

Probing multiple algorithms to calculate brain age: Examining reliability, relations with demographics, and predictive power

Eva Bacas¹  | Isabella Kahhalé^{1,2}  | Pradeep R. Raamana³ | Julian B. Pablo¹ | Apurva S. Anand¹ | Jamie L. Hanson^{1,2}

¹Learning, Research, and Development Center, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

²Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

³Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence

Jamie Hanson, Learning, Research, and Development Center, University of Pittsburgh, Pittsburgh, PA, USA.

Email: jamie.hanson@pitt.edu

Funding information

the National Science Foundation, Grant/Award Numbers: IIS-1912270, IIS-1636893, BCS-1734853

Abstract

The calculation of so-called “brain age” from structural MRIs has been an emerging biomarker in aging research. Data suggests that discrepancies between chronological age and the predicted age of the brain may be predictive of mortality and morbidity (for review, see Cole, Marioni, Harris, & Deary, 2019). However, with these promising results come technical complexities of how to calculate brain age. Various groups have deployed methods leveraging different statistical approaches, often crafting novel algorithms for assessing this biomarker derived from structural MRIs. There remain many open questions about the reliability, collinearity, and predictive power of different algorithms. Here, we complete a rigorous systematic comparison of three commonly used, previously published brain age algorithms (XGBoost, brainageR, and DeepBrainNet) to serve as a foundation for future applied research. First, using multiple datasets with repeated structural MRI scans, we calculated two metrics of reliability (intraclass correlations and Bland–Altman bias). We then considered correlations between brain age variables, chronological age, biological sex, and image quality. We also calculated the magnitude of collinearity between approaches. Finally, we used machine learning approaches to identify significant predictors across brain age algorithms related to clinical diagnoses of cognitive impairment. Using a large sample ($N = 2557$), we find all three commonly used brain age algorithms demonstrate excellent reliability ($r > .9$). We also note that brainageR and DeepBrainNet are reasonably correlated with one another, and that the XGBoost brain age is strongly related to image quality. Finally, and notably, we find that XGBoost brain age calculations were more sensitive to the detection of clinical diagnoses of cognitive impairment. We close this work with recommendations for future research studies focused on brain age.

KEY WORDS

aging, brain age, neuroscience, reliability, statistics

1 | INTRODUCTION

Aging is a process that involves multiple factors, with the causes of aging still poorly understood. As the human lifespan has increased, a greater proportion of the global population is beginning to show age-associated functional declines and disease (Roser et al., 2013). It is therefore critical for us as a society to delineate the pathways more richly to longevity, as well as age-associated deteriorations. To understand and predict these age-related differences in health and disease, novel approaches have emerged attempting to quantify an individual's "biological age" and its divergence from chronological age. These have included examination of telomeres, epigenetics, and other molecular and cellular elements (Aubert & Lansdorp, 2008; Fraga & Esteller, 2007; Mather et al., 2011; Pal & Tyler, 2016). Moving beyond molecular and cellular assays, calculation of so-called "brain age" is a novel biomarker that has recently emerged in aging research. This type of biological age is estimated from neuroimaging scans using data science and machine learning algorithms. Using large datasets where age and neuroimaging scans are available, age can be estimated on new participants only using neuroimaging data. Interestingly, these estimates, typically calculated from structural MRI scans, can diverge from participants' chronological age, suggesting potential alterations in someone's biological age. This is important since we know that the aging process does not affect people uniformly, both within and between individuals. Given the brain's central role in regulating behavior and multiple neuroendocrine processes, metrics of brain age may be a particularly predictive indicator of age-related mortality and morbidity. By understanding brain age, we may be able to build more comprehensive models of the biological aging process, as well as biomarkers to predict important clinical outcomes (Cole et al., 2019).

Surveying work on brain age to date, different research groups have focused on the discrepancies between brain age and chronological age. In this work, we refer to the difference between predicted brain age and chronological age as *brain age delta*, adopting the terminology used in Bashyam et al. (2021). Of note, other scholars use varying terms to refer to this difference, including brain age gap and brain-PAD (Cole et al., 2018; Kaufmann et al., 2019). This approach presumes that larger differences reflect poorer health, and suggestive evidence is growing to validate this idea. For example, Cole et al. (2018) found that greater brain age delta derived from structural MRIs was associated with a number of physiological measures related to senility, including weaker grip strength, poorer lung function, and longer times to walk a short distance. Particularly notable, greater brain age delta-derived from MRI scans of participants in their early 70s – was related to mortality years later. In this sample, older brain age was associated with reduced lifespan, with each additional year of brain aging being related to a ~6% increase in the likelihood of death between the ages of 72 and 80. In addition to aspects of senility and mortality, brain age has also been connected to several psychiatric and neurodegenerative conditions. Patients with serious psychiatric and neurological disorders show increased brain age, including patients with schizophrenia (Shahab et al., 2019), depression (Kaufmann et al., 2019), borderline personality disorder (Koutsouleris

et al., 2014), and Alzheimer's disease (Franke & Gaser, 2012, 2019; Gaser et al., 2013; Ly et al., 2020).

While promising, research deploying these approaches is still in its infancy. There are multiple unique algorithms to calculate brain age developed by pioneering groups. In previous brain age research, it is common to develop and apply newly developed algorithms in the same research report. From a technical perspective, the calculation of brain age from structural MRIs uses a wide variety of machine learning approaches, including Gaussian process regression, regularizing gradient boosting, and more recently, deep learning models. This has led to a host of debates in the field about how to conceptualize and validate these different algorithms (see Bashyam et al., 2021; Hahn et al., 2021). Adding to this complexity, the diverse aims and applications of brain age studies result in a variety of benchmarks. For example, one could examine the correlation between brain age and chronological age, different estimates of error in prediction (i.e., mean squared error), or the ability of algorithms to identify different age-related conditions and declines (e.g., early signs of Alzheimer's disease). Examined collectively, many open questions exist regarding the strengths and limitations of different algorithmic derivations of brain age.

Here, our aim was to provide a systematic comparison of three major brain age algorithms to serve as a point of reference for future applied research. To compare the algorithms, we looked at predictive power, reliability, and noise sensitivity. As previous research has examined relationships between brain age delta and age-related decline, we focused on predictive power as it relates to cognitive impairments. Reliability is a relevant point of comparison for researchers working with longitudinal data, and noise sensitivity is important for researchers working with noisy data or populations that typically produce noisy data (e.g. neuropsychiatric patients, children) (Wylie et al., 2014).

Related to intra-algorithm reliability, we use intraclass correlation and Bland–Altman bias metrics to compare algorithmic performance across repeated MRI scans of the same individuals. Given validation efforts completed during the initial development of these algorithms, we predicted high reliability of brain age estimates across each approach. We also examined correlations between different algorithms to see the consistency of results across approaches. To understand the influence of noise in MRI images, we examined relations between image quality and brain age calculation. We examined reliability and noise sensitivity in an older as well as a young sample of participants. This was motivated by the fact that brain age algorithms are typically developed in older age samples, but now are being applied to young participants (e.g. Keding et al., 2021). Given recent publications from our group (Gilmore et al., 2021), we predicted that image quality would be negatively related to brain age estimates, with greater brain age being found in lower quality scans. Finally, related to predictive power, we used a data-driven, machine learning approach to identify significant predictors across brain age algorithms related to a commonly used outcome, clinical diagnosis of cognitive impairment. For this question, we expected a greater brain age delta to be related to cognitive impairment. However, for each area of investigation, we

did not have specific predictions about the superiority of one algorithm over the others.

2 | METHODS

2.1 | Datasets

For our analyses, we examined three large open-access MRI datasets, including participants 19–100 years of age (*Analytic N* = 2557). Distribution of ages is visualized in Figure 1. When selecting datasets for analysis, we purposefully excluded certain datasets. A major concern in the development of statistical models is the issue of overfitting, when a model performs well on the training data but fails to generalize to new data (Ying, 2019). For this reason, we excluded all datasets used to train relevant brain age algorithms, excluding approximately 60 datasets in total. To examine the reliability of brain age metrics, we chose two datasets with repeated MRI scans: Amsterdam Open MRI Collection (AOMIC) and Open Access Series of Imaging Studies (OASIS). Many other open-access MRI datasets (e.g., Brain Genome Superstruct Project; Nathan Kline Institute Rockland Sample) do not have repeated structural MRI scans, or were excluded due to use in the development of relevant brain age algorithms. We also examined a third dataset that had reasonable variability in age (Human Connectome Project-Aging [HCP-A]). While HCP-A does not have repeated imaging assessments, this choice was motivated by the project's improved methods for data acquisition (as opposed to "standard" 3 T structural sequences). All data were collected on 3-Tesla (3 T) MRI scanners. We specifically focused on 3 T MRIs, as 3 T scanners are the most commonly-used research scanners and are also used in many ongoing public-access and large-scale neuroimaging initiatives (e.g., Human Connectome Project; ABCD Study®). Basic descriptions of the datasets are found in the following section. Additional details of the data collection are summarized in our supplemental materials.

2.2 | Amsterdam open MRI collection

AOMIC is an open-access neuroimaging dataset including structural and functional MRI scans (Snoek et al., 2021). Here, we analyze the "ID1000" subset of the data which comprised healthy young adults aged 19–26 ($N = 928$) scanned between 2010 and 2012 at University of Amsterdam. Participants were recruited from the general Dutch population, with efforts to recruit from a variety of educational backgrounds. Each participant was scanned three times in a single session using the same imaging parameters. Specifically, MR images were acquired with a Phillips Intera 3 T scanner. T1-weighted MR images were acquired using a sagittal 3D-MPRAGE sequence. In addition to MRI scanning, participants also completed many well-validated self-report scales and behavioral assessments, including measures of cognitive ability, personality, and motivation. Additional details of the data collection are summarized in our Supplemental Materials (S1).

2.3 | Open access series of imaging studies

OASIS is a multimodal neuroimaging project centralized at Washington University in St. Louis (LaMontagne et al., 2019). For our work, we used OASIS-3, a dataset of normally aging and Alzheimer's disease patients ($N = 1098$) aged 42–95. Six-hundred and five of these participants were neurologically healthy, while 493 participants presented with mild-cognitive impairment, Alzheimer's disease, and other neurological conditions of concern. Participants were recruited from other ongoing projects at Washington University in St. Louis focused on Alzheimer's and aging. For our work, we used data collected on Siemens TIM Trio 3 T scanners. A subset of OASIS were scanned at multiple timepoints months or years apart ("sessions") and some participants were scanned multiple times in a single session ("runs"). We included only participants with multiple runs per session. This eliminated 329 participants.

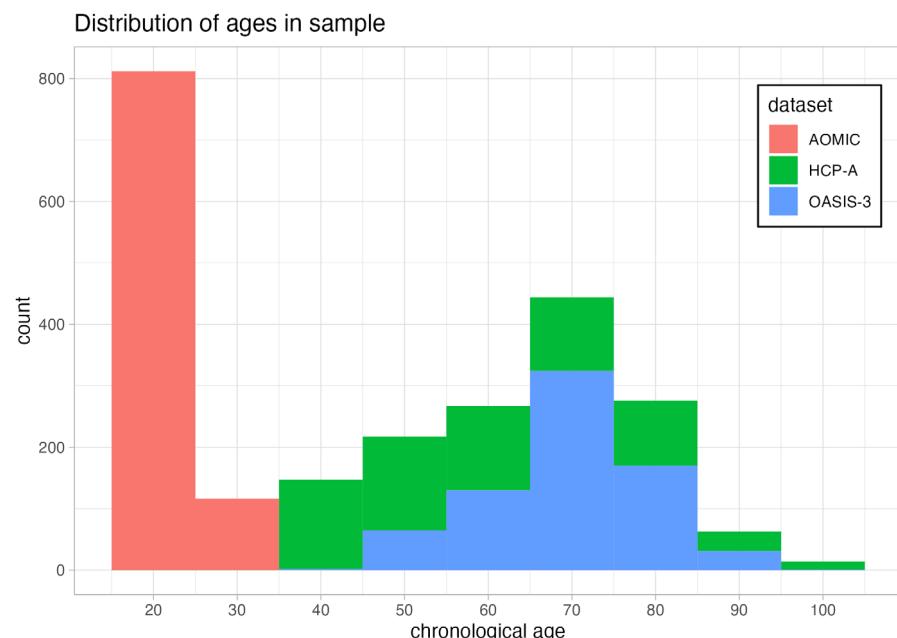


FIGURE 1 Distribution of participant age for the different projects we leveraged in our analyses. The horizontal axis depicts participant age in years, while the vertical axis shows the number of participants within a given age bin. Each dataset is shown in a different color, with AOMIC shown in red, HCP-A shown in green, and OASIS-3 shown in blue.

2.4 | Human connectome project in aging

HCP-A is an ongoing project collecting MRI data from 1200 healthy adults greater than 36 years of age (Bookheimer et al., 2019). Participants were recruited using flyers, advertisements, and outreach at community centers, with efforts to recruit a balanced number of participants across socioeconomic classes. Data collection was completed at 4 research sites: Massachusetts General Hospital, University of California at Los Angeles, University of Minnesota, Washington University in St. Louis. For this work, our sample was composed of 725 participants aged 36–100. MR images at all sites were collected using matched Siemens Prisma 3 T scanners. T1-weighted MR images were collected using a multi-echo sagittal 3D-MPRAGE sequence. Participants also completed many well-validated self-report scales and behavioral assessments, including measures of cognitive ability, personality, and mental health.

2.4.1 | Assessment of MRI image quality

Past work from our research group (Gilmore et al., 2021) has found that T1-weighted image quality is related to volumetric measures from commonly used morphometric tools suites (e.g. Freesurfer). We therefore assessed image quality to: (1) exclude particularly high-motion scans; and (2) investigate the impact of image quality on brain age estimates. To assess MRI quality, we generated a quantitative metric (“CAT12 score”) using the Computational Anatomy Toolbox 12 (CAT12). This metric considers four summary measures of image quality: noise-to-contrast ratio, coefficient of joint variation, inhomogeneity-to-contrast ratio, and root-mean-squared voxel resolution. CAT12 normalizes and combines these measures using a kappa statistic-based framework. The score is a value from 0 to 1, with 0 being the lowest quality and 1 being the highest quality. Informed by our past work (e.g., Gilmore et al., 2021), we excluded all lower quality scans, specifically with CAT12 scores <0.8.

2.4.2 | Analytic sample and brain age algorithms

After removing data with quality and preprocessing issues, our total analytic sample was 2557 participants, with 928 subjects from AOMIC, 705 subjects from HCP-A, and 584 subjects from OASIS. This allows for examination of a younger cohort (AOMIC; 19–26 years of age) and two middle-aged and older cohorts (over 36 years of age); however, and of note, there were no participants between the ages of 26 and 36. This was due to constraints on data selection, described previously in the *Datasets* section above.

We tested three brain age algorithms on our dataset: Kaufmann et al. (2019) (referred to as “XGBoost”), Cole et al. (2018) (referred to as “brainageR”), and Bashyam et al. (2021) (referred to as “DeepBrainNet”). We selected these three algorithms based on: open code, a large sample size in the training dataset (>2000 participants), diversity of machine learning models employed, and lastly, common use in

brain age research (as indexed by citations of the work). All of the papers connected to the algorithms have been cited more than 50 times per year since their year of publication, suggesting high adoption by neuroimaging researchers. DeepBrainNet and brainageR operate on raw T1-weighted MRI scans, and XGBoost requires pre-processing using Freesurfer (Fischl, 2012), an open source MRI processing software package. We provide brief summaries of the algorithms below. For detailed descriptions of model structure, please see the original papers cited here. Below, we briefly outline the varying benchmarks used in original publications, noting correlations between brain age and chronological age when available. Of note, past work has used many different statistical tests and validation approaches (e.g., mean squared error, accuracy), making exact comparisons between algorithms inaccurate or not possible. As all algorithms were built using training data from multiple datasets, sites, and scanners, we assumed that the models would be able to handle the effect of multiple sites. We confirmed this by running a linear model looking at the effects of site on brain age delta; this model can be found in the supplement (S4).

2.4.3 | XGBoost brain age algorithm

XGBoost uses gradient tree boosting to predict brain age based on 1118 features extracted using Freesurfer (Kaufmann et al., 2019). These features consist of thickness, area, and volume measurements from a multimodal parcellation of the cerebral cortex, cerebellum, and subcortex. Relevant code is available at: <https://github.com/tobias-kaufmann/brainage>. This algorithm was trained on a large and diverse sample ($N = 39,827$, female = 18,990). The sample was made up of healthy controls aged 3–89 drawn from 42 different datasets. All training data passed automatic quality control procedures. To account for potential variation, Kaufmann et al. trained separate models for male and female brain age. Using five-fold cross-validation, XGBoost produced strong correlations between brain age and chronological age (male: $r = 0.93$; female: $r = 0.94$). Kaufmann et al. tested the model on subjects with psychopathology, and calculated standardized mean difference (Cohen's d) for each group compared to a matched healthy control, finding significant differences for dementia, MCI, schizophrenia, and bipolar disorder. We deployed this algorithm by first completing standard processing approaches in Freesurfer 7.1 (<http://surfer.nmr.mgh.harvard.edu>). The technical details of this software suite are described in prior publications (Dale et al., 1999, p. 199; Fischl et al., 1999, 2002, 2004). Briefly, this processing includes motion correction and intensity normalization of T1-weighted images, removal of non-brain tissue (Ségonne et al., 2004), automated Talairach transformation, segmentation of white matter and gray matter volumetric structures, and derivation of cortical thickness. Freesurfer processing was implemented via Brainlife.io (brainlife/app-freesurfer), which is a free, publicly funded, cloud-computing platform for developing reproducible neuroimaging processing pipelines and sharing data (Avesani et al., 2019; Pestilli, 2018).

2.4.4 | brainageR brain age algorithm

brainageR uses Gaussian Process Regression to predict brain age based on raw, unprocessed, T1-weighted MR images (Cole et al., 2018). Relevant code is available at: <https://github.com/james-cole/brainageR>. This software uses SPM12 for segmentation and normalization with custom brain templates, and loads these images into R using the RNifti package. Gray matter, white matter and CSF vectors are then used to predict a brain age value with a model previously trained with kernlab. This algorithm was trained on a sample ($N = 2001$) of healthy adults aged 18–90, including scans from 14 different studies. Using ten-fold cross validation, brainageR produced a strong correlation between brain age and chronological age ($r = 0.92$; $MAE = 5.02$; $RMSE = 6.31$). The model was tested on a cohort of older adults ($N = 669$, mean age = 72.67 ± 0.73). In the test cohort, brain age delta did not correlate with chronological age ($r = -0.01$, $p = .79$), which Cole et al. believe reflects individual differences in aging. There was a significant relationship between brain age delta and mortality (hazard ratio = 1.061), but there was no significant relationship between brain age delta and cardiovascular disease, diabetes, or history of stroke.

2.4.5 | DeepBrainNet brain age algorithm

DeepBrainNet is a 2D convolutional neural network (CNN) built using the inception-resnetv2 framework (Bashyam et al., 2021). Notably, this model was initialized with random weights and trained exclusively on MRIs to create a brain-specific model. With this algorithm, raw, unprocessed, T1-weighted MR images are n4 bias corrected, skull-stripped, and affine registered to an MNI-template. This algorithm was implemented through the ANTsRNet package, an implementation of advanced normalization tools (ANTs) in the R programming language (Tustison et al., 2021). Relevant code for this algorithm is located at: <https://github.com/ANTsX/brainAgeR>. The algorithm was trained on a sample ($N = 11,729$) of healthy controls aged 3–95 drawn from 18 different datasets. All training data passed systematic quality control procedures. Bashyam et al. found a correlation of $r = 0.978$ between predicted brain age and chronological age. There was no significant difference in brain age delta between male and female subjects (male: $MAE = 3.68$; female: $MAE = 3.72$). The authors purposefully selected a “moderately fit” model over a loosely or tightly fit model. This was motivated by the belief that a moderately fit model would better reveal individual differences in pathology. To assist with the model selection process, Bashyam et al. tested the algorithm on subjects with psychopathology, including dementia, AD, schizophrenia, and depression. Similar to Kaufmann et al., the authors calculated standardized mean difference (Cohen's d) for each group compared to healthy controls, and found that the moderately fit model produced the largest Cohen's d values in comparison to the loosely and tightly fit models.

2.5 | Statistical analyses related to brain age reliability

To assess the reliability of brain age calculation by algorithm, we used two approaches of looking at reliability: intraclass correlation coefficient (ICC) and Bland–Altman analysis. Of note, for these analyses, we only used data from AOMIC and OASIS due to the repeated scans for each participant. ICC is a descriptive statistic indicating the degree of agreement between two or more sets of measurements. The statistic is similar to a bivariate correlation coefficient insofar as it has a range from 0 to 1 and higher values represent a stronger relationship. An ICC differs from the bivariate correlation in that it works on groups of measurements and gives an indication of the numerical cohesion across the given groups (McGraw & Wong, 1996). We calculated ICCs using the statistical programming language R, with the `icc` function from the package “`irr`” (Gamer et al., 2012). A two-way model with absolute agreement was used in order to investigate the exact estimate of brain age for each repeated scan. Although there are no definitive guidelines for precise interpretation of ICCs, results have frequently been binned into three (or four) quality groups where 0.0–0.5 is “poor”, 0.50–0.75 is “moderate”, 0.75–0.9 is “good” and 0.9–1.0 is “excellent” (Cicchetti, 1994; Koo & Li, 2016). Additionally, Bland–Altman analyses investigate reliability by considering the differences between paired groups of measurements. In our analysis, the paired groups are brain age predictions across two scans. As AOMIC contains three scans per participant, we compared across all pairings of scans (i.e., scan 1 versus scan 2, scan 2 versus scan 3, and scan 1 versus scan 3) and took the average difference metric. In addition to these raw difference scores (i.e., the difference across two instances of measurement), we also considered the proportion of difference, calculated by taking the difference divided by the mean value for a given pairing of measurements.

2.6 | Statistical analyses connected to image quality and sociodemographic factors

After reliability analyses, we reduced our analytic sample to include only the highest quality scan from each participant. This was done to maintain statistical independence of observations. Put another way, including repeated scans in standard regression models, but presuming independence, would violate fundamental statistical assumptions. The highest quality MRI scan was selected using CAT12 scores. Using these higher quality scans, we calculated bivariate correlations between brain age and real age. We also examined relations between brain age and multiple relevant variables, including image quality, and participant sex. This latter variable was included given suggestive sex differences in brain structure and cognitive decline (Levine et al., 2021; Ruigrok et al., 2014), likely contributing to differential estimates of “brain aging”. In these analyses, we looked at associations with both brain age and brain age delta in all three of our datasets. Of note, many brain age researchers choose to correct for age-related

bias using linear models or other methods (e.g., Le et al., 2018). This correction is often done in relation to group or individual difference variables. However, given that we were not focused on these differences and instead wanted to understand the full effects of potential confounds on the algorithms tested, we used uncorrected correlations in this portion of our analyses. Of note, XGBoost and DeepBrainNet excluded low-quality scans from training data, and none of the algorithms tested corrected training data for image quality.

2.7 | Statistical prediction using brain age variables

We used the OASIS dataset to look at the association between brain age variables and cognitive impairments. Based on previous findings, we would expect a greater brain age delta to be related to cognitive impairment caused by neurodegenerative conditions such as Alzheimer's disease. To investigate this hypothesis, participants were coded for presence of cognitive impairment (presence = 1; absence = 0). Cognitive impairment was determined using the Clinical Dementia Rating (CDR) score, which ranges from 0-3 (0 = no dementia; 0.5 = questionable dementia; 1 = MCI; 2 = Moderate Cognitive Impairment; 3 = Severe Cognitive Impairment). The presence of cognitive impairment was mapped to a CDR of > 1 , including subjects with mild to severe cognitive impairment, and excluding subjects with no or "questionable" dementia (CDR < 1). This was motivated by, and in keeping with, recent projects focused on replicating associations between brain age and mild cognitive impairment or dementia (e.g., Karim et al., 2022).

Additionally, combining groups increases statistical power, which is an issue given the modest sample of the OASIS project and the low incidence of occurrence of cognitive impairments in the sample (0: 544, 0.5: 139, 1: 36, 2: 4). However, even grouping together CDR scores > 1 , the cognitively impaired participants constituted only 5.5% of the data, which is far from ideal. For this reason, we implemented subsampling. Specifically, we used the Synthetic Minority Oversampling Technique ("SMOTE"), a method combining oversampling and under-sampling to achieve better classifier performance on imbalanced data (Chawla et al., 2002). We implemented SMOTE using the package *themis* in R (Hvitfeldt, 2022), using an over ratio of 0.5. Following subsampling, cognitively impaired participants accounted for 33% of the training data.

All models were created in the programming language R, using the *tidymodels* collection of packages (Kuhn et al., 2020). Individual logistic regression models were fit for each algorithm, and these models are detailed in the supplement. Presence of cognitive impairment was entered as the binary dependent variable (presence = 1, absence = 0). Brain age delta, chronological age, sex, and CAT12 scores were included as independent variables.

Given potential collinearity between brain age algorithms, we also fit an elastic net (EN) model to determine the best combination of predictors of clinical status. In brief, EN machine learning algorithms use multiple regression penalties (i.e., lasso [L1]; ridge [L2]) to prevent overfitting of the model, compromising between penalties by weighting the proportion of ridge and lasso penalties (α). To tune the penalty,

cross-validation identifies a second parameter, λ , which is the magnitude of the shrinkage penalty. This modeling and parameter identification was implemented using *tidymodels* and *glmnet* (Friedman et al., 2010). The presence of cognitive impairment was entered as the binary dependent variable (presence = 1, absence = 0). Potential independent variables included brain age deltas from all three algorithms, chronological age, sex, and CAT12 scores. The model was trained using ten-fold cross validation with ten repeats. We then compared the relative importance of the predictor variables.

3 | RESULTS

3.1 | Brain age reliability by algorithm

Given that brain age algorithms are being used in older (Cole et al., 2018), as well as younger (Keding et al., 2021), samples, we computed test-retest reliability in two different open-access neuroimaging projects with repeated scans. For the young adults in the AOMIC project, all three algorithms obtained ICCs greater than 0.9 (XGBoost: $r = 0.935$, brainageR: $r = 0.983$, DeepBrainNet: $r = 0.979$). The mean proportion of difference across all three scans was small. XGBoost had a mean of 6.535%, followed by brainageR with 2.449% and DeepBrainNet with 2.04%. These differences are depicted in Figure 2. The mean difference across all three scans was similarly small. XGBoost and brainageR had slightly negative mean differences of -0.139 ± 2.376 years and -0.098 ± 0.777 years, and DeepBrainNet had a mean difference of 0.011 ± 0.791 years.

For the older adult sample in OASIS, all three algorithms obtained ICCs greater than 0.95 (XGBoost: $r = 0.972$, brainageR: $r = 0.99$, DeepBrainNet: $r = 0.992$). Mean proportion of difference and mean difference across both scans were again small, with XGBoost showing a mean proportion of difference of 2.397%, followed by brainageR with 1.518% and DeepBrainNet with 1.259%. These differences are depicted in Figure 3. XGBoost had a small positive mean difference of 0.021 ± 1.865 years, and brainageR and DeepBrainNet had negative mean differences of -0.158 ± 1.483 and -0.086 ± 1.135 . Examined collectively, all three algorithms obtained high ICCs and low mean differences for both AOMIC and OASIS data, suggesting that these algorithms are highly reliable.

Figures 2 and 3 show density plots of the mean proportion of differences for both AOMIC and OASIS. The x-axes represent the proportion difference, ranging from 0 to 5, and the y-axes represent the probability density, calculated using a Gaussian kernel.

3.2 | Relations between brain age, chronological age, and sex brain age algorithms

Given that reliability metrics were similar in older and younger samples, we next examined correlations between brain age, brain age delta, and sociodemographic variables combining across these projects. When examining brain age and chronological age, there were strong correlations between these variables for each algorithm. All

FIGURE 2 Density Plot of Differences in 3 Brain Age Algorithms across the AOMIC sample. The horizontal axis shows the portion of differences between repeated scans (as a percentage). The vertical axis is the density (or frequency) of such bias. Each brain age algorithm is shown in a different color with XGBoost shown in light red, brainageR shown in light green, and DeepBrainNet shown in light blue.

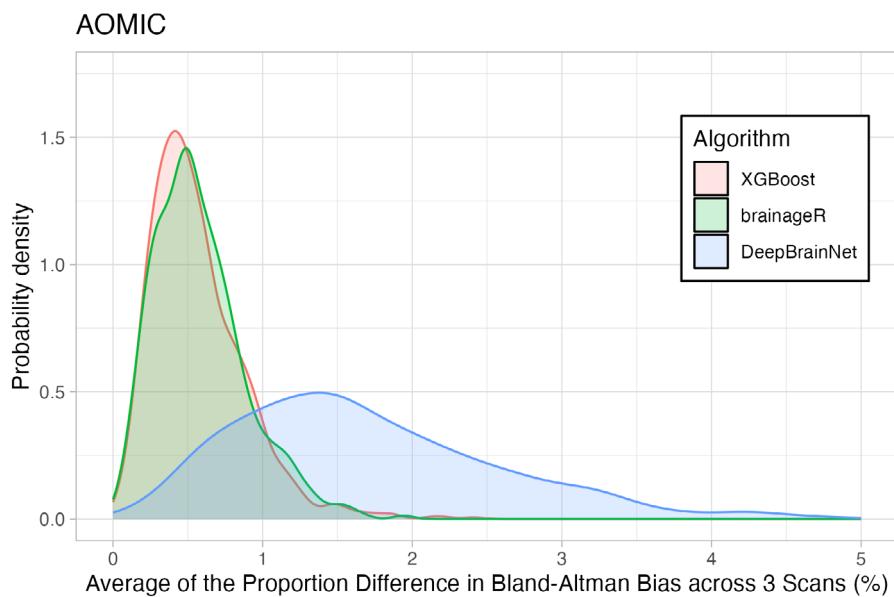
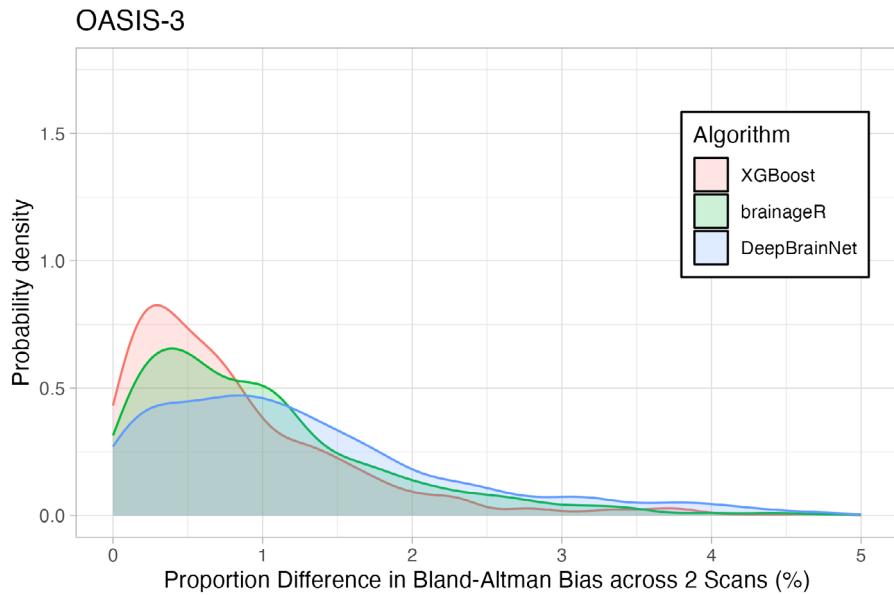


FIGURE 3 Density Plot of Differences in 3 Brain Age Algorithms across the OASIS sample. The horizontal axis shows the portion of differences between repeated scans (as a percentage). The vertical axis is the density (or frequency) of such bias. Each brain age algorithm is shown in a different color with XGBoost shown in light red, brainageR shown in light green, and DeepBrainNet shown in light blue.



correlations were greater than 0.9, with $r = 0.936$ for XGBoost, $r = 0.966$ for brainageR, and $r = 0.96$ for DeepBrainNet. As Kaufmann et al. separated models for male and female subjects, we also computed correlations for brain age and chronological age for each sex. With XGBoost, correlations were comparable across sex (female: $r = 0.932$; male: $r = .941$). This pattern was similar for brainageR (female: $r = 0.966$; male: $r = 0.968$) and DeepBrainNet (female: $r = 0.958$; male: $r = 0.963$). All correlations had p -values $<.001$. This relationship is visualized in Figure 4.

We additionally investigated associations between brain age delta and chronological age. Across the algorithms, there was a great variability in the relationship between brain age delta and chronological age: $r = -0.776$ for XGBoost, $r = -0.225$ for brainageR, and $r = -0.489$ for DeepBrainNet. All these correlations had p -values $<.001$.

All algorithms had significant negative correlations between brain age and image quality: XGBoost ($r = -0.381$, $p < .001$), brainageR

($r = -0.458$, $p < .001$), DeepBrainNet ($r = -0.464$, $p < .001$). In comparison to brain age and image quality, the relationship between brain age delta and image quality showed great variability in correlations. For XGBoost, there was modest correlation between image quality and brain age delta ($r = 0.36$, $p < .001$). brainageR also had a small negative correlation between brain age delta and quality ($r = -0.084$, $p < .001$). For DeepBrainNet, there was no significant relationship between brain age delta and image quality ($r = 0.033$, $p = .105$). These relationships are visualized in Figure 5.

3.3 | Brain age delta as a predictor of dementia status

To examine brain age delta as a predictor of dementia status, we built an elastic net (EN) model. The results are summarized in Table 1 and

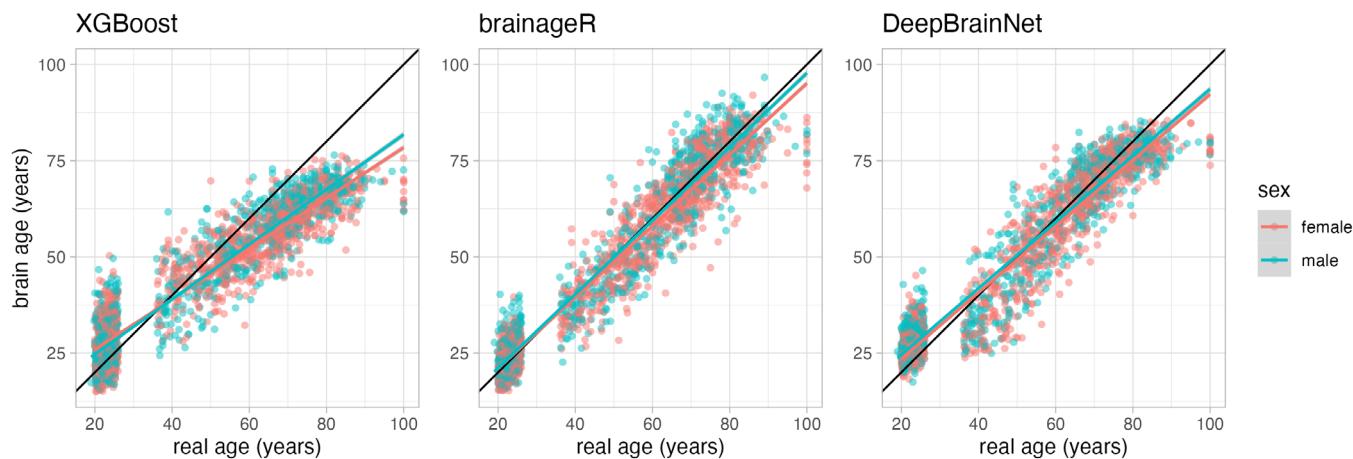


FIGURE 4 Scatterplots of the relationship between brain age and real age across all algorithms. There are three panels, each representing a different brain age algorithm—XGBoost is on the far-left panel, brainageR is in the middle, and DeepBrainNet is on the far-right panel. The horizontal axis shows participant chronological (real) age, while the vertical axis represents predicted brain age. In each panel, red dots represent female participants, and teal dots represent male participants.

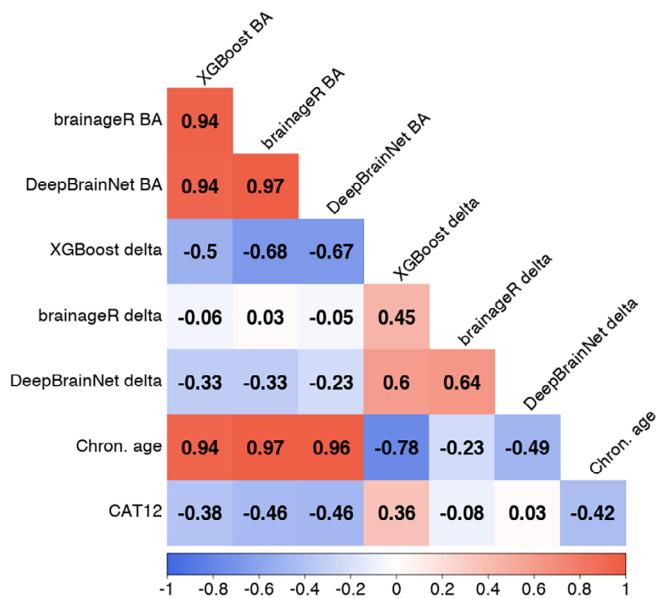


FIGURE 5 Correlation plot between brain age, brain age delta, chronological age, and CAT12 score across algorithms. There are rows and columns representing different relevant variables. The correlation between variables is shown at the confluence of a row and a column. The strength of a correlation is represented by the color of the background, and the exact value is written in black text. Negative correlations colored red, with strong negative correlations in dark red and weak negative correlations in light red. Positive correlations are colored blue, with strong positive correlations in dark blue and weak positive correlations in light blue. Weak or no correlation is colored white. BA = brain age.

Figures 6 and 7. Visualizations related to model tuning can be found in the supplement. Additionally, we built individual logistic regression models for each algorithm, which can be found in the supplement (S4).

EN models, similar to lasso regression models, select variables based on their predictive power. Additionally, for highly correlated

TABLE 1 A table depicting statistical parameter estimates from an elastic net logistic model (with clinical status as a binary dependent variable, and multiple brain age deltas [from different algorithms] and other covariates as independent variables)

Variable	β (std. coef.)
XGBoost - Brain age delta	0.530
brainageR - Brain age delta	0
DeepBrainNet - Brain age delta	0.415
Chronological age	0.733
Sex (male)	0
CAT12 score	-0.259

Note: Standardized coefficients and standard errors are shown in different columns, with statistics for each independent variable shown in each row.

variables, only the strongest predictor is included. This means not all independent variables are included in the final model. For our EN model, the final model after training uses XGBoost brain age delta, DeepBrainNet brain age delta, chronological age, and CAT12 score as predictors. Sex and brainageR brain age delta were excluded from the final predictors. The final model has an accuracy of 0.911.

Our model has a significant drawback. While it predicts cognitive impairment with a high degree of accuracy, the recall is only 0.45. This means for all positive cases, the model only correctly classifies 18/40 cases. A confusion matrix is provided below.

4 | DISCUSSION

Through the analysis of multiple large-scale, open-access neuroimaging datasets (Snoek et al., 2021; LaMontagne et al., 2019; Bookheimer et al., 2019), we investigated critical elements of the calculation of brain age. With an eye toward applied research, we examined the reliability, noise sensitivity, and predictive power of three commonly used brain age algorithms: XGBoost, brainageR, and DeepBrainNet.

FIGURE 6 Caption: ROC curve for the final model, trained across 10 repeats.

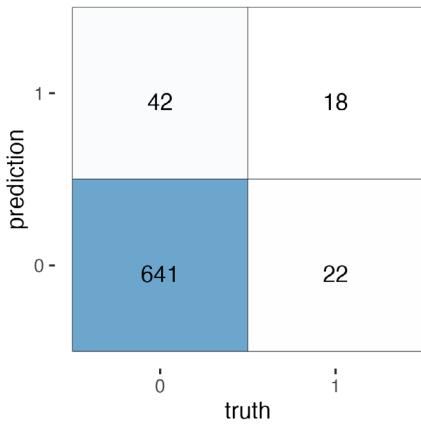
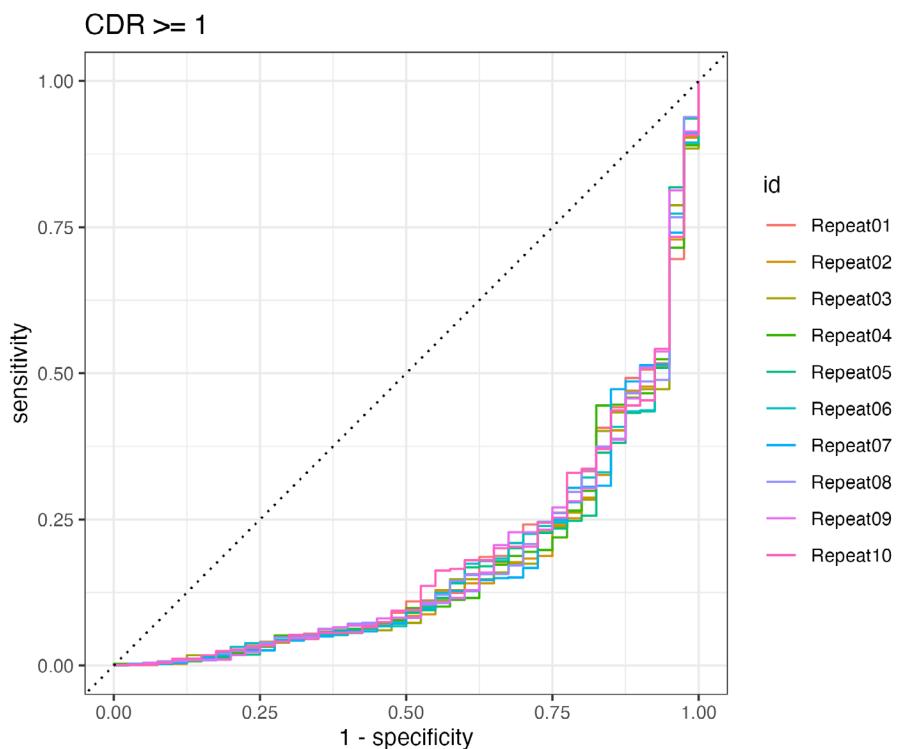


FIGURE 7 A confusion matrix for the EN predictions. True negative = 641. False negative = 22. False positive = 42. True positive = 18.

Regarding reliability, we found all brain age algorithms were highly reliable. This was assessed through ICC and Bland-Altman metrics in samples of older and younger participants with repeated MRI scans. Related to brain age calculation and demographic variables of interest, there were strong correlations between these variables for each algorithm. Calculated brain age for each algorithm strongly tracked with chronological age across males and females (r 's $> .9$). Connected to image quality, all algorithms had significant negative correlations between brain age and image quality. Notably, this was with brain age, and not brain age delta. Correlations between brain age delta and image quality were modest for XGBoost, but near zero for brainageR and DeepBrainNet. Chronological age, similar to brain age, had a

significant negative correlation with image quality. Turning to clinical prediction, individual logistic models suggested that brain age delta was a significant predictor of cognitive impairment (see supplement S4). In a penalized regression model more suited to deal with collinear variables, chronological age was the strongest predictor of cognitive impairment; however, while our model achieved high accuracy, the model performed poorly on identifying subjects with cognitive impairment (accuracy = 0.911, recall = 0.45). Brain age deltas derived from XGBoost and DeepBrainNet were also significant predictors, and brainageR's brain age delta was dropped from the final model.

Synthesizing across these different results, use of XGBoost may come with equal advantages and disadvantages. While all algorithms demonstrated excellent reliability as assessed by ICCs, it is notable that XGBoost had higher levels of bias (6.535%) in our younger sample of participants as assessed by Bland-Altman metrics. In the aggregate, this variation was small (-0.139 years) and that project (AOMIC) had a very narrow age range. However, in younger cohorts, this could lead to additional noise variance when examining this brain age algorithm in relation to important individual differences. Similarly, XGBoost's brain age delta had a modest correlation with image quality. This will be important to consider when selecting brain age algorithms in different cohorts. If the population is likely to have high levels of movement (i.e., children; individuals with significant cognitive deficits and impulsivity), this could create additional noise, and cloud relations between brain age delta and other variables of interest. Notably, past work (Ronan et al., 2016) has found that estimates of cortical volume and thickness decreased with greater motion. Given that typical aging is also associated with cortical atrophy (Scahill et al., 2003), structural MRI images of lower image quality and with greater motion would likely have lower estimates of cortical volume and

thickness, potentially biasing different algorithms that use these features in brain age calculation. For example, XGBoost uses derived morphometric parcels from Freesurfer and past work from our group has noted strong relations between Freesurfer outputs and image quality (Gilmore et al., 2021). Despite the poor recall of the EN model, it is still notable that XGBoost brain age delta was the most sensitive at differentiating cognitive impairment across the logistic and penalized models. Importantly, these models included a metric of image quality, suggesting that XGBoost's brain age derivation explained important variance above and beyond this nuisance variable.

Thinking about our findings in relation to past reports, similar patterns have been noted individually for each algorithm regarding reliability and correlations with chronological age. Publications detailing the development of each algorithm have reported similar ICCs and correlations. Our project, however, is the first to examine these elements across multiple algorithms. While no algorithm that we investigated was superior on reliability metrics, it will be critical for future projects focused on novel metrics of brain age to compare performance to other algorithms as relative benchmarks. Publications that simply report metrics of their newly derived algorithm's performance may be less useful to the field if this performance is not necessarily different or superior to previously developed approaches. As noted in our results section, brainageR and DeepBrainNet were highly correlated ($r = 0.97$), suggesting that these algorithms may be identifying similar patterns of advanced brain aging. It will again be critical for future projects focused on novel metrics of brain age to show that novel algorithms are identifying unique and additive variance in brain age.

Of note and important for future work is that we examined fairly significant clinical issues in thinking about prediction. There is ongoing work looking at different individual difference measures that span a more normative continuum of functioning. It could be particularly useful to see if brain ages from these different algorithms relate to these individual differences (e.g., stress exposure; general cognitive functioning; obesity, Ronan et al., 2016; Shokri-Kojori et al., 2021). Similarly, it will be critical for future investigations to probe multimodal MRI calculations of brain age. All of the algorithms examined here focused on T1-weighted images, either processed in Freesurfer (XGBoost) or in original NIfTI format (brainageR; DeepBrainNet). More recent work has leveraged diffusion imaging, often in concert with T1-weighted images for prediction of brain age (Beck et al., 2021; Richard et al., 2018). It is likely that brain ages calculated through multimodal MRI and with multiple algorithms could be more powerful in explaining age-associated functional declines and disease.

While we believe we advanced applied understanding of brain age calculation, our work is not without limitations. First, all of our data is cross-sectional in nature, and it will be important to think about estimation and validation of different performance metrics in participants with repeated MRI scans separated by long periods of time. By seeing levels of within- and between-person change in relation to different algorithms, we may be able to derive a particularly powerful window into age-associated functional declines and disease, and different clinically relevant issues. Second, we did not specifically focus on variations in MRI scanners, instead pooling across scanning

types. One cohort (HCP-A) had technically sophisticated MRI acquisition, potentially more sensitive than other "out of the box" neuroimaging scans. While we found similar results for HCP-A and OASIS (a dataset of similarly aged participants scanned with less sophisticated MR techniques), we did not specifically probe for variation across MRI scanners. Third, we tested three commonly used algorithms where code was publicly shared for mass implementation of brain age calculation. There are many in-press and preprinted manuscripts engineering new calculations of brain age. Such novel algorithms may exhibit superior performance and fewer limitations than the approaches we examined here. Finally, we were unable to discern the factors that might be driving variations in algorithmic performance. Brain age calculated via XGBoost uses Freesurfer parcels in its brain age calculation, while brainageR and DeepBrainNet both work off of less-processed NIfTI files. It may be possible to optimize elements of Freesurfer or other software to improve different metrics of reliability and prediction. Connected to this, our team is particularly interested in the effects of image quality on brain age calculation and how to probe different datasets where repeated MRI scans are acquired from the same individuals but there is intentional variability in motion-related artifacts. Tackling these and other open questions related to brain age could significantly advance our understanding of healthy, as well as accelerated, aging processes.

Limitations notwithstanding, additional research on "*biological age*" is imperative. Richer information about the brain and brain aging could be important for those focused on age-related mortality and morbidity. Here, we provide important information about multiple brain age algorithms for researchers to consider when they deploy this emerging biomarker. Thoughtful consideration about reliability, noise tolerance, and predictive power will be critical when making decisions about different brain age algorithms, especially with an ever-growing landscape of potential ways to calculate this variable.

FUNDING INFORMATION

This research used brainlife.io supported with grants from the National Science Foundation (IIS-1912270, IIS-1636893, BCS-1734853) to Dr. Franco Pestilli at the University of Texas at Austin.

DATA AVAILABILITY STATEMENT

OASIS-3 data is available on request at <https://www.oasis-brains.org/#access>. AOMIC ID-1000 data is available in OpenNeuro at 10.18112/openneuro.ds003097.v1.2.1, accession number ds003097. HCP-A data is available on request in NIMH Data Archive at <https://nda.nih.gov>. Access is restricted to researchers affiliated with an NIH-recognized institute with an active Federalwide Assurance.

ORCID

Eva Bacas  <https://orcid.org/0000-0002-8844-3257>

Isabella Kahhalé  <https://orcid.org/0000-0002-0963-9738>

REFERENCES

Aubert, G., & Lansdorp, P. M. (2008). Telomeres and aging. *Physiological Reviews*, 88(2), 557–579. <https://doi.org/10.1152/physrev.00026.2007>

Avesani, P., McPherson, B., Hayashi, S., Caiafa, C. F., Henschel, R., Garyfallidis, E., Kitchell, L., Bullock, D., Patterson, A., & Olivetti, E. (2019). The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. *Scientific Data*, 6(1), 1–13. <https://doi.org/10.1038/s41597-019-0073-y>

Bashyam, V. M., Shou, H., & Davatzikos, C. (2021). Reply: From 'loose fitting' to high-performance, uncertainty-aware brain-age modelling. *Brain*, 144, e32. <https://doi.org/10.1093/brain/awaa455>

Beck, D., de Lange, A.-M. G., Maximov, I. I., Richard, G., Andreassen, O. A., Nordvik, J. E., & Westlye, L. T. (2021). White matter microstructure across the adult lifespan: A mixed longitudinal and cross-sectional study using advanced diffusion models and brain-age prediction. *Neuroimage*, 224, 117441. <https://doi.org/10.1016/j.neuroimage.2020.117441>

Bookheimer, S. Y., Salat, D. H., Terpstra, M., Ances, B. M., Barch, D. M., Buckner, R. L., Burgess, G. C., Curtiss, S. W., Diaz-Santos, M., Elam, J. S., Fischl, B., Greve, D. N., Hagy, H. A., Harms, M. P., Hatch, O. M., Hedden, T., Hodge, C., Japardi, K. C., Kuhn, T. P., ... Yacoub, E. (2019). The lifespan human connectome project in aging: An overview. *NeuroImage*, 185, 335–348. <https://doi.org/10.1016/j.neuroimage.2018.10.009>

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>

Cole, J. H., Marioni, R. E., Harris, S. E., & Deary, I. J. (2019). Brain age and other bodily "ages": Implications for neuropsychiatry. *Molecular Psychiatry*, 24(2), 266–281. <https://doi.org/10.1038/s41380-018-0098-1>

Cole, J. H., Ritchie, S. J., Bastin, M. E., Valdés Hernández, M. C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S. E., Zhang, Q., Wray, N. R., Redmond, P., Marioni, R. E., Starr, J. M., Cox, S. R., Wardlaw, J. M., Sharp, D. J., & Deary, I. J. (2018). Brain age predicts mortality. *Molecular Psychiatry*, 23(5), 1385–1392.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I segmentation and surface reconstruction. *Neuroimage*, 9(2), 179–194. <https://doi.org/10.1006/nimg.1998.0395>

Fischl, B. (2012). FreeSurfer. *Neuroimage*, 62(2), 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., & Klaveness, S. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3), 341–355. [https://doi.org/10.1016/S0896-6273\(02\)00569-X](https://doi.org/10.1016/S0896-6273(02)00569-X)

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2), 195–207. <https://doi.org/10.1006/nimg.1998.0396>

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., & Kennedy, D. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1), 11–22. <https://doi.org/10.1093/cercor/bhg087>

Fraga, M. F., & Esteller, M. (2007). Epigenetics and aging: The targets and the marks. *Trends in Genetics*, 23(8), 413–418. <https://doi.org/10.1016/j.tig.2007.05.008>

Franke, K., & Gaser, C. (2012). Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych*, 25, 235–245. <https://doi.org/10.1024/1662-9647/a000074>

Franke, K., & Gaser, C. (2019). Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in Neurology*, 10, 789. <https://doi.org/10.3389/fneur.2019.00789>

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22. <https://doi.org/10.18637/jss.v033.i01>

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package "irr." Various Coefficients of Interrater Reliability and Agreement, 22.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., & Initiative, A. D. N. (2013). BrainAGE in mild cognitive impaired patients: Predicting the conversion to Alzheimer's disease. *PLoS One*, 8(6), e67346. <https://doi.org/10.1371/journal.pone.0067346>

Gilmore, A. D., Buser, N. J., & Hanson, J. L. (2021). Variations in structural MRI quality significantly impact commonly used measures of brain anatomy. *Brain Informatics*, 8(1), 7. <https://doi.org/10.1186/s40708-021-00128-2>

Hahn, T., Fisch, L., Ernstsing, J., Winter, N. R., Leenings, R., Sarink, K., Emden, D., Kircher, T., Berger, K., & Dannlowski, U. (2021). From "loose fitting" to high-performance, uncertainty-aware brain-age modelling. *Brain*, 144(3), e31.

Hvitfeldt, E. (2022). themis: Extra Recipes Steps for Dealing with Unbalanced Data. <https://github.com/tidymodels/themis>

Karim, H. T., Aizenstein, H. J., Mizuno, A., Ly, M., Andreescu, C., Wu, M., Hong, C. H., Roh, W. R., Park, B., Lee, H., Kim, N., Choi, J. W., Seo, W. S., Choi, S. H., Kim, E. J., Kim, B. C., Cheong, J. Y., Lee, E., Lee, D., ... & Son, S. J. (2022). Independent replication of advanced brain age in mild cognitive impairment and dementia: detection of future cognitive dysfunction. *Molecular Psychiatry*, 27(12), 5235–5243.

Kaufmann, T., van der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., Alnæs, D., Barch, D. M., Baur-Streubel, R., & Bertolino, A. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature Neuroscience*, 22(10), 1617–1623. <https://doi.org/10.1038/s41593-019-0471-7>

Keding, T. J., Heyn, S. A., Russell, J. D., Zhu, X., Cisler, J., McLaughlin, K. A., & Herringa, R. J. (2021). Differential patterns of delayed emotion circuit maturation in abused girls with and without internalizing psychopathology. *American Journal of Psychiatry*, 178(11), 1026–1036. <https://doi.org/10.1176/appi.ajp.2021.20081192>

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intra-class correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlander, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., & Reiser, M. (2014). Accelerated brain aging in schizophrenia and beyond: A neuro-anatomical marker of psychiatric disorders. *Schizophrenia Bulletin*, 40(5), 1140–1153. <https://doi.org/10.1093/schbul/sbt142>

Kuhn, et al. (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. <https://www.tidymodels.org>

LaMontagne, P. J., Benzinger, T. L. S., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A. G., Raichle, M. E., Cruchaga, C., & Marcus, D. (2019). OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for Normal aging and Alzheimer disease. *MedRxiv*. <https://doi.org/10.1101/2019.12.13.19014902>

Le, T. T., Kuplicki, R. T., McKinney, B. A., Yeh, H.-W., Thompson, W. K., Paulus, M. P., Aupperle, R. L., Bodurka, J., Cha, Y.-H., & Feinstein, J. S. (2018). A nonlinear simulation framework supports adjusting for age when analyzing BrainAGE. *Frontiers in Aging Neuroscience*, 10, 317. <https://doi.org/10.3389/fnagi.2018.00317>

Levine, D. A., Gross, A. L., Briceño, E. M., Tilton, N., Giordani, B. J., Sussman, J. B., Hayward, R. A., Burke, J. F., Hingtgen, S., Elkind, M. S. V., Manly, J. J., Gottesman, R. F., Gaskin, D. J., Sidney, S., Sacco, R. L., Tom, S. E., Wright, C. B., Yaffe, K., & Galecki, A. T. (2021). Sex differences in cognitive decline among US adults. *JAMA Network Open*, 4(2), e210169. <https://doi.org/10.1001/jamanetworkopen.2021.0169>

Ly, M., Gary, Z. Y., Karim, H. T., Muppudi, N. R., Mizuno, A., Klunk, W. E., Aizenstein, H. J., & Initiative, A. D. N. (2020). Improving brain age prediction models: Incorporation of amyloid status in Alzheimer's disease. *Neurobiology of Aging*, 87, 44–48. <https://doi.org/10.1016/j.neurobiolaging.2019.11.005>

Mather, K. A., Jorm, A. F., Parslow, R. A., & Christensen, H. (2011). Is telomere length a biomarker of aging? A review. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 66(2), 202–213. <https://doi.org/10.1093/gerona/glr180>

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30.

Pal, S., & Tyler, J. K. (2016). Epigenetics and aging. *Science Advances*, 2(7), e1600584. <https://doi.org/10.1126/sciadv.1600584>

Pestilli, F. (2018). Human white matter and knowledge representation. *PLoS Biology*, 16(4), e2005758. <https://doi.org/10.1371/journal.pbio.2005758>

Richard, G., Kolskår, K., Sanders, A.-M., Kaufmann, T., Petersen, A., Doan, N. T., Sánchez, J. M., Alnæs, D., Ulrichsen, K. M., & Dørum, E. S. (2018). Assessing distinct patterns of cognitive aging using tissue-specific brain age prediction based on diffusion tensor imaging and brain morphometry. *PeerJ*, 6, e5908. <https://doi.org/10.7717/peerj.5908>

Roser, M., Ortiz-Ospina, E. & Ritchie, H. (2013). Life expectancy Our World in Data. <https://ourworldindata.org/life-expectancy>

Ronan, L., Alexander-Bloch, A. F., Wagstyl, K., Farooqi, S., Brayne, C., Tyler, L. K., & Fletcher, P. C. (2016). Obesity associated with increased brain age from midlife. *Neurobiology of Aging*, 47, 63–70. <https://doi.org/10.1016/j.neurobiolaging.2016.07.010>

Ruigrok, A. N. V., Salimi-Khorshidi, G., Lai, M.-C., Baron-Cohen, S., Lombardo, M. V., Tait, R. J., & Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39, 34–50. <https://doi.org/10.1016/j.neubiorev.2013.12.004>

Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *Neuroimage*, 22(3), 1060–1075. <https://doi.org/10.1016/j.neuroimage.2004.03.032>

Scahill, R. I., Frost, C., Jenkins, R., Whitwell, J. L., Rossor, M. N., & Fox, N. C. (2003). A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. *Archives of neurology*, 60(7), 989–994.

Shahab, S., Mulsant, B. H., Levesque, M. L., Calarco, N., Nazeri, A., Wheeler, A. L., Foussias, G., Rajji, T. K., & Voineskos, A. N. (2019). Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls. *Neuropsychopharmacology*, 44(5), 898–906. <https://doi.org/10.1038/s41386-018-0298-z>

Shokri-Kojori, E., Bennett, I. J., Tomeldan, Z. A., Krawczyk, D. C., & Rypma, B. (2021). Estimates of brain age for gray matter and white matter in younger and older adults: Insights into human intelligence. *Brain Research*, 1763, 147431. <https://doi.org/10.1016/j.brainres.2021.147431>

Snoek, L., van der Miesen, M. M., Beemsterboer, T., van der Leij, A., Eigenhuis, A., & Steven Scholte, H. (2021). The Amsterdam open MRI collection, a set of multimodal MRI datasets for individual difference analyses. *Scientific Data*, 8(1), 85.

Tustison, N. J., Cook, P. A., Holbrook, A. J., Johnson, H. J., Muschelli, J., Devenyi, G. A., Duda, J. T., Das, S. R., Cullen, N. C., & Gillen, D. L. (2021). The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific Reports*, 11(1), 1–13.

Wylie, G. R., Genova, H., DeLuca, J., Chiaravalloti, N., & Sumowski, J. F. (2014). Functional magnetic resonance imaging movers and shakers: Does subject-movement cause sampling bias?: Subject motion and sampling bias. *Human Brain Mapping*, 35(1), 1–13. <https://doi.org/10.1002/hbm.22150>

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168, 022022. <https://doi.org/10.1088/1742-6596/1168/2/022022>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bacas, E., Kahhalé, I., Raamana, P. R., Pablo, J. B., Anand, A. S., & Hanson, J. L. (2023). Probing multiple algorithms to calculate brain age: Examining reliability, relations with demographics, and predictive power. *Human Brain Mapping*, 44(9), 3481–3492. <https://doi.org/10.1002/hbm.26292>