A U-NET MODEL OF LOCAL BRAIN-AGE

Sebastian G. Popescu[1,2*], **Ben Glocker**[1], **David J. Sharp**[2,3], **and James H. Cole**[4,5]

[1]Biomedical Image Analysis Group, Imperial College London
[2]Computational, Cognitive & Clinical Neuroimaging Laboratory, Imperial College London
[3]Care Research & Technology Centre, UK Dementia Research Institute
[4]Centre for Medical Image Computing, University College London
[5]Dementia Research Centre, University College London

ABSTRACT

We propose a new framework for estimating neuroimaging-derived "brain-age" at a local level within the brain, using deep learning. The local approach, contrary to existing global methods, provides spatial anatomical information on patterns of brain ageing. We trained a U-Net model on brain MRI scans from n=3463 healthy people to produce individualised 3D maps of brain-predicted age. Testing on n=692 healthy people resulted in a median (across subject) mean absolute error (within subject) of 9.0 years. Performance was more accurate (MAE around 7 years) in the prefrontal cortex and periventricular areas. We also introduce a new voxelwise method to reduce the age-bias when predicting local brain-age "gaps". To validate local brain-age predictions, we tested the model in people with mild cognitive impairment or dementia using data from OASIS3 (n=267). Different local brain-age patterns were clearly evident between healthy controls and people with mild cognitive impairment or dementia, particularly in subcortical regions, with the accumbens, putamen, pallidum, hippocampus and amygdala. Comparing groups based on mean local brain-age over regions-of-interest resulted in large effects sizes, with Cohen's d values >1.5, for example when comparing people with stable and progressive mild cognitive impairment.

***K*eywords** Brain age · Deep learning · Dementia · U-Net · Voxelwise

1 Introduction

Brain ageing is associated with cognitive decline and an increased risk of neurodegenerative disease, though these effects vary greatly between individuals. Brain atrophy, often measured using structural MRI, is commonly seen in many neurological diseases [1, 2] and in normal ageing. Even hippocampal atrophy, which is often thought to be characteristic of Alzheimer's disease can be seen in many other neurological and psychiatric conditions, respectively in normal ageing [3]. Evidently, both healthy ageing and dementia can affect the same brain regions [4]. This fact complicates research into the earliest stages of age-related neurodegenerative diseases, as determining what changes are 'normal' and what are pathological is challenging. Brain-Age offers a way of establishing if changes in the brain are abnormal or not for a given age. The difference between chronological age and "brain-predicted age" obtained from neuroimaging data has been provided insights into the relationship between brain ageing and diseases, and may be a useful biomarker for predicting clinical outcomes [5, 6, 7]. For example, in Alzheimer's Disease (AD), patients have previously been shown to have older-appearing brains, and that individuals with mild cognitive impairment (MCI) who had an older-appearing brain were increasingly likely to progress to dementia within three years [8, 9, 10, 11]. However, despite the large growing literature employing the brain-age paradigm [12, 13], current approaches tend to generate brain-age predictions at a global level, with a single value per brain image. While some efforts have been made to derive patterns of 'feature importance' or similar from brain-age models [14, 15, 16, 17], these patterns are at population-level, and do not apply to the individual.

Localized Brain Predicted Age Obtaining a finer grained picture of patterns of brain-ageing for a given brain disease is likely to provide several benefits. Firstly, neuroanatomical patterns should enable inferences to be made about mechanisms underlying the clinical manifestation of the disease. Secondly, better predictive discrimination

*s.popescu16@imperial.ac.uk

between clinical groups should be possible, as different groups are likely to be associated with different spatial patterns of age-related brain changes, even in the case where 'global' brain-age differences are similar. Thirdly, the local individualised maps should enable fine-grain characterisation of brain changes over time, as disease progresses or in response to treatment. Finally, spatial patterns of brain-age could be used to discover clinically-relevant subgroups in a data-driven manner, for example using clustering techniques.

Related work Limited prior work on local predictions of brain-age are available. Of note, is the early work of Cherubini et al. [18], who used linear regression models with voxel-level features derived from voxel-based morphometry and diffusion-tensor imaging to demonstrate reasonable prediction results in a small sample of healthy people (n=140). This approach of using a separate linear regression model for each voxel is limited as it does not incorporate contextual information from neighbouring voxels, and is insensitive to non-linear relationships. Other studies have provided 'patch' level information on brain-age, subsequently averaging predictions across brain regions to arrive at a global-level prediction [19, 20]. In Bintsi et al. [20], the authors use a ResNet [21] for each 3D block, reporting MAE values between 2.16 and 4.19 depending on block origin. While these approaches are promising, the size of the patch limits spatial resolution which results in less insightful inference in clinical settings. For example, semantic dementia is associated with a relatively localised spatial pattern of atrophy, often the left anterior and middle temporal lobe [22, 23], which could be overlooked by brain-age prediction models that lack spatial resolution.

Contributions The goal of this work was to develop an accurate local brain-age model, by incorporating voxelwise information using recent developments in deep learning. U-Nets [24], which are typically used for tumor [25] or organ [26] segmentation, provide an excellent framework for voxelwise predictions, as their specific architecture enables the inclusion of contextual information into individual predictions. Here, we introduce a deep learning algorithm that is trained to predict localised brain-age, producing high-resolution maps of brain-predicted age differences (brain-PAD maps) covering the entire brain (see Figure 1). We hypothesised that brain-PAD would smoothly vary across regions of the brain, and in dementia patients we would see higher values in regions previously associated with AD-related atrophy. We provide an in-depth analysis of the structural differences seen in people with MCI and AD patients. We also examine the reliability of local brain-age predictions, both within and between scanners.

2 Methods

2.1 Participants

To train, test and validate our local brain-age model, we collated the following T1-weighted MRI brain scan datasets. Unless otherwise mentioned, studies were reviewed and approved by the local ethics committees. All participant data were from publicly accessible databases.

2.1.1 Healthy datasets for training/testing

Healthy training dataset is composed of 2001 healthy individuals with a male/female ratio of 1016/985, with a mean age of 36.95±18.12, aged 18-90 years. This dataset was itself an amalgam of 14 individual datasets, as used in our previous brain-age research [27].

Dallas Lifespan Brain Study (DLBS) is a major effort designed to understand the antecedents of preservation and decline of cognitive function at different stages of the adult lifespan, with a particular interest in the early stages of a healthy brain's march towards Alzheimer Disease. For our purpose we have selected solely the T1-weighted MRI scans, totaling n=315 healthy participants aged 18-89 years.

Cambridge Centre for Ageing and Neuroscience (Cam-CAN) neuroimaging dataset is part of larger project which is trying to use epidemiological, behavioural and neuroimaging data to understand how individuals can best retain cognitive abilities into old age. The dataset consists of n=652 T1-weighted MRI scans from participants aged 18-88 years.

Southwest University Adult Lifespan Dataset (SALD) comprises a large cross-sectional sample (n = 494; age range = 19-80) undergoing a multi-modal (sMRI, rs-fMRI, and behavioral) neuroimaging. Only T1-weighted MRI were used here. The goals of the SALD are to give researchers the opportunity to map the structural and functional changes the human brain undergoes throughout adulthood and to replicate previous findings.

Wayne State The Wayne State longitudinal data set for the Brain Aging in Detroit Longitudinal Study, comprises 200 healthy individuals, with n=302 total anatomical scans across two waves of data collection. All the participants were

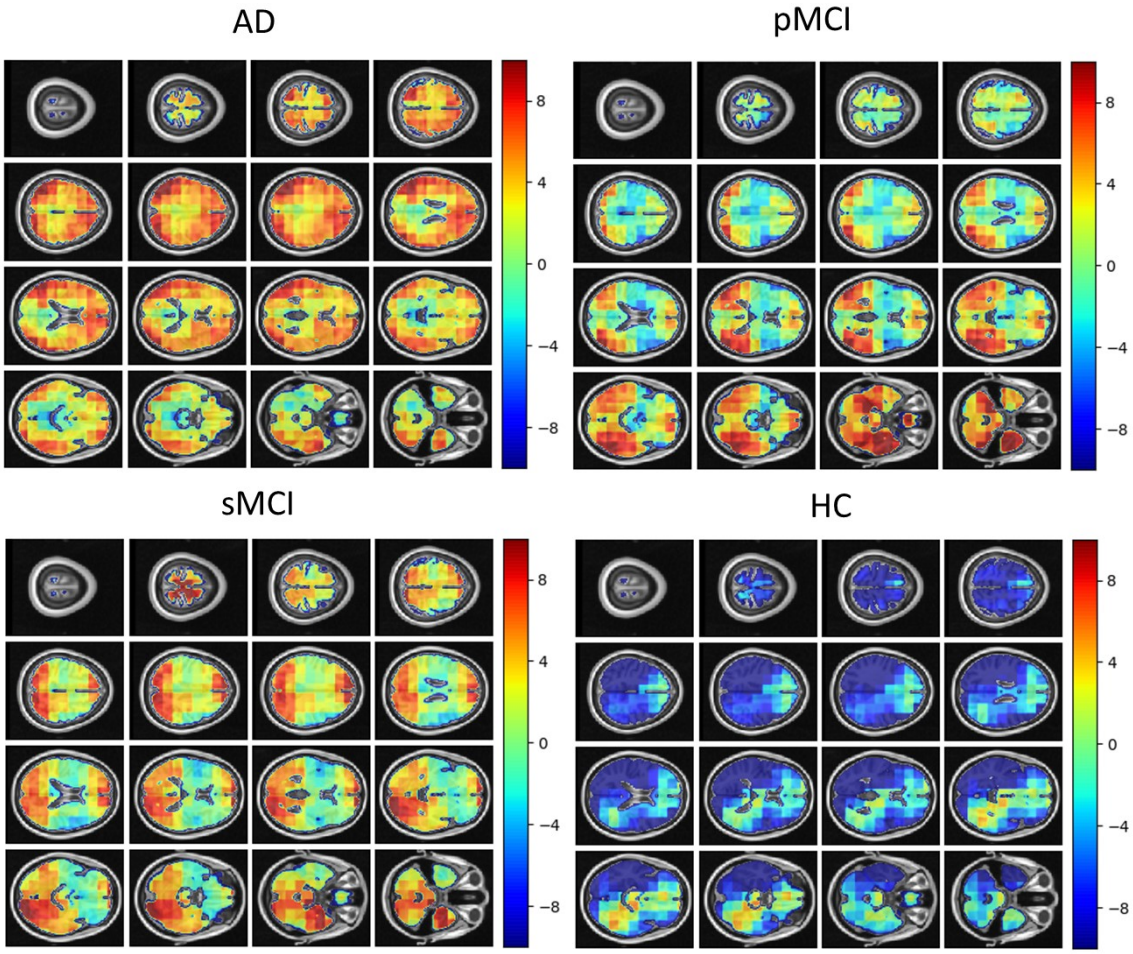


Figure 1: Local brain-PAD maps for randomly sampled subjects from clinical groups in cross-sectional OASIS3 dataset. Positive values indicate an increased pattern of local volume differences compared to healthy ageing patterns at the respective age. HC = Healthy Controls, pMCI = progressive MCI, sMCI = stable MCI, AD = Alzheimer's Disease.

screened by the local research centres to be free from neurological or psychiatric disorders according to well established protocols. All of the neuroimaging data were acquired either at 1.5T or 3T using standard T1-weighted sequences.

Within-scanner reliability dataset Here we used data from the Imperial College London project, STudy Of Reliability of MRI (STORM). The study comprises of 20 subjects with a male-female ratio of 12/8, with a mean age at the first scan undertaken of 34.05 ± 8.71 . The participants were scanned for the second time at an average distance of 28.35 ± 1.09 days. All participants were free from any neurological or psychiatric disorders. Data were acquired using a Siemens Verio 3T scanner.

Scanner calibration dataset 11 participants took place in this two-centre study, having a mean age at first scan of 30.88 ± 6.16 and with a male/female ration of 7/4. The two scanning sites were at Imperial College London, where a Siemens Verio 3T was used, whereas a Philips Ingenia 3T scanner was used at the Academic Medical Center Amsterdam. The average time interval between scans was of 68.17 ± 92.23 days.

2.1.2 Dementia datasets for testing clinical suitability of Local brain-age

OASIS3 is a retrospective compilation of data for >1000 participants that were collected across several ongoing projects through the WUSTL Knight ADRC over the course of 30 years. Participants include n=609 cognitively normal adults and n=489 individuals at with MCI or dementia ranging in age from 42-95 years. Using Clinical Dementia Rating

scale(CDR) scores, we classified participants as healthy control (HC), stable MCI, progressive MCI or AD, as detailed in Table 1. Follow-up CDR scores used to define MCI status were from at least 3 years after baseline assessments.

Characteristics	HC (n=128)	sMCI (n=29)	pMCI (n=29)	AD (n=78)
Males/Females, n	70/58	15/14	18/11	33/45
Age, mean (SD) years	68.14 (9.40)	76.44 (6.81)	75.72 (7.68)	75.02 (8.90)
Age, range years	42.66-97.11	59.2-94.44	49.38-93.93	50.35-95.58
Baseline CDR	0.0	0.5	0.5	≥ 1.0
Follow-up CDR	0.0	0.5	≥ 1.0	-

Table 1: Demographic characteristics for the OASIS3 dataset. CDR = Clinical Dementia Rating scale, HC = Healthy Controls, pMCI = progressive MCI, sMCI = stable MCI, AD = Alzheimer's Disease.

2.2 Data pre-processing

All T1-weighted brain MRI scans were pre-processed using the Statistical Parametric Mapping (SPM12) software package (`https://www.fil.ion.ucl.ac.uk/spm/software/spm12/`). This entailed tissue segmentation into grey matter (GM) and white matter (WM), followed by a nonlinear registration procedure using the DARTEL algorithm [28] to the Montreal Neurological Institute 152 (MNI152) space, subsequently followed by resampling to $1.5mm^3$ with a 4mm smoothing kernel.

2.3 Statistical analysis

Standard effect sizes To assess the magnitude of differences in local brain-PAD values between different groups, we used the standardised effect size Cohen's d:

$$d = \frac{m_1 - m_2}{\sqrt{\frac{(count_1-1)*var_1+(count_2-1)*var_2}{count_1+count_2-2}}} \tag{1}$$

where m_k is the mean, var_k represents the variance, whereas $count_k$ defines the number of subjects within group k. In all subsequent analysis using Cohen's d we perform the analysis at voxel-level between Brain-PAD scores from two different groups.

Intraclass Correlation Coefficient The intraclass correlation coefficient (ICC) is used to test the reproducibility of a certain quantitative measurement made by a specified number of observers which rate the same subject. The original formula is given as follows:

$$r = \frac{1}{Ns^2}\sum_{n=1}^{N}(x_{n,1} - \tilde{x})(x_{n,2} - \tilde{x}) \tag{2}$$

where $\tilde{x} = \frac{1}{2N}\sum_{n=1}^{N}(x_{n,1} + x_{n,2})$ and $s_2 = \frac{1}{2N}\{\sum_{n=1}^{N}(x_{n,1} - \tilde{x})^2 + \sum_{n=1}^{N}(x_{n,2} - \tilde{x})^2\}$

Here, we used ICC[2,1] as defined by Shrout and Fleiss [29]. The interval of values ranges from $[-1, 1]$ with values closer to 1 denoting that the observers (i.e., MRI scans or scanners) agree with each other.

2.4 Local "Brain-age" Prediction

We used a fully convolutional neural network (CNN) inspired by the U-Net architecture introduced in Ronneberger et al. [24]. Our network architecture is illustrated in Figure 2. Input images were the output from SPM12 pre-processing, representing voxelwise volume of GM and WM. These images were split into overlapping 3-dimensional blocks of size 52 voxels isotropic. The convolutional layers in our network used an isotropic 3x3x3 filter, convolved over the input image after which element-wise multiplication with the filter weights and subsequent summation was performed at each location. Subsequently, to allow for non-linear modelling, we passed the obtained values through an "activation function"; we used a LeakyReLu with alpha=0.2. $LeakyReLu(\alpha)$ are defined by the following equation $LeakyReLu(x) = max(x, 0) + min(x * \alpha, 0)$, thus allowing a small, non-zero gradient when the unit is not active.

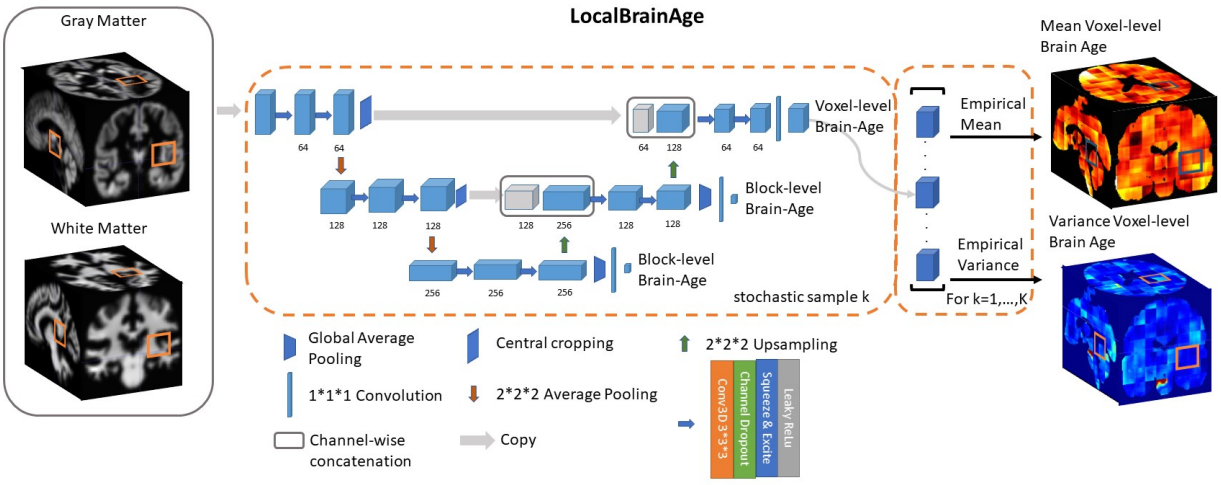


Figure 2: U-Net architecture for voxel-level brain-age prediction. Raw T1 scans are pre-processed through the DARTEL pipeline in SPM12, obtaining modulated gray and white matter segmentations registered to MNI152 template. Additional auxilliary block-level brain-age loss functions are added at each level of the U-Net to facilitate training.

The convolution operation is also controlled by its stride, which is how many pixels/voxels are skipped after every element-wise weight multiplication and summation. Downsampling increases the effective field of view or "receptive field" of layers higher in the hierarchy. We set the stride equal to 1 unless mentioned otherwise.

For the downsampling part of the U-Net we used at each scale two consecutive 3D 3x3x3 filter kernels with an initial number of channels = 64, which get multiplied by 2 as we progress down the downsampling path. For downsampling we used 2x2x2 average pooling.

For the upsampling part of the U-Net we inverted the downsampling architecture, with the downsampling layers being replaced by 2x2x2 upsampling layers. At each convolution we used a squeeze-and-excite unit. Squeeze & Excite networks were introduced in Hu et al. [30] and can be viewed as computationally less intensive method of performing attention over the channels of a given feature block.

Besides the voxel-level mean absolute error cost function on the output layer we introduced two additional cost functions at the two other scales of the architecture. We applied global average pooling followed by a dense layer to predict brain-age at block-level. During training, we observed that the addition of these auxiliary loss functions helped stabilise the learning process.

2.5 Removing bias in brain age at voxel level

Subtracting chronological age from estimated brain age provides a measure of the difference between an individual's predicted and chronological age, also known as the brain-age 'gap', brain-predicted age difference (brain-PAD) or brain-age 'delta'. A so-called 'regression dilution' has been commonly observed in brain-age prediction algorithms, caused by noise in the neuroimaging features leading to a greater under- or over-estimate of age, the further away a sample is from the training set mean age. In other words, the under-estimation of brain-predicted age for older subjects and over-estimation for younger subjects. Two main classes of approaches to fixing this problem have been reported:

$$\Delta = \alpha * Age + \beta \tag{3}$$

where Δ is the brain-age delta of a group of subjects from an external dataset that is used specifically for adjusting the bias. α and β are the parameters of a linear regression with the covariate Age representing chronological age.

Then, to obtain the bias-adjusted age we have the following equation:

$$\Delta_{adjusted} = \Delta - \alpha * Age + \beta \tag{4}$$

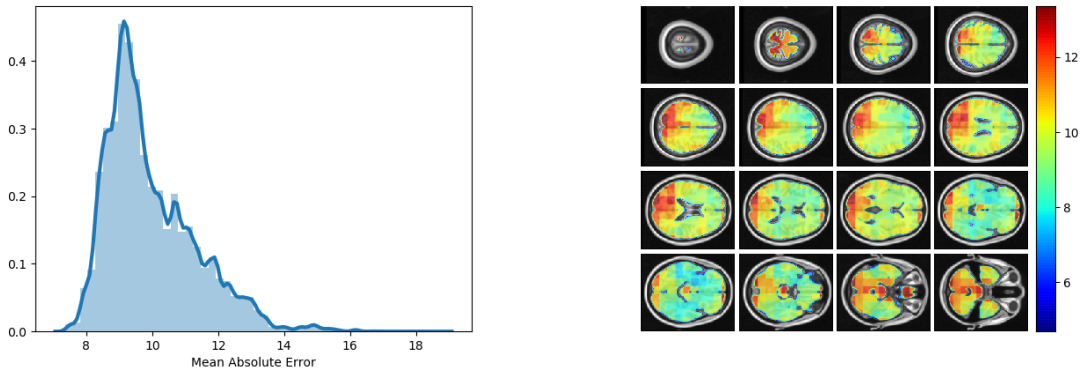
Another approach involves using the brain-predicted age in the linear regression. de Lange and Cole [31] showed that using either formulation results in the same statistical outcome.

Our approach involved using an independent dataset of participants, not present in either the training or testing set, which was stratified into different age bins, with an interval of 5 years. We then averaged their voxel-level brain-age delta values to obtain a 3D heatmap to use for de-biasing. We show additional results using the first approach in the supplementary material.

3 Results

3.1 Model performance in independent healthy test datasets

We tested the local brain-age model on healthy participants combined from the OASIS3 (n = 128), AIBL (n = 83) and Wayne State (n = 200) datasets. When first averaging the local MAE values within each individual to derive participant-level MAEs, the median MAE across subjects was 9.751 years (Figure 3c). The MAE of the model varied in different brain regions. We observed lower MAE values were observed locally in the prefrontal cortex and subcortical regions and higher MAE in the occipital lobe, cerebellum and brainstem (Figure 3b).

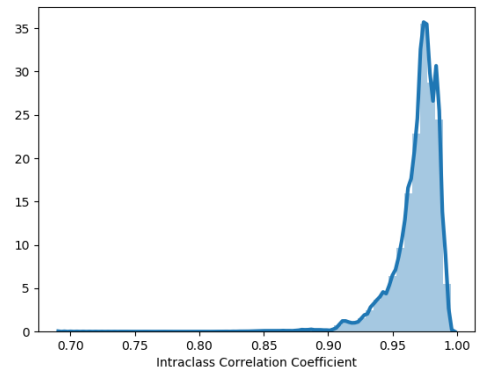


(a) (b)

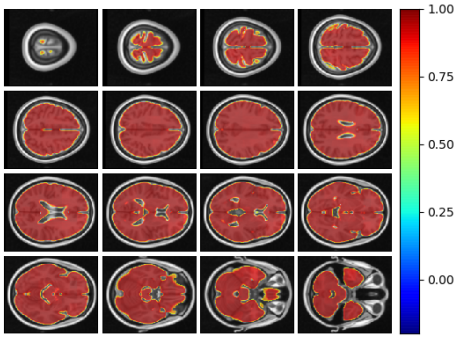
Figure 3: **Left** : Histogram of unadjusted, averaged MAE values across subjects for each voxel. **Right**: Axial slices showing the spatial heterogeneity in averaged across subjects MAE values.

3.2 Reliability of local brain-age

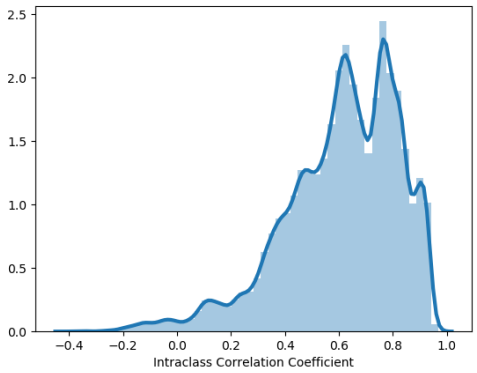
Using local brain-age values for both the test-retest dataset and between-scanner dataset, ICC was calculated per voxel. For the test-retest dataset, the vast majority of voxels were ICC<0.90, with a median of ICC=0.98. This indicated very high test-retest, or within-scanner, reliability, of local brain-age predictions. We observed comparatively lower ICC values at the extremities of the brain, evident in Figure 4b. This could potential be attributed to residual misregistration or partial volume effects. For the between-scanner analysis the reliability was lower, with median voxel-level ICC = 0.623. Interestingly, the pattern of ICC varied across the brain, with higher values observed in the prefrontal cortex and lower values in more inferior regions, particularly the brainstem and cerebellum (Figure 4d).



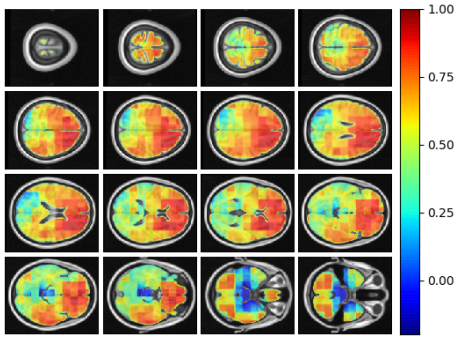
(a) Within-scanner variability



(b) Within-scanner variability



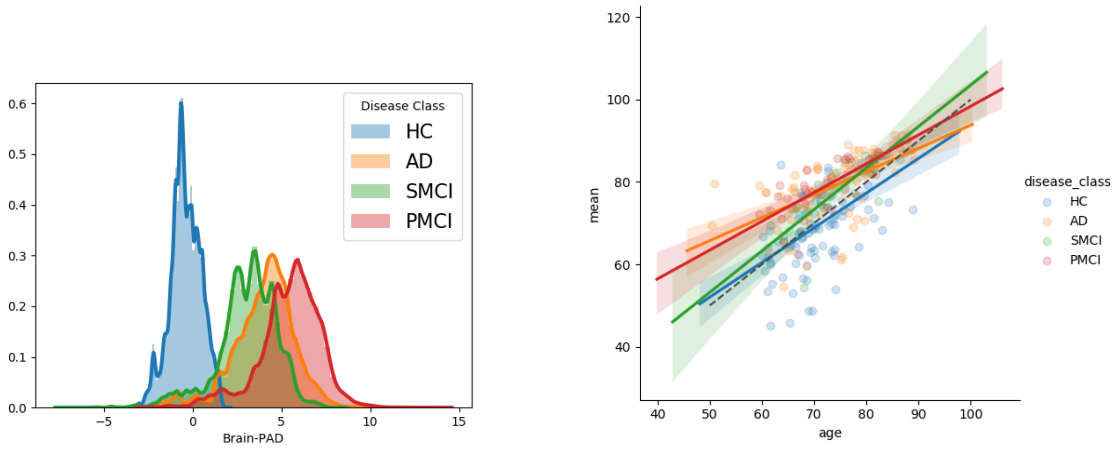
(c) Between-scanner variability



(d) Between-scanner variability

Figure 4: **Upper left**: Histogram of Intraclass Correlation Coefficients computed at voxel-level on STORM dataset. Values above 0.9 indicate strong agreements. **Upper right**: ICC values at different views on the axial plane on test-retest (i.e., within-scanner) dataset (n=20). **Bottom left**: Histogram of Intraclass Correlation Coefficients computed at voxel-level on between-scanner reliability dataset (n=11, Siemens and Philips scanners). **Bottom right**: ICC values at different axial slices from the between-scanner dataset.

3.3 Differences in local brain-age patterns between healthy controls, patients with mild cognitive impairment and dementia



(a) (b)

Figure 5: **Left - Voxel-level** : Histogram at voxel-level of Brain-PAD scores of certain clinical groups from OASIS3. Brain-PAD after applying the bias-adjustment scheme is calculated for every voxel and then aggregated to the mean across all subjects. Histograms in the plot are composed of the mean Brain-PAD values for all voxels in the brain; **Right - Global-level** : Adjusted predictions averaged across voxels for each subject; HC = Healthy controls, sMCI= stable MCI, pMCI = progressive MCI, AD = Alzheimer's disease

We examined patterns of local brain-age algorithm in the context of MCI and dementia using the cross-sectional OASIS3 dataset. Firstly, we investigated if the global level brain-predicted age corresponds to previously reported differences from models that direct predict global brain age. We averaged local brain-age (after bias correction) across voxels per individual to generate a global brain age and then calculate Brain-PAD. The median Brain-PAD values were: 0.95 years for healthy controls, 2.83 years for stable MCI (sMCI), respectively 4.94 years for progressive MCI (pMCI) and 4.63 years for AD patients.

Disease Groups	HC	sMCI	pMCI	AD
HC	-	«0.001	«0.001	«0.001
sMCI	0.0369	-	«0.001	«0.001
pMCI	0.0007	0.1611	-	«0.001
AD	0.0004	0.4141	0.3714	-

Table 2: Independent t-test results between disease groups in OASIS3 for global brain-age prediction in lower triangular part of the matrix. Paired t-test results results between disease groups in OASIS3 for voxel-level brain-age prediction in upper triangular part of the matrix.

To assess the significance of group differences we performed an independent two-sample t-test, with differences between cognitively impaired groups and healthy controls being significant (HC-AD p-value=0.0004; HC-sMCI p-value=0.0369; HC-pMCI=0.0007). Between groups with varying degrees of cognitive impairment, the lowest p-value was reported between sMCI-pMCI at 0.1611, whereas the other two combinations were also insignificant (AD-sMCI p-value=0.4141; AD-pMCI=0.3714). In comparison, we note that our proposed algorithm is capable to estimate a positive median local brain-PAD score for sMCI (3.266), thereby managing to differentiate between sMCI and HC. On the OASIS3 dataset, we get an estimated median local brain-PAD score of 7.7 years (5.663) for pMCI, respectively of 4.229 for AD (Figure 5). By comparing brain-age difference locally, differing patterns between between all disease groups are evident (Table 2).

From Figure 6 we can observe that the most striking differences are in the temporal lobe and subcortical regions when comparing AD to HC, which is in line with literature. For a more in-depth look at differences between disease groups, we expand the analysis to subcortical ROIs. The accumbens, putamen, pallidum and hippocampus were the most

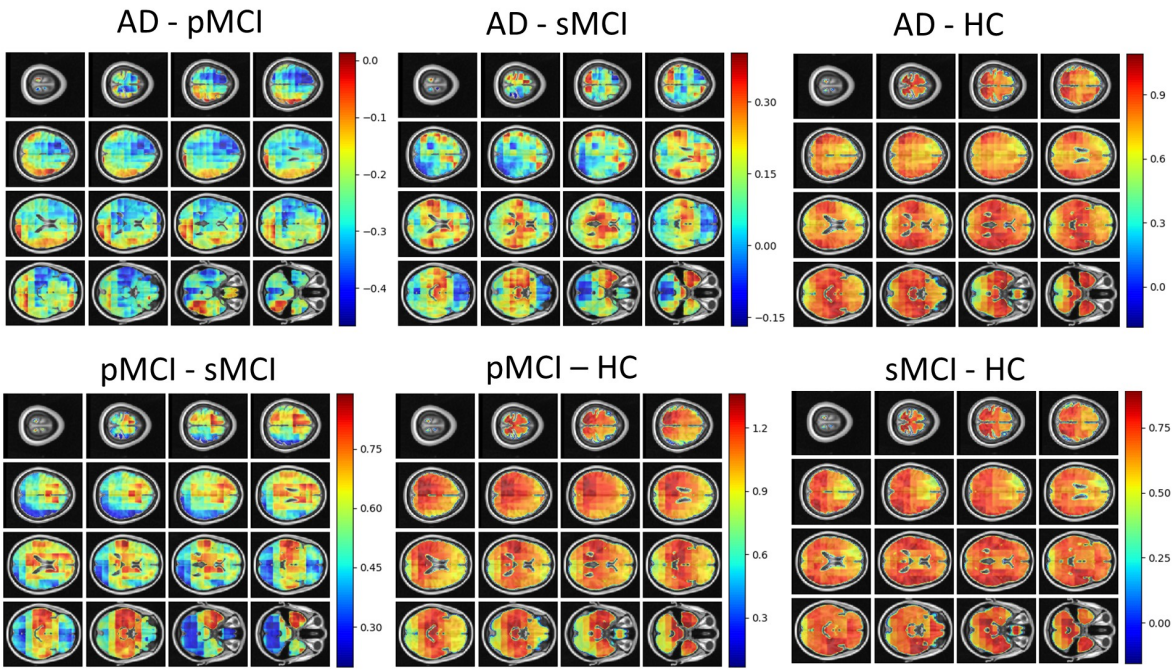


Figure 6: Cohen's d maps for different combinations of clinical groups in cross-sectional OASIS3. Positive values indicate a positive effect for the first group. HC=Healthy Controls; pMCI = progressive MCI; sMCI = stable MCI; AD=Alzheimer's Disease.

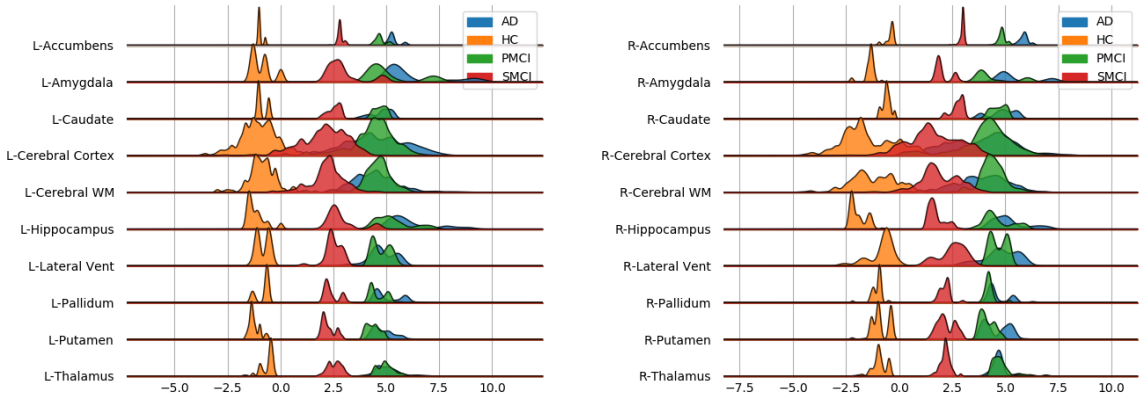
Subcortical ROI name	AD vs. HC Left	pMCI vs sMCI Left	AD vs. HC Right	pMCI vs sMCI Right
Brain-Stem	1.82	0.53	-	-
Amygdala	4.41	1.43	4.00	1.57
Caudate	4.69	1.69	4.41	1.47
Cerebral Cortex	1.98	0.81	1.93	0.89
Cerebral WM	5.29	2.48	4.32	2.56
Hippocampus	4.95	1.72	5.77	2.65
Lateral Vent	0.30	0.11	0.27	0.09
Pallidum	11.80	5.91	13.31	8.31
Putamen	15.59	6.82	10.66	5.13
Thalamus	4.02	1.60	4.80	2.11
Accumbens	30.02	9.60	31.00	9.80

Table 3: Cohen's d values for different subcortical ROIs from the Oxford-Harvard atlas. Higher values indicate a positive effect size for the first disease group specified. Values of above absolute value 0.2 are regarded as having a significant effect.

discriminative ROIs in terms of Cohen's d scores both for separating AD from HC and sMCI from pMCI (Table 3). We also provide histograms of the local brain-PAD scores for each disease group per subcortical ROI to get a better intuition being the Cohen's d scores (Figure 7). We notice that the high Cohen's d scores for the accumbens are due to the low variance in predictions occurring in that region.

4 Discussion

In this paper, we introduced an algorithm capable of reliably estimating neuroimaging derived age at a high resolution compared to existing models in literature. Having obtained a voxel-level distribution of mean absolute error values centred at 9.0 years, the accuracy of our proposed model is behind the current state of the art for global-level brain age prediction [32], which currently gravitates towards 3.0 years, or even in comparison with patch-level results [20] that are in an interval between 2.5 and 4.0 years, albeit using large blocks. Nevertheless, we provide results on the testing set



(a) Left hemisphere

(b) Right hemisphere

Figure 7: Subcortical ROI-based difference in voxel level Brain-PAD scores averaged across subjects from clinical groups from OASIS3. X axis registers Brain-PAD values within the given ROI; HC = Healthy controls, sMCI= stable MCI, pMCI = progressive MCI, AD = Alzheimer's disease

with no post-hoc bias-adjustment. Furthermore, it is worth mentioning that our model is assessed on a multi-site testing set with a flat chronological age histogram spanning ages between 18 and 90 as opposed to an age interval between 45 and 90, which is usually seen in papers assessed on UK Biobank [32, 20].

As biomarker of brain health, brain-age models may have clinical utility. To this end, we assessed the usefulness of local brain-age in a clinically-relevant setting, by exploring brain-ageing in dementia at a finer-grained scale then previous brain-age research. [33] highlights that in the early stages of AD, accelerated degeneration occurs in the medial temporal lobe (MTL), eventually affecting the amygdala, entorhinal and parahippocampal cortices [34, 35, 36]. In the case of discerning between pMCI versus sMCI, the authors reported that the amygdala and the inferior lateral ventricle were the most discriminative. Taking into account the Cohen's d values between local brain-PAD values within ROIs from the Oxford-Harvard atlas, we have observed that the accumbens, putamen, pallidum, hippocampus and amygdala were the most discriminative of the subcortical regions. Our findings are in line with previous work. For example, [37] showed that the highest discriminative power between AD and HC was present in the volume of the amygdala, with other MTL regions such as the hippocampus also highly discriminative, whereas [38] has observed a prominent loss of GM in the putamen for AD. As a prospective clinical application, the ROI-based BrainPAD values can be used by an additional classifier to distinguish between pMCI and sMCI. Results on cortical ROIs are provided in the supplementary material.

Previous work involving brain-age and dementia have obtained "brain-AGE" scores of -0.2 for sMCI, 6.2 for pCMI and 6.7 for AD on the ADNI dataset [9]. Our results are mostly in line with these previous findings, with the exception that for sMCI we obtained a positive median value of 2.83 years. However, we did not find any studies using OASIS3, so for our results we obtain a higher median local brain-PAD for pMCI (4.94 years) than for AD (4.63 years).

Biomarkers derived from neuroimaging have potential to provide clinically-relevant information about dementia and related conditions. While previous studies has reported standardised effect sizes from global brain-age, we used atlas ROIs to summarise regional values of local brain-PAD and generate Cohen's d values from pairwise group comparisons. Using conventional hippocampal volumetric measures, Henneman et al. [39] reported baseline effect size of 0.73 when comparing controls and MCI groups, and 0.33 when comparing MCI and AD patients. With our local brain-age framework, the control-MCI effect size for the hippocampus (average bilaterally) was $d = 5.45$ and the MCI-AD effect size was $d = 0.48$. Using voxel-based morphometry, [40] generated Cohen's d values for the hippocampus ($d = 0.6$) and amygdala ($d = 0.45$), when comparing sMCI and pMCI patients. Here, our local brain-age framework resulted in $d = 2.18$ for the hippocampus and $d = 1.5$ for the bilateral amygdala. This suggests that use of the brain-age paradigm to capture local age-related changes, relative to a healthy ageing model, could increase statistical power in experimental research and clinical trials, relative to conventional imaging biomarkers.

Out proposed U-Net local brain-age framework has some strengths and weaknesses to consider. Our model showed excellent test-retest reliability, giving confidence that the model could be applied longitudinally to assess individual patterns of brain-ageing changes. However, the between-scanner reliability was moderate, similar to our previous work using deep learning to predict brain age [27]. In the latter work, brain-age prediction was performed directly on raw

MRI scans, hence the deep learning model is picking up on site artefacts. One might expect that after preprocessing raw structural MRI scans through DARTEL these site effects might be partially eliminated. However, this is not the case as previous research [41] has demonstrated. Consequently, one drawback of the current algorithm is the requirement to have a healthy population from a given site, as site or scanner effects may result in the local brain-PAD distribution not being centred at zero. Future work should explore scanner harmonisation techniques to remove scanner effects [42].

In terms of model accuracy, local brain-age was not able to match its global-level models. However, this is to be expected as each predicted voxel has a field-of-view of 12^3 voxels, each of $1.5mm^3$, providing far less information than global models receive. Interestingly, regions such as the occipital lobe, brainstem and cerebellum has poorer accuracy than other brain areas. This may be due to local image noise or artefacts in the MRI data, more complex patterns of ageing in these areas not detected by our model, or potentially a lack of pronounced age-related changes to tissue volume in these regions.

Our results at testing time do not take into consideration bias correction as our proposed bias adjustment technique uses the actual chronological age for removing age-related bias, in opposition to the predicted age which is often used for improving predictive results. We note that we have attempted a bias-adjustment scheme reliant on a mass univariate linear regression which takes the predicted age for each voxel in part, but the results were only marginally improved as the voxel-level brain age delta distribution is noisy. Examples of some distributions and the fitted linear regressions are given in the supplementary material. Nevertheless, our bias adjustment technique is adequate for eliminating age related biases when the end goal is group comparison as opposed to predictive accuracy [31]. We noticed that age related negative correlations are eliminated at global level (taking the average over debiased voxel-level predictions). However upon closer inspection at voxel-level, there are slight issues for subjects between 18 and 55 years, with certain voxels exhibiting high brain age delta values (see supplementary materials). For subjects above 55 years, our bias adjustment technique is capable to eliminate age-related bias, thus enabling accurate statistical inference for clinical datasets involving older subjects.

By using a U-Net trained with the ground truth voxel-level regression objective given by a three-dimensional block filled with the chronological age, we encourage the network to emphasize the context encoded in its lower layers. As the individual voxel location we are aiming to obtain a prediction for is not necessarily related to the imposed ground truth output, the U-Net architecture is biased towards using the context information. Hence, in the worst case scenario where no voxel-level relationship is learned, the true resolution of our voxel-level predictions is actually blocks of 12^3 voxels. While this means that our resolution is not necessarily at the voxel-level, 12^3 voxels is still substantially higher resolution compared to existing models in literature. In the 3D block approach of Bintsi et al. [20], blocks are much larger, 64^3 voxels. Hence, any block-level age prediction will be biased towards the global-level brain age prediction as the blocks include a substantial portion of the overall brain. Moreover, in splitting the whole brain into blocks, naturally some blocks will include non-brain tissue or empty space, which will naturally reduce the amount of discriminative information present there, reducing the validity of results for regions within the respective block.

This paradigm can not only be applied to local-brain age prediction, but the final output can be composed of any other biomarker, such as polygenic risk score, MMSE scores or blood sample biomarkers. Whereas predicting deviations from normal ageing is well-motivated as one cannot easily evaluate the positive or negative accelerated ageing of the brain from a structural scan, all the aforementioned biomarkers can be easily obtained in a clinical setting, then the motivation stems mostly from gaining a mechanistic understanding of the relation between for example a structural MRI scan and the respective biomarker value. One potential research avenue for local brain-age is in disease subtyping, where one can use the local brain-age heatmaps to obtain clusters of different facets of accelerated brain ageing in a certain disease. Furthermore, another interest might be in exploring local brain-age's suitability in longitudinal datasets.

VBM constitutes an alternative to LocalBrainAge, being the de facto method to quantitatively assess differences between groups at voxel-level [43]. A common criticism of brain-age prediction in general is whether it brings added value compared to a standard VBM pipeline. We consider the two approaches to be in essence equivalent. In VBM pipelines, one assesses the statistical differences between a control group and a diseased group. BrainPAD implicitly measures this deviation of the diseases group from what constitutes a normative pattern of ageing, by placing the subject on a density of what a normative pattern constitutes for a given age. We leave for further work the comparison between VBM and LocalBrainAge.

Conclusion We have introduced a new algorithm that is capable of reliably estimating brain-age locally, providing high resolution maps revealing information on spatial patterns of age-related changes to brain volume. We were able to demonstrate the potential of this approach in clinical settings by mapping differences in local brain-PAD scores at patients with cognitive impairment and dementia.

Data and code availability statement The data used in these experiments are available on application to the relevant studies. The code used is available at https://github.com/SebastianPopescu/U-NET-for-LocalBrainAge-prediction alongside the pre-trained models.

References

[1] Abhijit Chaudhuri. Multiple sclerosis is primarily a neurodegenerative disease. *Journal of neural transmission*, 120(10):1463–1466, 2013.

[2] Valentina Lorenzetti, Nicholas B Allen, Alex Fornito, and Murat Yücel. Structural brain abnormalities in major depressive disorder: a selective review of recent mri studies. *Journal of affective disorders*, 117(1-2):1–17, 2009.

[3] MP Laakso, Kaarina Partanen, P Riekkinen, Maarit Lehtovirta, E-L Helkala, Merja Hallikainen, Tuomo Hanninen, Paula Vainio, and Hilkka Soininen. Hippocampal volumes in alzheimer's disease, parkinson's disease with and without dementia, and in vascular dementia an mri study. *Neurology*, 46(3):678–681, 1996.

[4] Samuel N Lockhart and Charles DeCarli. Structural imaging measures of brain aging. *Neuropsychology review*, 24(3):271–289, 2014.

[5] Johnny Wang, Maria J Knol, Aleksei Tiulpin, Florian Dubost, Marleen de Bruijne, Meike W Vernooij, Hieab HH Adams, M Arfan Ikram, Wiro J Niessen, and Gennady V Roshchupkin. Gray matter age prediction as a biomarker for risk of dementia. *Proceedings of the National Academy of Sciences*, 116(42):21213–21218, 2019.

[6] James H Cole, Stuart J Ritchie, Mark E Bastin, MC Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23 (5):1385–1392, 2018.

[7] James H Cole, Joel Raffel, Tim Friede, Arman Eshaghi, Wallace J Brownlee, Declan Chard, Nicola De Stefano, Christian Enzinger, Lukas Pirpamer, Massimo Filippi, et al. Longitudinal assessment of multiple sclerosis with the brain-age paradigm. *Annals of Neurology*, 2020.

[8] Katja Franke, Gabriel Ziegler, Stefan Klöppel, Christian Gaser, Alzheimer's Disease Neuroimaging Initiative, et al. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: exploring the influence of various parameters. *Neuroimage*, 50(3):883–892, 2010.

[9] Katja Franke, Eileen Luders, Arne May, Marko Wilke, and Christian Gaser. Brain maturation: predicting individual brainage in children and adolescents using structural mri. *Neuroimage*, 63(3):1305–1312, 2012.

[10] Christian Gaser, Katja Franke, Stefan Klöppel, Nikolaos Koutsouleris, Heinrich Sauer, Alzheimer's Disease Neuroimaging Initiative, et al. Brainage in mild cognitive impaired patients: predicting the conversion to alzheimer's disease. *PloS one*, 8(6):e67346, 2013.

[11] Sebastian Popescu, Alex Whittington, Roger N Gunn, Paul M Matthews, Ben Glocker, David J Sharp, and James H Cole. Nonlinear biomarker interactions in conversion from mild cognitive impairment to alzheimer's disease. *medRxiv*, page 19002378, 2019.

[12] James H Cole, Riccardo E Marioni, Sarah E Harris, and Ian J Deary. Brain age and other bodily 'ages': implications for neuropsychiatry. *Molecular psychiatry*, 24(2):266–281, 2019.

[13] Katja Franke and Christian Gaser. Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in neurology*, 10:789, 2019.

[14] A Erramuzpe, R Schurr, JD Yeatman, IH Gotlib, MD Sacchet, KE Travis, HM Feldman, and AA Mezer. A comparison of quantitative r1 and cortical thickness in identifying age, lifespan dynamics, and disease states of the human cortex. *Cerebral Cortex*, 2020.

[15] Nicola K Dinsdale, Emma Bluemke, Stephen M Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana IL Namburete. Learning patterns of the ageing brain in mri using deep convolutional networks. *NeuroImage*, 224: 117401, 2020.

[16] Arinbjörn Kolbeinsson, Sarah Filippi, Yannis Panagakis, Paul M Matthews, Paul Elliott, Abbas Dehghan, and Ioanna Tzoulaki. Accelerated mri-predicted brain ageing and its associations with cardiometabolic and brain disorders. *Scientific Reports*, 10(1):1–9, 2020.

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

[18] Andrea Cherubini, Maria Eugenia Caligiuri, Patrice Péran, Umberto Sabatini, Carlo Cosentino, and Francesco Amato. Importance of multimodal mri in characterizing brain tissue and its potential application for individual age prediction. *IEEE J. Biomedical and Health Informatics*, 20(5):1232–1239, 2016.

[19] Nick Pawlowski and Ben Glocker. Is texture predictive for age and sex in brain mri? *arXiv preprint arXiv:1907.10961*, 2019.

[20] Kyriaki-Margarita Bintsi, Vasileios Baltatzis, Arinbjörn Kolbeinsson, Alexander Hammers, and Daniel Rueckert. Patch-based brain age estimation from mr images. *arXiv preprint arXiv:2008.12965*, 2020.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[22] Ramon Landin-Romero, Rachel Tan, John R Hodges, and Fiona Kumfor. An update on semantic dementia: genetics, imaging, and pathology. *Alzheimer's research & therapy*, 8(1):1–9, 2016.

[23] Lorna Harper, Frederik Barkhof, Philip Scheltens, Jonathan M Schott, and Nick C Fox. An algorithmic approach to structural imaging in dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, 85(6):692–698, 2014.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[25] Xue Feng, Nicholas J Tustison, Sohil H Patel, and Craig H Meyer. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Frontiers in Computational Neuroscience*, 14: 25, 2020.

[26] Sulaiman Vesal, Nishant Ravikumar, and Andreas Maier. A 2d dilated residual u-net for multi-organ segmentation in thoracic ct. *arXiv preprint arXiv:1905.07710*, 2019.

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

[28] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[29] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[30] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

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

[32] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *BioRxiv*, 2019.

[33] Keith A Johnson, Nick C Fox, Reisa A Sperling, and William E Klunk. Brain imaging in alzheimer disease. *Cold Spring Harbor perspectives in medicine*, 2(4):a006213, 2012.

[34] Yanica Klein-Koerkamp, Rolf A Heckemann, Kylee T Ramdeen, Olivier Moreaud, Sandrine Keignart, Alexandre Krainik, Alexander Hammers, Monica Baciu, Pascal Hot, Alzheimer's disease Neuroimaging Initiative, et al. Amygdalar atrophy in early alzheimer's disease. *Current Alzheimer Research*, 11(3):239–252, 2014.

[35] Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.

[36] Clifford R Jack Jr, Heather J Wiste, Prashanthi Vemuri, Stephen D Weigand, Matthew L Senjem, Guang Zeng, Matt A Bernstein, Jeffrey L Gunter, Vernon S Pankratz, Paul S Aisen, et al. Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to alzheimer's disease. *Brain*, 133(11):3336–3348, 2010.

[37] Christian Ledig, Andreas Schuh, Ricardo Guerrero, Rolf A Heckemann, and Daniel Rueckert. Structural brain imaging in alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Scientific reports*, 8(1):1–16, 2018.

[38] Xiaojuan Guo, Zhiqun Wang, Kuncheng Li, Ziyi Li, Zhigang Qi, Zhen Jin, Li Yao, and Kewei Chen. Voxel-based assessment of gray and white matter volumes in alzheimer's disease. *Neuroscience letters*, 468(2):146–150, 2010.

[39] W. J. Henneman, J. D. Sluimer, J. Barnes, W. M. van der Flier, I. C. Sluimer, N. C. Fox, P. Scheltens, H. Vrenken, and F. Barkhof. Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. *Neurology*, 72(11):999–1007, Mar 2009.

[40] Shannon L Risacher, Andrew J Saykin, John D Wes, Li Shen, Hiram A Firpi, and Brenna C McDonald. Baseline mri predictors of conversion from mci to probable ad in the adni cohort. *Current Alzheimer Research*, 6(4): 347–361, 2009.

[41] Ben Glocker, Robert Robinson, Daniel C Castro, Qi Dou, and Ender Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*, 2019.

[42] Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. Unlearning scanner bias for mri harmonisation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 369–378. Springer, 2020.

[43] Juha Koikkalainen, Hanneke Rhodius-Meester, Antti Tolonen, Frederik Barkhof, Betty Tijms, Afina W Lemstra, Tong Tong, Ricardo Guerrero, Andreas Schuh, Christian Ledig, et al. Differential diagnosis of neurodegenerative diseases using structural mri data. *NeuroImage: Clinical*, 11:435–449, 2016.