFA and PPA-NOS. Finally, SD shows the lowest brain-PAD. SD and PPA-NOS have the lowest associated Std., implying more homogeneity across patients than in bvFTD and PNFA patient subgroups. The r-values and r^2-values describing the chronological and predicted ages for all disorder subtypes perform poorly throughout all diagnoses.

Table 10: The brainageR age prediction errors were assessed for each FTD diagnostic subtype.

Processing Method	MAE	Std.	R Value	R^2 Value	Data Length
bvFTD	8.94	5.89	0.4974	0.2474	431
PNFA	7.32	5.86	0.4289	0.1840	201
SD	6.10	4.84	0.5504	0.3030	202
PPA-NOS	7.58	4.50	0.4697	0.2206	43

The relationship between these subtype-specific brain-PADs were assessed for dissimilarity using the MWU statistical analysis, depicted in Table 11. Each subtype-specific dataset was compared against the rest of the FTD data with labelled subtypes. SD and bvFTD were found to be strongly statistically dissimilar from the rest of the FTD data, as shown by the very small p values. The difference between PPA-NOS and the rest of the data was found to have only a slight statistical dissimilarity and may be challenging to differentiate in practice. PNFA however, was not found to be statistically dissimilar from the rest of the dataset and was determined indistinguishable from the rest of the FTD data.

Table 11: The produced brainageR age prediction errors for each FTD diagnostic subtype processed with Pipeline 1 was assessed and compared with the rest of the FTD data using Mann-Whitney-U tests.

Dataset 1	Dataset 2	p Value
PPA-NOS	**Without PPA-NOS**	0.0455
SD	**Without SD**	3.2572E-07
PNFA	**Without PNFA**	0.0966
bvFTD	**Without bvFTD**	5.2692E-11

Similarly, the data with participant genetic mutation labels were individually assessed and considered. *C9orf72* and *MAPT* mutations were both found to have high MAE of 11.11 and 11.98 years in Table 12, relative to the average FTD MAE of 7.85 years shown in Table 8. In contrast, participants with the *GRN* mutation were found to have a lower MAE of 8.17 years, with a low Std. of 3.88 years relative to the average FTD Std. of 5.71 years.

Table 12: The brainageR age prediction errors were assessed for each FTD genetic mutation.

Processing Method	MAE	Std.	R Value	Data Length
C9orf72	11.11	6.37	0.5177	77
GRN	8.17	3.88	0.2193	44
MAPT	11.98	5.33	0.6231	87

To compare these relationships and assess statistical dissimilarity, a MWU statistically test was applied as shown in Table 13. This test compared each genetic mutation to the rest of the dataset and found the distribution of the brainageR output error for each genetic mutation to be statistically dissimilar from the rest of the dataset.

Table 13: The produced brainageR age prediction errors for each FTD genetic mutation processed with Pipeline 1 was assessed and compared with the rest of the FTD data using Mann-Whitney-U tests.

Dataset 1	Dataset 2	p Value
MAPT	**Without *MAPT***	0.0007
GRN	**Without *GRN***	1.2841E-07
C9orf72	**Without *C9orf72***	0.0434

To understand the difference between the three pipelines, the predicted ages were plotted against the actual participant age for the control data processed with each pipeline in Figure 7. The data which closely follows this line produces highly accurate age predictions, with less aligned data producing less accurate predictions. These results show that Pipeline 1 best predicts participants of different ages, having a consistent error rate throughout the participant data. Pipeline 2 consistently underpredicts participant ages, producing brainageR outputs which fall well below the actual participant age, as demonstrated in the data both below and to the right of the linear relationship line. The results of Pipeline 3 are shown to be age-dependant, with younger chronological ages showing a trend of overprediction and older ages showing a trend of underprediction.

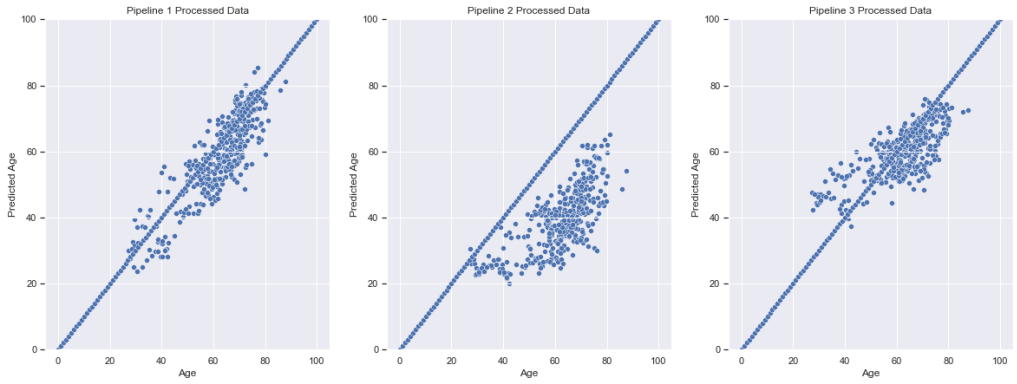

Figure 7: The relationship between the actual participant ages and predicted ages for all control UCL data was depicted for each pipeline, including the linear relationship representing perfect age prediction accuracy.

The initial relationships between chronological and predicted ages are shown to be inconsistent across pipelines in Figure 7. The relationship between the brain-PAD and chronological ages was investigated for chronological age-dependent error relationships using linear regressions for each pipeline, shown in Table 14. The conducted linear regressions show that the predicted age differences are dependent on the chronological age of the participant for all pipelines. Pipeline 1 showed the weakest relationship with a low r value while Pipeline 3 showed the strongest relationship with a large r-value, showing a large dependency on chronological age.

Table 14: The linear regression produced describing the actual participant ages and predicted ages for all control UCL data was produced for each pipeline.

Data	Slope	Intercept	r Value	p Value	Std. Err
Pipeline 1	-0.0919	1.8248	-0.1678	0.0010	0.0278
Pipeline 2	-0.4093	2.6069	-0.5893	1.01883E-36	0.0289
Pipeline 3	-0.5004	28.8353	-0.7466	1.5569E-68	0.0230

The relationship between the predicted age difference and chronological age and the associated linear regression were graphed and visually analysed in Figure 8. Even with this age dependency however, the MAE and Std. for each dataset did not differ very significantly between Pipelines 1 and 3.

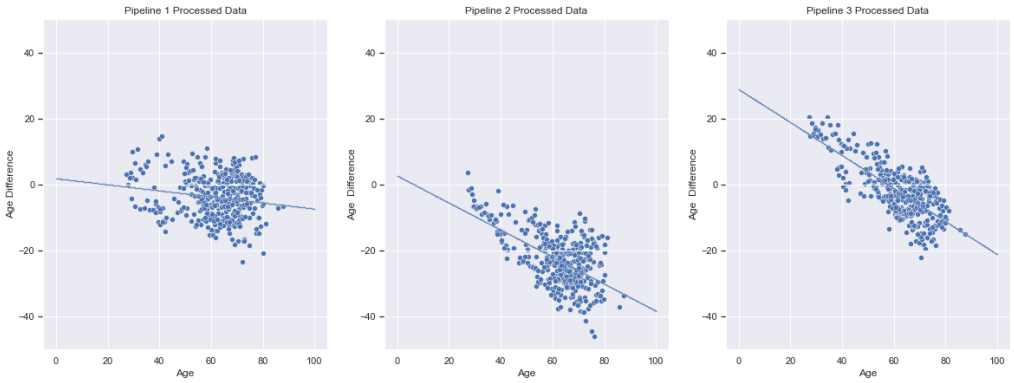

Figure 8: The relationship between the actual participant ages and predicted age error (PAD) for all control UCL data was depicted for each pipeline, including the linear regression of the relationship.

Post-processing techniques correcting Pipeline 3 to reduce chronological age-dependency were investigated. Pipeline 3 was selected for this investigation as this pipeline has the strongest dependency, and therefore should have the most impactful benefit from the post-processing if successful.

The absolute brain-PAD was used to produce MAE, Std, and r-values for the corrected and uncorrected data as discussed in Section 2.4. Pipeline Analysis Methods and shown in Table 15. The resulting values were quite similar, not significantly improving the age prediction accuracy in any categories.

Table 15: The Pipeline 3 control data was post-processed to correct for chronological age-dependency. The brainageR age prediction errors for both corrected and uncorrected data are depicted.

Processing Method	MAE	Std.	R Value
Uncorrected	6.44	5.06	0.7461
Corrected	6.92	4.90	0.7462

3.6. Longitudinal Analysis

To understand the change in brain-PAD over time, longitudinal Analysis was conducted using the multiple scans recorded for each patient. Of the 317 patients recorded with longitudinal scans, 68 participants were recorded using multiple scanners, with limited apparent impact on the brian-PAD as demonstrated by the lack of pattern in Figure 9 below.

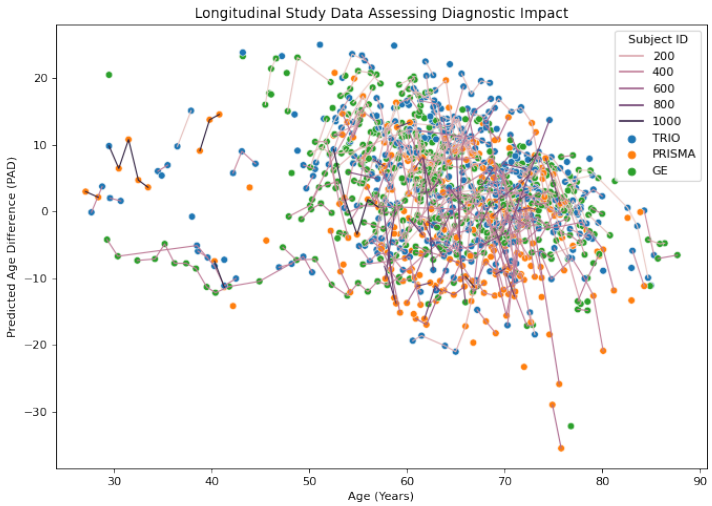

Figure 9: Longitudinal Assessment of recordings over time. Each recording datapoint is coloured according to type of MRI scanner, with the recordings for each patient connected by lines.

The brain-PAD over time was also graphed for each patient in Figure 10, with the data points identifying the diagnostic group. The slopes for each patient appear to be consistent across time. The intercepts of each patient, however, appear largely dependent on the control or FTD diagnosis.

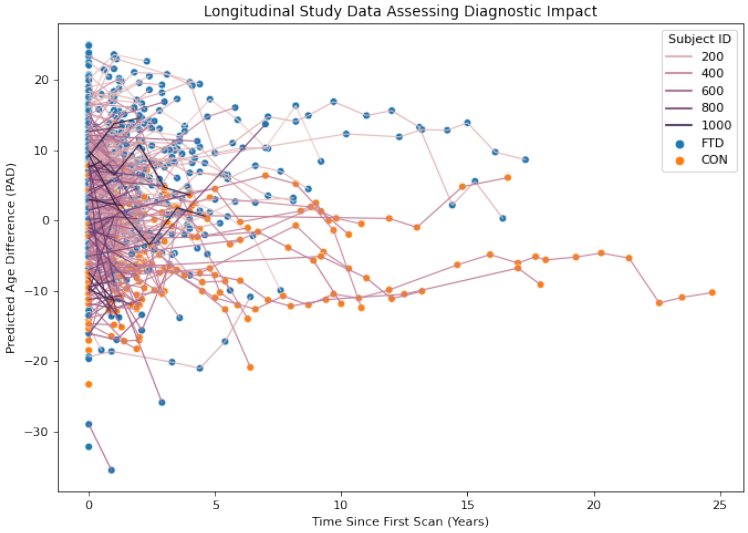

Figure 10: Longitudinal assessment of the control and FTD data for each patient is presented as the Time Since the First Scan and the associated brain-PAD. The data points are depicted as points and coloured according to group assignments of diseased (FTD) or control (CON). The lines each connect the data recordings for each subject, labelled as Subject ID in the legend and coloured in a gradient which darkens with a higher ID number.

The lines describing each patient in Figure 10 were separated by FTD subtype diagnostic groups and further assessed in Figure 11. All diagnostic groups present with data in the -10 to +10 brain-PAD range, however some groups such as bvFTD and PNFA additionally have much higher PAD values and likely contributed to the FTD high brain-PAD values trend in Figure 10. The bvFTD and control data show the longest recording timelines up to 20 years, potentially implying a longer lifespan associated with the diagnosis. The progression shown in bvFTD could also be due to its earlier age of onset, shown in Table 3 which provides an overview of the data characteristics.

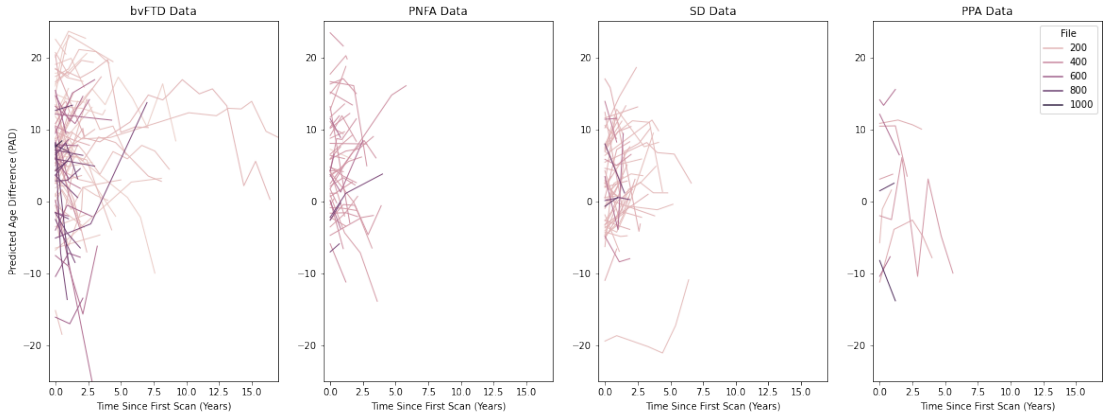

Figure 11: Line plots of each patient's brain-PAD and Time Since First Scan data, organized into diagnosis-specific subplots. The legend and y label both describe all data presented.

Confounding factors were graphically assessed in Figure 12 using the linear regression slopes describing the brain-PAD as a factor of time since the primary scan. The results depict similar medians and greater variance in the FTD slopes relative to the control slopes. These results are supported by the subtype-specific boxplot which demonstrates similar medians across each diagnostic but with a comparatively small control boxplot implying a higher level of agreement across the control data. Across the genetic mutations, the largest difference is seen with the comparatively large boxplot of the *GRN* genetic mutation, suggesting the mutation may impact prognosis through changes in atrophy.

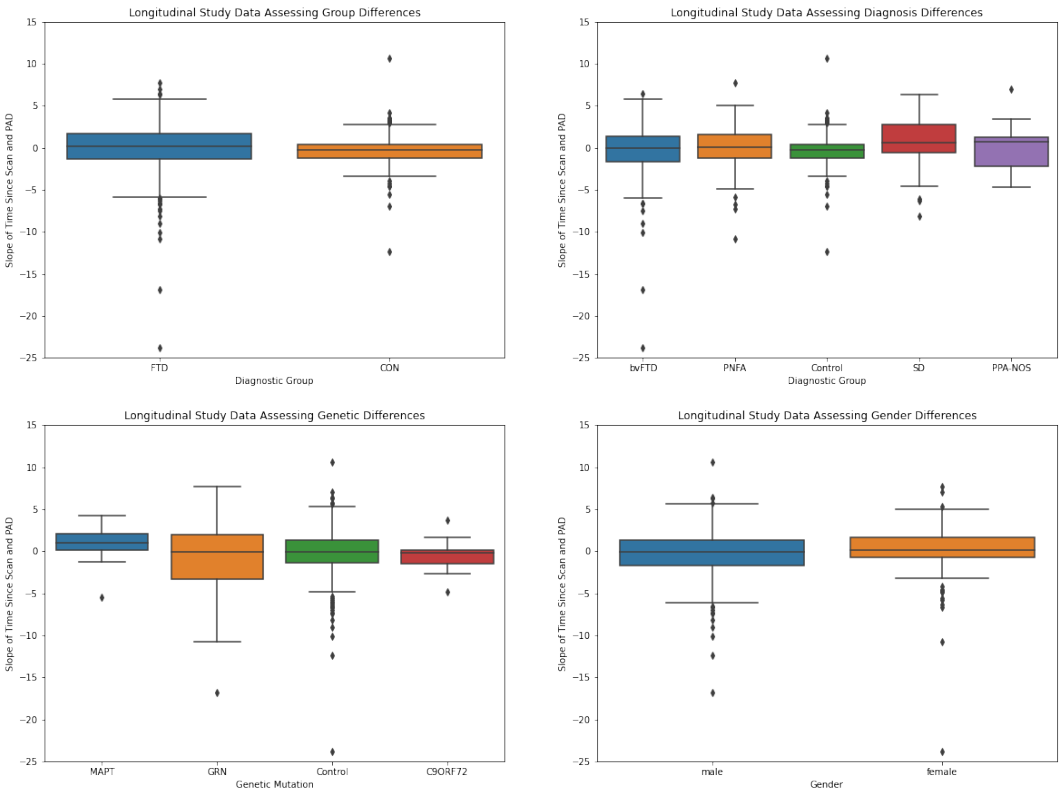

Figure 12: The linear regression slope describing the relationship between time and brain-PAD for each patient is depicted here for different subgroups to qualitatively depict patterns.

The linear regression intercepts associated with the above slopes better communicate different trends between the subgroups in Figure 13. The intercepts produced for each subject were shown to be higher in FTD than in control subjects, with bvFTD producing the highest intercepts. PPA was shown to produce a larger variance, with the shortest tails communicating large but consistent data variance. Additionally, sex was found to have a minor impact. The *C9orf72* genetic mutation also produced the highest intercept, however, had a similar median to the *MAPT* mutation data. Participants with the *GRN* mutation showed lower intercepts, however, demonstrated a larger variance, largely describing the lower intercept values.

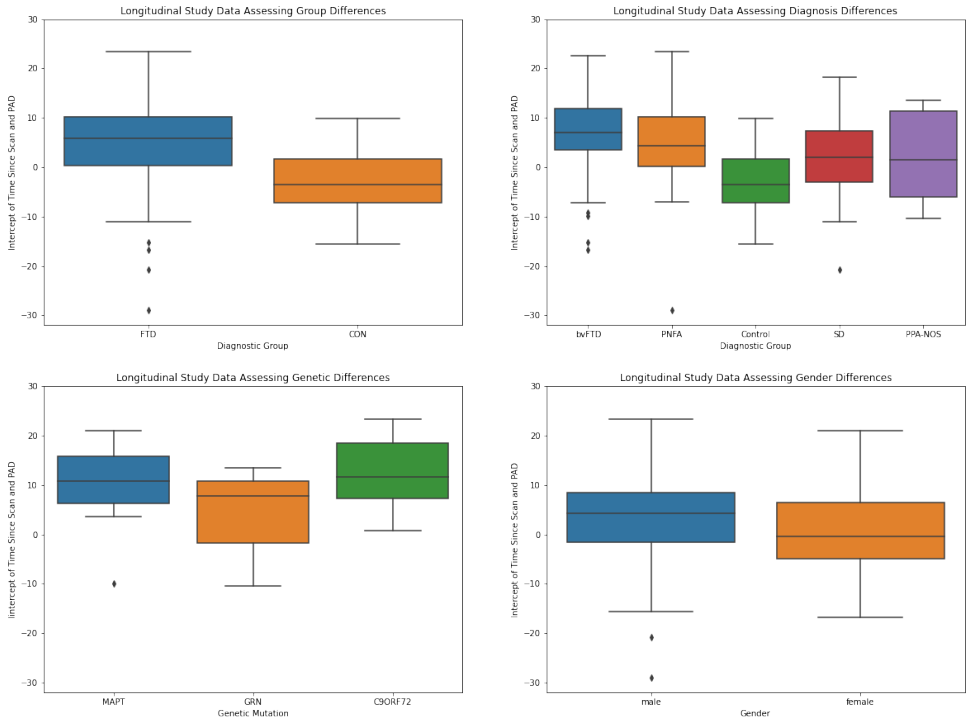

Figure 13: The linear regression intercept describing the relationship between time and brain-PAD for each patient is depicted here for different subgroups to qualitatively depict patterns.

The results from the longitudinal analysis were leveraged to produce a linear mixed effects model, shown in Table 16. Patients with a diagnosis of SD showed increased brain-age over time, demonstrated by the statistical relationship between time since the first scan and brain-PAD for each patient. Interestingly, no other statistically significant effects were shown between individual subgroups.

Table 16: The outputs of linear mixed effects models produced for each individual factor assessed describes the impact of each subgroup on the development of brain-PAD with time. The factors considered in analysis include diagnosis, mutation, sex, age, and group.

Factor Assessed	Subgroup	Coefficient	p Value
Group	Control	-0.101	0.111
	FTD	0.069	0.387
	PNFA	0.276	0.359
	PPA	-0.728	0.129
	SD	0.446	0.035
	bvFTD	0.013	0.881
Mutation	Sporadic	-0.086	0.131
	MAPT	0.244	0.072
	GRN	-0.604	0.346
	C9orf72	-0.023	0.880
Sex	Female	-0.017	0.797
	Male	-0.056	0.442
Chronological Age	-	-0.001	0.305

A linear mixed effects results which focused on the combined effect of the three primary factors, group (control or FTD label), sex, and chronological age, demonstrated more statistically impactful effects on the brain-PAD change over time. This combined analysis demonstrated in Table 17 showed much more impact from age, sex, and group than demonstrated in the individual subgroup assessments. Age, sex, and group were each determined to have a statistically significant effect on the brain-PAD over time. Age was shown to have the strongest impact on brain-PAD over time, with an increase in the chronological patient associated with a decrease in brain-PAD. This relationship is likely resulting from the chronological age dependency discussed in Section 4.2. Analysis of Chronological Age Dependency. Sex and group were both found to increase the brain-PAD score over time, with Male and FTD data producing higher brain-PAD scores than the respective Female and control data.

Table 17: Linear Mixed Effects Model produced using age, sex, and the group produced statistically significant results describing the change in brain-PAD over time.

Factor Assessed	Coefficient	p Value
Age	-0.012	0.015
Sex: Male	0.639	0.056
Sex: Female	0.718	0.031
Group	0.216	0.038

4.0. Discussion

The aim of this investigation was to test if image enhancement improved sensitivity to cross-sectional and longitudinal brain-age differences in FTD patients. These results were then leveraged to determine if brainageR is an appropriate biomarker to differentiate between healthy participants and patients with FTD diagnoses. As the cohort data included both 1.5T and 3.0T scans, various preprocessing methods were tested to optimize age-prediction accuracy on control data. It was hypothesized that the best-performing pipeline brain-PAD across control data would be statistically dissimilar from the brain-PAD produced for FTD data. Secondly, it was hypothesized that Pipeline 3 would most accurately predict control ages as SynthSR preprocessing is designed to normalize data across scan resolutions and the threshold was optimized for each dataset.

Using the difference between predicted and chronological age as a measure of validity, the error in age prediction for healthy subjects was compared across pipelines. Pipeline 1, without SynthSR preprocessing, most accurately predicted healthy participant age. Additionally, Pipeline 1 brainageR predictions for control and FTD data were statistically dissimilar, validating brainageR as a potential biomarker for FTD.

4.1. Analysis of Pipeline Comparison Results

Pipeline 1 best predicted the ages of healthy participants, with minor performance quality differences in the data collected using different scanners and acquisition parameters.

Graphical analysis of control data age predictions highlighted chronological age-dependent errors in all pipelines. Pipeline 2, which included SynthSR preprocessing, underpredicted the participant ages producing significant inaccuracy. Pipeline 3 improved these results, predicting ages with similar accuracy to Pipeline 1. Pipeline 3 however, was found to be

highly dependent on the chronological age of the participants. Pipeline 1 had a weak relationship between chronological age and brain-PAD, with an r-value describing of -0.17. This agrees with previous findings, which showed a significant relationship (r-value = -0.379) between chronological participant age and predicted ages when brainageR was applied to the Cam-CAN dataset (Cole, 2022). The difference in UCL and Cam-CAN age-dependency strength, however, may result from varying scanner resolutions or from unknown differences in the data. SynthSR preprocessing caused an increased chronological age dependency for the predicted ages, indicated by the change in r-value of -0.17 in Pipeline 1 to -0.59 in Pipeline 2. Pipeline 3 had the strongest dependency, an r-value of -0.75, implying additional processing worsened age dependency while improving overall accuracy. This large chronological age dependence likely accounts for the lower r-value of 0.7461 produced in Pipeline 3, depicting a weaker relationship between chronological and predicted age than in Pipeline 1 and 2 (respective r-values of 0.8594 and 0.7250). This difference could result from normalization of the data, which may have removed important elements of the signal input to brainageR, increasing the effect of the chronological age-based errors. In future investigations, this could be potentially avoided by removing the age-based dependency from brainageR under Pipeline 1 conditions.

The translation of Pipeline 3 development using the IXI dataset to the UCL dataset was shown to be effective, producing similar pipeline accuracy. Compared to the UCL dataset, Pipeline 3 applied to the IXI dataset produced a better r-value and similar MAE and Std. values. These lower r-value and higher Std. from the UCL dataset Pipeline 3 analysis produce a higher number of outliers, aligning with the actual ages with less precision. A larger discrepancy between the Pipeline 3 results was expected as IXI data was recorded with a magnetic field strength of 1T, which is expected by brainageR. In contrast, the UCL dataset

has varying resolutions and was therefore not anticipated to meet brainageR expectations. Future investigations should consider adding more preprocessing steps to Pipeline 3 to reduce the outliers present and further improve the brainageR predictions.

The UCL dataset includes data of various resolutions collected from different scanning technologies. Spatial resolution is a major obstacle in translating clinically recorded data to research investigations. As SynthSR preprocessing enables this translation through normalization, Pipeline 3 was expected to better predict ages after the normalization of down-sampled data. As the acquisition parameters with the largest spatial resolution, GE Signa in Table 4, also has a magnetic field strength of 1.5T, impact of the scanner or spatial resolution on the age prediction performance cannot be individually assessed. In future investigations, further down-sampling could be applied to the data so that the impact of down-sampling the data can be separately identified, and the impact of scan and spatial resolution can be separately quantified. This further understanding of each impact would enable preprocessing to be tailored to the exact data considered.

The threshold investigations for Pipeline 3 produced unexpected results when extrapolating results from the 60-sample dataset to the entire dataset. The optimal thresholds produced for the 60-sample dataset showed that for the maximum lower threshold and mean upper threshold options produced the most accurate brainageR predictions. The maximum lower threshold option and mean upper threshold options for both the 60-sample dataset and the entire dataset were tested on the full dataset and compared using the age predictions. The results of this test were expected to show that the threshold options from the entire dataset would produce optimal results for both resolution groups. Instead, the 1.5T control dataset performed best with the subset threshold options while the 3.0T control dataset

performed best with the entire dataset threshold options. This finding implies that the optimal threshold for the dataset was likely not identified properly with the subset of data, and a more sophisticated wavelet analysis should be used in the future.

4.2. Analysis of Chronological Age Dependency Investigation

To counteract chronological age dependency in Pipeline 3, post-processing was conducted which applied a correction to predicted ages to improve accuracy. This correction leveraged the linear regression model defining the relationship between chronological age and brain-PAD for healthy participants, to create an age-based error. These results removed age-dependency from the results yet did not significantly impact the resulting pipeline accuracy in MAE or Std. Additionally, Pipeline 1 continued to outperform the corrected Pipeline 3, implying that intervention must occur earlier in the process to increase overall age prediction accuracy. The corrected Pipeline 3 outputs, however, better predicted younger ages for healthy participants, providing a better algorithm for removing recording-based factors in younger patients. This improved algorithm could potentially better model participants assessed for early onset FTD, more accurately estimating brain-PAD associated with the patient state. For healthy participants, this would closely match the chronological age, whereas an older age would be estimated for FTD participants.

FTD subjects were expected to produce older brain age estimations, and therefore larger brain-PAD values, as the FTD pathology causes atrophy in the brain reminiscent of ageing. This hypothesis was confirmed, with the linear regression intercepts shown to be significantly higher for the FTD diagnostic group compared to the controls.

4.3. Analysis of Group Differences, Contributing Factors and Longitudinal Analysis

The analysis identified differences in the diagnostic groups, the FTD clinical subtypes, the genetic mutations, and the gender of the participants.

The linear regression describing the relationship between brain-PAD and time since first scan for each patient showed clear dependency on the diagnostic group. Although the slopes for the change in brain-PAD over time showed limited change across patients, the intercepts were shown to be significantly higher for patients diagnosed with FTD. This is likely due to the atrophy which has occurred in the brain of the patient, resulting in a higher age estimation than their chronological age and increasing the brain-PAD produced from all MRI scans recorded for each FTD participant.

The FTD clinical subtype investigation showed bvFTD to have the largest brain-PAD, with the earliest age of onset and longest disease duration. The analysis demonstrated that bvFTD and PFNA showed the greatest heterogeneity across the subgroups. For bvFTD, this aligns with research as bvFTD has been identified as very heterogeneous in terms of age at onset, duration, symptoms, and pathology (Miller and Llibre Guerra, 2019).

Participants with genetic FTD, caused by *GRN*, *MAPT*, or *C9orf72* mutations, were found to produce much higher brain-PAD values than the sporadic FTD cases. *C9orf72* was found to have the largest associated brain-PAD with the largest standard deviation, which is appropriate for the heterogeneity generally found in *C9orf72* patients (E. Devenney *et al.*, 2015). Similarly, patients with *GRN* mutations often present with homogeneous atrophy, appropriate for the small standard deviation associated with the *GRN* brain-PAD values. Patients presenting with *MAPT* mutations have similar homogeneity to patients with *GRN* mutations and would therefore expect a small brain-PAD standard deviation to reflect the

common temporal lobe atrophy associated with *MAPT*. The *MAPT* mutation brain-PAD MAE and Std. produced however, are large and closely resembles the values found in patients with *C9orf72* mutations, implying heterogeneity inconsistent with research. Future investigations should investigate the exact atrophy occurring in the *MAPT* mutations in the UCL dataset, assessing changes in regional atrophy and tissue volumes for each patient.

The linear regression slopes describing the change in brain-PAD over time for patients with *GRN* mutations were found to be relatively variable compared to the other mutations. This large variance could be impacted by heterogeneous atrophy in patients with *GRN* mutations. Interestingly, the linear regression intercept associated with *GRN* mutations are generally lower than those associated with *MAPT* or *C9orf72* mutations, implying that the atrophy in *GRN* brains begins closer to the actual patient age and the age estimation error quickly grows. This finding aligns with research, as genetic FTD with *GRN* mutations is the most rapidly progressing.

Age predictions were also found to be statistically dissimilar when compared across genetic mutations and FTD subtype diagnosis. Longitudinal analysis of the data recorded for each patient over time did not show statistically significant changes in predicted age over disease development. The identified contributing factors were found to heavily influence the intercept of the resulting age predictions but did not statistically impact the change in brain-PAD over time.

4.4. Limitations and Future Investigations

This investigation considered only one image quality parameter in quantifying image noise: histogram abnormalities. Including other important image assessments could have improved the pipeline. Two assessments that could be investigated in future analyses are

image integrity and image quality, quantified using Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). The PSNR determines the amount of data in a sample which is useful and conveys meaningful information for analysis relative to the amount of noise in the sample. Similarly, the SSIM assesses an image relative to an undistorted or original version of the same image. Expanding the image quality analysis and applying additional improvements to Pipeline 3 would further reduce the MAE and enable age predictions with higher accuracy and precision.

Pipeline 2 used a generalized SynthSR model which was designed to accept all scanner and spatial resolutions. This model in combination with brainageR did not perform well on healthy data, producing an MAE of 22.76 for the UCL dataset and 16.57 for the IXI dataset. In future investigations, it would be interesting to explore the application of a more precise, resolution-specific SynthSR model to the resolution-specific datasets to see if outcomes have improved accuracy. Additionally, other image enhancement methods compatible with MRI, such as IQT and iRAD, could also be evaluated for accuracy and compared to the SynthSR-based Pipelines 2 and 3.

This investigation only considered the brain age paradigm, ignoring other biomarkers which can be produced from MRI data. Future investigations could focus on quantifying and tracking changes including whole brain atrophy, region-specific atrophy, tissue-specific atrophy, or cortical thickness. For example, in addition to brain ages brainageR generates white matter, grey matter, and cerebrospinal fluid tissue volume approximations which could be leveraged in an investigation. These tissue volume assessments could be used individually or in combination with brain age estimations to improve to the diagnosis and prognosis of FTD.

This current investigation did not include FTD diagnoses with motor symptoms. Future FTD investigations conducted using brainageR should aim to include data from participants on the FTD-ALS spectrum. As FTD-ALS spectrum patients often have a *C9orf72* genetic mutation and there is significant symptom overlap across all FTD subtypes, including FTD-ALS data in analysis is a vital step towards assessing clinical applicability of brainageR. As FTD-ALS is associated with a significantly shorter lifespan than other FTD subtypes, accurately recognizing and quickly communicating this diagnosis is particularly important (Hodges *et al.*, 2003).

These results confirm the statistical dissimilarity of Pipeline 1 predicted ages for healthy and FTD data. This proves a difference between healthy and FTD data which is recognized by brainageR. To qualify the algorithm as a potential biomarker in FTD, prediction, diagnosis, and prognosis accuracy would need to be assessed, statistically correlating the brain-PAD with clinical performance. The next step to investigate the feasibility of this biomarker on an individual level is to explore the odds of a sample being correctly identified as a healthy sample or FTD sample. One method to apply these findings as individual diagnostic tools is to train a supervised machine learning model which would classify the data as healthy or FTD. This model would take the difference between the brainageR age predictions and the chronological age as the input, with the data labelled with the diagnosis of healthy or FTD. If highly accurate, a model such as the one described here would be used in tandem with Pipeline 1 to help diagnose FTD using 1.5T or 3.0T MRI data recorded at 1mm.

Any future models could be developed using the insight gathered from the longitudinal analysis of the data. Both the genetic groups and clinical subgroups of FTD were found to impact the resulting age prediction accuracy. Each genetic and clinical diagnostic produced a

unique strength in the relationship between the time since scan and brain-PAD, which could be expanded to predict disease development over time. Future investigations should further evaluate the impact of each subgroup, evaluating these findings on larger datasets. These investigations could identify new patterns occurring in the data, informing how changes in volumetric atrophy influence the resulting age predictions. For example, the low time and brain-PAD linear regression intercepts associated with the *GRN* genetic mutation highlight a low age prediction error rate relative to the *MAPT* and *C9orf72* mutations. This relationship was unexpected, as *GRN* mutations are generally associated with a fast rate of whole brain atrophy (Gossye, Van Broeckhoven and Engelborghs, 2019). This relationship could result from confounding variables, such as disease duration, and would require further investigation. To gain more understanding of how this theory relates to empirical data, these relationships should assess the impact of *GRN* on the CSF, White Matter, and Grey Matter tissue volumes produced by brainageR and assess the impact on age predictions.

As this Pipeline is impacted by chronological age throughout the pipelines, it works disproportionately poorly with older or younger participants. As FTD is the second most common young onset dementia, appropriately accommodating younger diagnostic data and removing age-dependency from the model is very important. Early diagnoses of Early Onset FTD are important for the treatment and management of the disorder and have the largest impact early in the disease. It is therefore vital to extend the model to more accurately predict ages for young ages which would be impacted by early onset to improve diagnostic feasibility.

The success of Pipeline 1 brainageR age prediction for 1.5T and 3.0T data enables future investigations to also leverage 1.5T and 3.0T-weighted data. This is essential for future

investigations into the feasibility of a brainageR-FTD biomarker as combining data of various resolutions into large datasets is expected to produce more insights which will better extrapolate to cross-sectional group differences.

5.0. Conclusion

The analysis undertaken as part of this study identified a preprocessing pipeline for MRI scans using brainageR to differentiate between longitudinal healthy and FTD MRI data.

Pipeline 1, which assesses raw data of various scanner and spatial resolutions, without SynthSR or additional preprocessing, was found to best predict healthy control participant ages. The longitudinal study revealed that brain-PAD is statistically impacted by overall diagnosis, genetic mutation and FTD subtype, but not time.

The confirmation of this pipeline validates the use of brainageR for larger datasets, combining data collected by different institutions recorded using 1T, 1.5T and 3.0T MRI scanner resolutions at 1 mm, 1.1 mm, 1.2 mm, and 1.5 mm slice thicknesses. These results enable the use of large datasets with similar scanners and acquisition parameters, such as the 3T scans from various manufacturers in the Genetic Frontotemporal Initiative dataset, in future investigations.

When applied to a joint dataset including healthy control data and the data of participants diagnosed with FTD, the diagnosis-dependent subgroups were found to be statistically dissimilar. This suggests the need to further explore diagnostic accuracy in predicting FTD subtypes using brainageR.

These results provide initial validation of brainageR age predictions to differentiate between control and FTD data. The findings support future investigations to further assess the efficacy of brainageR in identifying FTD subtypes from MRI scans. The statistical dissimilarity of control and FTD groups encourages the further development of brainageR as a diagnostic aid to label an MRI scan as diseased or healthy. The next steps toward an automated system

could include the development of a supervised machine learning algorithm which takes the difference between real and predicted ages as an input and the diagnosis as the label.

This project validated brainageR as a tool for differentiating between healthy and FTD data and identified statistically impactful factors which can be further evaluated in future investigations.

6.0. Bibliography

Alexander, D.C. *et al.* (2017) 'Image quality transfer and applications in diffusion MRI', *NeuroImage*, 152, pp. 283–298. Available at: https://doi.org/10.1016/j.neuroimage.2017.02.089.

Balchandani, P. and Naidich, T.P. (2015) 'Ultra-High-Field MR Neuroimaging', *AJNR: American Journal of Neuroradiology*, 36(7), pp. 1204–1215. Available at: https://doi.org/10.3174/ajnr.A4180.

Basodi, S. *et al.* (2021) 'Federation of Brain Age Estimation in Structural Neuroimaging Data', in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3854–3857. Available at: https://doi.org/10.1109/EMBC46164.2021.9629865.

van Beek, E.J.R. *et al.* (2019) 'Value of MRI in Medicine: More Than Just Another Test?', *Journal of magnetic resonance imaging : JMRI*, 49(7), pp. e14–e25. Available at: https://doi.org/10.1002/jmri.26211.

Beheshti, I. *et al.* (2019) 'T1-weighted MRI-driven Brain Age Estimation in Alzheimer's Disease and Parkinson's Disease', *Aging and Disease*, 11(3), pp. 618–628. Available at: https://doi.org/10.14336/AD.2019.0617.

Bhonsle, D., Chandra, V. and Sinha, G.R. (2012) 'Medical Image Denoising Using Bilateral Filter', *International Journal of Image, Graphics and Signal Processing*, 4(6), pp. 36–43. Available at: https://doi.org/10.5815/ijigsp.2012.06.06.

Buchanan, C.R. *et al.* (2021) 'Comparison of structural MRI brain measures between 1.5T and 3T: data from the Lothian Birth Cohort 1936'. medRxiv, p. 2021.04.23.21256000. Available at: https://doi.org/10.1101/2021.04.23.21256000.

Che, X.-Q. *et al.* (2018) 'Precision medicine of frontotemporal dementia from genotype to phenotype', *Frontiers in Bioscience*, 23(3), pp. 1144–1165. Available at: https://doi.org/10.2741/4637.

Cole, J. (2022) 'brainageR'. Available at: https://github.com/james-cole/brainageR (Accessed: 4 August 2022).

Cruts, M. *et al.* (2006) 'Null mutations in progranulin cause ubiquitin-positive frontotemporal dementia linked to chromosome 17q21', *Nature*, 442(7105), pp. 920–924. Available at: https://doi.org/10.1038/nature05017.

Devenney, E. *et al.* (2015) 'Clinical heterogeneity of the *C9orf72* genetic mutation in frontotemporal dementia', *Neurocase*, 21(4), pp. 535–541. Available at: https://doi.org/10.1080/13554794.2014.951058.

Devenney, Emma *et al.* (2015) 'Motor neuron disease-frontotemporal dementia: a clinical continuum', *Expert Review of Neurotherapeutics*, 15(5), pp. 509–522. Available at: https://doi.org/10.1586/14737175.2015.1034108.

Erkkinen, M.G., Kim, M.-O. and Geschwind, M.D. (2018) 'Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases', *Cold Spring Harbor Perspectives in Biology*, 10(4), p. a033118. Available at: https://doi.org/10.1101/cshperspect.a033118.

Espay, A.J. and Litvan, I. (2011) 'Parkinsonism and Frontotemporal Dementia: The Clinical Overlap', *Journal of Molecular Neuroscience*, 45(3), pp. 343–349. Available at: https://doi.org/10.1007/s12031-011-9632-1.

Ferrari, R. *et al.* (2014) 'Frontotemporal dementia and its subtypes: a genome-wide association study', *Lancet neurology*, 13(7), pp. 686–699. Available at: https://doi.org/10.1016/S1474-4422(14)70065-1.

Franke, K. *et al.* (2010) 'Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters', *NeuroImage*, 50(3), pp. 883–892. Available at: https://doi.org/10.1016/j.neuroimage.2010.01.005.

Friston, K.J. (ed.) (2007) *Statistical parametric mapping: the analysis of funtional brain images*. 1st ed. Amsterdam ; Boston: Elsevier/Academic Press.

FSLeyes — FSLeyes 1.4.6 documentation (no date). Available at: https://open.win.ox.ac.uk/pages/fsl/fsleyes/fsleyes/userdoc/ (Accessed: 4 August 2022).

Gijselinck, I. *et al.* (2015) 'Loss of TBK1 is a frequent cause of frontotemporal dementia in a Belgian cohort', *Neurology*, 85(24), pp. 2116–2125. Available at: https://doi.org/10.1212/WNL.0000000000002220.

Gorno-Tempini, M.L. *et al.* (2011) 'Classification of primary progressive aphasia and its variants', *Neurology*, 76(11), pp. 1006–1014. Available at: https://doi.org/10.1212/WNL.0b013e31821103e6.

Gossye, H., Van Broeckhoven, C. and Engelborghs, S. (2019) 'The Use of Biomarkers and Genetic Screening to Diagnose Frontotemporal Dementia: Evidence and Clinical Implications', *Frontiers in Neuroscience*, 13, p. 757. Available at: https://doi.org/10.3389/fnins.2019.00757.

Harris, J.M. *et al.* (2013a) 'Classification and pathology of primary progressive aphasia', *Neurology*, 81(21), pp. 1832–1839. Available at: https://doi.org/10.1212/01.wnl.0000436070.28137.7b.

Harris, J.M. *et al.* (2013b) 'Sensitivity and specificity of FTDC criteria for behavioral variant frontotemporal dementia', *Neurology*, 80(20), pp. 1881–1887. Available at: https://doi.org/10.1212/WNL.0b013e318292a342.

Hodges, J.R. *et al.* (2003) 'Survival in frontotemporal dementia', *Neurology*, 61(3), pp. 349–354. Available at: https://doi.org/10.1212/01.WNL.0000078928.20107.52.

Hornberger, M. *et al.* (2010) 'Orbitofrontal dysfunction discriminates behavioral variant frontotemporal dementia from Alzheimer's disease', *Dementia and Geriatric Cognitive Disorders*, 30(6), pp. 547–552. Available at: https://doi.org/10.1159/000321670.

Iglesias, J.E. *et al.* (2021) 'Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast', *NeuroImage*, 237, p. 118206. Available at: https://doi.org/10.1016/j.neuroimage.2021.118206.

iRAD 510(k) Premarket Notification (2021) *USA Food and Drug Administration*. Available at: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K212470 (Accessed: 9 August 2022).

'IXI Dataset – Brain Development' (no date). Available at: https://brain-development.org/ixi-dataset/ (Accessed: 4 August 2022).

Knopman, D.S. and Roberts, R.O. (2011) 'Estimating the Number of Persons with Frontotemporal Lobar Degeneration in the US Population', *Journal of molecular neuroscience : MN*, 45(3), pp. 330–335. Available at: https://doi.org/10.1007/s12031-011-9538-y.

Laird, A.R. (2021) 'Large, open datasets for human connectomics research: Considerations for reproducible and responsible data use', *NeuroImage*, 244, p. 118579. Available at: https://doi.org/10.1016/j.neuroimage.2021.118579.

Miller, B. and Llibre Guerra, J.J. (2019) 'Chapter 3 - Frontotemporal dementia', in V.I. Reus and D. Lindqvist (eds) *Handbook of Clinical Neurology*. Elsevier (Psychopharmacology of Neurologic Disease), pp. 33–45. Available at: https://doi.org/10.1016/B978-0-444-64012-3.00003-4.

Moore, K.M. *et al.* (2020) 'Age at symptom onset and death and disease duration in genetic frontotemporal dementia: an international retrospective cohort study', *The Lancet Neurology*, 19(2), pp. 145–156. Available at: https://doi.org/10.1016/S1474-4422(19)30394-1.

Morovic, S. *et al.* (2019) 'Possibilities of Dementia Prevention - It is Never Too Early to Start', *Journal of Medicine and Life*, 12(4), pp. 332–337. Available at: https://doi.org/10.25122/jml-2019-0088.

Obusez, E.C. *et al.* (2018) '7T MR of intracranial pathology: Preliminary observations and comparisons to 3T and 1.5T', *NeuroImage*, 168, pp. 459–476. Available at: https://doi.org/10.1016/j.neuroimage.2016.11.030.

Onyike, C.U. and Diehl-Schmid, J. (2013) 'The epidemiology of frontotemporal dementia', *International Review of Psychiatry*, 25(2), pp. 130–137. Available at: https://doi.org/10.3109/09540261.2013.776523.

Park, S.H. and Han, K. (2018) 'Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and

Prediction', *Radiology*, 286(3), pp. 800–809. Available at: https://doi.org/10.1148/radiol.2017171920.

Peelle, J.E. *et al.* (2008) 'Sentence comprehension and voxel-based morphometry in progressive nonfluent aphasia, semantic dementia, and nonaphasic frontotemporal dementia', *Journal of neurolinguistics*, 21(5), pp. 418–432. Available at: https://doi.org/10.1016/j.jneuroling.2008.01.004.

Rascovsky, K. *et al.* (2011) 'Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia', *Brain*, 134(9), pp. 2456–2477. Available at: https://doi.org/10.1093/brain/awr179.

Ratnavalli, E. *et al.* (2002) 'The prevalence of frontotemporal dementia', *Neurology*, 58(11), pp. 1615–1621. Available at: https://doi.org/10.1212/WNL.58.11.1615.

Rohrer, J.D. *et al.* (2015) 'Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Frontotemporal dementia Initiative (GENFI) study: a cross-sectional analysis', *The Lancet Neurology*, 14(3), pp. 253–262. Available at: https://doi.org/10.1016/S1474-4422(14)70324-2.

Rosen, H.J. *et al.* (2002) 'Patterns of brain atrophy in frontotemporal dementia and semantic dementia', *Neurology*, 58(2), pp. 198–208. Available at: https://doi.org/10.1212/WNL.58.2.198.

Sivasathiaseelan, H. *et al.* (2019) 'Frontotemporal Dementia: A Clinical Review', *Seminars in Neurology*, 39(2), pp. 251–263. Available at: https://doi.org/10.1055/s-0039-1683379.

Stonnington, C.M. *et al.* (2008) 'Interpreting scan data acquired from multiple scanners: A study with Alzheimer's disease', *Neuroimage*, 39(3), pp. 1180–1185. Available at: https://doi.org/10.1016/j.neuroimage.2007.09.066.

Sujitha, R. *et al.* (2017) 'WAVELET BASED THRESHOLDING FOR IMAGE DENOISING IN MRI IMAGE', 12(1), p. 10.

Swift, I.J. *et al.* (2021) 'Fluid biomarkers in frontotemporal dementia: past, present and future', *Journal of Neurology, Neurosurgery, and Psychiatry*, 92(2), pp. 204–215. Available at: https://doi.org/10.1136/jnnp-2020-323520.

Van Mossevelde, S. *et al.* (2016) 'Clinical features of TBK1 carriers compared with C9orf72, GRN and non-mutation carriers in a Belgian cohort', *Brain*, 139(2), pp. 452–467. Available at: https://doi.org/10.1093/brain/awv358.

Veldsman, M. and Egorova, N. (2017) 'Advances in Neuroimaging for Neurodegenerative Disease', *Advances in Neurobiology*, 15(Neurodegenerative Diseases), p. 28.

Virtanen, P. *et al.* (2020) 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nature Methods*, 17(3), pp. 261–272. Available at: https://doi.org/10.1038/s41592-019-0686-2.

Younes, K. and Miller, B.L. (2020) ‘Frontotemporal Dementia: Neuropathology, Genetics, Neuroimaging, and Treatments’, *Psychiatric Clinics of North America*, 43(2), pp. 331–344. Available at: https://doi.org/10.1016/j.psc.2020.02.006.

Yousaf, T., Dervenoulas, G. and Politis, M. (2018) ‘Advances in MRI Methodology’, in *International Review of Neurobiology*. Elsevier, pp. 31–76. Available at: https://doi.org/10.1016/bs.irn.2018.08.008.

van der Zee, J. *et al.* (2013) ‘A Pan-European Study of the C9orf72 Repeat Associated with FTLD: Geographic Prevalence, Genomic Instability, and Intermediate Repeats’, *Human Mutation*, 34(2), pp. 363–373. Available at: https://doi.org/10.1002/humu.22244.

Appendix 1: Pipeline 3 UCL Dataset Specifications

Pipeline 3 for the UCL Dataset was developed as described in Section 2.5. Image Enhancement Pipeline 3.

As predicted ages are only assumed to represent the chronological ages of the participants in healthy participants, Pipeline 3 was developed using only the control UCL data.

As the UCL dataset includes MRI data of varying resolution and acquisition parameters, the dataset-specific optimal preprocessing steps in Pipeline 3 were determined from a randomly selected subset where half the subset had been recorded with each scanner resolution. The scanner and spatial resolutions can be separated into two categories: the 1.5T Signa at 1.5mm thickness; and 3.0T Trio and 3.0T Prisma at 1.1mm thickness. Although both scanner and spatial resolution potentially contribute to subset differences, for simplification these categories are referred to solely by their magnetic field strength of 1.5T and 3.0T. Separately analysing each resolution-specific subset removed recording-based confounding factors and identified recording-specific group threshold differences. The same process described in 2.5. Image Enhancement Pipeline 3 was then repeated with two resolution-specific subsets of the control participant data in the UCL dataset, discussed in Appendix 2 Sample Size.

The threshold for the UCL dataset was determined using the same methods as discussed above for the IXI dataset: histograms of a subset of the data were leveraged to determine threshold options describing the dataset; these threshold options were tested with the best performing preprocessing extrapolated to the complete dataset. To determine the appropriate threshold values for this dataset, 60-scans of the control data with 30-scans of each resolution were used to optimize the brainageR prediction accuracy relative to the actual brain ages.

The optimal threshold values produced for the control values were applied to the entire dataset. As gold standard brain ages exist only for control participants, accuracy feedback was only available for the control data. Additionally, it was assumed that all data were similarly collected for each resolution, and therefore that experimentally caused noise is only validly determined in the control-based thresholds. Furthermore, it can be assumed that the recording noise in the FTD data is similar to the recording noise in the control data and that any other patterns in the data are produced by atrophy.

Similarity across the various datasets organized by resolution and diagnosis was assessed by applying the same spike and threshold wavelet analysis to compare sample features. The thresholds were applied to the resolution-specific datasets, and the thresholds most accurately predicting control subject brain ages were used in Pipeline 3 for the final pipeline comparison.

The best performing pipeline was then determined using the MAE, Std. and r-values, calculated from the brainageR results for each pipeline using the methodology discussed in Section 2.4. Pipeline Analysis Methods.

Appendix 2: Sample Size

Methods

Thirty randomly selected scans of the IXI dataset were used to avoid running these computationally expensive tests on the entire data. To ensure these scans accurately represented the full dataset, the distribution of the minimum and maximum threshold values produced for the subset and dataset were compared. The data for all considered datasets were assessed for normality using the Shapiro Wilk normality tests, described in Section 2.4. Pipeline Analysis Methods. The MWU test was then applied to compare distribution similarity across the datasets. If the p-value was below 0.4, larger datasets were used which were more representative of the parent dataset.

The same process was repeated using 30 randomly selected resolution- and diagnostic-specific randomly selected subsets of the UCL dataset.

IXI Dataset Sample Size Investigation Results

This was demonstrated by histograms demonstrating different sample distributions of the lower and upper threshold troughs in Figure 14. This reflects an inherent difference in the datasets which may influence the ability of thresholds determined for the 30-scan dataset to be successfully applied to the full dataset.

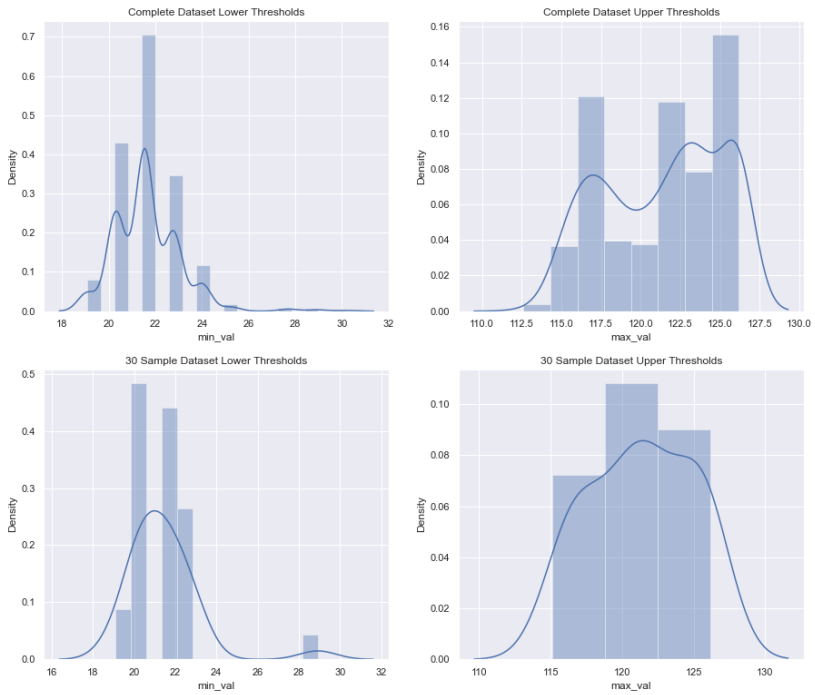

Figure 14: The histograms for the IXI dataset, 30 sample subset, and the IXI data not included in the 30-sample subset, were all used to assess the normality of the data.

To overcome this, 30 scans were added to the subset, improving the dataset similarity produced using a MWU test. The lower thresholds were very weakly represented by the 30-scan subset and better represented by the 60-scan lower thresholds, as shown by the p-values in Table 18. A sample size of 60 was used throughout the rest of the IXI investigations.

Table 18: Results from the Mann-Whitney-U test are depicted for the minimum and maximum threshold values for the 30 and 60 IXI sample subset and the rest of the IXI dataset for each subset.

Comparison Datasets	Comparison Values	p Value
30 IXI Scans to All Other Scans	**Lower Threshold Values**	0.139
	Upper Threshold Values	0.851
60 IXI Scans to All Other Scans	**Lower Threshold Values**	0.475
	Upper Threshold Values	0.823

Across the 60-scan subset and entire dataset produced similar minimum, maximum, mean, and medium values of the lower and upper thresholds. The spike values for the lower and upper spikes were consistent across the scans with negligible standard deviations. The mean

and median values for the minimum and maximum thresholds are also very similar across the considered datasets, demonstrating consistency in the data.

UCL Dataset Sample Size Investigation Results

The lower and upper thresholds for each dataset were determined to be non-parametric using the Shapiro-Wilk test and plotted as histograms. As the datasets are also independent, the sample size was evaluated for similarity using a MWU test, shown in Table 19. The 3.0T sample describes the overall 3.0T dataset, however with limited strength. 1.5T sample data shows weak similarity to all 1.5T data.

Table 19: The Mann-Whitney-U statistical comparison test was used to assess the similarity between the threshold options for each resolution-specific 30- scan subset of UCL control data to the respective remaining data for each subset.

Comparison Datasets	Comparison Values	p Value
30 Control 1.5T Scans to All Other Control 1.5T Scans	Lower Threshold Values	0.124
30 Control 1.5T Scans to All Other Control 1.5T Scans	Upper Threshold Values	0.325
30 Control 3.0T Scans to All Other Control 3.0T Scans	Lower Threshold Values	0.867
30 Control 3.0T Scans to All Other Control 3.0T Scans	Upper Threshold Values	0.447

All lower and upper thresholds of the subsets were graphed to qualitatively assess the similarity in distribution.

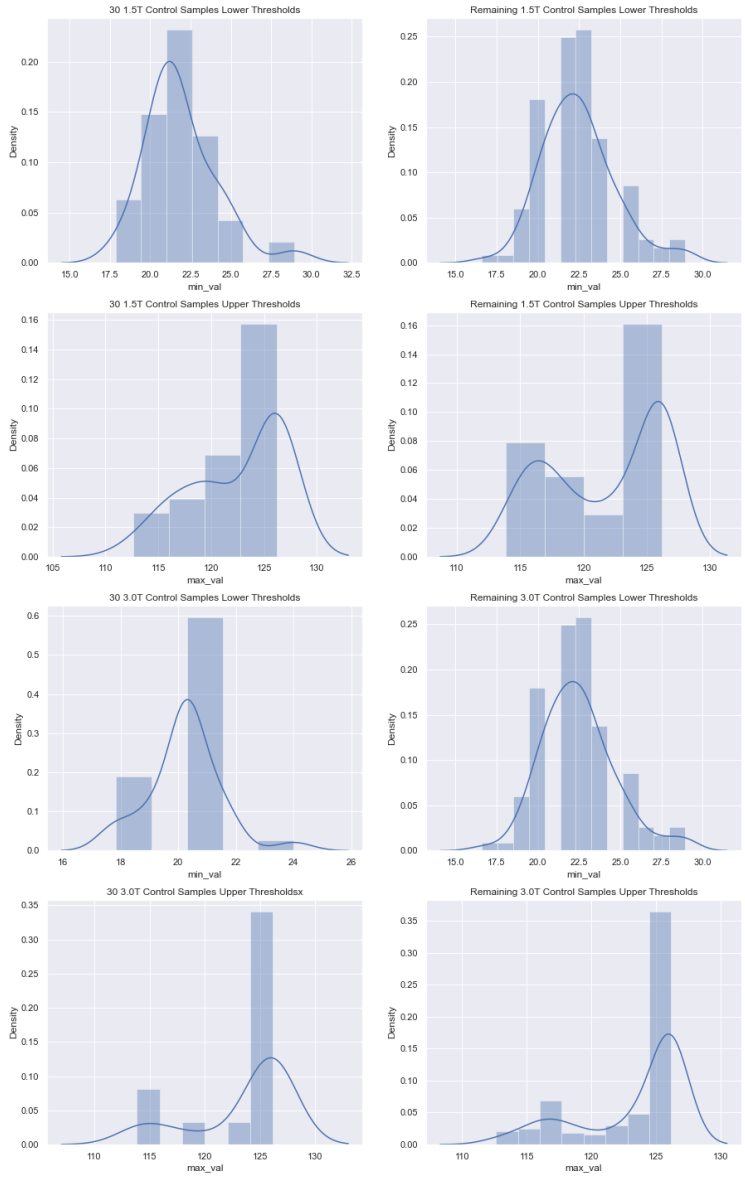

Figure 15: Histograms for the considered UCL 30 sample datasets were assessed for normality.

As the table results and histograms show similar outputs, the samples were considered to reasonably describe the larger dataset, even though the p-value of the MWU test was low. As the 1.5T control dataset is not very large, the 30-sample size was kept to avoid overfitting.

Appendix 3: Pipeline 3 Development with IXI

The IXI dataset was used to validate the Pipeline 3 development process before the same process was applied to the UCL dataset.

Further processing in Pipeline 3 is required to apply image enhancement in combination with brainageR, as demonstrated by the poor age prediction performance of Pipeline 2, shown in Table 20. The large MAE and Std., strong r-value, and acceptable r^2 value implies that the brainageR age predictions have precision but not accuracy, following a similar pattern to the actual participant ages but predicting the wrong year by a consistent margin.

Table 20: Pipeline 2 results when applied to the entire IXI dataset.

Processing Method	Processing Steps	MAE	Std.	R Value	R^2 Value
Pipeline 2	**Reoriented Raw, SynthSR and BrainAgeR**	16.98	9.50	0.8334	0.6945

Histograms of the data before and after SynthSR preprocessing show raw data with varying low-frequency spike locations, and preprocessed data instead show a consistent spike at 11.69, demonstrated in Figure 16 below. Wavelet analysis applied to the histograms of the preprocessed IXI dataset identified the lower and upper threshold troughs from the low and high-frequency spikes.

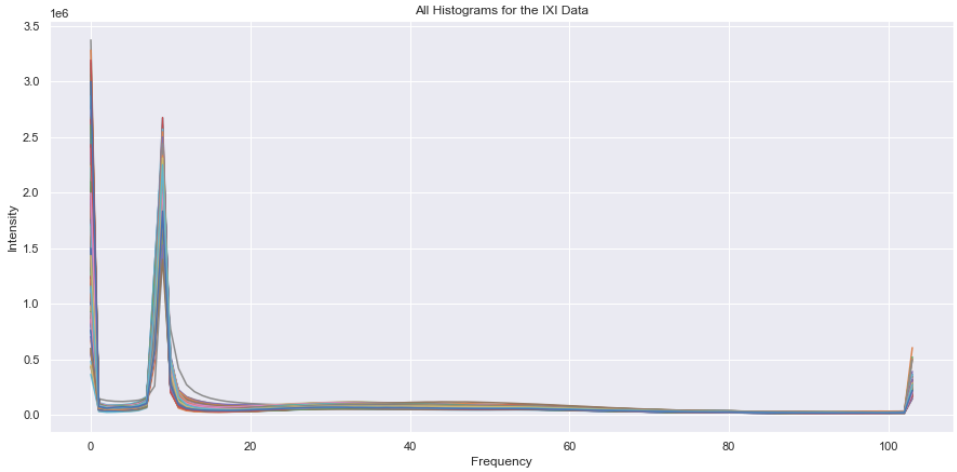

Figure 16: Histograms of all IXI data after the application of SynthSR preprocessing. These histograms depict the consistency of the noise-related spikes in the preprocessed data.

A subset of 30 randomly selected scans did not appropriately represent the larger dataset. An additional 30 scans were added to the subset to achieve better similarity with the complete dataset, as described in Appendix 2: Sample Size.

Each 60-scan threshold option was tested using brainageR, and the age prediction error analysis for each threshold pair is displayed in Table 21. The best performing combination of minimal MAE and Std., and maximal r-value was the maximum lower and maximum upper thresholds of 28.9 and 126.2.

Table 21: BrainageR age prediction accuracy for each threshold option assessed which had been applied to a 60-sample subset of the IXI dataset. The threshold options were produced using the 60-sample subset. The best performing threshold is bolded.

Lower Threshold Category	Lower Threshold	Upper Threshold Category	Upper Threshold	MAE	Std.	R Value	R^2 Value
Minimum	**19.1**	**Maximum**	**126.2**	7.34	5.56	0.8920	0.7957
Median	**21.5**	**Maximum**	**126.2**	6.92	5.17	0.8946	0.8003
Mean	**21.8**	**Maximum**	**126.2**	6.92	5.15	0.8942	0.7996
Maximum	**28.9**	**Maximum**	**126.2**	**6.78**	**4.86**	**0.8937**	**0.7988**
Maximum	**28.8**	**Mean**	**121.3**	9.31	7.49	0.8925	0.7966
Maximum	**28.9**	**Median**	**121.2**	9.38	7.55	0.8923	0.7962
Maximum	**28.9**	**Minimum**	**112.6**	14.04	10.45	0.8900	0.7921

The best performing threshold values of maximum lower and maximum upper threshold were applied to the full dataset, as shown in Table 22. The threshold option values used were produced from both the 60-scan subset and the full dataset and can be found in Appendix 5: Threshold Options. The 60-sample values produced slightly better MAE and Std, but a worse performing r-value describing the predicted and real ages. The better performing r-value was selected, and the thresholds of 30.2 to 126.2 were applied in Pipeline 3.

Table 22: The best thresholds from the 60-sample analysis were applied to the entire dataset in two different methods and individually analysed. Each method leveraged the winning threshold option category of maximum lower threshold and maximum upper threshold. The first method used the values identified in the 60-sample threshold options while the second method used the values identified in the entire IXI dataset threshold options.

Dataset Used to Determine Threshold Option	Lower Threshold	Upper Threshold	MAE	Std.	R Value	R^2 Value
60 Sample Subset	**28.9**	**126.2**	6.74	4.72	0.8334	0.8068
Full IXI Dataset	**30.2**	**126.2**	6.88	4.84	0.8950	0.8009

Appendix 4: Pipeline 3 Application to the UCL Dataset

The wavelet analysis produced similar threshold options for the combined 60 scans, the 1.5T scans, and the 3.0T scans. These threshold options were separately applied to the 30-scan resolution-specific subsets, with the results depicted in Table 23. These brainageR results produce age predictions with the highest accuracy when thresholds 23.9 to 124.4 and 24.0 to 123.5 are respectively applied to 1.5T and 3.0T data.

Table 23: BrainageR age prediction accuracy for each threshold option assessed which had been applied to the resolution-specific 30-sample subsets of the UCL dataset. The best performing thresholds are bolded. The Median Upper Threshold Category is not shown here as it was identical to the Maximum category for both 1.5T and 3.0T resolution data.

Data Resolution	Lower Threshold Category	Lower Threshold	Upper Threshold Category	Upper Threshold	MAE	Std.	R Value	R^2 Value
1.5T	Maximum	28.9	Maximum	126.2	9.3757	5.800	0.7227	0.5223
	Mean	21.9	Maximum	126.2	10.742	6.6037	0.7427	0.5516
	Median	21.5	Maximum	126.2	10.862	6.529	0.749	0.5616
	Minimum	17.8	Maximum	126.2	12.569	6.543	0.754	0.5689
	Maximum	28.9	Mean	122.4	**6.812**	**4.713**	**0.750**	**0.5632**
	Maximum	28.9	Minimum	112.6	6.864	6.436	0.792	0.6274
3.0T	Maximum	24.0	Maximum	126.2	9.457	5.558	0.749	0.5613
	Mean	20.3	Maximum	126.2	11.308	5.949	0.745	0.5545
	Median	20.2	Maximum	126.2	11.356	5.928	0.748	0.5588
	Minimum	17.8	Maximum	126.2	12.886	6.372	0.718	0.5153
	Maximum	24.0	Mean	123.5	**6.5796**	**4.848**	**0.752**	**0.5649**
	Maximum	24.0	Minimum	113.9	7.283	6.679	0.7195	0.5177

To compare the FTD and control thresholds and better understand the difference between the datasets, the optimal threshold values produced from wavelet analysis were determined for resolution- and diagnostic-specific subgroups. The maximum lower threshold and mean upper threshold values, as shown in Table 24, are different for all 1.5T control values, 1.5T FTD values, 3.0T control values, and 3.0T FTD values. The larger maximum and mean values in the FTD participant data highlight threshold outliers which

may impact brainageR results when optimal control data thresholds are applied to control and FTD data.

Table 24: The best threshold options for each resolution- and diagnostic- specific subset of the UCL data were considered. Each method leveraged the winning threshold option category from the 30-sample analysis of maximum lower threshold and mean upper threshold. This table depicts the threshold options assessed for all 1.5T control data, all 3.0T control data, all 1.5T FTD data, and all 3.0T FTD data. The FTD data is shown here purely for investigative purposes, and the FTD thresholds are not used in determining the ultimate pipeline.

Frequency Threshold	Successful Categories	1.5T Control Data	3.0T Control Data	1.5T FTD Data	3.0T FTD Data
Lower Threshold	**Maximum**	28.9	25.2	56.0	51.1
Upper Threshold	**Mean**	121.9	123.3	125.1	125.4

The maximum lower threshold and mean upper threshold calculated using the resolution-specific 30-scan threshold options and resolution-specific control data subsets were tested for all control data in Table 25. The results across thresholds tested were nearly identical for each resolution, yet the 30 sample and entire data threshold options performed slightly better respectively for the 1.5T data and 3.0T data. Although the optimal thresholds were found to be resolution specific, with the thresholds used only slightly differentiated across resolution and could likely be optimized for both datasets. Therefore, Pipeline 3 uses the best performing thresholds of 28.9 to 122.4 for 1.5T and 25.2 to 123.3 for 3.0T obtained with control data.

Table 25: The best thresholds from the 30-sample analyses were applied to the entire dataset in two different methods and were individually analysed for each resolution-specific subset of data. Each method leveraged the winning threshold option category of maximum lower threshold and mean upper threshold. The first method used the values identified in the 30-sample control data threshold options while the second method used the values identified using the threshold options from all control data.

Threshold Origin	Data Resolution	Lower Threshold	Upper Threshold	MAE	Std.	R Value	R^2 Value
30-Sample Control Data	1.5T	28.9	122.4	7.26	5.76	0.4952	0.2452
	3.0T	24.0	123.5	7.13	5.09	0.5679	0.3225
All Control Data	1.5T	28.9	121.9	7.31	5.82	0.4991	0.2491
	3.0T	25.2	123.3	7.05	5.05	0.5685	0.3233

Appendix 5: Threshold Options

Table 26: Lower and upper frequency threshold options and the associated spikes were determined from the histograms for the entire UCL and IXI dataset, and associated subsets.

Data Subset Considered	Frequency Threshold	Maximum	Mean	Median	Minimum	Standard Deviation	Spike Mean	Spike Standard Deviation
All IXI Data	Lower	30.2	21.6	21.5	19.1	1.42	11.7	1.78E-13
	Upper	126.2	121.4	122.5	112.6	3.72	127.4	6.68E-13
60 Sample IXI Subset	Lower	28.9	21.8	21.5	19.1	1.59	11.7	1.79E-15
	Upper	126.2	121.3	121.2	112.6	3.66	127.4	1.15E-13
All UCL Control Data	Lower	28.9	21.1	20.3	16.6	2.02	11.7	0.06
	Upper	126.2	122.7	126.2	112.6	4.37	127.4	1.57E-13
All UCL FTD Data	Lower	56.0	21.6	21.5	15.4	2.79	11.7	0.04
	Upper	126.2	125.3	126.2	111.4	2.72	127.4	2.80E-12
1.5T UCL Control Data	Lower	28.9	22.3	21.5	16.6	2.23	11.7	1.78E-15
	Upper	126.2	121.9	123.7	112.6	4.51	127.4	1.43E-13
3.0T UCL Control Data	Lower	25.2	20.2	20.3	17.8	1.32	11.7	0.08
	Upper	126.2	123.3	126.2	112.6	4.17	127.4	1.57E-13
60 Sample UCL Control Data	Lower	28.9	21.0	20.3	17.8	1.97	11.7	1.79E-15
	Upper	126.2	123.0	126.2	112.6	4.48	127.4	1.15E-13
1.5T 30-Sample UCL Control Data	Lower	28.9	21.9	21.5	17.8	2.21	11.7	1.81E-15
	Upper	126.2	122.4	126.2	112.6	4.47	127.4	4.34E-14
3.0T 30-Sample UCL Control Data	Lower	24.0	20.2	20.3	17.8	1.25	11.7	1.82E-15
	Upper	126.2	123.5	126.2	113.9	4.50	127.4	4.34E-14

Appendix 6: Shapiro Wilk Normality Results

Table 27: Shapiro Wilk normality tests were used to assess normality for various UCL threshold subsets processed with Pipeline 1.

Data Subset	Assessed Parameter	p Value	Normality
30 1.5T Control Scans	Lower Thresholds	7.6913e-03	not normal
30 1.5T Control Scans	Upper Thresholds	3.6854e-05	not normal
Remaining 1.5T Control Scans	Lower Thresholds	1.0206e-04	not normal
Remaining 1.5T Control Scans	Upper Thresholds	4.0355e-11	not normal
30 3.0T Control Scans	Lower Thresholds	4.437999e-04	not normal
30 3.0T Control Scans	Upper Thresholds	1.504657e-07	not normal
Remaining 3.0T Control Scans	Lower Thresholds	4.966649e-11	not normal
Remaining 3.0T Control Scans	Upper Thresholds	5.764166e-18	not normal

Table 28: Shapiro Wilk normality tests were used to assess the normality of brainageR age predictions for various UCL subsets processed with Pipeline 1.

Data Subset	p Value	Normality
Control on 1.5T	0.5792	normal
Control on 3.0T	0.4635	not normal
FTD on 1.5T	0.0005	not normal
FTD on 3.0T	0.0307	not normal
All Control	0.5160	normal
All FTD	0.0001	not normal