



Brain-predicted age difference score is related to specific cognitive functions: a multi-site replication analysis

Rory Boyle¹ · Lee Jollans² · Laura M. Rueda-Delgado¹ · Rossella Rizzo³ · Görsev G. Yener^{4,5,6} · Jason P. McMorrow^{7,8} · Silvin P. Knight^{8,9} · Daniel Carey^{9,10} · Ian H. Robertson^{1,11} · Derya D. Emek-Savaş^{4,11,12} · Yaakov Stern¹³ · Rose Anne Kenny^{8,9,14} · Robert Whelan^{1,11}

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Brain-predicted age difference scores are calculated by subtracting chronological age from ‘brain’ age, which is estimated using neuroimaging data. Positive scores reflect accelerated ageing and are associated with increased mortality risk and poorer physical function. To date, however, the relationship between brain-predicted age difference scores and specific cognitive functions has not been systematically examined using appropriate statistical methods. First, applying machine learning to 1359 T1-weighted MRI scans, we predicted the relationship between chronological age and voxel-wise grey matter data. This model was then applied to MRI data from three independent datasets, significantly predicting chronological age in each dataset: Dokuz Eylül University ($n = 175$), the Cognitive Reserve/Reference Ability Neural Network study ($n = 380$), and The Irish Longitudinal Study on Ageing ($n = 487$). Each independent dataset had rich neuropsychological data. Brain-predicted age difference scores were significantly negatively correlated with performance on measures of general cognitive status (two datasets); processing speed, visual attention, and cognitive flexibility (three datasets); visual attention and cognitive flexibility (two datasets); and semantic verbal fluency (two datasets). As such, there is firm evidence of correlations between increased brain-predicted age differences and reduced cognitive function in some domains that are implicated in cognitive ageing.

Keywords MRI · Brain ageing · Cognitive ageing · Cognitive function · Machine learning · Biomarkers

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11682-020-00260-3>) contains supplementary material, which is available to authorized users.

✉ Robert Whelan
robert.whelan@tcd.ie

¹ Trinity College Institute of Neuroscience, Trinity College Dublin, Lloyd Building, Dublin 2, Ireland

² Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, München, Germany

³ Physics Department, University of Calabria, Rende, CS, Italy

⁴ Department of Neurosciences, Institute of Health Sciences, Dokuz Eylül University, İzmir, Turkey

⁵ Department of Neurology, Dokuz Eylül University Medical School, İzmir, Turkey

⁶ Brain Dynamics Multidisciplinary Research Center, Dokuz Eylül University, İzmir, Turkey

⁷ Centre for Advanced Medical Imaging, St. James’s Hospital, Dublin 8, Ireland

⁸ School of Medicine, Trinity College Dublin, Dublin 2, Ireland

⁹ The Irish Longitudinal Study on Ageing, Trinity College Dublin, Dublin 2, Ireland

¹⁰ Department of Medical Gerontology, Trinity College Dublin, Dublin 2, Ireland

¹¹ Global Brain Health Institute, Trinity College Dublin, Lloyd Building, Dublin 2, Ireland

¹² Department of Psychology, Faculty of Letters, Dokuz Eylül University, İzmir, Turkey

¹³ Cognitive Neuroscience Division, Department of Neurology, Columbia University, New York, NY, USA

¹⁴ Mercer’s Institute for Successful Ageing, St. James’s Hospital, Dublin 8, Ireland

Introduction

Longitudinal neuropsychological testing in older adults can be used to detect cognitive decline. However, practice effects can obscure assessment of cognitive ability (Elman et al. 2018), and test performance is affected by subject-level factors such as the individual's level of comprehension, reading ability, self-efficacy, motivation, fatigue, and fluctuations in concentration (McCaffrey and Westervelt 1995). In contrast, objective biomarkers are not subject to such biases or patients' physical limitations (Jollans and Whelan 2016). An objective biomarker of cognitive ageing would therefore be useful for the timely identification of cognitive decline outside of age-related norms.

Ageing is a process with significant heterogeneity across individuals (McCrary and Kenny 2018). Consequently, chronological age is not the most accurate marker of an individual's rate of biological ageing (Sprott 2010). Ageing biomarkers have been developed that provide additional information about an individual's health status and life expectancy (Dean and Morgan 1988). For example, DNA methylation data can estimate epigenetic ageing ('epigenetic clocks'), reflecting the age of an individual's tissues or blood cells (Fiorito et al. 2019). Subtracting chronological age from the biological age results in a biologically informative summary score – the predicted age difference – for each individual, which reflects the deviation from typical lifespan trajectories (Richard et al. 2018). This approach has also been applied in neuroimaging, where machine learning can be used to quantify the relationship between structural MRI data and chronological age, in order to estimate an individual's 'brain age'. Subtracting chronological age from the estimated 'brain age' results in a brain predicted-age difference score (*brainPAD*, also referred to as brain age gap, brainAGE, Brain-Age Score; Beheshti et al. 2018; Franke et al. 2010; Schnack et al. 2016) which quantifies how a person's brain health differs from what would be expected for their chronological age.

BrainPAD is a promising biomarker of general brain ageing as it already satisfies several criteria for ageing biomarkers (Butler et al. 2004). BrainPAD is predictive of mortality and of age-sensitive physiological measures, including grip strength, lung function, walking speed and allostatic load (Cole et al. 2018). Moreover, brainPAD has been associated with cognitive impairment (Liem et al. 2017), is negatively correlated with fluid cognitive performance (Cole et al. 2018) and is significantly increased in Alzheimer's disease (AD) and mild cognitive impairment (MCI; Franke and Gaser 2012; Gaser et al. 2013; Löwe et al. 2016). As such, brainPAD could serve as a cognitive ageing biomarker. However, this potential use of brainPAD is currently limited by a lack of knowledge regarding the exact relationship between brainPAD and specific cognitive functions in healthy individuals.

Studies relating specific cognitive functions and brainPAD have been assessed in solely clinical samples (e.g., Cole et al.

2015; traumatic brain injury), or in mixed samples of clinical groups and healthy controls (e.g., Beheshti et al. 2018; AD, MCI, and healthy controls) and not samples comprised only of healthy adults. As such, the reported associations between brainPAD and specific domains of cognitive function in such studies (Beheshti et al. 2018; Cole et al. 2015) may be skewed towards significance by the inclusion of the clinical samples with typically higher brainPADs. Consequently, these findings may not represent the brainPAD-cognition relationship in normal ageing. For example, Le et al. (2018) reported a significant negative correlation between brainPAD and response inhibition and selective attention in a sample of individuals comprised of healthy controls and patients with mood or anxiety disorders, substance use disorder and/or eating disorders. However, significantly increased brainPADs have been reported in mood disorders such as major depression (Koutsouleris et al. 2014) and in substance use disorders such as alcohol dependence (Guggenmos et al. 2017). As both major depression and alcohol dependence are associated with cognitive impairments (Chanraud et al. 2007; McIntyre et al. 2013), the significant brainPAD-cognitive function correlations reported across samples including such populations could be driven by the inclusion of such clinical groups.

The relationship between specific cognitive functions and BrainPAD has also been somewhat obscured by statistical considerations. Recent work has empirically demonstrated that chronological age must be controlled for when testing relationships between brainPAD and cognitive functions (Le et al. 2018; Smith et al. 2019). Failure to correct for chronological age can result in false positive findings because some cognitive variables are correlated with chronological age – but *not* brain ageing – and brainPAD is typically correlated with chronological age (Le et al. 2018). In light of this recent work, it is difficult to interpret studies that did not control for chronological age when investigating the brainPAD-cognition relationship in healthy controls (Franke et al. 2013; Löwe et al. 2016). A second statistical issue is a failure to correct for multiple comparisons. Researchers testing the brainPAD-cognition relationship have tended to carry out multiple statistical tests of the correlation between brainPAD and various cognitive measures. The performance of multiple statistical tests can increase the Type I error and result in false positive findings (Ranganathan et al. 2016). However, some papers did not control for multiple comparisons when investigating the brainPAD-cognition relationship (Beheshti et al. 2018; Cole et al. 2017b). Other studies have investigated the relationship between brainPAD and specific domains of cognitive function while controlling for chronological age and multiple comparisons, but there are conflicting results for most cognitive domains. For example, a significant correlation between verbal fluency and brainPAD was reported by Franke et al. (2013) whereas Richard et al. (2018) found no association between verbal fluency

and brainPAD. We have summarized the brainPAD-cognition findings in Table 1.

The first step in generating a brainPAD score is creating a feature set of neuroimaging data which is correlated with chronological age. Neuroimaging data have high dimensionality, which can result in overfitting and overoptimistic predictions (Whelan and Garavan 2014). Brain age prediction models thus rely on feature engineering techniques such as principal components analysis (PCA; Franke et al. 2010; Gutierrez Becker et al. 2018) or even dot products of different features (e.g. vectors of GM and white matter (WM) voxels as in Cole et al. 2015; Cole et al. 2018; Cole et al. 2017b) in order to reduce the dimensionality (Mwangi et al. 2014). These techniques map the original variables onto a feature space (in effect, creating ‘new’ variables) typically using linear transformations in the case of dot products (Snyder et al. 2013), although non-linear transformation may also be used for kernel methods (Honeine and Richard 2009; Kwok and Tsang 2004). While these models create generalizable and accurate predictions, this may come at the cost of reduced interpretability of the contributions of the features (Bunea et al. 2011; Mateos-Pérez et al. 2018), which is important for assessing the neurobiological validity of the model (Woo et al. 2017) and to identify specific brain areas for further investigation (Scheinost et al. 2019). Due to the importance of interpretability in neuroimaging, unlike with other data (e.g. credit card transactions for fraud detection), the application of machine learning to MRI does not necessarily involve the goal of achieving the highest accuracy (Mateos-Pérez et al. 2018). While methods do exist for projecting the ‘new’ variables back from the feature space to the input space (Honeine and Richard 2009; Kwok and Tsang 2004; Snyder et al. 2013) thus enabling interpretability of models employing dot products, PCA or kernel methods, these methods are not always implemented and/or reported in brain-age papers (Cole et al. 2015, 2018; Cole et al. 2017a; Gaser et al. 2013; Gutierrez Becker et al. 2018; Nenadić et al. 2017). In contrast, penalized regression methods (e.g., the Elastic Net; Zou and Hastie 2005) do not require the back-projection of coefficients from feature space to input space and therefore have good interpretability, particularly when less complex feature sets are used (Luo et al. 2019). GM data is particularly well-suited for age prediction as GM volume linearly declines with age (but cf. Fjell et al. 2013) whereas WM volume has a less straightforward relationship with age, as it doesn’t decline significantly until middle age (Farokhian et al. 2017; Ge et al. 2002). The Elastic Net is a machine learning model well-suited to the high dimensionality and multicollinearity inherent in neuroimaging data as shown by the finding that it produced the most consistent predictions as compared to various other models over datasets with varying sample-, feature set-, and effect-sizes (Jollans et al. 2019).

A final challenge in the development of neuroimaging biomarkers, or neuromarkers, is ensuring the generalisability of the neuromarker to new data. For practical reasons, cross-validation, where a dataset is split into a training set and a test set (Varoquaux et al. 2017), is often used as an estimate of model accuracy for new data (Jollans and Whelan 2018; Scheinost et al. 2019). However, cross-validation accuracy estimates are often optimistically biased and can vary considerably (Varoquaux et al. 2017), particularly when preprocessing and feature selection are carried out on the entire dataset before splitting it into training and test sets (Dwyer et al. 2018; Woo et al. 2017). As such, the gold-standard for assessing the external validity and generalisability of a neuromarker is by testing how the model performs on a completely independent held-out dataset (Jollans and Whelan 2018). While various brainPAD studies have externally validated their models (Beheshti et al. 2018; Cole et al. 2015, 2018; Cole et al. 2017b; Franke et al. 2010; Gutierrez Becker et al. 2018; Lancaster et al. 2018; Liem et al. 2017; Madan and Kensinger 2018; Varikuti et al. 2018), only a few studies have reported model performance in terms of accuracy (i.e., correlation or mean absolute error between brain-predicted age and chronological age) on the external validation dataset (Cole et al. 2015; Lancaster et al. 2018; Liem et al. 2017; Madan and Kensinger 2018). This does not necessarily cast doubt on the validity of the models whose accuracy is reported in terms of internal cross-validation performance. However, not reporting the external validation performance limits the interpretation of the accuracy and generalisability of various brainPAD models as typically performance will be lower in the external validation dataset.

In order to clarify the unclear relationship between brainPAD and specific domains of cognitive function, we aimed to 1) establish an interpretable model of brainPAD using the Elastic Net with GM voxel-wise data, 2) externally validate this model in three independent datasets, and 3) to establish the domains of cognitive function that are reliably correlated with brainPAD across different datasets.

Methods

Study design

The present study used data from open-access repositories to form a training set in which a machine learning model was trained. Data from three separate datasets (Dokuz Eylül University (DEU); Cognitive Reserve/Reference Ability Neural Network Study (CR/RANN) and The Irish Longitudinal Study on Ageing (TILDA)) were then used to form three external validation sets in which the machine learning model was validated and the relationship between brainPAD and cognitive function was investigated. In all cases, the data were collected prior to conception and design

Table 1. Summary of findings on the relationship between brainPAD and cognitive function.

Cognitive Domain	Measure	Reference	Sample	n	Sig.	Sig. in HCs	Age adj.	MC corr.
General Cognitive Status	MOCA	(Richard et al., 2018)	HC	265	✗	✓	✓	✓
	MMSE	(Kaufmann et al., 2019)	MCI; DEM	921; 707	✓	✗	✓	✓
	MMSE	(Gaser et al., 2013)	MCI	195	✗	✗	✗	✗
	CDR				✓	✗	✗	✗
	ADAS				✓	✗	✗	✗
	MMSE	(Löwe et al., 2016)	APOE; Non APOE	219; 186	✓	✗	✗	? ^a
	CDR		APOE; Non APOE	219; 186	✓	✗	✗	? ^a
	ADAS		APOE; Non APOE	219; 186	✓	✗	✗	? ^a
	MMSE	(Beheshti et al., 2018)	AD; pMCI; sMCI; HC	147; 112; 102; 146	✓	✗	✗	✓ ^b
	CDR				✓	✗	✗	✓ ^b
	ADAS				✓	✗	✗	✓ ^b
Verbal Fluency	Composite measure ^c	(Cole et al., 2017b)	HIVp; HC	161; 102	✓	✓	✗ ^d	✓ ^b
	Composite measure ^e	(Cole et al., 2017b)	HIVp; HC	161; 102	✗	✓	✗ ^d	✗
	Composite measure ^f	(Richard et al., 2018)	HC	265	✗	✓	✓	✓
	<i>Semantic</i> (Category Fluency Test)	(Franke et al., 2013)	DM2; HC	98; 87	✓	✓	✓	✓
	<i>Phonemic</i> (Letter Fluency Test)	(Cole et al., 2015)	TBI	89	✓	✗	✓	✓
Processing Speed	Composite measure ^g	(Cole et al., 2017b)	HIVp; HC	161; 102	✓	✓	✗ ^d	✗
	Composite measure ^h	(Richard et al., 2018)	HC	265	✗	✓	✓	✓
	TMT-A	(Cole et al., 2015)	TBI	90	✓	✗	✓	✓
	CRT		TBI	66	✓	✗	✓	✓
Processing Speed, Visual Attention, and Cognitive Flexibility	Composite measure ⁱ	(Cole et al., 2017b)	HIVp; HC	161; 102	✓	✓	✗ ^d	✗
	TMT-B	(Cole et al., 2015)	TBI	90	✓	✗	✓	✓
Visual Attention and Cognitive Flexibility	TMT-B minus TMT-A	(Cole et al., 2015)	TBI	90	✓	✗	✓	✓
Response Inhibition and Selective Attention	D-KEFS CWIT Composite measure ^j	(Richard et al., 2018)	HC	265	✗	✓	✓	✓
	D-KEFS CWIT (Inhibition vs Color Naming – scaled)	(Le et al., 2018)	HC, MOOD/ANX, SUD, ED	489	✗	✗	✓	✓
	D-KEFS CWIT (Inhibition/Switching)		TBI	89	✓	✗	✓	✓
	D-KEFS CWIT (Inhibition/Switching minus Baseline Stroop performance)		TBI	89	✗	✗	✓	✓
Sustained Attention	Composite measure ^l	(Cole et al., 2017b)	HIVp; HC	161; 102	✗	✓	✗ ^b	✗
Verbal Episodic Memory	Composite measure ^k (<i>General</i>)	(Cole et al., 2017b)	HIVp; HC	161; 102	✓	✓	✗ ^b	✗
	CVLT Immediate Recall, CVLT Delayed Recall, CVLT Learning 1–5, (<i>all tested separately</i>)	(Richard et al., 2018)	HC	265	✗	✓	✓	✓
	People Test (<i>Immediate</i>)	(Cole et al., 2015)	TBI	90	✓	✗	✓	✓
Working Memory	Composite measure ^l	(Richard et al., 2018)	HC	265	✗	✓	✓	✓
	Blocked Verbal N-back Task	(Scheller et al., 2018)	HC	34	✗	✓	✗	n/a ^m
Motor Function	Composite measure ⁿ	(Cole et al., 2017b)	HIVp; HC	161; 102	✗	✓	✗ ^b	✗
Intelligence	WASI Similarities (<i>Abstract verbal reasoning</i>)	(Cole et al., 2015)	TBI	90	✗	✗	✓	✓
	WASI Matrix Reasoning (<i>Non-verbal reasoning</i>)		TBI	88	✗	✗	✓	✓

Table 1. (continued)

Cognitive Domain	Measure	Reference	Sample	n	Sig.	Sig. in HCs	Age adj.	MC corr.
	Composite measure ^o (Fluid-type intelligence)	(Cole et al., 2018)	HC	669	✓	✓	✓	✗
	Moray House Test (Childhood IQ)				✗	✓	✗	✗

Sig.: results were statistically significant; *Sig. in HC*: results were statistically significant in healthy controls; *Age adj.*: results were adjusted for age; *MC corr.*: results were corrected for multiple comparisons. *MOCA* = Montreal Cognitive Assessment; *MMSE* = Mini Mental State Examination, *MCI* = Mild Cognitive Impairment, *DEM* = Dementia; *CDR* = Clinical Dementia Rating Scale, *AD* = Alzheimer's Disease, *ADAS* = Alzheimer's Disease Assessment Scale, *HC* = Healthy Controls, *APOE* = APOE e4 carrier, *Non APOE* = APOE e4 non-carrier, *sMCI* = Stable MCI, *pMCI* = Progressive MCI, *HIVp* = HIV-positive, *DM2* = Diabetes Mellitus Type 2, *TBI* = Traumatic Brain Injury, *TMT-A* = Trail Making Test A (Time to complete), *CRT* = Choice Reaction Time Task (Median reaction time), *TMT-B* = Trail Making Test B (Time to complete), *D-KEFS CWIT* = Delis-Kaplan Executive Function System D Color-Word Interference Test, *MOOD/ANX* = Mood/Anxiety Disorder, *SUD* = Substance use Disorder, *ED* = Eating Disorder, *CVLT* = California Verbal Learning Test, *WASI* = Weschler Abbreviated Scale of Intelligence

^a No response from authors – Multiple comparison correction not outlined in relation to brainPAD-cognition tests but used elsewhere in paper.

^b Finding not corrected for multiple comparison but likely would have survived Bonferroni correction so not affected by lack of correction.

^c Average of average standardised t-scores (adjusted for age, sex, education) across domains of verbal fluency, processing speed, executive function, memory, attention, and motor function

^d T-scores controlled for effect of age on cognitive scores, but relationship between brainPAD and age was not controlled for, so not fully adjusted for age.

^e Average of standardised t-scores (adjusted for age, sex, education) from Category Fluency and Letter Fluency tests

^f Cluster measure combining Phonological Flow and Semantic Flow measures from CABPad (Willer, Pedersen, Forchhammer, & Christensen, 2016)

^g Average of standardised t-scores (adjusted for age, sex, education) from TMT-A, WAIS-III Digit Symbol and Symbol Search, and Stroop Colour-Word Test

^h Cluster measure combining processing speed parameters based on the Theory of Visual Attention obtained from test battery using CABPad

ⁱ Average of standardised t-scores (adjusted for age, sex, education) from TMT-B and WCST (Number of total errors, perseverative errors and responses)

^j Cluster measure combining scores from the Colour-Naming, Reading, Inhibition, and Inhibition/Switching trails of the D-KEFS CWIT

^k Average of standardised t-scores (adjusted for age, sex, education) from Rey Auditory Verbal Learning test and WMS-IV Visual Reproduction

^l Cluster measure combining measures from working memory test of CABPad

^m Only one test of brainPAD-cognition relationship conducted so multiple comparison correction not necessary

ⁿ Average of standardised t-scores (adjusted for age, sex, education) from Grooved Pegboard and Finger Tapping tasks

^o Index derived from a principal components analysis of WASI-III Letter-number sequencing, digit span backwards, matrix reasoning, block design, digit symbol coding, symbol search

of the present study. The target population were healthy adults.

Participants

Training set. The data were comprised of MRI scans from 1359 healthy adults (mean age 40.04 years, SD = 17.78 years, range = 18.00–88.36 years; 855 females) drawn from various open-access data repositories (see Table S1 in Supplementary Info). Inclusion criteria for the training cohort were: over 18 years old, age and gender data available, and not diagnosed with any neurological, psychiatric or major medical conditions.

Test Set 1 –DEU. The first test set was comprised of 175 community-dwelling adults (mean age = 68.95 years, SD = 8.59 years; range = 47.56–93.51 years; 104 females) recruited as part of a study conducted at Dokuz Eylül University, Izmir, Turkey. Exclusion criteria included history of neurological or

psychiatric diseases, use of psychotropic drugs including cholinesterase inhibitors, traumatic brain injury, history of stroke, drug and/or alcohol addiction and uncontrolled systemic diseases.

Test Set 2 – CR/RANN. The second test set was comprised of 380 community-dwelling adults (mean age = 52.41 years, SD = 17.09 years; range = 19–80 years; 210 females) who participated in the CR/RANN study (Stern et al. 2018; Stern et al. 2014). These participants were screened for MRI contraindications, hearing and visual impairments, medical or psychiatric conditions, and dementia and MCI. Further inclusion criteria were a score of over 135 on the Mattis Dementia Rating Scale (Jurica et al. 2001), a reading level at least equivalent to the US 4th grade, and minimal complaints of functional impairment.

Test Set 3 –TILDA. The third test set was comprised of an MRI subset of a nationally representative longitudinal cohort study

of community-dwelling adults in Ireland (Whelan and Savva 2013). From an initial subset of 502 participants, participants were excluded due to missing a portion of the cerebellum ($n = 2$), a history of Parkinson's disease, stroke, or transient ischemic attack ($n = 11$) and no cognitive data ($n = 2$). The final test set was comprised of MRI data from 487 participants (mean age = 68.6 years, SD = 7.21 years; range = 50–88 years; 260 females).

MRI data acquisition

Training Set. A range of T1-weighted MRI scans from different scanners and using different protocols were used as the training set (see Table S1 in Supplementary Info).

Test Set 1 – DEU. DEU participants underwent a 10 min T1 scan in a 1.5 T Philips Achieva scanner as part of a larger 20-min MRI battery. Two separate protocols were used for scans included here. The Alzheimer's Disease Neuroimaging Initiative (ADNI) T1 protocol was followed for 126 scans using the turbo field echo sequence with the following parameters: number of slices = 166, FOV = 240 mm³, matrix size = 256 × 256, slice thickness = 1 mm, slice gap = 0 mm, TR = 9 ms, TE = 4 ms. For 49 scans, a local protocol using a gradient echo sequence was followed with the following parameters: FOV = 230 mm³, matrix size = 400 × 512, slice thickness = 1 mm, slice gap = 0 mm, TR = 25 ms, TE = 6 ms.

Test Set 2 – CR/RANN. CR/RANN participants underwent a 5 min T1 MPRAGE scan in a 3 T Philips Achieva scanner as part of a larger 2-h imaging battery. The following parameters were used: FOV = 256 × 256 × 180 mm, matrix size = 256 × 256, slice thickness = 1 mm, slice gap = 0 mm, TR = 6.5 ms, TE = 3 ms.

Test Set 3 – TILDA. TILDA participants underwent a 5 min 24 s T1 MPRAGE scan in a 3 T Philips Achieva scanner as part of a larger 45-min MRI battery. The following parameters were used: FOV = 240 × 240 × 162 mm³, matrix size = 288 × 288, slice thickness = 0.9 mm, slice gap = 0 mm, TR = 6.7 ms, TE = 3.1 ms.

MRI pre-processing

All images were preprocessed using SPM12 (University College London, London, UK). Prior to processing, all scans were automatically approximately reoriented (see Supplemental Information; MRI pre-processing) to a canonical SPM template. All scans were then visually inspected for good orientation and gross artefacts before preprocessing. In the test set, badly oriented scans were manually reoriented before preprocessing. In both training and test sets, each individual dataset was preprocessed in a separate batch. Bias

correction was applied to images which were then segmented into GM, WM, and CSF. Segmented GM images were non-linearly registered to a custom template, using SPM's *DARTEL*. Images were then affine registered to MNI space (1 mm³) and resampled with modulation to preserve the total amount of signal from each voxel. Images were smoothed with a 4 mm full-width at half maximum Gaussian kernel. Finally, images were visually inspected for accurate segmentation. The code used to auto-reorient and preprocess the MRI data is available at <https://github.com/roxytboyle/brainPAD>.

Machine learning

Data preparation GM images were resized to 2 mm³ voxels and individual voxel values were extracted from each image. A threshold was applied such that a voxel was retained if it had GM density > 0.2 in that voxel across all 1359 training set images. After thresholding, the training data consisted of 1359 images, each with 54,869 voxels.

Machine learning model The goal of the training phase was to construct a generalizable model that could predict chronological age from GM data. In order to increase generalizability, a data resampling ensemble approach was used. That is, 500 participants, with a 50:50 gender ratio, were randomly sampled without replacement from the training data to form a nested training set. This process was repeated 25 times, creating 25 nested training sets. Each nested training set (500 participants × 54,869 voxels), was used as the input to a regularized linear regression model (Elastic Net), with 10-fold cross-validation (CV), to predict the chronological age of each participant (see Supplementary Info for further information on the machine learning model). The performance of the model was quantified using the mean of each of the 25 nested models' Pearson's correlation between chronological age and predicted age (r), mean absolute error (MAE), and the weighted MAE. The weighted MAE is equal to the MAE divided by the age range of the sample tested and is a more suitable metric for comparing the MAE of brainPAD models across studies as it accounts for the impact of a sample's age range on prediction accuracy (Cole et al. 2019). A lower weighted MAE reflects greater accuracy.

Application to independent test sets First, the average coefficient value for each voxel across all folds in all 25 training models was calculated, resulting in a vector of length 54,869. For each independent test set, the mean coefficient values were multiplied by the voxels' GM density values and the product was summed to create a brain-age prediction for each participant. To correct for the proportional bias in the model, the prediction was added to the intercept of the training set, and the result was then divided by the slope of the training set. This correction does not affect the relationship between

brainPAD and outcome measures but scales the data correctly so that brainPAD scores can be interpreted in units of years proportional to a person's chronological age. Similar corrections have been applied in other brainPAD models (Cole et al. 2018). BrainPAD was calculated by subtracting chronological age from the corrected predicted age, hence, a positive brainPAD value indicates a brain-predicted age that exceeds the participant's chronological age, suggesting accelerated brain ageing. The code used to make brain-age predictions and calculate brainPAD scores for independent test sets is available at <https://github.com/rorytboyle/brainPAD>.

Cognitive function measures

Each of the three datasets contained a wide range of cognitive measures. However, as the three datasets were completely independent of one another, and all data collection was completed prior to conception and design of the present study, different cognitive measures were used across the datasets. For the purposes of the present study, a cognitive measure was selected for analysis if it assayed a cognitive domain that was assessed in at least one other dataset. For example, the AMNART and NART assessed premorbid intelligence in CR/RANN and TILDA respectively so both measures were selected for analysis and considered as 'comparable' measures. The cognitive domains assessed by each measure were decided with reference to the literature as outlined in the Supplementary Information. Across all three datasets, 17 common cognitive domains were identified (see Table 2 for list of cognitive domains and cognitive measures used and Supplementary Information for detailed descriptions of each cognitive measure).

Statistical analysis

The statistical analysis was conducted using the following procedure:

1. **Correlate.** Within each independent test set, partial Spearman's rank order correlations were conducted between brainPAD scores and cognitive measures, controlling for chronological age and sex. Sex was adjusted for to account for a significant sex difference in brainPADs ($p < 0.0001$).
2. **Replicate.** For findings replicated in multiple datasets, the probability of obtaining p -values by chance was calculated by random-label permutation (see Supplementary Methods for further detail). Briefly, this involved randomly shuffling brainPAD scores, conducting Spearman's partial correlations between randomly shuffled brainPAD scores and the cognitive dependent variables, controlling for age and sex. This process was repeated one million times. The number of times in which all random p -values were more extreme (i.e. smaller) than the actual p -values was summed and divided by one million to obtain the

probability of the finding replicating across multiple datasets by chance. Replicated findings were deemed significant if this probability was less than .05.

3. **Correct for multiple comparisons.** All other correlations were then corrected for multiple comparisons, while allowing for correlations among dependent cognitive variables, using a maximum statistic approach (see Supplementary Methods for further detail). Briefly, in each test set, brainPAD scores were randomly shuffled and then Spearman's partial correlations were conducted between the randomly shuffled brainPAD scores and the cognitive dependent variables, controlling for age and sex. This process was repeated one million times and the maximum rho value was stored each time. Correlations between actual brainPAD scores and cognitive variables were deemed significant if they exceeded the 95th percentile of the maximum rho values.

Results

Brain age prediction

Training set

The model accurately predicted chronological age ($r = 0.85$, $MAE = 7.28$ years, weighted $MAE = 0.10$, $p < 0.0001$). As with other brain PAD models (e.g., Cole et al. 2018), a proportional bias was observed in this model where chronological age correlated with prediction error ($r = -0.4452$, $p = 1.1036e-10$).

Independent test sets

The model accurately predicted chronological age in each of the three independent test sets (see Table 3).

Sex differences in brainPAD

Mean brainPAD differed significantly by sex in all datasets, Welch's $t(1009.55) = -5.79$, $p < .0001$. Males ($M = -1.81$, $SD = 9.92$) had significantly lower brainPADs than females ($M = 1.81$, $SD = 10.23$; see Fig. 1). Within individual test sets, males had significantly lower brainPADs, compared to females, in in CR/RANN ($p < .0001$) and TILDA ($p < .0001$) but not in DEU ($p = 0.148$; see Fig. 2).

Brain regions involved in brain age prediction

The voxel-wise method used here to predict brain age resulted in individual coefficient values for each voxel. Voxels with positive coefficient values contributed to older brain age predictions and voxels with negative coefficient values

Table 2 Cognitive measures available across each dataset in comparable cognitive domains.

Cognitive Domain(s)	DEU Measure (<i>N</i>)	CR/RANN Measure (<i>N</i>)	TILDA Measure (<i>N</i>)
General Cognitive Status	MMSE (total score) (172)	DRS (370)	MMSE (485)
Premorbid Intelligence	n/a	AMNART (362)	NART (486)
Phonemic Verbal Fluency	KAS Test (137)	CFL Test (360)	n/a
Semantic Verbal Fluency	Animals Test (175)	Animals Test (361)	Animals Test (487)
Processing Speed	TMT A (93)	TMT A (361)	CTT 1 (487)
Processing Speed, Visual Attention, Cognitive Flexibility	TMT B (84)	TMT B (357)	CTT 2 (482)
Visual Attention, Cognitive Flexibility	TMT B minus TMT A (84)	TMT B minus TMT A (357)	CTT 2 minus CTT 1 (482)
Cognitive Flexibility	WCST Perseverative Error Percentage (50)	WCST Perseverative Error Percentage (327)	n/a
Response Inhibition, Selective Attention	Stroop (Turkish Capa version; Emek-Savaş, Yerlikaya, Yener, & Öktem, 2019) Interference Score - Time (150)	Stroop (Golden version; Golden, 1978) Interference Score - Words (359)	n/a
Sustained Attention (Errors of Commission)	n/a	PVT False Alarms (176)	SART Errors of Commission (482)
Sustained Attention (Reaction Time)	n/a	PVT Median Reaction Time (176)	SART Coefficient of Variation in Reaction Time (479)
Verbal Episodic Memory (Immediate)	OVMPT Immediate Recall (175)	SRT Total Score (360)	Immediate Recall (487)
Verbal Episodic Memory (Delayed)	OVMPT Delayed Recall (175)	SRT Delayed Recall (360)	Delayed Recall (487)
Verbal Episodic Memory (Learning)	OVMPT Total Learning Score (175)	SRT Consistent Long Term Retrieval (360)	n/a
Working Memory	WMS-R Digit Span Forward Test (171) WMS-R Digit Span Backward Test (170)	WAIS-III Letter Number Sequencing Test (360)	n/a
Visuospatial Ability	BLOT (80)	WAIS-III Block Design Test (356)	n/a

See Table S.4 for full information on each measure; MMSE = Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975); DRS Total Score = Mattis Dementia Rating Scale-2 – Total Score (Jurica et al., 2001); NART = National Adult Reading Test (Nelson & Willinson, 1982); AMNART = American National Adult Reading Test (Grober & Sliwinski, 1991); CTT = Colour Trails Test (D’Elia, Satz, Uchiyama, & White, 1996); TMT = Trail Making Test (Reitan, 1955); WCST = Wisconsin Card Sorting Test (Heaton, Chelune, Talley, Kay, & Curtiss, 1993); SART = Sustained Attention to Response Test (Robertson, Manly, Andrade, Baddeley, & Yiend, 1997); PVT = Psychomotor Vigilance Task (Dorrian, Rogers, & Dinges, 2005); OVMPT = Öktem Verbal Memory Processes Test (Öktem, 1992); SRT = Selective Reminding Test (Buschke & Fuld, 1974); WMS-R = Wechsler Memory Scale (Wechsler, 1987); WAIS-III = Wechsler Adult Intelligence Scale – Third Edition (Wechsler, 1997); BLOT = Benton’s Judgement of Line Orientation Test (Benton, Varney, & Hamsher, 1978)

contributed to younger brain age predictions. Fig. 3 shows all voxels with binarised negative and positive coefficient values, respectively. However, as it is difficult to visualize a 3D object containing a large number of small voxels, a .nii file of the regression coefficients can be downloaded here: <https://osf.io/5n6t8/>. Overlaying this .nii file in a viewer such as mriCGL will allow for a more

accurate view of the regression coefficients. Moreover, an .xlsx file containing the absolute value of each regression coefficient, the sign direction of that coefficient (i.e. positive/negative), the coefficient rank in terms of absolute values, as well as the MNI coordinates and anatomical labels of the coefficients, is available here: <https://osf.io/dkz67/>.

Table 3. Results of application of trained model parameters to 3 independent test sets.

Test Set	Pearson’s <i>r</i>	Mean brainPAD	SD brainPAD	MAE	Weighted MAE
Test Set 1 – DEU	0.78*	+6.60	6.44	7.60	0.17
Test Set 2 – CR/RANN	0.87*	+6.39	8.57	8.56	0.14
Test Set 3 - TILDA	0.65*	–6.97	7.52	8.42	0.22

* = $p < 10^{-37}$. Pearson’s *r* between brain age and chronological age. Weighted MAE = MAE divided by age range

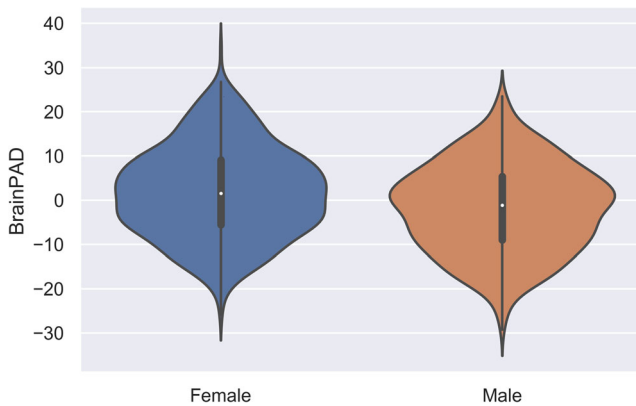


Fig. 1 Violin plot comparing distributions of brainPADs between sexes across all datasets

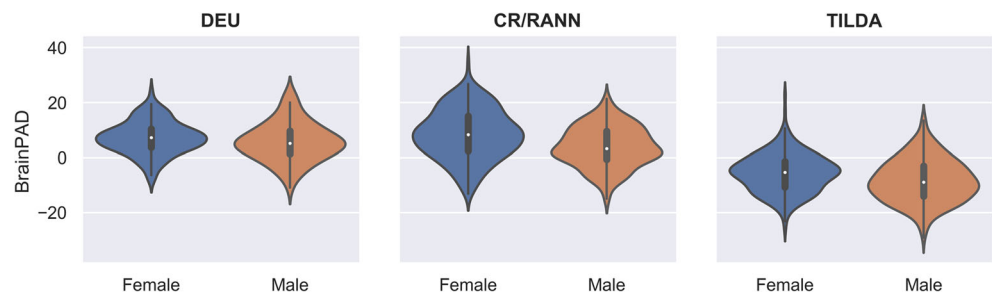
BrainPAD and cognitive function

Across multiple datasets, higher brainPAD scores were significantly correlated with reduced performance on measures of general cognitive status, semantic verbal fluency, processing speed, cognitive flexibility, and visual attention (see Fig. 4 and Table 4).

Discussion

A penalized regression approach was able to produce accurate brain-age predictions from T1 MRI data in three independent datasets. In non-demented adults, brainPAD, calculated by subtracting these brain-age predictions from chronological age, was negatively correlated with general cognitive status, semantic verbal fluency, processing speed, visual attention, and cognitive flexibility; and visual attention and cognitive flexibility in multiple datasets. BrainPAD was significantly correlated with phonemic verbal fluency, premorbid intelligence, verbal episodic memory (learning score), and visuospatial ability in single datasets after controlling for multiple comparisons; however, these correlations were not replicated in another dataset so we do not have strong evidence here in support of these relationships. BrainPAD was not significantly correlated with processing speed, cognitive flexibility, response inhibition and selective attention, sustained attention, verbal episodic memory (immediate recall or delayed recall), or working memory in any dataset.

Fig. 2 Violin plots comparing distributions of brainPADs between sexes within datasets

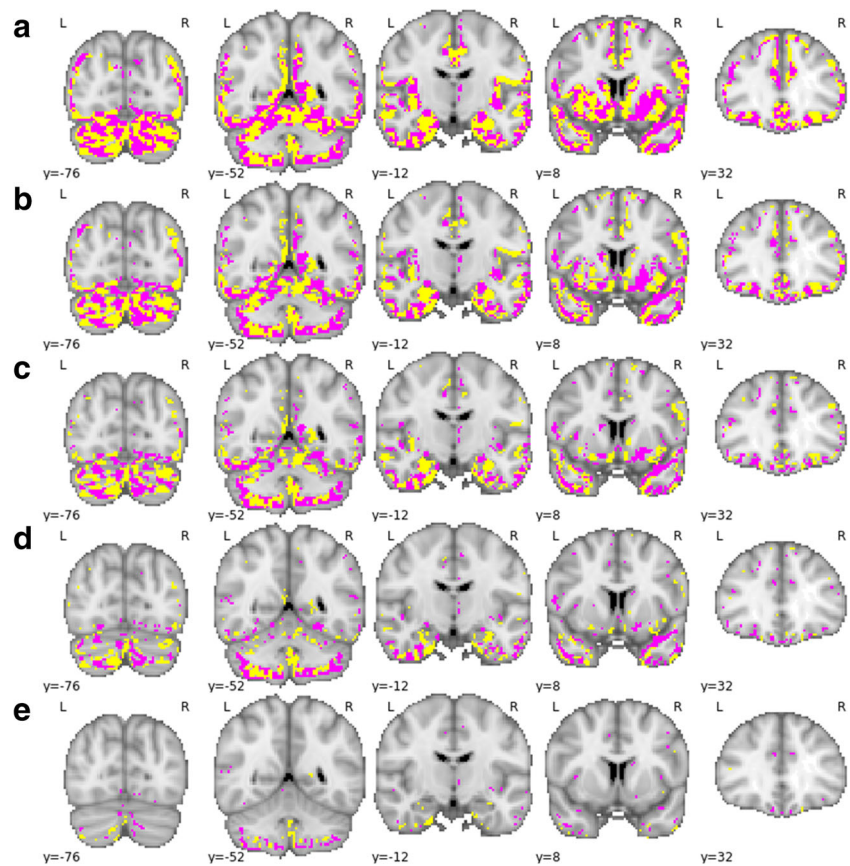


BrainPAD and cognition

General cognitive status

BrainPAD was negatively correlated with general cognitive status, as measured using the MMSE and DRS, in DEU and CR/RANN, and the replication of this result across both datasets was statistically significant. However, brainPAD was not significantly correlated with the MMSE in TILDA. Nonetheless, given the statistically significant replication across two of the three datasets, there is reliable evidence in support of the correlation between brainPAD and general cognitive status in healthy older adults. Previous studies have reported that brainPAD is related to general cognitive status, albeit in samples including individuals with MCI, AD, or dementia (Beheshti et al. 2018; Kaufmann et al. 2019), and without adjusting for the effect of age or controlling for multiple comparisons (Beheshti et al. 2018; Cole et al. 2017b; but see Table 1 Footnote 4 for information about adjusting for age in Cole et al. 2017b). In contrast to our findings, Gaser et al. (2013) reported that brainPAD was correlated with the CDR and ADAS but not the MMSE in an MCI sample. However, Gaser et al. (2013) did not account for the effect of age. While Löwe et al. (2016) reported that brainPAD was negatively correlated with the MMSE across mixed samples of APOE e4 carriers and non-carriers (including healthy controls, MCI, and AD), it was not significantly correlated with the MMSE within healthy control and MCI subgroups. Sample sizes within these subgroups were relatively small, ranging from 14 to 81 participants. Consequently, the correlations between brainPAD and the MMSE in these participants may not have been adequately powered to reach significance. Our study is the first to report a relationship between brainPAD and measures of general cognitive status in healthy adults while controlling for the effects of age and correcting for multiple comparisons. This is also the first study to investigate the relationship between brainPAD and DRS score. Our findings provide strong support for the existence of a significant negative relationship between brainPAD and general cognitive status. As such, this finding provides some preliminary support in favour of brainPAD as an objective measure of general cognitive function given that brainPAD is not subject to the various biases and effects (e.g. low reliability, practice effects) that

Fig. 3. Binarised regression coefficients (positive coefficients shown in pink, negative coefficients shown in yellow) overlaid on 5 coronal slices. A: No threshold applied; B: thresholded at 25th percentile of absolute value of regression coefficients; C: thresholded at 50th percentile of absolute value of regression coefficients; D: thresholded at 75th percentile of absolute value of regression coefficients; E: thresholded at 95th percentile of absolute value of regression coefficients



limit the MMSE (Galasko et al. 1993; Pfeffer et al. 1984; Tombaugh and McIntyre 1992) and the DRS (Emery et al. 1996; Green et al. 1995).

Semantic verbal fluency

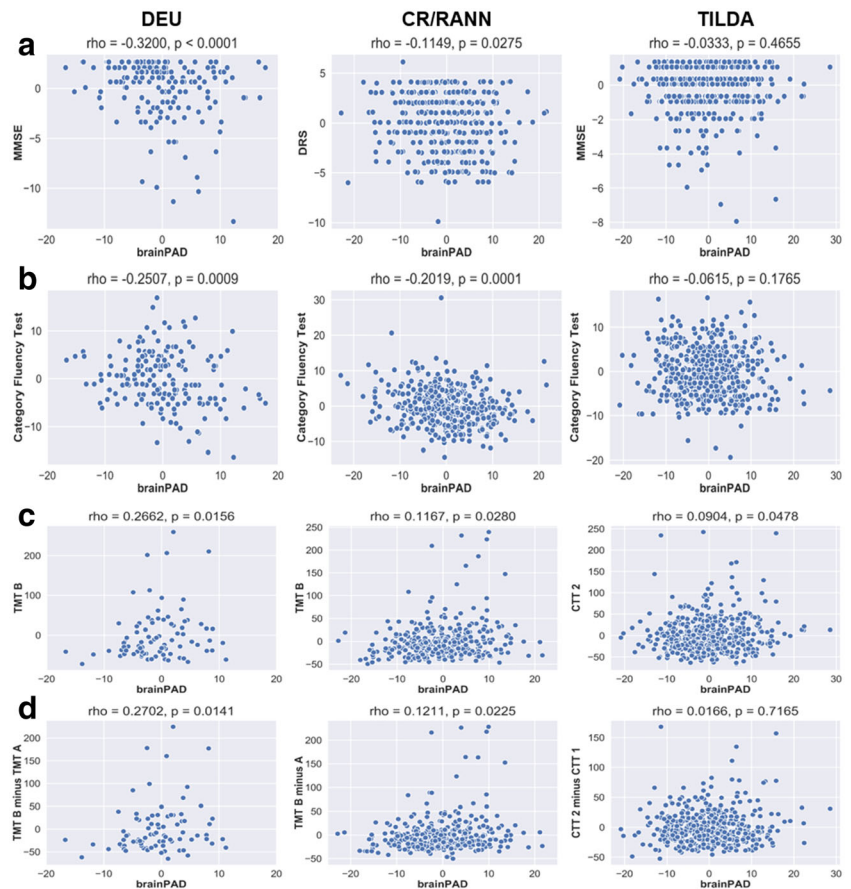
BrainPAD was significantly negatively correlated with semantic verbal fluency, as measured using the Animals task, in both DEU and CR/RANN but not in TILDA. Regardless, the replication of this result across both DEU and CR/RANN was statistically significant. This finding contradicts non-significant correlations between brainPAD and composite measures of semantic and phonemic verbal fluency (Cole et al. 2017b; Richard et al. 2018), although the former study used age-adjusted t-scores to control for the age-cognition relationship rather than adding age as a covariate to the brainPAD-fluency analysis (cf. Le et al. 2018). As semantic verbal fluency is associated with age (Clark et al. 2009; Santos Nogueira et al. 2016), the failure to adjust for age may have obscured a significant effect. Alternatively, these previously reported non-significant correlations could be explained by the use of composite measures of both phonemic and semantic fluency as we did not find strong evidence for a relationship between phonemic verbal fluency and brainPAD (although it was significant in DEU, this correlation was not replicated in CR/RANN). Therefore, it is possible

that a non-significant relationship between phonemic fluency and brainPAD in the Cole et al. (2017a) and Richard et al. (2018) study may have diluted a possible significant relationship between semantic fluency and brainPAD. In a study controlling for age, brainPAD was found to significantly negatively correlate with semantic verbal fluency (Franke et al. 2013). Although the Animals task has been described as an optimal test of neuropsychological function (Ardila et al. 2006), scores on this task are affected by various factors, including scoring and administration procedures (Woods et al. 2016) and practice effects (Cooper et al. 2001; Harrison et al. 2000; Wilson et al. 2000). As such, brainPAD, as an objective marker of general brain health and global cognitive function, could be a viable alternative to the Animals task. In sum, our results provide further evidence in support of a correlation between brainPAD and semantic verbal fluency.

Processing speed, visual attention, and cognitive flexibility

Across all three datasets, brainPAD was negatively correlated with processing speed, visual attention, and cognitive flexibility as measured by trail-making tests (TMT B or CTT 2). The TMT B is a relatively sensitive measure of cognitive decline: completion times were shown to be significantly different between healthy controls, MCI, and AD (Ashendorf et al. 2008).

Fig. 4 Scatterplots of replicated correlations between the residuals of brainPAD and cognitive measures after regressing brainPAD on age and sex, and each cognitive measure on age and sex. A: General cognitive status; B: Semantic verbal fluency; C: Processing speed, visual attention, and cognitive flexibility; D: Visual attention and cognitive flexibility. For scatterplots of non-replicated correlations, see Supplementary Info, Fig. S.4



Likewise, the CTT 2 is sensitive to cognitive decline, with differences between AD and healthy controls (Lin et al. 2014), and between healthy controls, MCI, and AD (Guo et al. 2010). Therefore, it is no surprise that processing speed, visual attention, and cognitive flexibility were also negatively correlated with an index of accelerated brain ageing. Indeed, previous studies have reported similar results for trail-making versus brainPAD; however, these studies did not correct for multiple comparisons (Cole et al. 2017b) or used clinical samples (TBI; Cole et al. 2015). Our data therefore augment these findings by replicating this result across three independent datasets. This evidence could suggest that brainPAD may be a potential objective measure of cognitive decline as it is not subject to the same factors which bias trail-making performance, including to practice effects (Bartels et al. 2010), rater effects (Feeney et al. 2016) and participant literacy (Vaucher et al. 2014).

Visual attention and cognitive flexibility

BrainPAD was also negatively correlated with visual attention and cognitive flexibility (TMT B minus A), in DEU and CR/RANN, but not in TILDA (CTT 2 minus 1). Replication of this finding in DEU and CR/RANN suggests a modest association between visual attention and cognitive flexibility. The relationship between brainPAD and TMT B minus A was only

investigated in one previous study, in a TBI sample, (Cole et al. 2015) where a significant positive correlation was reported. Although the TMT B minus A can distinguish between stable and progressive MCI on a group level (Zanetti et al. 2006), and is associated with reduced mobility, increased mortality risk (Vazzana et al. 2010) and slower walking speed (Ble et al. 2005), as a derived measure of the TMT, the TMT B minus A index is similarly affected by the various factors that can limit interpretation of the TMT B scores. Therefore, given the correlation shown here between TMT B minus A and brainPAD, brainPAD may be a potential objective measure of general cognitive function.

It is notable that several significant brainPAD-cognition relationships were observed in the DEU and CR/RANN datasets, but not in TILDA. We tentatively offer some suggestions for this pattern of results. Confounding factors obscuring the brainPAD-general cognitive status relationship may have been uniquely present in TILDA. Whereas the DEU and CR/RANN cohorts were part of neuroimaging research studies, which have typically strict inclusion criteria, the TILDA MRI sample were a subset of a large nationally representative longitudinal study encompassing health, economic and social research (Whelan and Savva 2013). TILDA therefore had few inclusion criteria: being at least 50 years old, having a residential address, and absence of dementia at baseline (Kearney et al. 2011; Savva et

Table 4. Results of Spearman's partial correlations between brainPAD and 17 cognitive domains.

Cognitive Domain	DEU			CR/RANN			TILDA			Probability of replicating by chance	Sig. by max statistic correction (where finding not replicated)
	rho	df	p	rho	df	p	rho	df	p		
General Cognitive Status	-0.3199	168	<0.0001	-0.1449	366	0.0275	-0.0333	481	0.4655	<0.00001	n/a
Premorbid Intelligence	n/a			-0.2322	358	<0.0001	0.0485	482	0.2873	n/a	CR/RANN
Phonemic Verbal Fluency	-0.3259	134	0.0001	-0.0771	356	0.1454	n/a			n/a	DEU
Semantic Verbal Fluency	-0.2507	171	0.0009	-0.2019	357	0.0001	-0.0615	483	0.1765	<0.00001	n/a
Processing Speed	0.1232	89	0.2448	0.0595	357	0.2610	0.1208	483	0.0077	n/a	None
Processing Speed, Visual Attention, Cognitive Flexibility	0.2662	80	0.0156	0.1167	353	0.0279	0.0904	478	0.0478	0.00005	n/a
Visual Attention, Cognitive Flexibility	0.2702	80	0.0141	0.1211	353	0.0225	0.0166	478	0.7165	0.00097	n/a
Cognitive Flexibility	0.0722	46	0.6258	0.0429	323	0.4411	n/a			n/a	None
Response Inhibition, Selective Attention	0.0854	146	0.3019	-0.1755	355	0.0009	n/a			n/a	None
Sustained Attention (Errors of Commission)	n/a			0.0203	172	0.7902	0.0499	478	0.2752	n/a	None
Sustained Attention (Reaction Time)	n/a			-0.0212	172	0.7813	0.0436	475	0.3425	n/a	None
Verbal Episodic Memory (Immediate)	0.2194	171	0.0037	-0.0407	356	0.4428	-0.0347	483	0.4114	n/a	None
Verbal Episodic Memory (Delayed)	0.2797	171	0.0002	0.0343	356	0.5173	0.0122	483	0.7887	n/a	None
Verbal Episodic Memory (Learning)	-0.3196	171	<0.0001	0.0657	356	0.2151	n/a			n/a	DEU
Working Memory	-0.1310	167	0.0895 ^a	-0.0469	360	0.3759	n/a			n/a	None
	-0.2974	166	0.0001 ^b								
Visuospatial Ability	-0.0809	76	0.4815	-0.1824	352	0.0006	n/a			n/a	CR/RANN

^a Digit Span Forwards; ^b Digit Span Backwards

al. 2013). TILDA's MRI sample were screened for MRI contraindications and were on average healthier than the full sample, but it is likely that the TILDA sample included participants who might normally be excluded from neuroimaging research studies (e.g., those using psychotropic or cardiovascular medication). Moreover, the range of some cognitive measures in TILDA was also smaller than DEU and CR/RANN in some cases (see Supplemental Information: Table S5): notably for general cognitive status, and visual attention and cognitive flexibility, where the brainPAD-cognition correlations were not replicated within TILDA. Restricted range of scores on these measures in TILDA may have contributed to smaller correlation coefficients (Bland and Altman 2011; Mendoza and Mumford 1987). Additionally, the age range within TILDA was smaller than both DEU and CR/RANN which may have reduced the statistical power of the brainPAD-cognition correlations within TILDA as range restriction on covariates has also been shown to reduce power (Miciak et al. 2016) and decrease the magnitude of correlation coefficients (Sackett and Yang 2000).

The smaller age range within TILDA (38 years) as compared to DEU (45.95 years) and CR/RANN (70 years) might also have contributed to the weaker correlation between chronological age and brain-predicted age in TILDA, as *range*

restriction will reduce the size of correlation coefficients (Goodwin and Leech 2006). Moreover, a negative mean brainPAD was reported in TILDA (-6.97 years) whereas both DEU and CR/RANN had positive mean brainPADs, +6.6 and +6.39 years respectively. Various factors, including intelligence, educational attainment, and environmental factors, have been proposed to affect brain ageing (Irimia et al. 2015). TILDA had significantly higher levels of education versus both CR/RANN and DEU (see Supplementary Results). Steffener et al. (2016) reported that brainPAD was significantly related to education, with higher education associated with younger brains (or smaller/more negative brainPADs). This association with education could be one reason why much lower mean brainPADs were observed for the TILDA dataset. As the cohorts are each from different countries, there could be various other environmental factors that could further explain this relationship.

Model evaluation

We evaluated our model based on its predictive accuracy in three independent test sets, as proposed by Madan and Kensinger (2018). While internal cross-validation is a

valuable and widely used technique that can attenuate overfitting (Arlot and Celisse 2010); the use of cross-validation in certain situations and when it is not implemented correctly, can result in overestimated prediction accuracy and overfitting (Saeb et al. 2016; Skocik et al. 2016; Varoquaux et al. 2017). For brainPAD to be considered for clinical use, it must perform accurately with MRIs acquired in different scanners and under different protocols. However, in most instances of cross-validation, while the test set is split and held completely independent from the training set, factors common to both sets, such as scanner and protocol, could influence model performance. As such, the gold-standard evaluation for brainPAD should be accurate performance on independent external datasets.

The significant correlations between chronological age and brain-predicted age in all three external datasets shows that our model is accurate and generalizable (0.65, 0.78, and 0.87 for external datasets). Although the magnitude of these correlations is lower than correlations reported elsewhere, ranging from 0.91 to 0.94 (Cole et al. 2015; Cole et al. 2017a; Franke et al. 2010; Lancaster et al. 2018; Liem et al. 2017), it exceeds other externally validated brain-predicted age studies, ranging from 0.65 to 0.85 (Beheshti et al. 2018; Madan and Kensinger 2018; Varikuti et al. 2018).

With respect to mean absolute error (MAE), our model did not perform as well as other externally validated studies, ranging from 4.28 to 7.5 years (Beheshti et al. 2018; Cole et al. 2018; Franke et al. 2010; Lancaster et al. 2018; Madan and Kensinger 2018). As a result, it could be possible that we may have lost some precision by not integrating WM information as input in the model, as was done by Cole et al. (2018), for example. Another potential reason is that other studies centered the age predictions using the mean of the ages from the test set. Although this correction is typically not explicitly described in method sections, Madan and Kensinger (2018) note that this is a standard correction in brain age prediction. Moreover, some studies also match the variance in predicted age in the test set with the variance of the training data (Madan and Kensinger 2018). Both corrections are principled and acceptable methods of correcting for the regression to the mean artefact in brain age predictions but they result in biased age predictions in the test set. These corrections also limit the use of brainPAD to make single subject predictions, as both the test set mean and variance are used in the prediction. Our method used only training set information and therefore produced slightly less accurate but less biased predictions. Finally, our model may also appear to be less precise in terms of MAE as an artefact of the greater age range of our sample in comparison to most brainPAD studies. An alternative metric, the weighted MAE (calculated by dividing the MAE by the age range of the sample), may enable better comparisons across studies with different age ranges (Cole et al. 2019). While our weighted MAE is higher than some studies, ranging

from 0.072 to 0.087 (Lancaster et al. 2018; Liem et al. 2017), the lowest weighted MAE in our sample (0.14 in CR/RANN) outperformed this metric when calculated for other studies, 0.178 (Beheshti et al. 2018), and 0.18 (Varikuti et al. 2018) and is comparable to 0.139 (Franke et al. 2010, ‘Test 4’ external test set). As such, the predictive accuracy of our model is comparable to the rest of the literature and is arguably less biased as only training set information is used.

Sex differences in brainPAD

There were significantly higher mean brainPADs in females in two of the three datasets in this study (TILDA and CR/RANN). There is mixed evidence in relation to sex differences in other brain age prediction studies, with some studies reporting significantly higher mean brainPADs in males (Cole et al. 2018; Franke et al. 2013; Luders et al. 2016), some reporting no significant sex differences (Azor et al. 2019; Cruz-Almeida et al. 2019; Franke et al. 2014; Han et al. 2019), and another study, with a notably large sample size of 19,000, reporting higher mean brainPADs in females (Smith et al. 2019). Even studies using the same training sets have contrasting results in terms of sex effects. For example, the same training set resulted in significantly higher male brainPADs in two studies (Franke et al. 2013; Luders et al. 2016) but no sex differences in another study (Franke et al. 2014); however, this divergence could be due to the likely mean centering of both brainPADs in both sexes in the latter study (i.e., male and female groups had mean brainPADs of 0 years). This was also the case in another training set used in multiple studies, with one study reporting significantly higher male brainPADs (Cole et al. 2018) but another reporting no significant differences (Azor et al. 2019). As such, it is likely that sex differences in brainPAD reflect the characteristics of the test sample. This is apparent in the present study with two out of the three datasets showing higher mean female brainPAD but one dataset showing no significant differences. We therefore recommend that future brain age studies report sex differences.

Model interpretation

Model interpretability

The interpretability of machine learning models is an important and widely discussed problem (Doshi-Velez and Kim 2017), and although it is poorly defined (Lipton 2018) it has been described as “*the ability to explain or to present in understandable terms to a human*” (Doshi-Velez and Kim 2017, p. 2) and elsewhere as the ability to “*understand the contribution of individual features in the model*” (Lou et al. 2012, p. 1). Additionally, Lipton (2018) argued that for a model to be considered truly interpretable, it should possess the following

three properties: *algorithmic transparency* (i.e. it should be possible to understand the mechanism by which the model works), *decomposability* (each part of the model, such as the model input and parameters, should have an intuitive explanation), and *simulatability* (a person should be able to consider the entire model at once). We contend that our model possesses these three properties as well as conforming to the definitions proposed above. First, our model possesses algorithmic transparency in that the Elastic Net is a penalized linear regression. Second, our model possesses decomposability. The inputs to the model were GM voxel density values and the parameters, or beta coefficient values, weighted the contribution of each individual value to the model output, which is brain predicted age. Third, our model possesses simulatability as the entire model can be considered as follows: summing the multiplication of GM voxel density values by the average contribution of these voxels to the prediction of chronological age in the training set (i.e., the beta coefficient values) resulted in a prediction of a new individual's brain age.

Biological interpretability

Our statistical model of brain age contains many adjacent voxels that have opposite signs. The negative weights represent those areas with less volume (associated with older age). The positive weights show areas that have more volume associated with older age, which may seem counterintuitive: we propose that this is because GM in these areas represents a shift away from the cortex or periventricular regions (i.e., younger participants would have WM or CSF in those regions). An example of a similar result can be seen in an Alzheimer's disease classification study (Dubois et al. 2014), which also used penalized regression.

Limitations

While we argue that the current model has good biological interpretability, this could be further improved by forcing sparsity to limit the number of voxels making significant contributions to brain age predictions. Modified Elastic Net algorithms, such as Enet-BETA (Liu and Li 2017), can obtain sparser models which would reduce the number of predictive voxels, thereby further improving interpretability. However, as the Elastic Net's prediction accuracy can increase with feature set size (Jollans et al. 2019), further limiting the feature set size could reduce model accuracy. As such, it might be difficult to achieve the right balance between interpretability and accuracy. An alternative approach could be to incorporate a penalty such as Total Variation within the Elastic Net in order to take into account the spatial structure of MRI data and produce weight maps that show the predictive voxels clustered in regions rather than dispersed across the brain (Dubois et al. 2014). These algorithms have been shown to produce models with greater

biological interpretability (i.e. spatially organized weight maps) and comparable predictive accuracy to regular Elastic Net models for classification problems (Dubois et al. 2014). However, the technical implementation of such algorithms can be difficult and computationally expensive, although solutions such as early stopping and feature screening, have been proposed (Dohmatob et al. 2015).

Another possible limitation of the current model is that it uses only voxel-wise GM density data and thus our model may have lower accuracy due to this restricted feature set. Other brain age models have used feature sets including combinations of cortical and subcortical GM regional volumes (Steffener et al. 2016); combinations of GM voxel density values, cortical thickness, and regional volume data (Gutierrez Becker et al. 2018); combinations of cortical thickness, cortical surface area, subcortical volume, and functional connectivity information (Liem et al. 2017); and combinations of GM and WM voxel-wise density information (Cole et al. 2015, 2018; Cole et al. 2017b). Notably, diffusion tensor imaging metrics and cortical thickness have been related to the cognitive domains assessed here, including semantic verbal fluency (Eastman et al. 2013; Rodríguez-Aranda et al. 2016) and processing speed, visual attention, and cognitive flexibility (Ciulli et al. 2016; Dickerson et al. 2008). More complex feature sets, which combined different feature types and imaging modalities, resulted in higher predictive accuracy versus single feature sets (Liem et al. 2017). As such, it is possible that a more accurate model using a more complex feature set would strengthen the brainPAD-cognition correlations reported here. However, such feature sets typically require dimension reduction such as PCA (Gutierrez Becker et al. 2018) or even dot products to combine GM and WM data (Cole et al. 2015, 2018; Cole et al. 2017b). These steps can reduce the interpretability of the relationship between the original feature and brain age (Mateos-Pérez et al. 2018), although methods exist for making such feature sets interpretable (Honeine and Richard 2009; Kwok and Tsang 2004; Snyder et al. 2013). However, our aim was to produce an interpretable model with a relatively straight-forward method, an aim which required a simple feature set. While this approach may have limited our model's accuracy as larger and more complex feature sets often produce more accurate predictions (Scheinost et al. 2019), our model's accuracy is still comparable to other models reported to-date in the literature.

The major limitation of our study is that for the majority of the cognitive domains investigated here, we used different cognitive measures to assess the putatively same cognitive processes. For example, although we considered the CTT 2 as a direct 'culture-free' analogue of the TMT B, as it is widely described (Elkin-Frankston et al. 2007; Messinis et al. 2011), the CTT 2 has different stimuli (shapes and colors vs numbers and letters) and takes longer because it has more stimuli (Mitrushina et al. 2005). Consequently, some have argued,

based on findings of significant difference in mean scores on CTT 2 and TMT B, that the tests are not direct equivalents (Dugbartey et al. 2000; Strauss et al. 2006). However, mean scores for both measures are calculated as time to completion and thus a difference in means between both measures reflects a difference primarily in test length. A more appropriate measure of test equivalence would be correlations between mean scores, and various studies report significant correlations between both measures (Dugbartey et al. 2000; Elkin-Frankston et al. 2007; Lee et al. 2000; Messinis et al. 2011). Similar arguments might be made for the other tests (e.g. the MMSE and DRS) that we used to assess the same cognitive constructs (e.g. general cognitive status). While it would be preferable to use the identical measures across datasets, our study used existing data and was designed after data collection. As a result, this approach was not possible here. Nonetheless, the measures used here were broadly comparable in that they are apparent measures of the same underlying cognitive constructs and it is these constructs which we are most interested in, more so than the actual measures.

Conclusions

The brain age model presented here is accurate and generalizable as it significantly predicts chronological age in three independent datasets. Furthermore, this model is interpretable and biologically plausible as older brain age is driven by decreased GM density in voxels that have been previously shown to be vulnerable to GM atrophy and volume loss. Finally, brainPAD scores, calculated using this model, are associated with reduced cognitive performance within the domains of general cognitive status; semantic verbal fluency; processing speed, visual attention, and cognitive flexibility; and visual attention and cognitive flexibility. The replication of these correlations in multiple datasets demonstrates that the relationship between brainPAD and these domains of cognitive function is robust to cultural- and site/scanner effects. As such, given that brainPAD is also not limited by task effects which can hinder neuropsychological assessment, these findings provide support for the use of brainPAD as an objective measure of general cognitive function with applications as a general measure of brain health and cognitive performance in the clinic and as a summary outcome measure for intervention studies in research settings.

Acknowledgements The authors would like to thank all participants who participated in the various studies which are used here.

Author's contributions Author contributions included conception and study design (RW and RB), data collection or acquisition (GGY, DDES, YS, JPM, SPK, DC, and RAK), preprocessing and quality control of MRI data (RR and RB), statistical analysis (LMRD, LJ, RW, and RB), interpretation of results (RW and RB), drafting the manuscript work (RW and

RB), revising the manuscript critically for important intellectual content (DC, DDES, IHR, LMRD, LJ, YS, RAK, RW and RB) and approval of final version to be published and agreement to be accountable for the integrity and accuracy of all aspects of the work (All authors).

Funding information RB is supported by the Irish Research Council under grant number EPSPG/2017/277. LMRD and RW are supported by the Science Foundation Ireland under grant number 16/ERC/3797. RR is supported by a PhD scholarship funded by the Region Calabria. Data collection in Dokuz Eylul University, managed and supervised by GGY and DDS, was partially supported by the Turkish National Science and Research Council (TUBITAK, Grant number: 112S459) and the Dokuz Eylul University Scientific Research Projects (Grant number: 2018.KB.SAG.084). The Irish Longitudinal Study on Ageing is funded by core grants from the Health Research Board, Atlantic Philanthropies and Irish Life. MRI data collection in TILDA was supported by the Centre for Advanced Medical Imaging (CAMI) at St. James' Hospital, Dublin. IHR thanks The Atlantic Philanthropies for their grant to the Global Brain Health Institute. YS is supported by NIA RF1 AG038465 and R01 AG026158. The funding agencies had no involvement in the conduct of the research or preparation of the article.

Compliance with ethical standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

Conflict of interest The authors report no conflict of interest.

References

- Ardila, A., Ostrosky-Solis, F., and Bernal, B. (2006). Cognitive testing toward the future: The example of semantic verbal fluency (ANIMALS). *International Journal of Psychology*, 41(5), 324–332. <https://doi.org/10.1080/00207590500345542>.
- Arlot, S., & Celsis, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4(0), 40–79. <https://doi.org/10.1214/09-SS054>.
- Ashendorf, L., Jefferson, A. L., O'connor, M. K., CHAISSON, C., Green, R. C., Stern, R. A., et al. (2008). Trail making test errors in normal aging, mild cognitive impairment, and dementia. *Archives of Clinical Neuropsychology*, 23(2), 129–137. <https://doi.org/10.1016/j.acn.2007.11.005>.
- Azor, A. M., Cole, J. H., Holland, A. J., Dumba, M., Patel, M. C., Sadlon, A., Goldstone, A. P., & Manning, K. E. (2019). Increased brain age in adults with Prader-Willi syndrome. *Neuroimage Clin.*, 21, 101664. <https://doi.org/10.1016/j.nicl.2019.101664>.
- Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, 11, 118. <https://doi.org/10.1186/1471-2202-11-118>.
- Beheshti, I., Maikusa, N., & Matsuda, H. (2018). The association between “brain-age score” (BAS) and traditional neuropsychological screening tools in Alzheimer's disease. *Brain and Behavior: A Cognitive Neuroscience Perspective*, 8(8), e01020. <https://doi.org/10.1002/brb3.1020>.
- Benton, A. L., Varney, N. R., & Hamsher, K. D. (1978). Visuospatial judgment. A clinical test. *Arch. Neurol.*, 35(6), 364–367 Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/655909>.

- Bland, J. M., & Altman, D. G. (2011). Correlation in restricted ranges of data. *BMJ*, *342*, d556. <https://doi.org/10.1136/bmj.d556>.
- Ble, A., Volpato, S., Zuliani, G., Guralnik, J. M., Bandinelli, S., Lauretani, F., Bartali, B., Maraldi, C., Fellin, R., & Ferrucci, L. (2005). Executive function correlates with walking speed in older persons: The InCHIANTI study. *Journal of the American Geriatrics Society*, *53*(3), 410–415. <https://doi.org/10.1111/j.1532-5415.2005.53157.x>.
- Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., & Cohen, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *NeuroImage*, *55*(4), 1519–1527. <https://doi.org/10.1016/j.neuroimage.2010.12.028>.
- Buschke, H., & Fuld, P. A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology*, *24*(11), 1019–1025. <https://doi.org/10.1212/WNL.24.11.1019>.
- Butler, R. N., Sprott, R., Warner, H., Bland, J., Feuers, R., Forster, M., et al. (2004). Aging: The reality: Biomarkers of aging: From primitive organisms to humans. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *59*(6), B560–B567. <https://doi.org/10.1093/geronol/59.6.B560>.
- Chanraud, S., Martelli, C., Delain, F., Kostogianni, N., Douaud, G., Aubin, H. J., Reynaud, M., & Martinot, J. L. (2007). Brain Morphometry and cognitive performance in detoxified alcohol-dependents with preserved psychosocial functioning. *Neuropsychopharmacology*, *32*(2), 429–438. <https://doi.org/10.1038/sj.npp.1301219>.
- Ciulli, S., Citi, L., Salvadori, E., Valenti, R., Poggesi, A., Inzitari, D., et al. (2016). Prediction of impaired performance in trail making test in MCI patients with small vessel disease using DTI data. *IEEE J. Biomed. Health*, *20*(4), 1026–1033. <https://doi.org/10.1109/JBHI.2016.2537808>.
- Clark, L. J., Gatz, M., Zheng, L., Chen, Y.-L., McCleary, C., & Mack, W. J. (2009). Longitudinal verbal fluency in Normal aging, preclinical, and prevalent Alzheimer's disease. *American Journal of Alzheimer's Disease and Other Dementias*, *24*(6), 461–468. <https://doi.org/10.1177/1533317509345154>.
- Cole, J. H., Franke, K., & Cherbuin, N. (2019). Quantification of the biological age of the brain using neuroimaging. In A. Moskalev (Ed.), *Biomarkers of human aging* (pp. 293–328). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-24970-0_19.
- Cole, J. H., Leech, R., & Sharp, D. J. (2015). Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of Neurology*, *77*(4), 571–581. <https://doi.org/10.1002/ana.24367>.
- Cole, J. H., Poudel, R. P. K., Tsagkrasoulis, D., Caan, M. W. A., Steves, C., Spector, T. D., & Montana, G. (2017a). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, *163*, 115–124. <https://doi.org/10.1016/j.neuroimage.2017.07.059>.
- 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. <https://doi.org/10.1038/mp.2017.62>.
- Cole, J. H., Underwood, J., Caan, M. W. A. A., De Francesco, D., van Zoest, R. A., Leech, R., et al. (2017b). Increased brain-predicted aging in treated HIV disease. *Neurology*, *88*(14), 1349–1357. <https://doi.org/10.1212/WNL.0000000000003790>.
- Cooper, D. B., Epker, M., Lacritz, L., Weiner, M., Rosenberg, R. N., Honig, L., & Cullum, C. M. (2001). Effects of practice on category fluency in Alzheimers disease*. *The Clinical Neuropsychologist*, *15*(1), 125–128. <https://doi.org/10.1076/clin.15.1.125.1914>.
- Cruz-Almeida, Y., Fillingim, R. B., Riley, J. L. I. I., Woods, A. J., Porges, E., Cohen, R., & Cole, J. (2019). Chronic pain is associated with a brain aging biomarker in community-dwelling older adults. *PAIN*, *160*(5), 1119–1130. <https://doi.org/10.1097/j.pain.0000000000001491>.
- D'Elia, L. F., Satz, P., Uchiyama, C. L., & White, T. (1996). *Color trials test. Professional manual*. Odessa: Psychological Assessment Resources.
- Dean, W., & Morgan, R. F. (1988). In defense of the concept of biological aging measurement—current status. *Archives of Gerontology and Geriatrics*, *7*(3), 191–210. [https://doi.org/10.1016/0167-4943\(88\)90002-7](https://doi.org/10.1016/0167-4943(88)90002-7).
- Dickerson, B. C., Fenstermacher, E., Salat, D. H., Wolk, D. A., Maguire, R. P., Desikan, R., Pacheco, J., Quinn, B. T., van der Kouwe, A., Greve, D. N., Blacker, D., Albert, M. S., Killiany, R. J., & Fischl, B. (2008). Detection of cortical thickness correlates of cognitive performance: Reliability across MRI scan sessions, scanners, and field strengths. *NeuroImage*, *39*(1), 10–18. <https://doi.org/10.1016/j.neuroimage.2007.08.042>.
- Dohmatob, E., Eickenberg, M., Thirion, B., & Varoquaux, G. (2015). Speeding-up model-selection in Graphnet via early-stopping and Univariate feature-screening. In *2015 International Workshop on Pattern Recognition in Neuroimaging* (pp. 17–20). IEEE. <https://doi.org/10.1109/PRNI.2015.19>.
- Dorrian, J., Rogers, N., & Dinges, D. (2005). Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In C. A. Kushida (Ed.), *Sleep deprivation: Clinical issues, pharmacology, and sleep loss effects* (Vol. 193, pp. 39–70). Boca Raton: CRC Press.
- Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Retrieved from <http://arxiv.org/abs/1702.08608>
- Dubois, M., Hadj-Seleem, F., Lofstedt, T., Perrot, M., Fischer, C., Frouin, V., & Duchesnay, E. (2014). Predictive support recovery with TV-elastic net penalty and logistic regression: An application to structural MRI. In *Proceedings - 2014 International Workshop on Pattern Recognition in Neuroimaging, PRNI 2014* (pp. 1–4). Tubingen. <https://doi.org/10.1109/PRNI.2014.6858517>.
- Dugbartey, A. T., Townes, B. D., & Mahurin, R. K. (2000). Equivalence of the color trails test and trail making test in nonnative English-speakers. *Archives of Clinical Neuropsychology*, *15*(5), 425–431. [https://doi.org/10.1016/S0887-6177\(99\)00034-7](https://doi.org/10.1016/S0887-6177(99)00034-7).
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>.
- Eastman, J. A., Hwang, K. S., Lazaris, A., Chow, N., Ramirez, L., Babakchian, S., et al. (2013). Cortical thickness and semantic fluency in Alzheimer's disease and mild cognitive impairment. *Am. J. Alzheimers Dis. (Columbia)*, *1*(2), 81–92. <https://doi.org/10.7726/ajad.2013.1006>.
- Elkin-Frankston, S., Lebowitz, B. K., Kapust, L. R., Hollis, H. H., & O'Connor, M. G. (2007). The use of the color trails test in the assessment of driver competence: Preliminary report of a culture-fair instrument. *Archives of Clinical Neuropsychology*, *22*(5), 631–635. <https://doi.org/10.1016/j.acn.2007.04.004>.
- Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., Gustavson, D. E., Franz, C. E., Hatton, S. N., Jacobson, K. C., Toomey, R., McKenzie, R., Xian, H., Lyons, M. J., & Kremen, W. S. (2018). Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. *Alzheimers Dement. (Amst)*, *10*, 372–381. <https://doi.org/10.1016/J.DADM.2018.04.003>.
- Emek-Savaş, D., Yerlikaya, D., Yener, G., & Öktem, Ö. (2019). Validity, reliability and norm scores of the Stroop test Çapa version. *Türk Psikiyatri Dergisi*, in press. <https://doi.org/10.5080/u23549>.
- Emery, V. O. B., Gillie, E. X., & Smith, J. A. (1996). Reclassification of the vascular dementias: Comparisons of infarct and noninfarct vascular dementias. *International Psychogeriatrics*, *8*(1), 33–61. <https://doi.org/10.1017/S1041610296002475>.

- Farokhian, F., Yang, C., Beheshti, I., Matsuda, H., & Wu, S. (2017). Age-related gray and White matter changes in Normal adult brains. *Aging and Disease*, 8(6), 899–909. <https://doi.org/10.14336/AD.2017.0502>.
- Feeney, J., Savva, G. M., O'Regan, C., King-Kallimanis, B., Cronin, H., & Kenny, R. A. (2016). Measurement error, reliability, and minimum detectable change in the mini-mental state examination, Montreal cognitive assessment, and color trails test among community middle-aged and older adults. *Journal of Alzheimer's Disease*, 53(3), 1107–1114. <https://doi.org/10.3233/JAD-160248>.
- Fiorito, G., McCrory, C., Robinson, O., Carmeli, C., Rosales, C. O., Zhang, Y., et al. (2019). Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: A multi-cohort analysis. *Aging*, 11(7), 2045–2070. <https://doi.org/10.18632/aging.101900>.
- Fjell, A. M., Westlye, L. T., Grydeland, H., Amlien, I., Espeseth, T., Reinvang, I., et al. (2013). Critical ages in the life course of the adult brain: Nonlinear subcortical aging. *Neurobiology of Aging*, 34(10), 2239–2247. <https://doi.org/10.1016/j.neurobiolaging.2013.04.006>.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3), 189–198. [https://doi.org/10.1016/0022-3956\(75\)90026-6](https://doi.org/10.1016/0022-3956(75)90026-6).
- Franke, K., & Gaser, C. (2012). Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych*, 25(4), 235–245. <https://doi.org/10.1024/1662-9647/a000074>.
- Franke, K., Gaser, C., Manor, B., & Novak, V. (2013). Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Frontiers in Aging Neuroscience*, 5, 90. <https://doi.org/10.3389/fnagi.2013.00090>.
- Franke, K., Ristow, M., Gaser, C., & Alzheimer's Disease Neuroimaging Initiative. (2014). Gender-specific impact of personal health parameters on individual brain aging in cognitively unimpaired elderly subjects. *Frontiers in Aging Neuroscience*, 6, 94. <https://doi.org/10.3389/fnagi.2014.00094>.
- Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *Neuroimage*, 50(3), 883–892. <https://doi.org/10.1016/j.neuroimage.2010.01.005>.
- Galasko, D., Abramson, I., Corey-Bloom, J., & Thal, L. J. (1993). Repeated exposure to the mini-mental state examination and the information-memory-concentration test results in a practice effect in Alzheimer's disease. *Neurology*, 43(8), 1559–1563. <https://doi.org/10.1212/wnl.43.8.1559>.
- Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., & Alzheimer's Disease Neuroimaging Initiative. (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>.
- Ge, Y., Grossman, R. I., Babb, J. S., Rabin, M. L., Mannon, L. J., & Kolson, D. L. (2002). Age-related total gray matter and white matter changes in normal adult brain. Part I: Volumetric MR imaging analysis. *AJNR Am. J. Neuroradiol.*, 23(8), 1327–1333. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12223373>.
- Golden, C. J. (1978). Stroop color and word test: A manual for clinical and experimental uses. Chicago: Stoelting. <https://doi.org/10.1002>
- Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education*, 74(3), 249–266. <https://doi.org/10.3200/JEXE.74.3.249-266>.
- Green, R. C., Woodard, J. L., & Green, J. (1995). Validity of the Mattis dementia rating scale for detection of cognitive impairment in the elderly. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 7(3), 357–360. <https://doi.org/10.1176/jnp.7.3.357>.
- Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology*. <https://doi.org/10.1080/01688639108405109>.
- Guggenmos, M., Schmack, K., Sekutowicz, M., Garbusow, M., Sebold, M., Sommer, C., et al. (2017). Quantitative neurobiological evidence for accelerated brain aging in alcohol dependence. *Transl. Psychiatry*, 7(12), 1279. <https://doi.org/10.1038/s41398-017-0037-y>.
- Guo, Q. H., Cao, X. Y., Zhou, Y., Zhao, Q. H., Ding, D., & Hong, Z. (2010). Application study of quick cognitive screening test in identifying mild cognitive impairment. *Neuroscience Bulletin*, 26(1), 47–54. <https://doi.org/10.1007/s12264-010-0816-4>.
- Gutierrez Becker, B., Klein, T., & Wachinger, C. (2018). Gaussian process uncertainty in age estimation as a measure of brain abnormality. *Neuroimage*, 175, 246–258. <https://doi.org/10.1016/J.NEUROIMAGE.2018.03.075>.
- Han, L. K. M., Dinga, R., Hahn, T., Ching, C., Eyler, L., Aftanas, L., et al. (2019). Brain aging in major depressive disorder: Results from the ENIGMA major depressive disorder working group. *BioRxiv*, 560623. <https://doi.org/10.1101/560623>.
- Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test-retest reliability. *Br. J. Clin. Psychol.*, (2), 181–191. <https://doi.org/10.1348/014466500163202>.
- Heaton, R. K. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). Wisconsin card sorting test manual: Revised and expanded. Psychological reports. Odessa, FL: Psychological Assessment Resources. <https://doi.org/10.2466/pr0.1995.76.2.623>.
- Honeine, P., & Richard, C. (2009). Solving the pre-image problem in kernel machines: A direct method. In *2009 IEEE International Workshop on Machine Learning for Signal Processing* (pp. 1–6). <https://doi.org/10.1109/MLSP.2009.5306204>.
- Irimia, A., Torgerson, C. M., Goh, S.-Y. M., & Van Horn, J. D. (2015). Statistical estimation of physiological brain age as a descriptor of senescence rate during adulthood. *Brain Imaging and Behavior*, 9(4), 678–689. <https://doi.org/10.1007/s11682-014-9321-0>.
- Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivieres, S., Grigis, A., et al. (2019). Quantifying performance of machine learning methods for neuroimaging data. *Neuroimage*. <https://doi.org/10.1016/J.NEUROIMAGE.2019.05.082>.
- Jollans, L., & Whelan, R. (2016). The clinical added value of imaging: A perspective from outcome prediction. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, 1(5), 423–432. <https://doi.org/10.1016/j.bpsc.2016.04.005>.
- Jollans, L., & Whelan, R. (2018). Neuromarkers for mental disorders: Harnessing population neuroscience. *Frontiers in Psychiatry*, 9, 242. <https://doi.org/10.3389/fpsy.2018.00242>.
- Jurica, P. J., Leitten, C. L., & Mattis, S. (2001). DRS-2: Dementia rating Scale-2: Professional manual. Psychological assessment resources. Retrieved from <https://books.google.ie/books?id=tovFPwAACAAJ>
- Kaufmann, T., van der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., et al. (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>.
- Kearney, P. M., Cronin, H., O'Regan, C., Kamiya, Y., Savva, G. M., Whelan, B., & Kenny, R. (2011). Cohort profile: The Irish longitudinal study on ageing. *International Journal of Epidemiology*, 40(4), 877–884. <https://doi.org/10.1093/ije/dyr116>.
- Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H. J., Reiser, M., Pantelis, C., & Meisenzahl, E. (2014). Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. *Schizophrenia Bulletin*, 40(5), 1140–1153. <https://doi.org/10.1093/schbul/sbt142>.

- Kwok, J. T., & Tsang, I. W. (2004). The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, *15*(6), 1517–1525. <https://doi.org/10.1109/TNN.2004.837781>.
- Lancaster, J., Lorenz, R., Leech, R., & Cole, J. H. (2018). Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Front. Aging Neurosci.*, *10*(FEB), 1–10. <https://doi.org/10.3389/fnagi.2018.00028>.
- Le, T. T., Kuplicki, R. T., McKinney, B. A., Yeh, H.-W., Thompson, W. K., Paulus, M. P., & Tulsa 1000 Investigators, T. 1000. (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>.
- Lee, T. M. C., Cheung, C. C. Y., Chan, J. K. P., & Chan, C. C. H. (2000). Trail making across languages. *Journal of Clinical and Experimental Neuropsychology*, *22*(6), 772–778. <https://doi.org/10.1076/jcen.22.6.772.954>.
- Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian Masouleh, S., Huntenburg, J. M., Lampe, L., Rahim, M., Abraham, A., Craddock, R. C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M. L., Witte, A. V., Villringer, A., & Margulies, D. S. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage*, *148*, 179–188. <https://doi.org/10.1016/j.neuroimage.2016.11.005>.
- Lin, Y. C., Shih, Y. C., Tseng, W. Y. I., Chu, Y. H., Wu, M. T., Chen, T. F., et al. (2014). Cingulum correlates of cognitive functions in patients with mild cognitive impairment and early Alzheimer's disease: A diffusion Spectrum imaging study. *Brain Topography*, *27*(3), 393–402. <https://doi.org/10.1007/s10548-013-0346-2>.
- Lipton, Z. C. (2018). The mythos of model interpretability. *ACM Queue*, *16*(3), 30:31–30:57. <https://doi.org/10.1145/3236386.3241340>.
- Liu, W., & Li, Q. (2017). An efficient elastic net with regression coefficients method for variable selection of Spectrum data. *PLoS One*, *12*(2), e0171122. <https://doi.org/10.1371/journal.pone.0171122>.
- Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 150–158). New York: ACM. <https://doi.org/10.1145/2339530.2339556>.
- Löwe, L. C., Gaser, C., & Franke, K. (2016). The effect of the APOE genotype on individual BrainAGE in Normal aging, mild cognitive impairment, and Alzheimer's disease. *PLoS One*, *11*(7), e0157514. <https://doi.org/10.1371/journal.pone.0157514>.
- Luders, E., Cherbuin, N., & Gaser, C. (2016). Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners. *Neuroimage*, *134*, 508–513. <https://doi.org/10.1016/j.neuroimage.2016.04.007>.
- Luo, Y., Tseng, H.-H., Cui, S., Wei, L., Ten Haken, R. K., & El Naqa, I. (2019). Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR|Open*. <https://doi.org/10.1259/bjro.20190021>.
- Madan, C. R., & Kensinger, E. A. (2018). Predicting age from cortical structure across the lifespan. *The European Journal of Neuroscience*, *47*(5), 399–416. <https://doi.org/10.1111/ejn.13835>.
- Mateos-Pérez, J. M., Dadar, M., Lacalle-Aurioles, M., Iturria-Medina, Y., Zeighami, Y., & Evans, A. C. (2018). Structural neuroimaging as clinical predictor: A review of machine learning applications. *Neuroimage Clin.*, *20*, 506–522. <https://doi.org/10.1016/j.nicl.2018.08.019>.
- McCaffrey, R. J., & Westervelt, H. J. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review*, *5*(3), 203–221. <https://doi.org/10.1007/BF02214762>.
- McCrory, C., & Kenny, R. A. (2018). Rebuking the concept of ageing as a disease. *Lancet Diabetes Endocrinol.*, *6*(10), 768. [https://doi.org/10.1016/S2213-8587\(18\)30266-3](https://doi.org/10.1016/S2213-8587(18)30266-3).
- McIntyre, R. S., Cha, D. S., Soczynska, J. K., Woldeyohannes, H. O., Gallagher, L. A., Kudlow, P., et al. (2013). Cognitive deficits and functional outcomes in major depressive disorder: Determinants, substrates, and treatment interventions. *Depression and Anxiety*, *30*(6), 515–527. <https://doi.org/10.1002/da.22063>.
- Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, *12*(3), 282. <https://doi.org/10.2307/1164688>.
- Messinis, L., Malegiannaki, A.-C., Christodoulou, T., Panagiotopoulos, V., & Papatanasopoulos, P. (2011). Color trails test: Normative data and criterion validity for the Greek adult population. *Archives of Clinical Neuropsychology*, *26*(4), 322–330. <https://doi.org/10.1093/acclin/acr027>.
- Miciak, J., Taylor, W. P., Stuebing, K. K., Fletcher, J. M., & Vaughn, S. (2016). Designing intervention studies: Selected populations, range restrictions, and statistical power. *J. Res. Educ. Eff.*, *9*(4), 556–569. <https://doi.org/10.1080/19345747.2015.1086916>.
- Mitrushina, M. N., Boone, K. B., Razani, J. L., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (second). New York: Oxford University Press. Retrieved from <https://global.oup.com/academic/product/handbook-of-normative-data-for-neuropsychological-assessment-9780195169300?cc=ie&lang=en&>
- Mwangi, B., Tian, T. S., & Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229–244. <https://doi.org/10.1007/s12021-013-9204-3>.
- Nelson, H. E., & Willinson, J. (1982). The National Adult Reading Test (NART): Test manual. Windsor, UK: NFER: Nelson.
- Nenadić, I., Dietzek, M., Langbein, K., Sauer, H., & Gaser, C. (2017). BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. *Psychiatry Research: Neuroimaging*, *266*, 86–89. <https://doi.org/10.1016/j.psychres.2017.05.006>.
- Öktem, O. (1992). A verbal test of memory processes: A preliminary study. *Noro Psikiyatri Arsivi*, *29*(4), 196–206.
- Pfeffer, R. I., Kurosaki, T. T., Chance, J. M., Filos, S., & Bates, D. (1984). Use of the mental function index in older adults: Reliability, validity, and measurement of change over time. *American Journal of Epidemiology*, *120*(6), 922–935. <https://doi.org/10.1093/oxfordjournals.aje.a113963>.
- Ranganathan, P., Pramesh, C. S., & Buyse, M. (2016). Common pitfalls in statistical analysis: The perils of multiple testing. *Perspectives in Clinical Research*, *7*(2), 106–107. <https://doi.org/10.4103/2229-3485.179436>.
- Reitan, R. M. (1955). The relation of the trail making test to organic brain damage. *Journal of Consulting Psychology*. <https://doi.org/10.1037/h0044509>.
- Richard, G., Kolskår, K., Sanders, A.-M., Kaufmann, T., Petersen, A., Doan, N. T., Monereo Sánchez, J., Alnæs, D., Ulrichsen, K. M., Dørum, E. S., Andreassen, O. A., Nordvik, J. E., & Westlye, L. T. (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>.
- Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). 'oops!': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*(6), 747–758. [https://doi.org/10.1016/S0028-3932\(97\)00015-8](https://doi.org/10.1016/S0028-3932(97)00015-8).
- Rodríguez-Aranda, C., Waterloo, K., Johnsen, S. H., Eldevik, P., Sparr, S., Wikran, G. C., Herder, M., & Vangberg, T. R. (2016). Neuroanatomical correlates of verbal fluency in early Alzheimer's disease and normal aging. *Brain and Language*, *155–156*, 24–35. <https://doi.org/10.1016/J.BANDL.2016.03.001>.
- Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *The Journal of Applied Psychology*, *85*(1), 112–118. <https://doi.org/10.1037/0021-9010.85.1.112>.

- Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2016). Voodoo machine learning for clinical predictions. *BioRxiv*, 059774. <https://doi.org/10.1101/059774>.
- Santos Nogueira, D., Azevedo Reis, E., & Vieira, A. (2016). Verbal fluency tasks: Effects of age, gender, and education. *Folia Phoniatr. Logo.*, 68(3), 124–133. <https://doi.org/10.1159/000450640>.
- Savva, G. M., Maty, S. C., Setti, A., & Feeney, J. (2013). Cognitive and physical health of the older populations of England, the United States, and Ireland: International comparability of the Irish longitudinal study on ageing. *Journal of the American Geriatrics Society*, 61, S291–S298. <https://doi.org/10.1111/jgs.12196>.
- Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. M., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D. S., Yip, S. W., Rosenberg, M. D., & Constable, R. T. (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage*, 193, 35–45. <https://doi.org/10.1016/J.NEUROIMAGE.2019.02.057>.
- Scheller, E., Schumacher, L. V., Peter, J., Lahr, J., Wehrle, J., Kaller, C. P., ... Klöppel, S. (2018). Brain aging and APOE ϵ 4 interact to reveal potential neuronal compensation in healthy older adults. *Frontiers in Aging Neuroscience*, 10, 74. <https://doi.org/10.3389/fnagi.2018.00074>.
- Schnack, H. G., van Haren, N. E. M., Nieuwenhuis, M., Hulshoff Pol, H. E., Cahn, W., & Kahn, R. S. (2016). Accelerated brain aging in schizophrenia: A longitudinal pattern recognition study. *The American Journal of Psychiatry*, 173(6), 607–616. <https://doi.org/10.1176/appi.ajp.2015.15070922>.
- Skocik, M., Collins, J., Callahan-Flintoft, C., Bowman, H., & Wyble, B. (2016). I tried a bunch of things: The dangers of unexpected overfitting in classification. *BioRxiv*, 078816. <https://doi.org/10.1101/078816>.
- Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., & Miller, K. L. (2019). Estimation of brain Age Delta from brain imaging. *Neuroimage*. <https://doi.org/10.1016/j.neuroimage.2019.06.017>.
- Snyder, J. C., Mika, S., Burke, K., & Müller, K.-R. (2013). Kernels, Pre-images and Optimization. In B. Schölkopf, Z. Luo, & V. Vovk (Eds.), *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (pp. 245–259). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_21.
- Sprott, R. L. (2010). Biomarkers of aging and disease: Introduction and definitions. *Experimental Gerontology*, 45(1), 2–4. <https://doi.org/10.1016/J.EXGER.2009.07.008>.
- Steffener, J., Habeck, C., O'Shea, D., Razlighi, Q., Bherer, L., & Stern, Y. (2016). Differences between chronological and brain age are related to education and self-reported physical activity. *Neurobiology of Aging*, 40, 138–144. <https://doi.org/10.1038/nn.3945>. [Dopaminergic](https://doi.org/10.1038/nn.3945).
- Stern, Y., Gazes, Y., Razlighi, Q., Steffener, J., & Habeck, C. (2018). A task-invariant cognitive reserve network. *Neuroimage*, 178, 36–45. <https://doi.org/10.1016/J.NEUROIMAGE.2018.05.033>.
- Stern, Y., Habeck, C., Steffener, J., Barulli, D., Gazes, Y., Razlighi, Q., et al. (2014). The reference ability neural network study: Motivation, design, and initial feasibility analyses. *Neuroimage*, 103, 139–151. <https://doi.org/10.1016/J.NEUROIMAGE.2014.09.029>.
- Strauss, E. H., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests; administration norms and commentary* (3rd ed.). New York: Oxford University Press. <https://doi.org/10.1016/j.jvolgeores.2008.06.015>.
- Tombaugh, T. N., & McIntyre, N. J. (1992). The mini-mental state examination: A comprehensive review. *Journal of the American Geriatrics Society*, 40(9), 922–935. <https://doi.org/10.1111/j.1532-5415.1992.tb01992.x>.
- Varikuti, D. P., Genon, S., Sotiras, A., Schwender, H., Hoffstaedter, F., Patil, K. R., Jockwitz, C., Caspers, S., Moebus, S., Amunts, K., Davatzikos, C., & Eickhoff, S. B. (2018). Evaluation of non-negative matrix factorization of grey matter in age prediction. *Neuroimage*, 173(March), 394–410. <https://doi.org/10.1016/j.neuroimage.2018.03.007>.
- Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*, 145, 166–179. <https://doi.org/10.1016/J.NEUROIMAGE.2016.10.038>.
- Vaucher, P., Herzig, D., Cardoso, L., Herzog, M. H., Mangin, P., & Favrat, B. (2014). The trail making test as a screening instrument for driving performance in older drivers: a translational research. *BMC Geriatrics*, 14(1), 123. <https://doi.org/10.1186/1471-2318-14-123>.
- Vazzana, R., Bandinelli, S., Lauretani, F., Volpato, S., Lauretani, F., Di Iorio, A., et al. (2010). Trail making test predicts physical impairment and mortality in older persons. *Journal of the American Geriatrics Society*, 58(4), 719–723. <https://doi.org/10.1111/j.1532-5415.2010.02780.x>.
- Wechsler, D. (1987). *Wechsler memory scale - revised manual*. San Antonio, Texas: Psychological Corporation.
- Wechsler, D. (1997). *WAIS-III administration and scoring manual*. The Psychological Corporation, San Antonio, Texas. <https://doi.org/10.1177/1073191102009001003>.
- Whelan, B. J., & Savva, G. M. (2013). Design and methodology of the Irish longitudinal study on ageing. *Journal of the American Geriatrics Society*, 61, S265–S268. <https://doi.org/10.1111/jgs.12199>.
- Whelan, R., & Garavan, H. (2014). When optimism hurts: Inflated predictions in psychiatric neuroimaging. *Biological Psychiatry*, 75(9), 746–748. <https://doi.org/10.1016/j.biopsych.2013.05.014>.
- Willer, L., Pedersen, P. M., Forchhammer, H. B., & Christensen, H. (2016). Cognitive assessment at bedside for iPad: A preliminary validation of a novel cognitive test for stroke patients. *European Stroke Journal*, 1(4), 294–301. <https://doi.org/10.1177/2396987316665233>.
- Wilson, B. A., Watson, P. C., Baddeley, A. D., Emslie, H., & Evans, J. J. (2000). Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. *J. Int. Neuropsychol. Soc.*, 6(4), 469–479. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10902416>.
- Woo, C. W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*, 20(3), 365–377. <https://doi.org/10.1038/nn.4478>.
- Woods, D. L., Wyma, J. M., Herron, T. J., & Yund, E. W. (2016). Computerized analysis of verbal fluency: Normative data and the effects of repeated testing, simulated malingering, and traumatic brain injury. *PLoS One*, 11(12), e0166439. <https://doi.org/10.1371/journal.pone.0166439>.
- Zanetti, M., Ballabio, C., Abbate, C., Cutaia, C., Vergani, C., & Bergamaschini, L. (2006). Mild cognitive impairment subtypes and vascular dementia in community-dwelling elderly people: A 3-year follow-up study. *Journal of the American Geriatrics Society*, 54(4), 580–586. <https://doi.org/10.1111/j.1532-5415.2006.00658.x>.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.