NeuroImage 163 (2017) 115–124

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker

James H. Cole [a], Rudra P.K. Poudel [b], Dimosthenis Tsagkrasoulis [c], Matthan W.A. Caan [d], Claire Steves [e], Tim D. Spector [e], Giovanni Montana [b,c,*]

[a] Computational, Cognitive & Clinical Neuroimaging Laboratory, Division of Brain Sciences, Imperial College London, London, UK
[b] Department of Biomedical Engineering, King's College London, London, UK
[c] Department of Mathematics, Imperial College London, London, UK
[d] Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
[e] Department of Twin Research & Genetic Epidemiology, King's College London, London, UK

ARTICLE INFO

Keywords:
Brain ageing
Neuroimaging
Reliability
Heritability
Biomarker
Deep learning
Convolutional neural networks
Gaussian processes

ABSTRACT

Machine learning analysis of neuroimaging data can accurately predict chronological age in healthy people. Deviations from healthy brain ageing have been associated with cognitive impairment and disease. Here we sought to further establish the credentials of 'brain-predicted age' as a biomarker of individual differences in the brain ageing process, using a predictive modelling approach based on deep learning, and specifically convolutional neural networks (CNN), and applied to both pre-processed and raw T1-weighted MRI data.

Firstly, we aimed to demonstrate the accuracy of CNN brain-predicted age using a large dataset of healthy adults (N = 2001). Next, we sought to establish the heritability of brain-predicted age using a sample of monozygotic and dizygotic female twins (N = 62). Thirdly, we examined the test-retest and multi-centre reliability of brain-predicted age using two samples (within-scanner N = 20; between-scanner N = 11). CNN brain-predicted ages were generated and compared to a Gaussian Process Regression (GPR) approach, on all datasets. Input data were grey matter (GM) or white matter (WM) volumetric maps generated by Statistical Parametric Mapping (SPM) or raw data.

CNN accurately predicted chronological age using GM (correlation between brain-predicted age and chronological age r = 0.96, mean absolute error [MAE] = 4.16 years) and raw (r = 0.94, MAE = 4.65 years) data. This was comparable to GPR brain-predicted age using GM data (r = 0.95, MAE = 4.66 years). Brain-predicted age was a heritable phenotype for all models and input data ($h^2 \geq 0.5$). Brain-predicted age showed high test-retest reliability (intraclass correlation coefficient [ICC] = 0.90–0.99). Multi-centre reliability was more variable within high ICCs for GM (0.83–0.96) and poor-moderate levels for WM and raw data (0.51–0.77).

Brain-predicted age represents an accurate, highly reliable and genetically-influenced phenotype, that has potential to be used as a biomarker of brain ageing. Moreover, age predictions can be accurately generated on raw T1-MRI data, substantially reducing computation time for novel data, bringing the process closer to giving real-time information on brain health in clinical settings.

1. Introduction

The human brain changes across the adult lifespan. This process of *brain ageing* occurs in accord with a general decline in cognitive performance, *cognitive ageing*. Although the changes associated with brain ageing are not explicitly pathological, with increasing age comes increasing risk of neurodegenerative disease and dementia (Abbott, 2011). However, the wide range of onset ages for age-associated brain diseases indicates that the effects of ageing on the brain vary greatly between individuals. Thus, advancing our understanding of brain ageing and identifying biomarkers of the process are vital to help improve detection of early-stage neurodegeneration and predict age-related

Abbreviations: CNN, convolutional neural network; GM, grey matter; MAE, mean absolute error; WM, white matter.

* Corresponding author. Department of Biomedical Engineering, King's College London, St Thomas' Hospital, The Rayne Institute 3rd Floor, Lambeth Wing St Thomas' Hospital, London SE1 7EH, UK.
E-mail address: giovanni.montana@kcl.ac.uk (G. Montana).

https://doi.org/10.1016/j.neuroimage.2017.07.059
Received 25 March 2017; Accepted 28 July 2017
Available online 29 July 2017

cognitive decline.

One promising approach to identifying individual differences in brain ageing derives from the research showing that neuroimaging data can be used to accurately predict chronological age in healthy individuals, using machine learning (Dosenbach et al., 2010; Franke et al., 2010). By 'learning' the correspondence between patterns in structural or functional neuroimaging data and an age 'label', machine-learning algorithms can formulate massively high-dimensional regression models, fitting large neuroimaging datasets as independent variables to predict chronological age as the dependent variable. The resulting brain-based age predictions are generally highly accurate, particularly when algorithms learn from large training datasets and are applied to novel or 'left-out' data (i.e., test datasets).

Neuroimaging-derived age predictions have been explored in the context of different brain diseases. By training models on healthy individuals, brain-based predictions of age can then be made in independent clinical samples. If 'brain-predicted age' is greater than an individual's chronological age, this is thought to reflect some aberrant accumulation of age-related changes to the brain. The degree of this 'added' brain ageing can be simply quantified by subtracting chronological age from brain-predicted age. This approach is being used more frequently and has demonstrated increased brain-predicted age in adults with mild cognitive impairment who progress to Alzheimer's (Franke and Gaser, 2012; Gaser et al., 2013), after traumatic brain injury (Cole et al., 2015), in schizophrenia (Koutsouleris et al., 2013; Schnack et al., 2016), HIV (Cole et al., 2017c), epilepsy (Pardoe et al., 2017), Down's syndrome (Cole et al., 2017a) and diabetes (Franke et al., 2013). At the same time, brain-predicted age has been used to demonstrate protective influences on brain ageing, including meditation (Luders et al., 2016) and increased levels of education and physical exercise (Steffener et al., 2016). Evidently, the extent to which one's brain resembles the typical structure or function appropriate for one's age can be affected by both positive and negative influences. By conceptualising brain ageing in this manner, highly-complex multivariate datasets and statistical procedures can be reformulated into an intuitively straightforward and widely-applicable biomarker. However, the practicality of using such a marker clinically, its reliability and relevance for normal variation in brain ageing need to be further demonstrated.

One hindrance to clinical applications for neuroimaging generally is the time needed for image 'post-processing' after acquisition (referred to as 'pre-processing' by neuroimagers), which can take hours or days, while clinical decisions often need to occur in minutes or less. Regardless of learning algorithm, previous brain-predicted age studies have required several pre-processing stages. Such steps are typically a sequence of data transformations that produce a representation of the original images that is sufficiently structured, compact and informative to support machine learning. These include the removal of non-brain tissue (i.e., skull stripping or brain extraction), affine or non-linear image registration, interpolation and smoothing. While pre-processing may reduce noise and permit voxelwise inter-individual statistical comparisons, there are numerous additional assumptions required for any pre-processing pipeline. These assumptions are often not met, particularly when analysing brain images containing gross pathology (Avants et al., 2008; Liu et al., 2015) and can even be an increased source of error. Recently, however, modelling methods that require little or no image pre-processing have become available, including so-called 'deep learning'.

The resurgence of interest in artificial neural networks for learning data representations, deep learning, offers a new way of approaching statistical modelling in neuroimaging, thanks to improvements in computing infrastructure. When sufficiently large volumes of data are available, no 'hand-engineering' (i.e., manually selecting *a priori* which features should be used as input) is needed as the deep learning algorithm is able to infer a compact representation of the data, starting only with raw images as input, which is optimally tailored for the particular predictive modelling task at hand. In this respect, deep learning offers several practical advantages for high-dimensional prediction tasks, that should enable the learning of both physiologically-relevant representations and latent relationships (Plis et al., 2014). Of particular interest to us is the potential for deep learning techniques, such as convolutional neural networks (CNN), to make predictions from raw, unprocessed neuroimaging data, thus obviating the reliance on time-consuming pre-processing and improving the clinical applicability of models of brain ageing.

Beyond improving clinical applicability, a biomarker of brain ageing needs to relate to naturally occurring variation, such as that caused by genetic factors. Many aspects of brain ageing and susceptibility to age-related brain disease are thought to be under genetic influence (Lee and Sachdev, 2014; Lu et al., 2004; Peters, 2006; Teter and Finch, 2004). Therefore, demonstrating a brain ageing biomarker is sensitive to genetic influences gives some external, genetic, validity to the measure. Furthermore, if a neuroimaging biomarker is heritable, this motivates further research into specific candidate genes, or sets of genes, that may affect this aspect of brain ageing. These candidate genes can then, in turn, provide biological targets for pharmacological interventions which aim to improve brain health in older adults.

Another important facet of any biomarker is reliability. If a biomarker is to be evaluated longitudinally, in clinical trials or research settings, to track change over time, establishing test-retest reliability is vital. Furthermore, as many neuroimaging studies are now international collaborative efforts, data collection often takes place across multiple scanning sites. Therefore, between-scanner reliability, which indicates that a method of obtaining a biomarker is generalizable to data acquired from other sites, is of increasing importance.

In this work, we sought to establish the credentials of CNN-predicted age as a potential biomarker of brain ageing in three different ways: 1) Demonstrate that CNNs can accurately predict age using structural neuroimaging data and compare predictions using pre-processed and 'raw' input data; 2) Establish the heritability of brain-predicted age using a sample of monozygotic and dizygotic twins; 3) Assess both the test-retest (i.e., within-scanner) and multi-centre (i.e., between-scanner) reliability of brain-predicted age.

2. Materials and methods

2.1. Datasets

All neuroimaging data used in the study were T1-weighted MRI scans. Details of the participants in the specific samples and the respective acquisition parameters used are outlined below:

2.1.1. Brain-predicted age evaluation cohort

The evaluation of the accuracy of age modelling using neuroimaging was conducted using the Brain-Age Healthy Control (BAHC) dataset. This cohort consisted of N = 2001 healthy individuals (male/female = 1016/985, mean age = 36.95 ± 18.12, age range 18–90 years). These data were compiled from 14 publicly-available sources (see Supplementary Material Table S1), made available via various data-sharing initiatives. All participants were screened to be free from major neurological or psychiatric diagnoses, according to local study protocols. All data were acquired at either 1.5T or 3T using standard T1-weighted sequences (full details in supplementary material). Each contributing study was ethically approved, as was subsequent data-sharing. Informed consent was obtained at each local study site in accordance with local guidelines.

2.1.2. Heritability assessment sample

Participants for heritability assessment were individuals from the UK Adult Twin Registry (TwinsUK), who were invited to take part in a neuroimaging sub-study. A total of 62 female individuals were scanned (mean age = 61.86 ± 8.36), including 27 monozygotic twin pairs and 4 dizygotic twin pairs. All participants were free from major neurological or psychiatric diagnoses and contraindications to MRI scanning. A Philips Achieva 3T was used to acquire T1-weighted 3D turbo field echo (TFE)

MRI with the following parameters: TE = 3.21 ms, TR = 6.89 ms, flip angle = 8°, field-of-view = 240 mm, 133 slices of 1.2 mm thickness, in-plane resolution = 1.07 × 1.07 mm. Each participant provided written and informed consent for academic use of the data. Experiments were approved by the National Research Ethics Service (NRES) Committee London - Westminster.

2.1.3. Within-scanner reliability sample

A total of 20 participants (male/female = 12/8, mean age at first scan = 34.05 ± 8.71) took part in the STudy Of Reliability of MRI (STORM) at Imperial College London. Participants were scanned an average of 28.35 ± 1.09 days apart. All participants were free from major neurological or psychiatric diagnoses. A Siemens Verio 3T scanner was used to acquire magnetisation-prepared rapid gradient-echo (MPRAGE) images as follows: TE = 2.98 ms, TR = 2300 ms, TI = 900 ms, flip angle = 9°, field-of-view = 256 mm, 160 slices of 1.0 mm thickness, in-plane resolution = 1.0 × 1.0 mm. The study was approved by the West London NRES Committee and informed, written consent was obtained from each participant before taking part in the research.

2.1.4. Between-scanner reliability sample

This dataset comprised 11 participants (male/female = 7/4, mean age at first scan = 30.88 ± 6.16), scanned at two different sites (Imperial College London, Academic Medical Centre Amsterdam). The average interval between each scan 68.17 ± 92.23 days, with eight participants being scanned in Amsterdam first, three in London first. High-resolution T1-weighted MRIs were acquired as follows: London Siemens Verio 3T; magnetisation-prepared rapid gradient-echo (MPRAGE), TE = 2.98 ms, TR = 2300 ms, TI = 900 ms, flip angle = 9°, field-of-view = 256 mm, 160 slices of 1.0 mm thickness, in-plane resolution = 1.0 × 1.0 mm. Amsterdam Philips Ingenia 3T; sagittal Turbo Field Echo (T1-TFE), TE = 3.1 ms, TR = 6.6 ms, flip angle = 9°, field-of-view = 270 mm, 170 slices of 1.2 mm thickness, in-plane resolution = 1.1 × 1.1 mm. The study was approved by the West London NRES and the Academic Medical Centre Amsterdam institutional review board respectively. Written consent was obtained from each participant before taking part in the research.

2.2. Neuroimaging processing

The T1-MRI data for all datasets were processed to generate normalised brain volume maps and 'raw' data appropriate for analysis.

2.2.1. Normalised brain volume maps

We followed the protocol as previously outlined (Cole et al., 2017a, 2017b, 2017c) to generate volumetric maps for use as features in our analysis. Grey matter (GM) and white matter (WM) images were analysed together, to generate a whole-brain predicted age, as well as age predictions for each tissue. In brief, all images were pre-processed using SPM12 (University College London, London, UK) to segment raw T1 images according to tissue classification (e.g. GM, WM or cerebrospinal fluid). Thorough visual quality control was conducted to ensure accuracy of segmentation and any motion-corrupted images were excluded. Normalised 3D maps of GM and WM volume were then generated in MNI152 space. Normalization used DARTEL for non-linear registration and resampling included modulation and 4 mm smoothing. This process was applied independently to images from all four datasets described in section 2.1, resulting in normalised maps with voxelwise correspondence for all participants.

2.2.2. Raw data

While the study aimed to use data in rawest possible form, some minimal pre-processing was carried out to facilitate comparison across different data sources. This included converting from DICOM to Nifti format, using dcm2nii from micron (Rorden and Brett, 2000), to be compatible with our in-house software. Raw Nifti files then underwent a rigid registration (i.e. six degrees-of-freedom) to MNI152 space, (FMRIB Software Library [FSL] FLIRT, Jenkinson and Smith, 2001) to ensure consistency of orientation (Right, Posterior, Inferior [RPI]). The images were resampled, using cubic spline interpolation, to common voxel sizes and dimensions (1 mm^3, 182 × 218 × 182), as the different contributing studies had acquired data at different dimensions. While not technically in 'raw' form, we assert that this is the rawest form possible for multi-site datasets, and that the assumptions used here are minimal and uncontroversial. Examples of the different data types used in the study are shown in Fig. 1.

2.3. Machine learning brain age modelling methods

2.3.1. Convolutional neural networks

Since their first appearance, CNNs (Lecun et al., 1998) have been very actively investigated, especially in more recent years. Several different network architectures have been proposed, which have enabled to reach state-of-the-art predictive performance in many computer vision and speech recognition tasks (Chen and Lin, 2014; Deng and Yu, 2013). Our hypothesis was that a CNN would provide an appropriate architecture to infer imaging features, from both processed and un-processed brain MRI scans, that optimally predict brain age. When properly trained, CNNs have been shown to be invariant to several variability sources, such as rotation or contrast (Krizhevsky et al., 2012), an aspect that makes them particularly appealing for our application. Given the nature of MR imaging, we have developed a network architecture that uses 3D convolutions (Ji et al., 2013), which are appropriate when dealing with fully volumetric images. Recently, 3D convolutional neural networks have been also proposed for Alzheimer's disease classification (Payan and Montana, 2015; Sarraf and Tofighi, 2016), brain lesion segmentation (Kamnitsas et al., 2017) and skull stripping (Kleesiek et al., 2016).

Our proposed 3D CNN architecture uses MRI volumes of a size (z × h × w) as inputs. The specific dimensions, in our applications, are 182 × 218 × 182 when using raw data and 121 × 145 × 121 when using registered GM/WM data. The output to be predicted is a single scalar representing the biological age. A schematic illustration of the 3D CNN architecture is given in Fig. 2. The architecture contains repeated 5 blocks of a (3 × 3 × 3) convolutional layer (with stride of 1), a rectified linear unit (ReLU), a (3 × 3 × 3) convolutional layer (with stride of 1), a 3D batch-normalization layer (Ioffe and Szegedy, 2015), a ReLU and finally a (2 × 2 × 2) max-pooling layer (with stride of 2). The number of feature channels was set to eight in the first block, and was doubled after each max-pooling layer to infer a sufficiently rich representation of the brain. The final age prediction is obtained by using a fully connected layer, which maps the output of the last block to a single output value. The total number of parameters for 5 blocks were 1992, 10,464, 41,664, 166,272, and 663,808 respectively and last fully connected layer had 5760 number of parameters. Hence, overall total number of parameters of our model was 889,960. We had used mini-batch of size 28. For the brain-predicted age using both GM and WM data, we first pre-trained the two individual networks using only GM and WM input data, and then created a single architecture where the highest level blocks of these two networks were joined. A final fully connected layer was then added to predict age using both inputs.

In each application, the network weights were trained by minimizing the mean absolute error (MAE) using a stochastic gradient descent optimisation algorithm with momentum (Sutskever et al., 2013). Back-propagation was used to compute the gradient of the objective function with respect to all parameters of the model. At the training phase, all datasets were augmented by generating additional artificial training images to prevent model over-fitting. The data augmentation strategy consisted of performing translation (±10 pixels) and rotation (±40°), and was found empirically to yield better performance compared to no data augmentation.

All the results reported in Section 2.4 refer to the best out of 3 experiments in which the models were initialised with random parameters

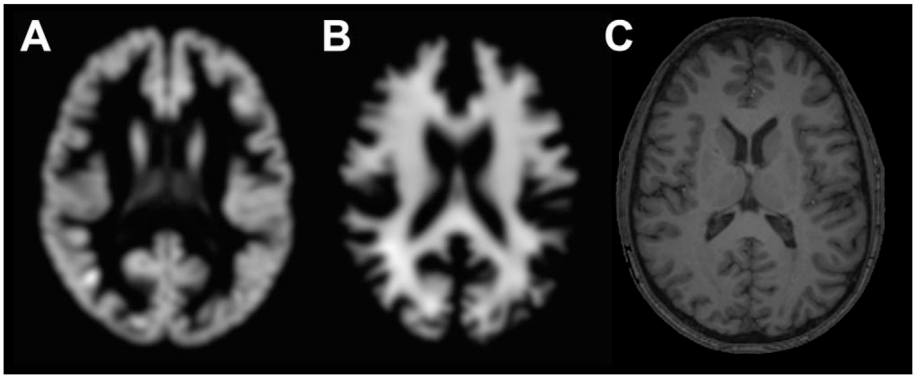

Fig. 1. Examples of neuroimaging input data for use in age prediction models. A) Grey matter volumetric map, normalised to MNI152 space using SPM DARTEL for non-linear registration, 4 mm smoothing and modulation, in axial view. B) White matter volumetric map, normalised to MNI152 space, in axial view. C) Raw, or minimally-processed, T1-weighted MRI, rigidly-registered to MNI152 space and resampled to a common voxel space.

and trained end-to-end. The best results were achieved using a learning rate of 0.01 with constant decay of 3% after each epoch, a momentum of 0.9 and weight decay of 0.00005. Training the CNN architectures with only GM or WM input, combined GM and WM input, and raw data input took 18, 42, and 83 h of training time, respectively, using four GPUs (NVidia TitanX). Importantly, however, testing time in all cases ranged between only 290–940 ms, depending on input type, on a single GPU. All software was written in Torch, a scientific computing framework with support for machine learning algorithms and GPU computing.

2.3.2. Gaussian processes regression

To contextualise the age-prediction performance of CNN, Gaussian Processes Regression (GPR) was used for comparison, as it has previously shown high accuracy in predicting chronological age from T1-MRI data (Cole et al., 2015, 2017a, 2017b, 2017c). A Gaussian Processes (GP) can be thought of as a function that extends the multivariate Gaussian distribution, that can be applied over an infinite number of variables. The assumption in GPs is that any finite subset of the data has a multivariate Gaussian distribution. The prior belief about the relationship between variables is informed by definition of these (unlimited number of) multivariate Gaussians in order to generate a model that represents the observed variance. As the multivariate Gaussians can reflect local patterns of covariance between individual points, the combination of multiple Gaussians in a GP can readily model non-linear relationships and is more flexible than conventional parametric models, which rely on fitting global models. GPs can be applied either to categorical data (for GP classification) or continuous data (the GPR approach).

The GPR method was implemented using the Pattern Recognition for Neuroimaging Toolbox (PRoNTo v2.0, www.mlnl.cs.ucl.ac.uk/pronto). Normalised volume images were converted to vectors and the resulting GM and WM vectors concatenated for each subject. A linear kernel representation of these data was then derived by calculating an N-by-N similarity matrix, where each point in the matrix was the dot (scalar) product of two subjects' image vectors. This step retains all the original image variance in a more compact representation, greatly reducing subsequent computation time. A GPR function was defined, with chronological age as the dependent variable and the image data (in similarity matrix form) as the independent variables, to build a model of healthy structural brain ageing across the adult lifespan. The model was then trained and tested to assess prediction accuracy, using a cross-validation process as outlined below.

2.4. Statistical analysis

2.4.1. Machine learning age prediction evaluation

Both the CNN and GPR methods were used to predict chronological age using structural neuroimaging as input data. The input data took four different forms; three using normalised brain volume maps (GM only, WM only, GM and WM combined [i.e. concatenated vectors]) and one using raw T1 data. Each learning method was evaluated with each data type, resulting in eight accuracy assessments.

The BAHC dataset (N = 2001) was used for this stage, and was randomly split into a large training (N = 1601), a validation set (N = 200) and a test set (N = 200). Age distributions of each sample can be seen in Supplementary material (Fig. S2). All accuracy assessments reported used predictions made on the test set. Model accuracy was expressed as the correlation between age and predicted age (Pearson's r), total variance explained (R^2), MAE and root mean squared error (RMSE).

2.4.2. Heritability analysis

Assessment of the heritability of brain-predicted age utilised the TwinsUK sample (N = 62 females). Using the models trained on the BAHC dataset (N = 1601), unbiased age predictions were made for the TwinsUK participants, to generate a brain-predicted age score for each individual. Heritability estimation was performed using Structural Equation Modelling (SEM), as implemented in the OpenMx software

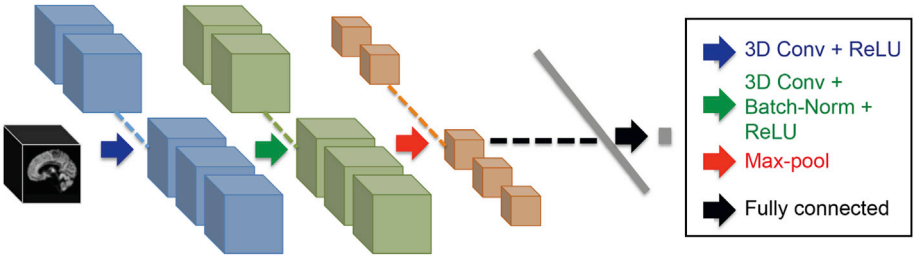

Fig. 2. Overview of the 3D convolutional neural network architecture. 3D boxes represent input and feature maps. The arrows represent network operations: blue arrow indicates 3D convolutional operation and a rectified linear unit (ReLU), green arrow indicates 3D convolutional operation, 3D batch normalization and ReLU, red arrow indicates max-pooling operation. Our brain age prediction architecture contains repeated 5 blocks of 3D convolution, ReLU, 3D convolution, 3D batch normalization, ReLU and max-pooling operations and one fully connected layer at the end, which generates the regression model to output brain-predicted age.

(Boker et al., 2011). Heritability in SEM is estimated as the proportion of phenotypic variance explained by genetic factors. This analysis was done on the generated brain-predicted ages (unadjusted model) and on age-adjusted values (age-corrected model). Age correction was carried out by partialling out chronological age from brain-predicted age in a linear regression model. The resulting residuals were then used as phenotypes for heritability analysis. SEM evaluates which combination of additive (A) genetic, common (C) environmental and unique (E) environmental variance components can best explain the observed phenotypic variance and covariance of monozygotic and dizygotic twin data. The importance of individual variance components is assessed by dropping components sequentially from the set of nested models: ACE→AE→E. In choosing between sub-models, variance components are excluded from the selection process if there is no significant deterioration in model fit after the component is dropped, as assessed by the Akaike Information Criterion (Akaike, 1974). The E component represents random error and must be retained in all models (Rijsdijk and Sham, 2002). A measure of reliability regarding the heritability results can be acquired by assessing model fit, using a log-likelihood ratio test, between a structured SEM model and a saturated model where no structure is imposed on the phenotypic covariance. The null model of the ratio test corresponds to the SEM model, while the alternative is the saturated model. Low values of the likelihood ratio test mean that the result is less likely to occur under the null model as compared to the alternative. On the other hand, high values of the statistic mean that the result is as likely to occur under the null as the alternative, and the null model cannot be rejected. The test statistic is asymptotically chi-squared distributed with degrees of freedom equal to the difference in the number of parameters between the two models. Heritability estimates for the AE models are calculated using the formula $h^2 = \frac{a^2}{a^2+e^2}$, where a and e are the path coefficients of the A and E variance components in the SEM model. Further details on SEM can be found in the Supplementary Text.

2.4.3. Reliability analysis

To calculate the reliability both within- and between-scanner, the intraclass correlation coefficient (ICC) was used. Specifically, this was ICC [2,1] according to Shrout and Fleiss' (1979) notation, to assess absolute agreement between single raters (e.g., scanners). Again using the models trained on the BAHC training set (N = 1601), unbiased age predictions were made for each participants' scans in the within-scanner and between-scanner reliability datasets. By subtracting chronological age (at time of scan) from brain-predicted age, a brain-predicted age difference (Brain-PAD) score was calculated. ICC of brain-PAD score was calculated comparing data from scans approximately four weeks apart (within-scanner sample) and comparing data from a Siemens scanner in London and a Philips scanner in Amsterdam (between-scanner sample).

3. Results

3.1. Convolutional neural networks accurately predict age using neuroimaging

Analysis showed that our CNN method could accurately predict the chronological age of healthy adults, using either processed volumetric maps or raw T1-MRI data (see Table 1). Prediction accuracy was similar for GPR. The lowest MAE achieved was using GM data and CNN analysis (MAE = 4.16 years), though other predictions were generally comparable. Using single tissues (i.e., GM or WM) did not appreciably alter the prediction accuracy compared to using all available input data for each subject (i.e., GM+WM or raw). The different prediction methods and input data combinations all provided highly accurate estimates. Three example predictions on the test set (N = 200) are shown in Fig. 3.

Further analysis indicated that brain-PAD varied as a function as age, with brain-PAD being significantly negatively correlated with chronological age (Pearson's r between −0.2 and −0.4) for all prediction methods (see Supplementary Material Table S2). It appears that there was a systematic over-estimation of brain-predicted age in younger individuals, and an under-estimation in older individuals. Individuals towards the mean age of the training sample were more accurately estimated.

Table 1
Chronological age prediction accuracy.

Method	Input data	MAE (years)	r	R^2	RMSE
CNN	GM	4.16	0.96	0.92	5.31
	WM	5.14	0.94	0.88	6.54
	GM+WM	4.34	0.96	0.91	5.67
	Raw	4.65	0.94	0.88	6.46
GPR	GM	4.66	0.95	0.89	6.01
	WM	5.88	0.92	0.84	7.25
	GM+WM	4.41	0.96	0.91	5.43
	Raw	11.81	0.57	0.32	15.10

MAE = mean absolute error, r = Pearson's r from correlation between chronological age and brain-predicted age, RMSE = root mean squared error, GM = Grey Matter, WM = White Matter.

3.2. Brain-predicted age is moderately heritable

Two sets of heritability analyses were performed. First, we estimated the heritability of brain-predicted age using the predictions from the various models. Subsequently, we partialled out the effect of chronological age from brain-predicted age, using a linear regression model. Estimates of the heritability of the age-corrected measurements, (i.e., the residuals from the regression model) were then made.

For all prediction methods, the AE models had the best fit, in comparison to ACE and E models (i.e., lowest AIC). Table 2 includes the heritability estimates for the AE models, along with their standard errors, computed by propagating the standard error of the model coefficients a and e, and finally log-likelihood ratio test p-values. For all AE models, the ratio test statistic was asymptotically chi-squared distributed with 7 degrees of freedom.

Brain-predicted age heritability estimates were consistently above 0.5, irrespective of the data and predictive model employed. Ratio test p-values were above 0.05, indicating that the data were consistent with the AE models. Similar estimates were acquired from both CNN and GPR predicted ages. The highest estimates, in both unadjusted and age-corrected cases, were produced for the CNN-predicted ages using GM+WM data, while the heritability of brain-predicted age was reduced after partialling out chronological age.

Heritability estimates are given by $h^2 = \frac{a^2}{a^2+e^2}$, where a and e are the path coefficients of the A and E variance components in the SEM model, ± the standard errors of the estimates. GM = grey matter, WM = white matter, CNN = convolutional neural network, GPR = Gaussian processes regression.

3.3. Brain-predicted age difference is highly reliable

Brain-PAD scores were generally highly reproducible, whether using CNN or GPR to generate brain-predicted ages. This was the case for both within-scanner (i.e., scanner test-retest) reliability and between-scanner (i.e., multi-site) reliability (see Table 3, Figs. 4 and 5). All the different combinations of input data (GM, WM, GM+WM, raw) and prediction method (CNN, GPR) resulted in a significant ICC for reliability ($p < 0.05$), with the exceptions of WM using CNN and raw data using GPR. Broadly speaking, the within-scanner reliability estimates were higher than the between-scanner estimates, and this difference was more pronounced for CNN brain-PAD than GPR brain-PAD. For the latter, between-scanner reliability for GM and GM+WM was as high as within-scanner reliability. Notably, the within-scanner reliability for raw data using CNN was very high (ICC = 0.94), though substantially reduced when comparing estimates from two scanners (ICC = 0.66).

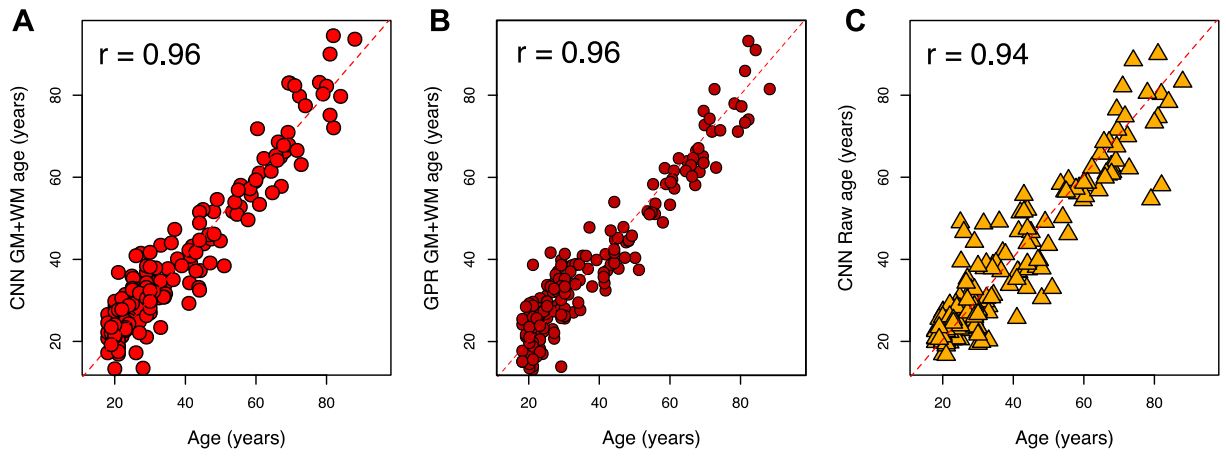

Fig. 3. Accuracy of CNN and Gaussian Processes Regression for age prediction. Scatterplots depict chronological age (x-axis) and brain-predicted age (y-axis) on the test-set subjects from the BAHC dataset (N = 200). A) Brain-predicted ages derived used GM maps as input data for the CNN method. B) Brain-predicted ages derived using GM maps as input data for the Gaussian Processes Regression (GPR) method. C) Brain-predicted ages derived using raw T1-MRI as input for the CNN method. *r*-values in all plots are the Pearson's correlation coefficient of brain-predicted age with chronological age.

4. Discussion

Using 3D convolutional neural networks, we accurately estimated chronological age from raw T1-weighted MRI brain scans of healthy adults. The accuracy of CNN for age prediction was also high when using processed GM and WM voxelwise images, and was comparable with age estimations made using GPR. Brain-predicted age estimates were significantly heritable and showed high levels of within-scanner and between-scanner reliability. These findings support the idea that deep learning methods can generate a viable biomarker of brain ageing: brain-predicted age.

This study is the first illustration that 3D convolutional neural networks can be used to accurately estimate chronological age from neuroimaging data. CNN-based predictions were similar compared to a previously employed method, GPR (Cole et al., 2015), and the predictions themselves were highly correlated with each. When using voxelwise images representing GM and WM volume, both methods were able to predict age with less than 5 years MAE. Importantly, CNN-based age prediction accuracy was equally high when using raw (or minimally pre-processed) T1 images as input and are in line with the standard in the field (Ashburner, 2007; Cole et al., 2017c; Erus et al., 2014; Franke et al., 2010; Konukoglu et al., 2013; Mwangi et al., 2013; Su et al., 2013). This means that sufficient age-related features can be extracted from a 3D image to enable accurate out-of-sample prediction and obviates the necessity of image pre-processing. This brings two key benefits, specifically: 1) removal of additional assumptions that are required to pre-process image data; 2) increasing the feasibility of such an approach for use in real-time (or near real-time), to aid clinical decision making.

Data pre-processing is almost ubiquitous in neuroimaging, including in previous studies using brain-based age predictions (Cole et al., 2015; Franke et al., 2010; Irimia et al., 2014; Schnack et al., 2016; Steffener et al., 2016). Multiple different options are available from different software packages for each stage of pre-processing, including bias-field correction, removal of non-brain tissue, tissue classification, motion correction, artefact removal, linear registration, non-linear registration, target image (e.g., atlas, average template), interpolation method and smoothing kernel. While we opted to use SPM here, the relative merits of different approaches are keenly debated (e.g., Barnes et al., 2008; Klein et al., 2010; Mohammadi et al., 2010; Ribeiro et al., 2015; Valverde et al., 2015). In the absence of a consensus, the ability to model outcome variables without conducting any of these steps is attractive. As the choice of pre-processing pipeline will undoubtedly influence any derived measures, using raw data for prediction removes a key source of variance. Moreover, as the assumptions underlying many of the different pre-processing steps are often not met when dealing with clinical populations containing individuals with atypical brain morphology, using raw data also removes additional confounds and potential biases. Nevertheless, our approach will require further validation on such atypical inputs.

A key goal of neuroimaging research is to make tools for clinical application, that can provide objective and reliable information that clinicians can use to help when treating brain diseases. One element of this goal is in producing real-time methods that generate interpretable outputs from imaging data for immediate use in clinical decision making. Image pre-processing can take more than 24 h, hence removing this step represents a substantial acceleration of the pipeline necessary to deliver information to the clinician. Admittedly, the training phase of a deep artificial neural network is computationally-intensive and time-consuming. However, once trained, the model can be applied to new data in a matter of seconds. Given the right software implementation,

Table 2
Heritability estimates from the AE SEM models for different brain-predicted age methods.

Method	GM	WM	GM+WM	Raw
Unadjusted				
CNN	0.74 ± 0.09 (p-value:0.584)	0.78 ± 0.07 (p-value:0.506)	0.84 ± 0.05 (p-value:0.083)	0.62 ± 0.10 (p-value:0.063)
GPR	0.78 ± 0.07 (p-value:0.57)	0.81 ± 0.06 (p-value:0.502)	0.82 ± 0.06 (p-value:0.443)	0.64 ± 0.10 (p-value:0.178)
With age-correction				
CNN	0.55 ± 0.11 (p-value:0.277)	0.65 ± 0.10 (p-value:0.776)	0.66 ± 0.09 (p-value:0.132)	0.50 ± 0.12 (p-value:0.075)
GPR	0.55 ± 0.11 (p-value:0.124)	0.60 ± 0.10 (p-value:0.837)	0.58 ± 0.11 (p-value:0.254)	0.64 ± 0.10 (p-value:0.23)

Table 3
Within-scanner and between-scanner reliability estimates of brain-predicted age difference.

Method	Dataset	GM	WM	GM+WM	Raw
CNN	Within	0.90 [0.76, 0.96]	0.97 [0.90, 0.99]	0.90 [0.77, 0.96]	0.94 [0.86, 0.98]
	Between	0.83 [0.49, 0.95]	0.51 [-0.08, 0.84]	0.85 [0.55, 0.96]	0.66 [0.17, 0.89]
GPR	Within	0.96 [0.90, 0.98]	0.98 [0.94, 0.99]	0.97 [0.92, 0.99]	0.99 [0.97, 0.99]
	Between	0.96 [0.88, 0.99]	0.77 [0.12, 0.94]	0.92 [0.74, 0.98]	0.56 [-0.02, 0.86]

All figures in the table are intraclass correlation coefficients (ICC) and 95% confidence intervals. GM = grey matter, WM = white matter, CNN = convolutional neural network, GPR = Gaussian processes regression.

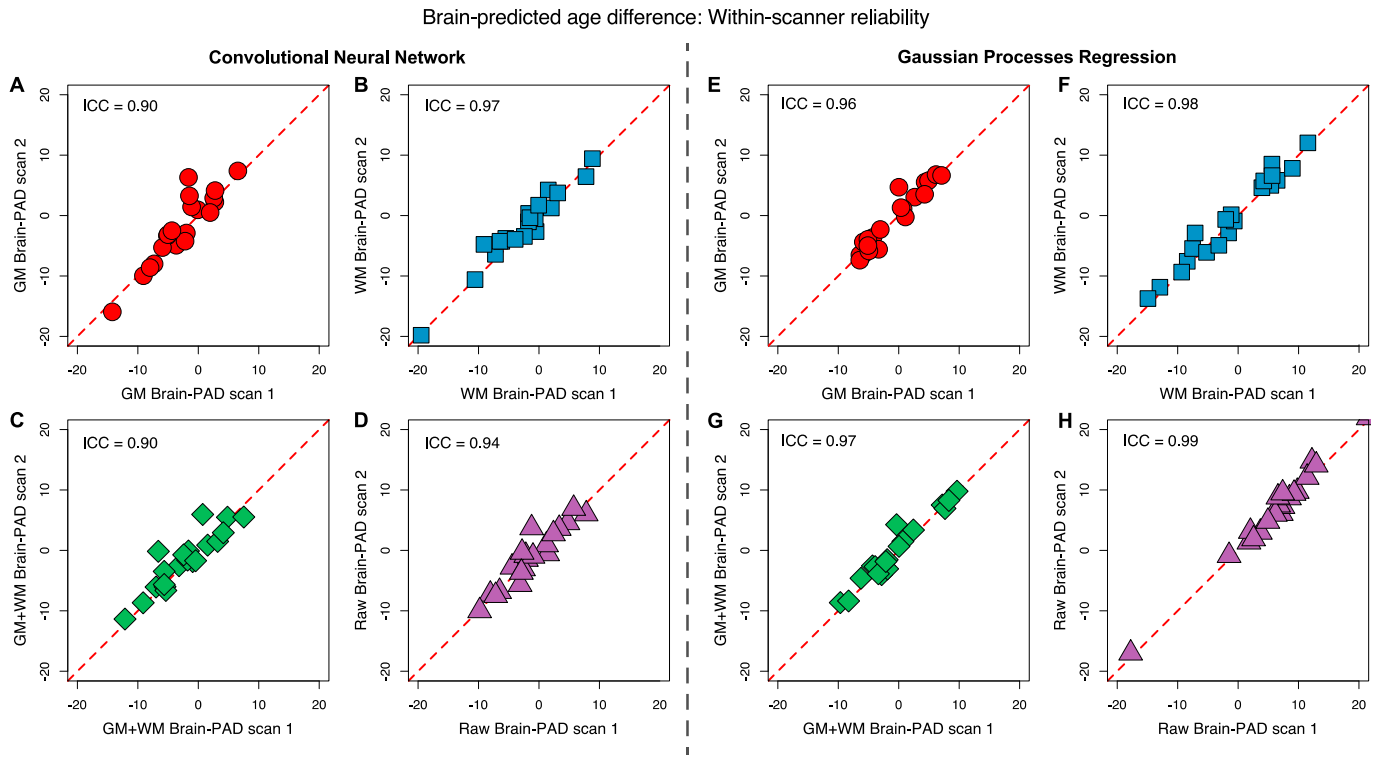

Fig. 4. Within-scanner reliability for Convolutional Neural Networks and Gaussian Processes Regression. Figure shows the correspondence between brain-predicted age difference (Brain-PAD) based on scans acquired four weeks apart on the same scanner (Siemens Verio 3T) on N = 20 individuals, with scan 1 on the x-axis and scan 2 (after four weeks) on the y-axis for all plots. A) Brain-PAD score based on GM maps using CNN. B) Brain-PAD score based on WM maps using CNN. C) Brain-PAD scored based on GM and WM maps combined using CNN. D) Brain-PAD scored based on raw T1-MRI using CNN. E) Brain-PAD score based on GM maps using Gaussian Processes Regression (GPR). F) Brain-PAD score based on WM maps using GPR. G) Brain-PAD score based on GM and WM maps combined using GPR. H) Brain-PAD score based on raw T1-MRI using GPR. The red dashed line in all plots is the line of identity.

brain-predicted age data could be made available to a clinician while the patient is still in the scanner. In our study, minimal processing was used when training/testing the CNN algorithm, only to ensure consistent image orientation and voxel dimensions between images. These processes require very limited assumptions and could readily be automated into MR scanner software.

The heritability of CNN brain-predicted age was consistently above 0.5, as were all prediction methods, indicating that moderate levels of genetic relatedness influence brain-predicted age. This is in line with previous research, that estimated the heritability of volumetric measures of brain structure to be between 0.45 and 0.9 (Baaré et al., 2001; Batouli et al., 2014a; Kremen et al., 2010; Winkler et al., 2010). This evidence of heritability is important as it provides a degree of external validity to the measure of brain-predicted age. If brain-predicted age were merely a reflection of disease-related atrophy or driven by noise, then the additive genetic models would not significantly account for the observed data. This supports further use of brain-predicted age as a biomarker of brain ageing. Moreover, as previous reports have indicated that brain-predicted age relates to measures of cognitive performance (Cole et al., 2015; Gaser et al., 2013), it could potentially be used to predict risk of future cognitive decline and risk of subsequent dementia. That our measure of brain ageing is under some genetic control corroborates research indicating that cognitive ageing is also influenced by genetic factors (Harris and Deary, 2011; McClearn et al., 1997; Tucker-Drob et al., 2014). Intuitively it follows that brain ageing (i.e., underlying anatomical changes) and cognitive ageing (i.e., manifest behavioural changes) must be linked. Therefore, our findings, along with previous demonstrations of the heritability of brain structure (Baaré et al., 2001; Batouli et al., 2014b; Kremen et al., 2010; Winkler et al., 2010), motivate research into specific genes which may influence rates of brain and cognitive ageing. Such genes have the potential to offer novel targets for pharmacological interventions aimed at reducing the risk of age-associated neurodegeneration and cognitive decline, as well as even slowing the rate of brain ageing itself.

In order to evaluate whether chronological age affects the heritability of brain age, we performed two different sets of heritability analyses, both before and after controlling for chronological age. The difference in estimated heritabilities between the two cases is a strong indicator that age does play an important role. This is in line with previous research into the heritability of brain volumes (Batouli et al., 2014a, 2014b) and cognitive function (Bouchard, 2013; Lee et al., 2010). This certainly has implications for genetic studies of brain ageing, including how studies are designed and samples are selected. However, the limited number of twins and their age distribution in the current study did not permit an exhaustive analysis of how heritability changes with age.

Brain-predicted age was highly reproducible. Reliability estimates varied for different combinations of input data and algorithm, however within-scanner test-retest reliability was high (ICC $\geq$ 0.90) for all analysis, even using raw data. This is crucial for any measure to be used in longitudinal studies, and has an important bearing on the sample sizes required to detect significant effects on repeated measures (Guo et al., 2013). The test-retest reliability of T1-MRI derived brain structural measures has been demonstrated (Morey et al., 2010; Nugent et al., 2013), and our results are consistent with these estimates. This high reproducibility supports the use of brain-predicted age in longitudinal research or potentially clinical settings.

Regarding between-scanner reliability for brain-predicted age, the results were more contrasting. GPR between-scanner reliability was generally higher than for CNN, and GM was generally better than WM. This agrees broadly with previous studies investigating the reliability of T1-MRI measures in multi-centre settings (Jovicich et al., 2006; Schnack et al., 2004). Between-scanner reliability was substantially reduced for raw data. Potential explanations for this include differences in contrast-to-noise ratio observed between different vendors on T1-MRI, or

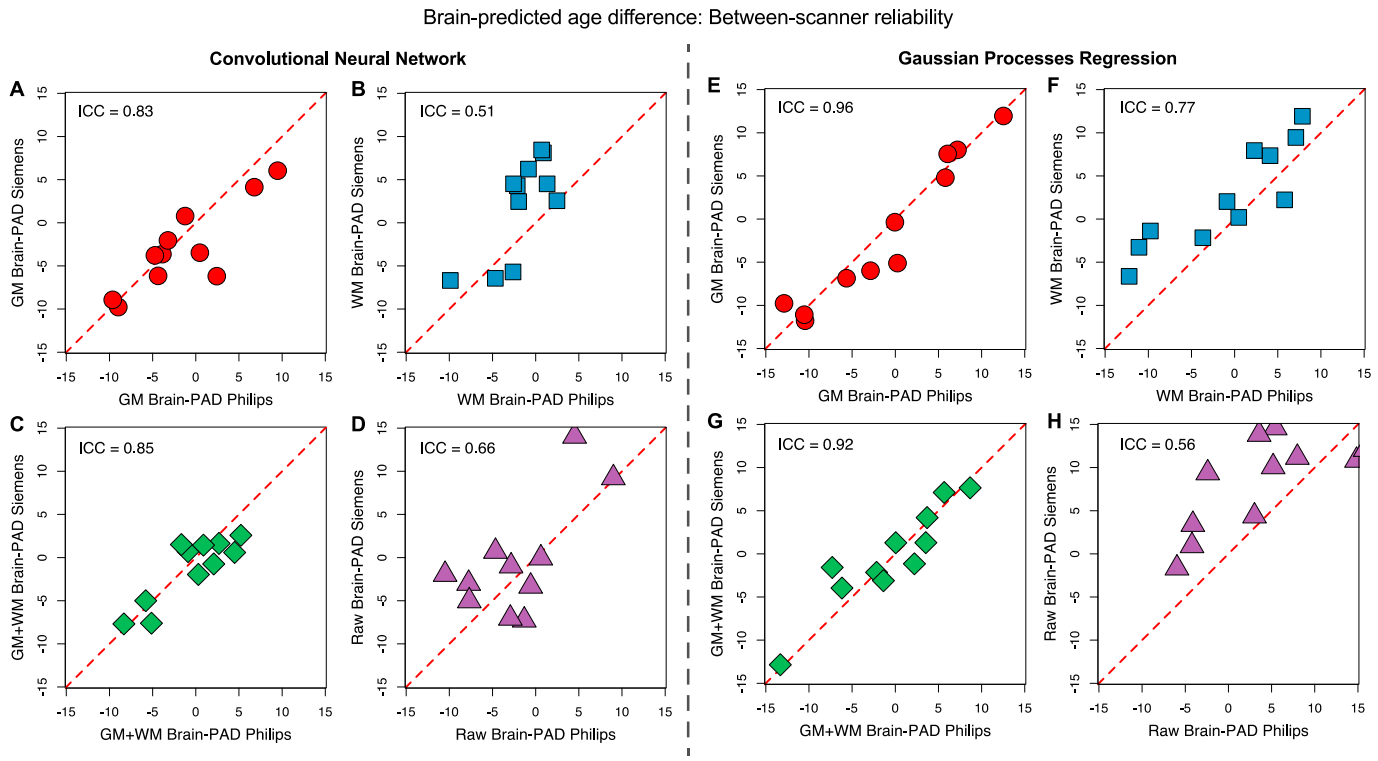

Fig. 5. Between-scanner reliability for Convolutional Neural Networks and Gaussian Processes Regression. Figure shows the correspondence between brain-predicted age difference (Brain-PAD) scores based on scans acquired on two different scanner systems (Siemens Verio 3T and Philips Intera 3T) in N = 11 individuals, with the Philips scan on the x-axis and Siemens scan on the y-axis for all plots. A) Brain-PAD score based on GM maps using CNN. B) Brain-PAD score based on WM maps using CNN. C) Brain-PAD scored based on GM and WM maps combined using CNN. D) Brain-PAD scored based on raw T1-MRI using CNN. E) Brain-PAD score based on GM maps using Gaussian Processes Regression (GPR). F) Brain-PAD score based on WM maps using GPR. G) Brain-PAD score based on GM and WM maps combined using GPR. H) Brain-PAD score based on raw T1-MRI using GPR. The red dashed line in all plots is the line of identity.

differences in shimming effectiveness between different scanners. The pre-processing steps used to generate normalised GM and WM images largely remove the effects of inconsistent gradient distortions, by carrying out bias-field correction and estimating tissue probabilities. Conversely, the CNN architecture may be characterising these explanatory features at a given level of the model. Currently, it would seem that deep learning models using raw data are most appropriate for longitudinal studies of brain ageing on the same scanner, where the issue of image heterogeneity due to inter-scanner variability is not present. Therefore, there are likely still benefits to data pre-processing when pooling data from multiple scanners, as would be the case in a multi-centre study, to remove clearly identifiable sources of technical variability that are unrelated to brain ageing.

There are some limitations of the study to consider. The sample size for the heritability estimates was small, particularly regarding the numbers of dizygotic twins included and the sample was composed of females, hence we cannot readily extrapolate to males. However, the standard errors take sample size into account and while the precision behind the estimates could be greatly improved by increasing numbers, the implication that at least some of the variance in brain-predicted age is moderated by genetic relatedness remains valid. Another limitation is the absolute accuracy of the age-prediction models. With a MAE of ~4–5 years, clearly the precision of the estimates in an individual case are insufficient for making clinically-meaningful estimates. Further research is necessary to further reduce MAE. One avenue for this would be an in-depth exploration of different data augmentation methods to optimise the CNN analysis, including manipulating the image contrast-to-noise ratio or using anatomical priors. A further limitation is that our between-scanner reliability analysis only used data from two scanners, with the same field strength. To build up a comprehensive picture of the influence of scanner system on brain-predicted age, further varieties of scanner should be analysed. Another issue is that our analysis currently provides no neuroanatomical specificity regarding which features are used in the prediction of chronological. There are two reasons for not addressing this currently. Firstly, brain ageing is generally speaking a global phenomenon, with all brain regions being affected (Fjell et al., 2009). Also, the 'key' regions or features may vary at different stages of the lifespan, as brain volume changes have been shown to be spatially-varying and non-linear (Fjell et al., 2013). Secondly, the technical challenges underlying the representation of features from CNNs in particular are highly complex. While efforts are underway to produce individual 'saliency' maps highlighting feature importance, this is will be addressed in future research.

It should also be noted that our analysis not account for in-scanner motion. Recent evidence suggests chronological age relates to subject motion and that subject motion can influence brain volume predictions (Pardoe et al., 2016). While such artefacts may potentially have influenced the current study, Pardoe et al., found that voxel-based measures were not associated with motion parameters when only including images that passed visual quality assurance. While we were unable to explore this issue explicitly due to a lack of data on level of motion, we can be confident that the increased noise this would cause would have caused a general error, rather than a bias, and would not have inflated the accuracy, heritability or reliability results. The issue of motion influencing brain volumetrics is important, particular as clinical groups appear more associated with scanner motion. Future studies could improve their experimental analysis by explicitly measuring and statistical accounting for any individual variability that is motion, not disease, related.

In addition, the accuracy of all predictive methods could be improved, and an influence of chronological age on predictions was evident. It appeared that estimated tended towards the training set mean age, with minor over-estimations in younger adults and under-estimations in older adults. While the reasons for this are unclear, one speculation is that distribution of noise is not consistent across the lifespan, as there are no

examples to learn from under the age of 18 or over 90 years. Future work could look to reduce this proportional error, perhaps through additional regularisation methods, and thus improve model accuracy still further.

5. Conclusions

Deep learning models based on T1-MRI can accurately predict chronological age in healthy individuals. This can be achieved using raw MRI data, with a minimum of processing necessary to generate an accurate age prediction. These estimates of brain-predicted age are also considerably heritable, giving external, genetic, validity to the measure and motivating its use in genetic studies of brain ageing. Finally, our analysis showed the brain-predicted age is highly reliable and thus appropriate for use in both longitudinal and multi-centre studies, though preprocessing appears necessary to achieve high between-scanner reliability. Brain-predicted age has the potential to be used as a biomarker to investigate the brain ageing process and how this relates to cognitive ageing, neurodegeneration and age-associated brain diseases.

Acknowledgements

The TwinsUK study was funded by the Wellcome Trust, Medical Research Council, European Commision's Seventh Framework Program (FP7/2007-2013, GA No 259749). The study also receives support from the National Institute for Health Research (NIHR), BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. The STudy Of Reliability of MRI (STORM) was funded by the NIHR Biomedical Research Centre based at Imperial College London. James Cole is funded by a research grant to Imperial College London from the Medical Research Council (MR/L022141/1). At the time of the study, Dimosthenis Tsagkrasoulis was funded by an EPSRC Doctoral Departmental Scholarship to Imperial College London.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2017.07.059.

References

Abbott, A., 2011. Dementia: a problem for our age. Nature 475, S2–S4.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19, 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705.
Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38, 95–113. http://dx.doi.org/10.1016/j.neuroimage.2007.07.007.
Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41. http://dx.doi.org/10.1016/j.media.2007.06.004.
Baaré, W.F.C., Hulshoff Pol, H.E., Boomsma, D.I., Posthuma, D., De Geus, E.J.C., Schnack, H.G., Van Haren, N.E.M., Van Oel, C.J., Kahn, R.S., 2001. Quantitative genetic modeling of variation in human brain morphology. Cereb. Cortex 11, 816–824.
Barnes, J., Foster, J., Boyes, R.G., Pepple, T., Moore, E.K., Schott, J.M., Frost, C., Scahill, R.I., Fox, N.C., 2008. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. NeuroImage 40, 1655–1671. http://dx.doi.org/10.1016/j.neuroimage.2008.01.012.
Batouli, S.A.H., Sachdev, P.S., Wen, W., Wright, M.J., Ames, D., Trollor, J.N., 2014a. Heritability of brain volumes in older adults: the older Australian twins study. Neurobiol. Aging 35, 937. http://dx.doi.org/10.1016/j.neurobiolaging.2013.10.079 e935-937.e918.
Batouli, S.A.H., Trollor, J.N., Wen, W., Sachdev, P.S., 2014b. The heritability of volumes of brain structures and its relationship to age: a review of twin and family studies. Ageing Res. Rev. 13, 1–9. http://dx.doi.org/10.1016/j.arr.2013.10.003.
Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., Mehta, P., Fox, J., 2011. OpenMx: an open source extended structural equation modeling framework. Psychometrika 76, 306–317. http://dx.doi.org/10.1007/s11336-010-9200-6.
Bouchard, T.J., 2013. The Wilson Effect: the increase in heritability of IQ with age. Twin Res. Hum. Genet. 16, 923–930. http://dx.doi.org/10.1017/thg.2013.54.
Chen, X.W., Lin, X., 2014. Big data deep learning: challenges and perspectives. IEEE Access 2, 514–525. http://dx.doi.org/10.1109/ACCESS.2014.2325029.
Cole, J.H., Annus, T., Wilson, L.R., Remtulla, R., Hong, Y.T., Fryer, T.D., Acosta-Cabronero, J., Cardenas-Blanco, A., Smith, R., Menon, D.K., Zaman, S.H., Nestor, P.J., Holland, A.J., 2017a. Brain-predicted age in Down Syndrome is associated with β-amyloid deposition and cognitive decline. Neurobiol. Aging. http://dx.doi.org/10.1016/j.neurobiolaging.2017.04.006.
Cole, J.H., Leech, R., Sharp, D.J., for the Alzheimer's Disease Neuroimaging, I, 2015. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. Ann. Neurol. 77, 571–581. http://dx.doi.org/10.1002/ana.24367.
Cole, J.H., Ritchie, S.J., Bastin, M.E., Valdes Hernandez, M.C., Munoz 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., 2017b. Brain age predicts mortality. Mol. Psychiatry. http://dx.doi.org/10.1038/mp.2017.62.
Cole, J.H., Underwood, J., Caan, M.W.A., De Francesco, D., van Zoest, R.A., Leech, R., Wit, F.W.N.M., Portegies, P., Geurtsen, G.J., Schmand, B.A., Schim van der Loeff, M.F., Franceschi, C., Sabin, C.A., Majoie, C.B.L.M., Winston, A., Reiss, P., Sharp, D.J., 2017c. Increased brain-predicted aging in treated HIV disease. Neurology 88, 1349–1357. http://dx.doi.org/10.1212/wnl.0000000000003790.
Deng, L., Yu, D., 2013. Deep learning: methods and applications. Found. Trends Signal Process. 7, 197–387. http://dx.doi.org/10.1561/2000000039.
Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W., Feczko, E., Coalson, R.S., Pruett, J.R., Barch, D.M., Petersen, S.E., Schlaggar, B.L., 2010. Prediction of individual brain maturity using fMRI. Science 329, 1358–1361. http://dx.doi.org/10.1126/science.1194144.
Erus, G., Battapady, H., Satterthwaite, T.D., Hakonarson, H., Gur, R.E., Davatzikos, C., Gur, R.C., 2014. Imaging patterns of brain development and their relationship to cognition. Cereb. Cortex. http://dx.doi.org/10.1093/cercor/bht425.
Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M., 2009. One-year brain atrophy evident in healthy aging. J. Neurosci. 29, 15223–15231. http://dx.doi.org/10.1523/JNEUROSCI.3252-09.2009.
Fjell, A.M., Westlye, L.T., Grydeland, H., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Holland, D., Dale, A.M., Walhovd, K.B., 2013. Critical ages in the life course of the adult brain: nonlinear subcortical aging. Neurobiol. Aging 34, 2239–2247.
Franke, K., Gaser, C., 2012. Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's Disease. GeroPsych. J. Gerontopsychology Geriatr. Psychiatry 25, 235–245.
Franke, K., Gaser, C., Manor, B., Novak, V., 2013. Advanced BrainAGE in older adults with type 2 diabetes mellitus. Front. Aging Neurosci. 5 (90) http://dx.doi.org/10.3389/fnagi.2013.00090.
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, 883–892.
Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., 2013. BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. PLoS One 8.
Guo, Y., Logan, H.L., Glueck, D.H., Muller, K.E., 2013. Selecting a sample size for studies with repeated measures. BMC Med. Res. Methodol. 13, 1–8. http://dx.doi.org/10.1186/1471-2288-13-100.
Harris, S.E., Deary, I.J., 2011. The genetics of cognitive ability and cognitive ageing in healthy older people. Trends Cognitive Sci. 15, 388–394. http://dx.doi.org/10.1016/j.tics.2011.07.004.
Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. Comput. Res. Repos. 448-456.
Irimia, A., Torgerson, C.M., Goh, S.-Y.M., Horn, J.D., 2014. Statistical estimation of physiological brain age as a descriptor of senescence rate during adulthood. Brain Imaging Behav. 9, 678–689. http://dx.doi.org/10.1007/s11682-014-9321-0.
Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.
Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35, 221–231. http://dx.doi.org/10.1109/TPAMI.2012.59.
Jovicich, J., Czanner, S., Greve, D., Haley, E., Van Der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., Fischl, B., Dale, A., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. NeuroImage 30, 436–443.
Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.
Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A., 2016. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. NeuroImage 129, 460–469. http://dx.doi.org/10.1016/j.neuroimage.2016.01.024.
Klein, A., Ghosh, S.S., Avants, B., Yeo, B.T.T., Fischl, B., Ardekani, B., Gee, J.C., Mann, J.J., Parsey, R.V., 2010. Evaluation of volume-based and surface-based brain image registration methods. NeuroImage 51, 214–220. http://dx.doi.org/10.1016/j.neuroimage.2010.01.091.
Konukoglu, E., Glocker, B., Zikic, D., Criminisi, A., 2013. Neighbourhood approximation using randomized forests. Med. Image Anal. 17, 790–804. http://dx.doi.org/10.1016/j.media.2013.04.013.
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., 2013. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophr. Bull. 40, 1140–1153. http://dx.doi.org/10.1093/schbul/sbt142.

Kremen, W.S., Prom-Wormley, E., Panizzon, M.S., Eyler, L.T., Fischl, B., Neale, M.C., Franz, C.E., Lyons, M.J., Pacheco, J., Perry, M.E., Stevens, A., Schmitt, J.E., Grant, M.D., Seidman, L.J., Thermenos, H.W., Tsuang, M.T., Eisen, S.A., Dale, A.M., Fennema-Notestine, C., 2010. Genetic and environmental influences on the size of specific brain regions in midlife: the VETSA MRI study. NeuroImage 49, 1213–1223. http://dx.doi.org/10.1016/j.neuroimage.2009.09.043.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 2012.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324. http://dx.doi.org/10.1109/5.726791.

Lee, T., Henry, J.D., Trollor, J.N., Sachdev, P.S., 2010. Genetic influences on cognitive functions in the elderly: a selective review of twin studies. Brain Res. Rev. 64, 1–13. http://dx.doi.org/10.1016/j.brainresrev.2010.02.001.

Lee, T., Sachdev, P., 2014. The contributions of twin studies to the understanding of brain ageing and neurocognitive disorders. Curr. Opin. Psychiatry 27, 122–127. http://dx.doi.org/10.1097/yco.0000000000000039.

Liu, X., Niethammer, M., Kwitt, R., Singh, N., McCormick, M., Aylward, S., 2015. Low-rank atlas image analyses in the presence of pathologies. IEEE Trans. Med. Imaging 34, 2583–2591. http://dx.doi.org/10.1109/TMI.2015.2448556.

Lu, T., Pan, Y., Kao, S.Y., Li, C., Kohane, I., Chan, J., Yankner, B.A., 2004. Gene regulation and DNA damage in the ageing human brain. Nature 429, 883–891. http://dx.doi.org/10.1038/nature02661.

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. http://dx.doi.org/10.1016/j.neuroimage.2016.04.007.

McClearn, G.E., Johansson, B., Berg, S., Pedersen, N.L., Ahern, F., Petrill, S.A., Plomin, R., 1997. Substantial genetic influence on cognitive abilities in twins 80 or more years old. Science 276, 1560–1563. http://dx.doi.org/10.1126/science.276.5318.1560.

Mohammadi, S., Möller, H.E., Kugel, H., Müller, D.K., Deppe, M., 2010. Correcting eddy current and motion effects by affine whole-brain registrations: evaluation of three-dimensional distortions and comparison with slicewise correction. Magn. Reson. Med. 64, 1047–1056. http://dx.doi.org/10.1002/mrm.22501.

Morey, R.A., Selgrade, E.S., Wagner 2nd, H.R., Huettel, S.A., Wang, L., McCarthy, G., 2010. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. Hum. Brain Mapp. 31, 1751–1762. http://dx.doi.org/10.1002/hbm.20973.

Mwangi, B., Hasan, K.M., Soares, J.C., 2013. Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: a machine learning approach. NeuroImage 75, 58–67. http://dx.doi.org/10.1016/j.neuroimage.2013.02.055.

Nugent, A.C., Luckenbaugh, D.A., Wood, S.E., Bogers, W., Zarate, C.A., Drevets, W.C., 2013. Automated subcortical segmentation using FIRST: test-retest reliability, interscanner reliability, and comparison to manual segmentation. Hum. Brain Mapp. 34, 2313–2329. http://dx.doi.org/10.1002/hbm.22068.

Pardoe, H.R., Cole, J.H., Blackmon, K., Thesen, T., Kuzniecky, R., 2017. Structural brain changes in medically refractory focal epilepsy resemble premature brain aging. Epilepsy Res. 133, 28–32. http://dx.doi.org/10.1016/j.eplepsyres.2017.03.007.

Pardoe, H.R., Kucharsky Hiess, R., Kuzniecky, R., 2016. Motion and morphometry in clinical and nonclinical populations. NeuroImage 135, 177–185. http://dx.doi.org/10.1016/j.neuroimage.2016.05.005.

Payan, A., Montana, G., 2015. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. Comput. Res. Repos. arXiv:1502.02506.

Peters, R., 2006. Ageing and the brain. Postgrad. Med. J. 82, 84–88. http://dx.doi.org/10.1136/pgmj.2005.036665.

Plis, S.M., Hjelm, D.R., Slakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H., Paulsen, J., Turner, J., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. http://dx.doi.org/10.3389/fnins.2014.00229.

Ribeiro, A.S., Nutt, D.J., McGonigle, J., 2015. Which Metrics Should Be Used in Non-linear Registration Evaluation?, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 388–395.

Rijsdijk, F.V., Sham, P.C., 2002. Analytic approaches to twin data using structural equation models. Briefings Bioinforma. 3, 119–133.

Rorden, C., Brett, M., 2000. Stereotaxic display of brain lesions. Behav. Neurol. 12, 191–200.

Sarraf, S., Tofighi, G., 2016. DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. bioRxiv 70441. http://dx.doi.org/10.1101/070441.

Schnack, H.G., Haren, N.E.M.v., Nieuwenhuis, M., Pol, H.E.H., Cahn, W., Kahn, R.S., 2016. Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study. Am. J. Psychiatry 173, 607–616. http://dx.doi.org/10.1176/appi.ajp.2015.15070922.

Schnack, H.G., Van Haren, N.E.M., Hulshoff Pol, H.E., Picchioni, M., Weisbrod, M., Sauer, H., Cannon, T., Huttunen, M., Murray, R., Kahn, R.S., 2004. Reliability of brain volumes from multicenter MRI acquisition: a calibration study. Hum. Brain Mapp. 22, 312–320. http://dx.doi.org/10.1002/hbm.20040.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

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. Neurobiol. Aging 40, 138–144. http://dx.doi.org/10.1016/j.neurobiolaging.2016.01.014.

Su, L., Wang, L., Hu, D., 2013. Predicting the Age of Healthy Adults from Structural MRI by Sparse Representation, pp. 271–279.

Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13). JMLR Workshop and Conference Proceedings, pp. 1139–1147.

Teter, B., Finch, C.E., 2004. Caliban's heritance and the genetics of neuronal aging. Trends Neurosci. 27, 627–632. http://dx.doi.org/10.1016/j.tins.2004.08.005.

Tucker-Drob, E.M., Reynolds, C.A., Finkel, D., Pedersen, N.L., 2014. Shared and unique genetic and environmental influences on aging-related changes in multiple cognitive abilities. Dev. Psychol. 50, 152–166. http://dx.doi.org/10.1037/a0032468.

Valverde, S., Oliver, A., Cabezas, M., Roura, E., Lladó, X., 2015. Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. J. Magn. Reson. Imaging 41, 93–101. http://dx.doi.org/10.1002/jmri.24517.

Winkler, A.M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P.T., Duggirala, R., Glahn, D.C., 2010. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. NeuroImage 53, 1135–1146. http://dx.doi.org/10.1016/j.neuroimage.2009.12.028.