Prediction of brain age using structural magnetic resonance imaging: A comparison of clinical validity of publicly available software packages

Ruben P. Dörfel[1,2,3]*, Brice Ozenne[2,4], Melanie Ganz[2,3], Jonas Svensson[1], Pontus Plavén-Sigray[1,2]

1. Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, & Stockholm Health Care Services, Region Stockholm, Sweden
2. Neurobiology Research Unit, Copenhagen University Hospital Rigshospitalet
3. Department of Computer Science, University of Copenhagen
4. Section of Biostatistics, University of Copenhagen

* *Corresponding author:* ruben.dorfel@ki.se

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

Abstract

Predicting brain age from structural magnetic resonance images is commonly used as a biomarker of biological aging and brain health. Ideally, as a clinically valid biomarker, brain age should indicate the current state of health and also be predictive of future disease onset and detrimental changes in brain biology. In this preregistered study, we evaluated and compared the clinical validity, i.e., diagnostic and prognostic abilities, of six publicly available brain age prediction packages using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Baseline brain age significantly differed between groups consisting of individuals with normal cognitive function, mild cognitive impairment, and Alzheimer's Disease for all packages, but with comparable performance to gray matter volume. Further, estimates were not centered around zero for cognitively normal subjects and showed considerable variation between packages. Finally, brain age did not show strong correlation with disease onset, memory decline, nor gray matter atrophy within four years from baseline in individuals without neurodegenerative disease. The substantial bias and variability in cross-sectional predictions, combined with the weak associations between brain age and longitudinal changes in memory performance or grey matter volume, suggested limited validity of current brain age estimates as a biomarker for biological aging.

Introduction

Biological aging is commonly defined as the accumulation of molecular and cellular defects that predispose humans to chronic diseases and physical deterioration, including neurodegeneration [1,2]. In the brain, these age-related changes manifest as increased atrophy and an accelerated decline in cognitive function[3]. According to the geroscience hypothesis, aging-related decline might be delayed or even prevented by pharmacological interventions [4]. However, prior to conducting clinical trials with the purpose of testing such putative geroprotective compounds, it is necessary to develop reliable biomarkers that can serve as surrogate endpoints of treatment efficacy [5,6].

In the brain, aging is associated with a range of structural changes. Magnetic resonance imaging (MRI) studies have shown that increasing age is associated with a decline in gray and white matter volume, and an enlargement of the ventricles [7,8]. The availability of large MRI datasets, in combination with advances in machine learning, has allowed the development of models to predict the apparent biological age, or *brain age,* of an individual [9].

Commonly, the primary interest is not the predicted brain age itself but rather its deviation from the individual's actual chronological age. This is often termed the *predicted age deviation* (PAD). A PAD close to zero suggests that the brain appears typical for its age. Conversely, a positive PAD, where the predicted brain age exceeds the chronological age, suggests that the brain appears older than the norm, indicating that the individual may age faster. In other words, the PAD should be indicative of an individual's aging trajectory, specifying whether the individual is aging faster or slower than the norm. By this assumption, individuals with high a high PAD should show: 1) accelerated atrophy, 2) increased cognitive decline, and 3) faster conversion to neurodegenerative state within the coming years.

Due to its apparent simplicity, brain age prediction based on structural MRI has become a popular area of research. Numerous groups have developed algorithms and applied them to clinical data in order to assess the potential of brain age as a marker for biological aging and brain health [10]. Many researchers have made their trained machine learning models

available for public use [10–15]. In a previous study, we compared the test-retest reliability of six such publicly available brain-age prediction software packages [16] and demonstrated that a subset of packages are both highly accurate and reliable in predicting chronological age.

For brain age to serve as a clinically valid indicator of an individual's aging trajectory, it should correlate with future changes in brain biology and age-related phenotypes, especially decline in certain cognitive abilities, such as memory-function. Current literature presents conflicting evidence on the association between PAD and longitudinal changes in brain biology and memory performance. For example, Elliott et al., 2021[17] and Gautherot et al., 2021 [18] show an association between PAD and longitudinal changes in the brain, while Vidal-Pineiro et al., 2021 [19] Korbmacher et al., 2024 [20] suggest the opposite.

In this study, we evaluate and compare the clinical validity of pre-trained, ready-to-use brain age prediction models, in both cross-sectional and longitudinal contexts. First, we investigate whether the baseline PAD differs between diagnostic groups, whether the PAD is associated with memory function, and whether the PAD predicts disease onset within a four-year period. Additionally, we investigate whether baseline brain age, and changes in brain age over time, are associated with memory decline and gray matter atrophy.

Methods (Online)

Study data

We analyzed data from 2330 subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu), a longitudinal, multi-site study with comprehensive neuroimaging and neuropsychological testing data. The baseline cohort comprised cognitively normal individuals (CN, n=857, mean age=72.24±6.63 years), individuals with mild cognitive impairment (MCI, n=1070, mean age=72.87±7.56 years), and Alzheimer's Disease patients (AD, n=403, mean age=74.83±7.8 years). For longitudinal analyses, we included yearly follow-ups up to four years from baseline. Demographic characteristics included age and sex, and years of education at baseline. Memory performance was quantified using the ADNI-Mem score, a composite measure derived from memory-related

neuropsychological tests [21]. An overview of the dataset and baseline characteristics is provided in Table 1.

Table 1. Participant demographics and scan counts by diagnostic group. This table presents the number of scans acquired at baseline and follow-up visits (M000, M012, M024, M036, and M048) alongside baseline clinical and demographic measures. Values for age, years of education, memory performance (ADNI-Mem), and average edge strength (AES) are expressed as mean (SD), while sex is reported as the count of females to males. The groups include cognitively normal (CN), mild cognitive impairment (MCI), Alzheimer's disease (AD), and the total sample.

	Scans					Baseline				
Group	**M000**	**M012**	**M024**	**M036**	**M048**	**Age [years] mean (SD)**	**Sex F/M**	**Education [years] mean (SD)**	**ADNI-Mem mean (SD)**	**AES mean (SD)**
CN	857	464	522	203	284	72.24 (6.63)	487/370	16.53 (2.51)	1.04 (0.58)	2.9 (0.26)
MCI	1070	753	501	237	217	72.87 (7.56)	447/623	15.96 (2.76)	0.19 (0.67)	2.92 (0.25)
AD	403	385	326	120	117	74.83 (7.8)	177/226	15.25 (2.91)	-0.86 (0.54)	3.04 (0.27)
Total	2330	1602	1349	560	618	72.98 (7.33)	1111/1219	16.05 (2.74)	0.32 (0.9)	2.93 (0.26)

Image acquisition, processing and quality control

All 3D T1-weighted MRI images were acquired at 1.5 Tesla and 3 Tesla. Images were corrected for gradient warping, B1 non-uniformity, and N3 non-uniformity by ADNI. The Clinica pipeline [22] was used to curate the ADNI data into the Brain Imaging Data Structure [23]. We used FreeSurfer[24] (v7.2) to derive cortical and subcortical measures for all scans. The intracranial volume and gray matter volume were computed using the segmentation pipeline from SPM12[25]. Quality control of the raw MRI images was performed using MRIQC [26], and quality control of the FreeSurfer output was performed using FSQC [27]. Scans with measures outside the 1.5 interquartile range were flagged for visual inspection. If artifacts were apparent, the scan was discarded. The average edge strength (AES) [28] was computed as a quality metric that reflects motion within MRI scans and has been validated against expert radiologist ratings [29]. The AES was computed on the brain-masked structural MRI image.

Brain age prediction packages

In line with our previous study [16], we used six brain age prediction packages that estimate brain age from structural MRI data: brainageR [30], DeepBrainNet [11], brainage [10], ENIGMA [14], pyment [15], and mccqrnn [13]. These encompass a wide set of different algorithms and input features. For more details, please refer to our first study or the original publications[10,11,13–15,30]. Additionally, we added the gray matter volume normalized by the subject's intracranial volume as a seventh baseline "model". An increase in gray matter volume is directly linked to neurodegeneration and accelerated aging, providing an anatomical estimate of brain-related aging decline that serves as a reference for more complex predictive approaches.

Statistical Analysis

We performed seven statistical analyses to assess the clinical validity of the six described brain age prediction packages. The analyses were pre-registered (AsPredicted #168021). For all inferential tests, significance levels were set to 0.05 and all p-values and confidence intervals were reported without adjustment for multiple testing. If not otherwise specified, the following covariates were included [31,32]: age, sex, and AES. We deviated from the preregistered analyses and included years of education as covariate, when the outcome of interest was ADNI-Mem. Age, sex, and years of education were included as present at baseline, while AES was computed for each scan. The statistical analysis was performed in R (4.4.1), and visualizations were created using Python (v3.9).

Convergent validity

To assess the convergent validity among brain age packages, i.e., the extent to which outcomes purporting to measure the same construct are associated with each other, we correlated the PADs for each package with the PADs from the other packages. Pearson's correlation and 95% confidence intervals are reported. Higher positive correlations indicate greater convergence between packages.

Differentiation between clinical groups at baseline and association with memory performance

To assess if the models could differentiate between different diagnostic groups, we used a linear model with age, sex, and AES as covariates to test for significant differences in the mean of the PADs between CN, MCI, and AD groups. Effect sizes were computed using *Cohen's d*. Comparisons were made for CN vs. MCI, CN vs. AD, and MCI vs. AD. We further assess the association between PAD and memory performance, using a linear model. The association was tested separately for CN, MCI, and AD groups.

Association between PAD at baseline and change of diagnostic group

To assess whether PAD at baseline is associated with the conversion from CN to MCI or AD within four years, we used logistic regression with inverse probability of censoring weighting (IPCW) [33]. IPCW accounted for non-random study dropout. Only subjects that were CN at baseline were included. The IPCW weights were computed with the same covariates as the logistic model. The R package *riskRegression* [34] was used to compute the IPCW weights, Brier score ad area under the curve (AUC) of the receiver operating characteristic curve used a dynamic definition[35]. After the model was fitted, we computed the probability of conversion for a subject with a PAD or normalized gray matter volume of {x=μ-SD, x=μ, x=μ +SD}. The covariates were fixed at their respective means for this calculation.

Association between (change in) PAD and longitudinal change in memory performance and gray matter volume

To assess the association between baseline PAD, and the change in gray matter volume or memory performance in subsequent years, we used a linear mixed model to estimate the covariance matrix between those measures, which was then used to estimate the correlation between the PAD and the change in normalized gray matter volume or ADNI-Mem over 4 years. More specifically, the R package *mmrm* [36] was used for an initial estimate of the variance-covariance matrix. The estimated variance-covariance matrix was then used to initialize a linear mixed model from the R package *LMMstar* [37]. The procedure was motivated because *mmrm* has a more stable optimizer while *LMMstar* facilitated the estimation of the

confidence intervals. The same approach was used to estimate the association between change in PAD and change in normalized gray matter volume or ADNI-Mem. A change in brain biology and memory performance should be tracked by a change in PAD. We do not apply this analysis to assess the association between a change in gray matter volume with itself, as this will be one.

Results

Convergent validity

To assess to what extent the predictions among packages are correlated to each other, we computed Pearson's correlation for the baseline PADs between packages. Additionally, we computed the correlation of the PAD for each package with the normalized gray matter volume. All six packages showed limited agreement across PADs for each package, as shown in Figure 1. The correlations between the packages range from 0.41, CI95% [0.38, 0.44] (brainageR, ENIGMA) to 0.71, CI95% [0.68, 0.73] (brainage, ENIGMA). Further, the correlation between PAD and normalized gray matter volume was considerably weaker between -0.1, CI95% [-0.14, -0.06] (pyment) and -0.29, CI95% [-0.33, -0.26] (brainageR).

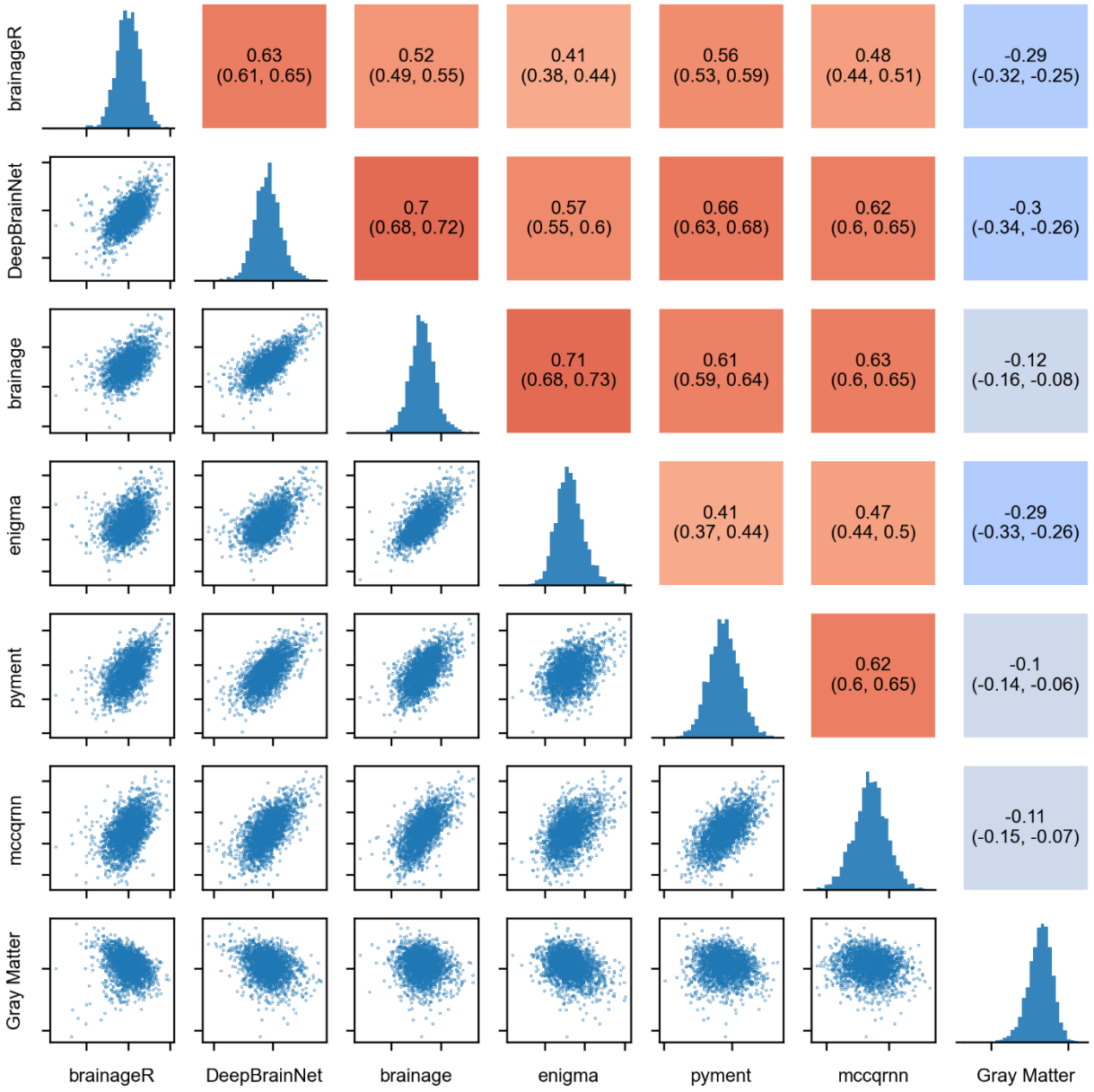

Figure 1: Scatterplot, histogram, and Pearson's correlation of the PADs at baseline across packages and normalized gray matter volume.

Differentiation between clinical groups at baseline

Overall, the mean PAD increased progressively from CN to MCI to AD groups, e.g. -2.17 years (95% CI =[-2.60, -1.74]), 0.09 years (95% CI = [-0.36, 0.53]), and 2.8 years (95% CI = [2.08, 3.53]) for brainageR. However, all models showed a systematic bias, with PAD values for CN

participants not centered around zero (first column in Table 2). Also, predictions had varying degrees of variance across packages, with enigma having the largest variance in predictions, and pyment the lowest (Figure 2A).

Comparing CN and AD groups, we observed the largest effect for the package brainage (Cohen's d = -0.95, 95% CI [-1.07, -0.82], $p < 0.001$), with AD patients showing PAD values .31 ± 0.31 years (mean ± SE) higher than CN participants. The smallest effect was found for the package mccqrnn (Cohen's d = -0.38, 95% CI [-0.50, -0.26], $p < 0.001$; $\beta = 3.04 \pm 0.27$ years).

For CN versus MCI, effect sizes were consistently smaller but remained significant across all packages. DeepBrainNet and brainage showed the strongest differentiation (DeepBrainNet: Cohen's d = -0.45, 95% CI [-0.54, -0.36], $p < 0.001$; brainage: Cohen's d = -0.49, 95% CI [-0.58, -0.40], $p < 0.001$). The differentiation between MCI and AD groups was largest in brainage (Cohen's d = -0.45, 95% CI [-0.56, -0.33], $p < 0.001$), while mccqrnn showed the weakest effect (Cohen's d = -0.11, 95% CI [-0.23, 0.00], $p < 0.001$).

For comparison, normalized gray matter volume showed similar or slightly larger effect sizes than the brain age packages, particularly between CN and AD groups (Cohen's d = 1.01, 95% CI [0.89, 1.14], $p < 0.001$).

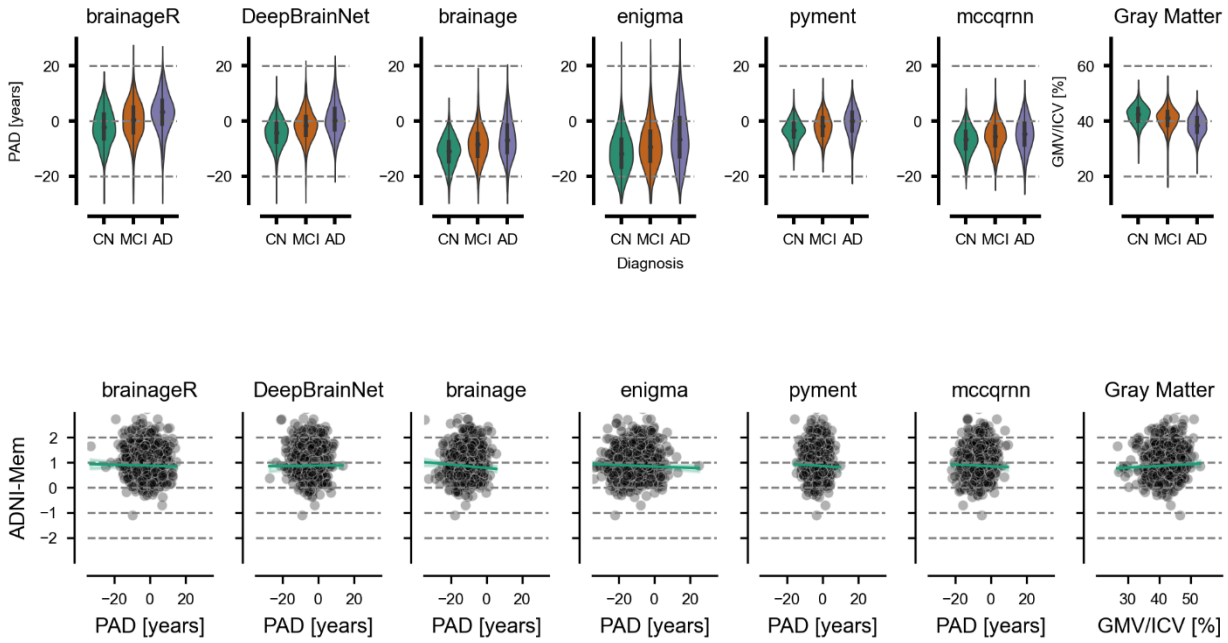

Figure 2: Cross-Sectional analysis of PAD. A) PAD and gray matter volume across cognitively normal and clinical groups. Cross sectional comparison of the predicted age deviation (PAD) across packages for cognitively normal (CN) individuals, patients with mild cognitive impairment (MCI), and patients with Alzheimer's disease (AD) for six brain age prediction packages and gray matter volume (used as a reference). Gray matter volume was normalized by the intracranial volume. **B)** Baseline PAD vs memory performance (ADNI-Mem) in CN individuals.

Table 2: Comparison of PAD and gray matter volume normalized by the intracranial volume between cognitively normal (CN), mild cognitively impaired (MCI), and Alzheimer's disease (AD) subjects. We report the mean PAD of the CN group. Further, we report the coefficient (Beta) of the linear model fitted with the covariates age, sex and AES. We further provide the effects size (Cohen's d) p-value of the mean difference (Beta). The mean and the effects size (Cohen's d) include the 95% confidence interval (95%CI), while we report the standard error (SE) for the Beta.

	CN	**CN vs AD**			**CN vs MCI**			**MCI vs AD**		
	Mean (95%CI)	**Beta (SE)**	**Cohen's d (95%CI)**	**p-value**	**Beta (SE)**	**Cohen's d (95%CI)**	**p-value**	**Beta (SE)**	**Cohen's d (95%CI)**	**p-value**
brainageR	-2.17 (-2.6, -1.74)	4.89 (0.41)	-0.73 (-0.86, -0.61)	<0.001	1.96 (0.32)	-0.32 (-0.41, -0.23)	<0.001	2.61 (0.43)	-0.37 (-0.48, -0.25)	<0.001
Deep BrainNet	-4.33 (-4.67, -3.99)	5.71 (0.3)	-0.78 (-0.9, -0.66)	<0.001	2.87 (0.23)	-0.45 (-0.54, -0.36)	<0.001	2.92 (0.29)	-0.33 (-0.45, -0.22)	<0.001
brainage	-11.17 (-11.54, -10.8)	5.31 (0.31)	-0.95 (-1.07, -0.82)	<0.001	2.45 (0.23)	-0.49 (-0.58, -0.4)	<0.001	2.86 (0.28)	-0.45 (-0.56, -0.33)	<0.001
enigma	-11.66 (-12.17, -11.15)	3.44 (0.24)	-0.71 (-0.83, -0.58)	<0.001	1.50 (0.18)	-0.35 (-0.44, -0.26)	<0.001	1.82 (0.24)	-0.33 (-0.45, -0.22)	<0.001
pyment	-3.38 (-3.63, -3.13)	6.96 (0.48)	-0.71 (-0.83, -0.59)	<0.001	2.93 (0.35)	-0.35 (-0.44, -0.26)	<0.001	3.96 (0.49)	-0.35 (-0.46, -0.23)	<0.001
mccqrnn	-6.93 (-7.25, -6.6)	3.04 (0.27)	-0.38 (-0.5, -0.26)	<0.001	1.50 (0.2)	-0.27 (-0.36, -0.18)	<0.001	1.39 (0.26)	-0.11 (-0.23, 0.0)	<0.001
Gray Matter	42.16 (41.91, 42.4)	-3.30 (0.22)	1.01 (0.89, 1.14)	<0.001	-1.03 (0.16)	0.35 (0.26, 0.44)	<0.001	-2.22 (0.22)	0.62 (0.5, 0.73)	<0.001

Association between memory performance and PAD at baseline

We further tested for an association between PAD and ADNI-Mem at baseline, hypothesizing that increased PAD would correlate with lower memory performance. We used linear models with age, sex, image quality, and years of education as covariates (Table 2, Figure 2B). CN participants, no significant associations were found between baseline PAD and memory performance across any of the brain age prediction packages (all $p>0.06$). However, significant associations between PAD and memory performance were observed in both MCI and AD groups (respectively all $p<0.001$ and all $p<0.039$). Gray matter volume normalized by intracranial volume showed a similar pattern to the brain age prediction models in all groups.

Table 3: Association between PAD and memory performance for cognitively normal (CN), mild cognitively impaired (MCI), and Alzheimer's disease (AD) subjects. We report the coefficient (Beta) and the corresponding p-value of the linear model fitted with the covariates age, sex, AES, and years of education. Gray matter volume normalized by the intracranial volume was added as reference.

	CN		MCI		AD	
	Beta (SE)	**p-value**	**Beta (SE)**	**p-value**	**Beta (SE)**	**p-value**
brainageR	-0.002 (0.003)	0.415	-0.02 (0.003)	<0.001	-0.011 (0.004)	0.005
DeepBrainNet	0.001 (0.004)	0.736	-0.042 (0.004)	<0.001	-0.033 (0.007)	<0.001
brainage	-0.007 (0.004)	0.06	-0.035 (0.004)	<0.001	-0.04 (0.007)	<0.001
enigma	-0.008 (0.005)	0.125	-0.043 (0.005)	<0.001	-0.026 (0.007)	<0.001
pyment	-0.002 (0.003)	0.496	-0.018 (0.002)	<0.001	-0.013 (0.003)	<0.001
mccqrnn	-0.003 (0.004)	0.489	-0.032 (0.004)	<0.001	-0.014 (0.007)	0.039
Gray Matter	0.008 (0.005)	0.161	0.033 (0.005)	<0.001	0.034 (0.007)	<0.001

Association between PAD and disease conversion

Among the 875 CN subjects, 65 converted to MCI or AD during this timeframe. We investigated whether baseline PAD could predict conversion from CN to MCI or AD within a four-year follow-up period using an IPCW logistic regression. The odds ratio (OR) for a unit-increase in PAD or normalized gray matter volume varied across methods, with three packages showing significant associations with conversion risk (Table 4). Among the evaluated packages, the strongest association was observed for brainage (OR = 1.15, 95% CI [1.04, 1.26], p = 0.005), followed by pyment (OR = 1.14, 95% CI [1.03, 1.27], p = 0.012). For these methods, each unit increase in PAD was associated with 14-15% higher odds of conversion from CN to MCI or AD. The remaining packages showed similar effects but did not reach statistical significance. For comparison, normalized gray matter volume showed a significant inverse association with conversion risk (OR = 0.89, 95% CI [0.81, 0.99], p = 0.028).

To contextualize these findings, we estimated conversion probabilities for subjects with PAD and gray matter volume at one standard deviation below the mean (μ -SD), at the mean (μ),

and one standard deviation above the mean (μ +SD). The general probability curves are presented in Figure 3A. Subjects with PAD = μ -SD showed estimated conversion probabilities ranging from 7-11% across methods. In contrast, subjects with PAD = μ +SD had higher conversion probabilities of 17-24%, with brainage showing the largest difference between low and high PAD values (7% versus 24%). Subjects with mean PAD values showed consistent conversion probabilities across methods (13-14%). The probability of converting based on gray matter volume followed a similar, although inverted trend (more gray matter is good, while a higher PAD is supposedly bad), with conversion probabilities of 19% at μ -SD, 13% at the mean, and 9% at μ +SD.

The discriminative ability of the models and gray matter volume was moderate, with AUCs ranging from 0.68 (brainageR, mccqrnn) to 0.73 (brainage). Brier scores were consistent across all methods (approximately 0.12), indicating similar calibration performance.

Table 4: Association between PAD and longitudinal disease conversion within four years from baseline. The odds-ratio and corresponding p-value are given for the coefficient. P(x) provides the probability of converting from CN to MCI or AD when a subject has a PAD of x ϵ { μ -SD, μ, μ +SD} at baseline. Further, the Brier score and the AUC are provided as overall performance measures.

Model	Odds (95%CI)	p-value (Odds)	P(x=-SD) (95%CI)	P(x=μ) (95%CI)	P(x=SD) (95%CI)	AUC (95%CI)	Brier (95%CI)
brainageR	1.04 (0.98,1.11)	0,188	0.11 (0.06,0.2)	0.14 (0.09,0.21)	0.18 (0.11,0.27)	0.68 (0.6,0.76)	0.12 (0.09,0.15)
DeepBrainNet	1.08 (0.96,1.21)	0,2	0.09 (0.04,0.19)	0.13 (0.08,0.19)	0.18 (0.1,0.3)	0.69 (0.61,0.78)	0.12 (0.09,0.15)
brainage	1.15 (1.04,1.26)	0,005	0.07 (0.03,0.13)	0.13 (0.09,0.2)	0.24 (0.14,0.39)	0.73 (0.65,0.81)	0.12 (0.09,0.14)
enigma	1.07 (1.02,1.12)	0,006	0.08 (0.05,0.15)	0.13 (0.09,0.2)	0.2 (0.12,0.3)	0.71 (0.63,0.79)	0.12 (0.09,0.14)
pyment	1.14 (1.03,1.27)	0,012	0.08 (0.04,0.15)	0.13 (0.08,0.19)	0.2 (0.12,0.3)	0.72 (0.65,0.8)	0.11 (0.09,0.14)
mccqrnn	1.06 (0.98,1.14)	0,162	0.11 (0.06,0.19)	0.14 (0.09,0.2)	0.17 (0.11,0.27)	0.68 (0.6,0.76)	0.12 (0.1,0.15)
Gray Matter	0.89 (0.81,0.99)	0,028	0.19 (0.12,0.28)	0.13 (0.09,0.2)	0.09 (0.05,0.17)	0.69 (0.61,0.77)	0.12 (0.09,0.14)

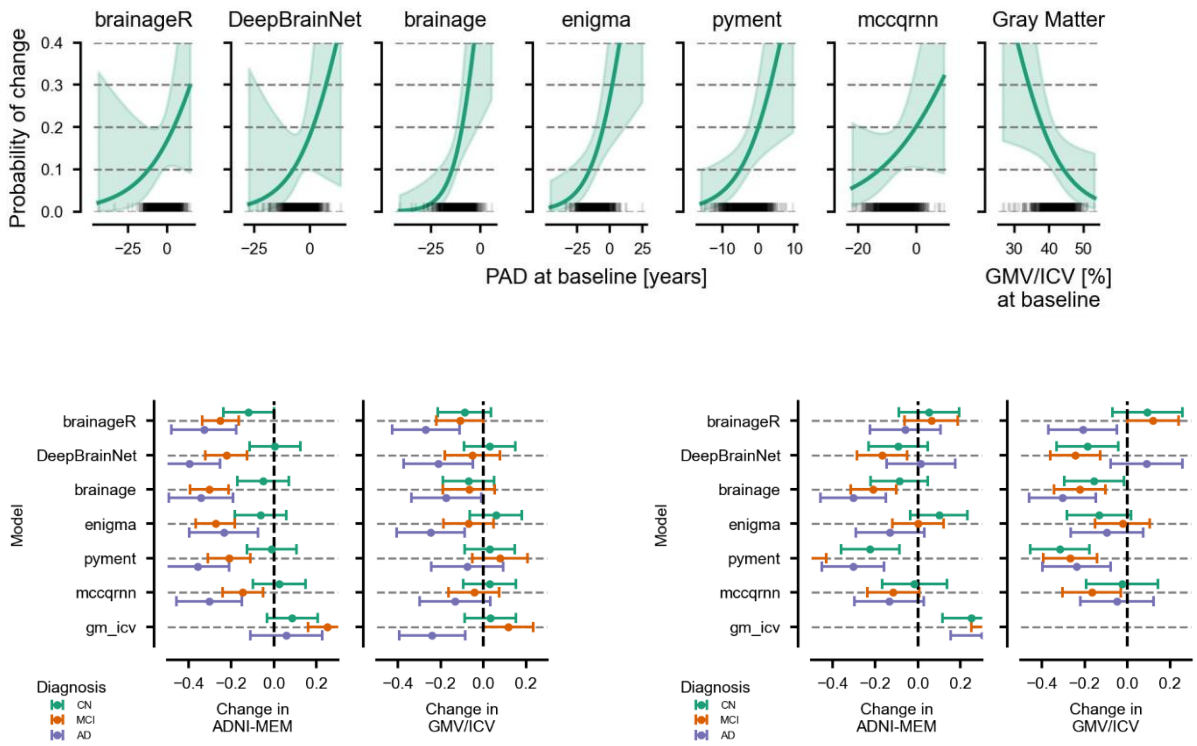

Figure 3: Longitudinal Analysis of PAD. A) The probability of converting from CN to MCI or AD within four years from baseline when having a specific PAD. The curves are computed with the covariates fixed to their respective means, and sex was set to female. The shaded area visualizes the 95% confidence interval. **B)** The association between PAD at baseline and change within four years in normalized gray matter volume or ADNI-Mem. **C)** The association between change in PAD and change in normalized gray matter volume or ADNI-Mem within four years.

Association between PAD and rate of decline in gray matter volume and memory function

To investigate the clinical validity of brain age predictions, we examined associations between PAD at baseline and subsequent changes in memory performance and normalized gray matter volume across diagnostic groups (Fig 3B).

The association between the PAD at baseline and memory decline varies substantially by diagnostic group and model. In CN individuals, associations with memory decline were small, with brainageR showing the strongest correlation among the packages (r = -0.12, 95% CI [-0.24, 0.0], p = 0.05). In MCI patients, all models showed significant, but modest, correlations with future memory decline, ranging from r = -0.30 (brainage) to r = -0.15 (mccqrnn). The strongest associations were observed in AD patients, where DeepBrainNet

achieved the highest correlation ($r = -0.40$, 95% CI [-0.54, -0.25], $p < 0.001$), and enigma the lowest ($r = -0.24$, 95% CI: [-0.4, -0.08], $p < 0.001$).

Baseline PAD showed no significant associations with gray matter atrophy in CN or MCI groups across all models. In AD patients, significant but modest correlations were observed for brainageR ($r = -0.27$, 95% CI [-0.45, -0.09], $p = 0.009$), enigma ($r = -0.25$, 95% CI: [-0.42, -0.08], $p = 0.008$), and DeepBrainNet ($r = -0.21$, 95% CI [-0.39, -0.03], $p = 0.026$).

To contextualize these findings, we examined baseline normalized gray matter volume as a reference measure, showing a similar, though inverted, pattern of associations with memory decline compared to the brain age models in CN subjects ($r = 0.09$, 95% CI [-0.03, 0.20], $p = 0.158$), and MCI patients ($r = 0.25$, 95% CI [0.16, 0.34], $p < 0.001$). There was no significant association in AD patients ($r = 0.06$, 95% CI: [-0.11, 0.22], $p = 0.493$). Further, normalized gray matter volume at baseline showed a similar pattern of association with gray matter atrophy compared to brainageR, and other models, in CN subjects and MCI and AD patients.

Association between changes in the PAD and decline in gray matter volume and memory function

Further, the analyses from the previous subsection were repeated using the change in PAD within four years (instead of baseline PAD) to investigate the association between changes in the PAD and changes in both memory performance and normalized gray matter volume across diagnostic groups.

The association between a change in the PAD and changes in memory performance varied markedly across diagnostic groups and models. The strongest associations were found using pyment, which showed significant, but again only modest, correlations across all diagnostic groups (CN: $r = -0.22$, 95% CI [-0.36, -0.09], $p = 0.002$; MCI: $r = -0.52$, 95% CI [-0.61, -0.43], $p < 0.001$; AD: $r = -0.30$, 95% CI [-0.45, -0.16], $p < 0.001$). The package brainage also demonstrated consistent negative correlations, particularly in AD ($r = -0.30$, 95% CI [-0.46, -0.15], $p = 0.001$) and MCI groups ($r = -0.21$, 95% CI [-0.32, -0.10], $p < 0.001$).

For the change in normalized gray matter volume, associations varied markedly across diagnostic groups and models as well. Pyment again demonstrated associations across all diagnostic groups (AD: $r = -0.24$, 95% CI [-0.40, -0.08], $p = 0.006$; CN: $r = -0.32$, 95% CI [-0.46, -0.18], $p < 0.001$; MCI: $r = -0.27$, 95% CI [-0.40, -0.14], $p < 0.001$). DeepBrainNet and brainage also showed significant, but small, correlations, particularly in the MCI group ($r = -0.24$, 95% CI [-0.36, -0.13], $p < 0.001$ and $r = -0.22$, 95% CI [-0.34, -0.10], $p < 0.001$, respectively).

The reference model, which directly quantifies gray matter volume changes, demonstrated significant positive correlations with memory decline across all groups (AD: $r = 0.31$, 95% CI [0.15, 0.46], $p = 0.001$; CN: $r = 0.25$, 95% CI [0.12, 0.38], $p = 0.001$; MCI: $r = 0.37$, 95% CI [0.25, 0.48], $p < 0.001$).

Discussion

Our evaluation of six publicly available brain age prediction packages showed substantial limitations when it comes to clinical validity. While the assessed packages showed some ability to distinguish clinical groups and predict disease conversion, we found low convergent validity across different packages and generally only weak to modest associations with established markers of brain aging, such as memory decline and gray matter atrophy. Further, our analysis demonstrated that simple normalized gray matter volume measurements performed comparably to the more complex brain age prediction packages, putting the added value of existing brain age models as descriptive, diagnostic or prognostic tools into question.

Brain age prediction from structural MRI has recently received great interest, with numerous research groups making their packages publicly available. Initial results showing that brain age could differ between clinical groups [10] and was predictive of mortality [12] suggested that outcomes could potentially be developed into useful biomarkers of brain aging and general health. However, neither the concept nor the tools have been thoroughly validated in a comparative setting, and increasing evidence suggests that current brain age estimates are not strongly indicative of the rate of biological aging of a person [19,20]. Here, we expand our

previous comparative reliability-assessment of brain age prediction packages[16] and attempted to investigate clinical validity of brain age packages in a large longitudinal dataset with both cognitively normal individuals, and individuals with age-related diseases (MCI, AD).

Brain age prediction packages vary in their preprocessing steps (registration to a common space, segmentation, surface-based analysis, etc.), the model architecture (statistical learning vs. deep learning), the type of input features (voxel-based vs regional, whole brain or restricted to gray matter), and the distribution of the input data (age, gender, ethnicity, etc.). Despite this heterogeneity, they are commonly referred to as "brain age" without any distinction. Here we show that there is limited convergent validity between packages, with the largest correlation between packages being r=0.71 (95% CI [0.68, 0.73]), suggesting that this notion is not entirely correct. Most likely, the low association between predictions from different packages is due to differences in training features and the distribution of the training data.

Further, the substantial variation and systematic bias (i.e. a non-zero mean) observed in brain age predictions and PADs among CN individuals challenges the interpretability and utility of PAD as a biomarker. Brain age prediction aims to provide information on the deviation from population norms, where a PAD of zero should theoretically represent age-appropriate brain health. Our analysis shows that the PAD, while correctly ranking the clinical groups for increasing disease severity, is set on arbitrary scales. More specifically,a PAD of zero does not indicate a cognitively intact or "healthy" brain. Further, all packages have different degrees of variance, making an interpretation of the PAD difficult: for instance, a PAD of ten years may fall within normal variation when calculated using enigma, but could indicate significant deviation from the norm when derived from pyment. These limitations question the validity of brain age as an interpretable, accessible, and meaningful biomarker for aging and brain health.

PAD showed no significant association with memory performance in CN individuals. While PAD correlated with decreased memory performance in MCI and AD groups for many of the evaluated packages, it should ideally also be indicative of memory performance in

individuals with no known neurological conditions. Of note, the composite score ADNI-Mem was not age-normalized [21]. PAD at baseline in CN individuals was however associated with conversion to MCI or AD, but the AUC was low (68% to 73%). These results indicate that there is littleprognostic gain to be had from brain age models, above the covariates of age, sex, and image quality.

PAD at baseline was not significantly associated with memory decline or gray matter atrophy within the four-year follow-up period in healthy individuals. If brain age is to function as a marker biological aging, we would expect individuals with an increased PAD - those who age supposedly faster – to show accelerated memory decline and gray matter atrophy. In the MCI and AD groups, the associations were larger and statistically significant, consistent with the assumption that these subjects are, by definition, on a neurodegenerative trajectory. However, the weak non-significant associations between brain age and decline in gray matter volume and memory performance in non-diseased individuals aligns well with the results in Vidal-Pineiro et al. 2021 [19] and Korbmacher et al. 2024 [20], who showed a similar trend. However, the pyment package did display that changes in PAD were associated with changes in gray matter volume and memory performance across all groups. The association between change in PAD and changes in gray matter volume or memory performance suggest that PAD is related to the current health state of the brain, although not associated with change in the future. As suggested by Smith et al. 2025 [38], it is difficult, or even impossible, to disentangle variations in PAD caused by biological aging, measurement noise, and congenital factors in cross-sectional data using the conventional framework of predicting chronological age from brain features. Therefore, focusing on longitudinal changes in gray matter, rather than relying solely on cross-sectional data, may be advantageous for future development of brain age models with higher prognostic value [39,40].

Gray matter volume showed comparable performance to more complex brain age prediction models across all our analyses. Thus, instead of using a constructed biomarker of aging, normative models using gray matter volume might be clinically useful, especially since gray matter volume directly reflects brain biology. Gray matter volume has been linked to adverse outcomes and has been used as exploratory endpoint in clinical trials [41], since more gray

matter is generally perceived as being advantageous, while accelerated atrophy is linked to neurodegeneration. Normative models of structural decline enable explicit assessment of deviations from population norms [7], whereas brain age as a concept only captures this implicitly. Moreover, deviations in normative models in metric units (liter) are directly interpretable, unlike brain age predictions where the clinical significance of a specific PAD value remains unclear. However, the brain age framework does offer the advantage to use non-linear models on large imaging datasets that not only focus on gray matter volume but also account for more complex patterns of aging in the brain.

Several limitations of our study should be noted. We focused on a period up to four years for the longitudinal analysis, and in, especially in CN individuals, a longer time interval might be necessary to detect changes in gray matter volume and memory performance. However, for brain age to be useful as a potential surrogate biomarker in clinical trials of, e.g., neuroprotective or preventive interventions, they should ideally be able to detect changes within four years. Another limitation in the available data was that only a few subjects converted from CN to MCI or AD within the four-year period, which might have contributed to the poor performance of brain age in predicting conversion. Further, our reported performance metrics likely represent an upper bound of the out-of-sample performance since we did not compute the AUC using on hold-out validation set. We expect the difference to out-of-sample estimates to be small as the number of parameters in the predictive model was small compared to the number of observations and events (respectively 6, 875, and 65). We also selected to focus specifically on memory function by using the ADNI-Mem composite score. Hence, the lack of significant associations between ADNI-Mem and PAD estimates can't be extrapolated to cognition in general. Finally, we only include a subset of existing brain age estimation packages in this analysis, as new packages are developed and published on a regular basis. We therefore cannot claim to provide an all-encompassing comparison of the clinical validity of all existing brain age models in the field. Still, the implemented packages provide a wide variety of pre-processing pipelines, input features, and model architecture. Since no package consistently outperformed the other packages, we suggest that there could be more fundamental considerations - such as the use of cross-

sectional data in general - to address for the field, rather than choice of model architecture and input features. We also did not retrain any of the algorithms on similar training data to make a full head-to-head comparison of the different model architectures but rather compared the clinical validity of existing publicly available packages when used as “of-the-shelve” estimation methods.

In conclusion, baseline brain age performs similarly to grey matter volume in terms of prognostic performance and does not strongly correlate with clinical outcomes such as memory decline or gray matter atrophy. One alternative for future development efforts of image-based biomarkers of aging would be to model longitudinal trajectories, rather than relying solely on cross-sectional data for predictions.

Acknowledgement

This work was supported by a Longevity Impetus Grant from the Norm Group. PPS was supported by a grant from the Swedish Brain Foundation (PD2024-0444). Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

Competing interests

The authors declare no competing interests.

References

1. Mattson, M. P. & Arumugam, T. V. Hallmarks of Brain Aging: Adaptive and Pathological Modification by Metabolic States. *Cell Metab.* **27**, 1176–1199 (2018).
2. Niccoli, T. & Partridge, L. Ageing as a Risk Factor for Disease. *Curr. Biol.* **22**, R741–R752 (2012).
3. Hou, Y. *et al.* Ageing as a risk factor for neurodegenerative disease. *Nat. Rev. Neurol.* **15**, 565–581 (2019).
4. Kennedy, B. K. *et al.* Geroscience: Linking Aging to Chronic Disease. *Cell* **159**, 709–713 (2014).
5. Higgins-Chen, A. T., Thrush, K. L. & Levine, M. E. Aging biomarkers and the brain. *Semin. Cell Dev. Biol.* **116**, 180–193 (2021).
6. Moqri, M. *et al.* Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* **186**, 3758–3775 (2023).
7. Bethlehem, R. A. I. *et al.* Brain charts for the human lifespan. *Nature* (2022) doi:10.1038/s41586-022-04554-y.
8. Rutherford, S. *et al.* Charting brain growth and aging at high spatial precision. *eLife* **11**, e72904 (2022).
9. Franke, K., Ziegler, G., Klöppel, S. & Gaser, C. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage* **50**, 883–892 (2010).
10. Kaufmann, T. *et al.* Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat. Neurosci.* **22**, 1617–1623 (2019).
11. Bashyam, V. M. *et al.* MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain* **143**, 2312–2324 (2020).
12. Cole, J. H. *et al.* Brain age predicts mortality. *Mol. Psychiatry* **23**, 1385–1392 (2018).

13. Hahn, T. *et al*. An uncertainty-aware, shareable, and transparent neural network architecture for brain-age modeling. *Sci. Adv.* **8**, eabg9471 (2022).

14. Han, L. K. M. *et al*. Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group. *Mol. Psychiatry* **26**, 5124–5139 (2021).

15. Leonardsen, E. H. *et al*. Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage* **256**, 119210 (2022).

16. Dörfel, R. P. *et al*. Prediction of brain age using structural magnetic resonance imaging: A comparison of accuracy and test–retest reliability of publicly available software packages. *Hum. Brain Mapp.* **44**, 6139–6148 (2023).

17. Elliott, M. L. *et al*. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. *Mol. Psychiatry* **26**, 3829–3838 (2021).

18. Gautherot, M. *et al*. Longitudinal Analysis of Brain-Predicted Age in Amnestic and Non-amnestic Sporadic Early-Onset Alzheimer's Disease. *Front. Aging Neurosci.* **13**, 729635 (2021).

19. Vidal-Pineiro, D. *et al*. Individual variations in 'brain age' relate to early-life factors more than to longitudinal brain change. *eLife* **10**, e69995 (2021).

20. Korbmacher, M. *et al*. Cross-sectional brain age assessments are limited in predicting future brain change. 2024.09.11.612523 Preprint at https://doi.org/10.1101/2024.09.11.612523 (2024).

21. Crane, P. K. *et al*. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging Behav.* **6**, 502–516 (2012).

22. Routier, A. *et al*. Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies. *Front. Neuroinformatics* **15**, (2021).

23. Gorgolewski, K. J. *et al*. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **3**, 160044 (2016).

24. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012).

25. Ashburner, J. SPM: A history. *NeuroImage* **62**, 791–800 (2012).

26. Esteban, O. *et al.* MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. 21 (2017).

27. Deep-MI/fsqc. Deep Medical Imaging Lab (PI Reuter) (2024).

28. Zacà, D., Hasson, U., Minati, L. & Jovicich, J. Method for retrospective estimation of natural head movement during structural MRI. *J. Magn. Reson. Imaging* **48**, 927–937 (2018).

29. Marchetto, E., Eichhorn, H., Gallichan, D., Schnabel, J. A. & Ganz, M. Agreement of Image Quality Metrics with Radiological Evaluation in the Presence of Motion Artifacts. Preprint at https://doi.org/10.48550/arXiv.2412.18389 (2024).

30. Cole, J. james-cole/brainageR: brainageR v2.1. (2019) doi:10.5281/ZENODO.3476365.

31. Bacas, E. *et al.* Probing multiple algorithms to calculate brain age: Examining reliability, relations with demographics, and predictive power. *Hum. Brain Mapp.* hbm.26292 (2023) doi:10.1002/hbm.26292.

32. Moqadam, R., Dadar, M. & Zeighami, Y. Investigating the impact of motion in the scanner on brain age predictions. *Imaging Neurosci.* **2**, 1–21 (2024).

33. Blanche, P. F., Holt, A. & Scheike, T. On logistic regression with right censored data, with or without competing risks, and its use for estimating treatment effects. *Lifetime Data Anal.* **29**, 441–482 (2023).

34. Gerds, T. A., Ohlendorff, J. S. & Ozenne, B. riskRegression: Risk regression models and prediction scores for survival analysis with competing risks. (2023).

35. Blanche, P. *et al.* Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* **71**, 102–113 (2015).

36. Bove, D. S. *et al.* mmrm: Mixed Models for Repeated Measures. (2024).

37. Ozenne, B. & Forman, J. LMMstar: Repeated Measurement Models for Discrete Times. (2024).

38. Smith, S. M., Miller, K. L. & Nichols, T. E. Characterising ongoing brain aging from cross-sectional data. 2025.01.21.634044 Preprint at https://doi.org/10.1101/2025.01.21.634044 (2025).

39. Roe, J. M. *et al*. Brain change trajectories in healthy adults correlate with Alzheimer's related genetic variation and memory decline across life. *Nat*. *Commun*. **15**, 10651 (2024).

40. Kim, H. *et al*. Learning-based inference of longitudinal image changes: Applications in embryo development, wound healing, and aging brain. *Proc*. *Natl*. *Acad*. *Sci*. **122**, e2411492122 (2025).

41. Edison, P. *et al*. Evaluation of liraglutide in the treatment of Alzheimer's disease. *Alzheimers Dement*. **17**, e057848 (2021).