NeuroImage 249 (2022) 118871

Contents lists available at ScienceDirect

NeuroImage

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

Accurate brain-age models for routine clinical MRI examinations

David A. Wood[a], Sina Kafiabadi[b], Ayisha Al Busaidi[b], Emily Guilhem[b], Antanas Montvila[b], Jeremy Lynch[b], Matthew Townend[c], Siddharth Agarwal[a], Asif Mazumder[d], Gareth J. Barker[e], Sebastien Ourselin[a], James H. Cole[e,f,g], Thomas C. Booth[a,b,1,*]

[a] School of Biomedical Engineering and Imaging Sciences, King's College London, Rayne Institute, 4th Floor, Lambeth Wing, London SE17 7EH, United Kingdom
[b] King's College Hospital NHS Foundation Trust, United Kingdom
[c] Wrightington, Wigan and Leigh NHSFT, United Kingdom
[d] Guy's and St Thomas' NHS Foundation Trust, United Kingdom
[e] Department of Neuroimaging, Institute of Psychiatry, Psychology, & Neuroscience, King's College London, United Kingdom
[f] Dementia Research Centre, Institute of Neurology, University College London, United Kingdom
[g] Centre for Medical Image Computing, Department of Computer Science, University College London, United Kingdom

ARTICLE INFO

Keywords:
Brain age
Deep learning
Convolutional neural networks
Brain-PAD
T2-weighted
Diffusion-weighted

ABSTRACT

Convolutional neural networks (CNN) can accurately predict chronological age in healthy individuals from structural MRI brain scans. Potentially, these models could be applied during routine clinical examinations to detect deviations from healthy ageing, including early-stage neurodegeneration. This could have important implications for patient care, drug development, and optimising MRI data collection. However, existing brain-age models are typically optimised for scans which are not part of routine examinations (e.g., volumetric T1-weighted scans), generalise poorly (e.g., to data from different scanner vendors and hospitals etc.), or rely on computationally expensive pre-processing steps which limit real-time clinical utility.

Here, we sought to develop a brain-age framework suitable for use during routine clinical head MRI examinations. Using a deep learning-based neuroradiology report classifier, we generated a dataset of 23,302 'radiologically normal for age' head MRI examinations from two large UK hospitals for model training and testing (age range = 18–95 years), and demonstrate fast (< 5 s), accurate (mean absolute error [MAE] < 4 years) age prediction from clinical-grade, minimally processed axial T2-weighted and axial diffusion-weighted scans, with generalisability between hospitals and scanner vendors (Δ MAE < 1 year). The clinical relevance of these brain-age predictions was tested using 228 patients whose MRIs were reported independently by neuroradiologists as showing atrophy 'excessive for age'. These patients had systematically higher brain-predicted age than chronological age (mean predicted age difference = +5.89 years, 'radiologically normal for age' mean predicted age difference = +0.05 years, $p < 0.0001$).

Our brain-age framework demonstrates feasibility for use as a screening tool during routine hospital examinations to automatically detect older-appearing brains in real-time, with relevance for clinical decision-making and optimising patient pathways.

1. Introduction

Convolutional neural networks (CNN) can accurately predict chronological age in healthy individuals from structural MRI brain scans. When applied in independent samples, deviations between an individual's brain-predicted age and their chronological age - the so-called 'brain predicted age difference' (brain-PAD), also known as brain-age gap, or delta - can be used to quantify deviations from healthy ageing (Cole and Franke, 2017b). Having a brain that more closely resembles that of an older healthy person (i.e., positive brain-PAD) has been associated with a number of neuropsychiatric diseases, including Alzheimer's disease (Franke and Gaser, 2012), mild cognitive impairment (Gaser et al., 2013), schizophrenia (Koutsouleris et al., 2014) and epilepsy (Pardoe et al., 2017); a positive brain-PAD has also been associated with cognitive impairment following traumatic brain injury (Cole et al., 2015), an increased risk of subsequent dementia (Biondo et al., 2020), and a greater risk of mortality (Cole et al., 2018a). These findings support the use of MRI-derived brain-age measures as a

* Corresponding author.
E-mail addresses: thomasbooth@nhs.net, thomas.booth@kcl.ac.uk (T.C. Booth).
[1] Department of Neuroradiology, Ruskin Wing, King's College Hospital NHS Foundation Trust, London, SE5 9RS, UK

https://doi.org/10.1016/j.neuroimage.2022.118871.
Received 15 September 2021; Received in revised form 26 November 2021; Accepted 3 January 2022
Available online 5 January 2022.

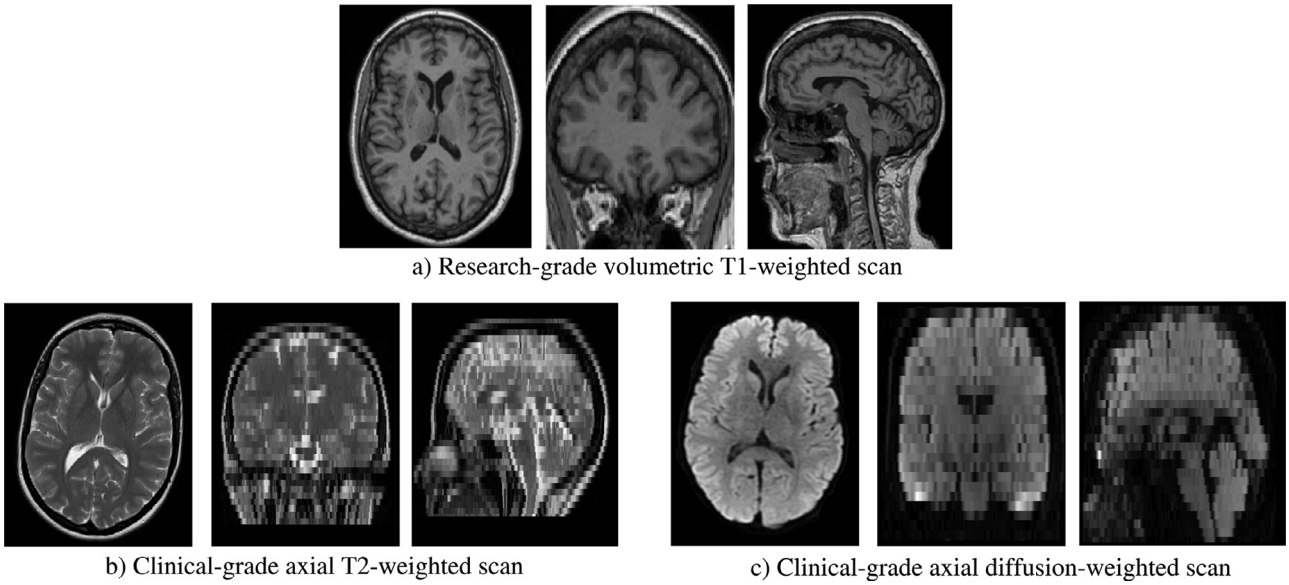

Fig. 1. Comparison of candidate structural MRI sequences for brain-age prediction. Previous studies have overwhelmingly used research-grade volumetric T1-weighted scans (a) for brain-age prediction. However, these scans are typically not part of routine hospital head MRI examinations; instead, axial T2-weighted (b) and axial diffusion-weighted (c) scans are typically acquired.

screening tool (opportunistically or otherwise) for these and other diseases to identify people at higher risk of poor health outcomes.

Potentially, brain-age could be predicted during routine clinical examinations to detect deviations from healthy ageing that may represent early-stage neurodegeneration. This in turn could help improve the patient pathway by expediting care (including intervention where available), and possibly accelerate the development of disease-modifying drugs through improved clinical trial enrolment. Brain-age could also feasibly guide adaptive MRI sequence acquisition (Cole et al., 2018b), enabling patients with older-appearing brains to undergo additional, more targeted, imaging (e.g., 'dementia protocol') whilst still in the scanner. To date, however, several key challenges have prevented clinical adoption:

1) Previous brain-age models have overwhelmingly used high-resolution volumetric T1-weighted scans, with isotropic or near-isotropic voxels, for brain-age prediction (Lemaitre et al., 2012)(Franke et al., 2010)(Franke and Gaser, 2012)(Gaser et al., 2013)(Koutsouleris et al., 2014)(Cole et al., 2017a)(Jónsson et al., 2019). However, these scans are typically not part of routine clinical examinations as they are time-consuming to acquire. Instead, anisotropic (e.g., low axial resolution) T2-weighted and diffusion-weighted images are considerably more common (ACR, 2019) (Fig. 1).
2) Most brain-age models have been trained and evaluated using scans obtained from curated, open-access research databases (e.g., OASIS, UK Biobank, IXI, Cam-CAN) which follow precise imaging protocols and participant inclusion criteria (Peterson et al., 2010). Hospital scans, by contrast, are more heterogeneous, with variable scanner vendors, imaging protocols, and patient populations between different sites likely to result in poorer model performance when applied in real-world clinical settings (Hosny et al., 2018)(Kocac et al., 2020)(Futoma et al., 2020).
3) Current state-of-the-art brain-age frameworks commonly rely on computationally expensive pre-processing such as bias-field correction, skull-stripping and spatial normalisation, which limits real-time clinical utility and introduces additional modelling assumptions.

A solution is to train brain-age models directly on 'raw' (i.e., minimally pre-processed, image-space) hospital head MRI scans, thereby ensuring that models are trained on data from a range of scanner vendors and acquisition protocols, drawn from a clinically-representative population. However, identifying sufficiently large training cohorts of patients with normal scans for deep learning is challenging, since archived hospital images are rarely stored with accompanying categorical labels (i.e., 'radiologically normal for age' or 'radiologically abnormal for age'). In recent years, however, breakthroughs in natural language processing (NLP) have made it feasible to derive accurate radiological labels from free-text radiology reports (Vaswani et al., 2017)(Devlin et al., 2019)(Wood et al., 2021b). This in turn enables the automatic grouping of large hospital databases of head MRI examinations according to these labels, facilitating downstream computer vision model development at scale (Wood et al., 2021a).

The purpose of this study was to build on these breakthroughs and develop a brain-age framework suitable for use during routine clinical head MRI examinations. We hypothesized that training at scale on clinically-representative data would result in generalisable models which are robust to variations in scanner vendors, imaging protocols and patient populations between different hospitals. We examined model generalisability by training and testing with different subsets of the available data from each participating hospital. In a separate experiment, we examined generalisability by training and testing with different subsets of available data from each scanner vendor. We also hypothesized that training at scale would obviate the need for pre-processing steps such as skull-stripping and spatial registration, since with sufficient training data a CNN should learn to focus on features relevant for brain-age prediction (i.e., brain parenchyma) and ignore irrelevant features such as extra-cranial tissue and absolute image position (LeCun et al., 2015); we sought to verify this by interrogating model decisions through interpretability methods. Finally, we hypothesized that deviation from our brain-age predictions could be used to detect atrophy in routine hospital examinations. We tested our model on scans of patients described independently by the radiologist during routine reporting as having atrophy 'excessive for age'; this way, we were able to evaluate brain-age predictions against expert radiological reports.

We focused primarily on brain-age prediction from axial T2-weighted scans as this sequence was performed in >90% of examinations in our UK NHS datasets. This is broadly similar to what is seen throughout the United States (ACR, 2019). The next most common sequence

Table 1
Dataset of scans reported as 'radiologically normal for age', used for brain-age model development.

Hospital	Axial T2-weighted scans	Axial DWI scans	Age (mean ± standard deviation)	Age range	Unique patients	Male/ female
KCH	9496	6592	43.7 ± 15.9	18 - 95	7425	4498/ 4998
GSTT	13,806	4806	43.4 ± 14.9	18 - 94	9419	6213/ 7593
Pooled	23,302	11,398	43.5 ± 15.3	18 - 95	16,844	10,711/ 12,591

was axial diffusion-weighted scans, performed in ~50% of examinations; therefore, as a secondary goal we investigated the use of axial diffusion-weighted scans for brain-age prediction. More advanced sequences (e.g. T1-weighted contrast enhanced and susceptibility-weighted scans, as well as volumetric scans) were obtained in under 10% of examinations and were therefore outside the scope of the current study.

2. Materials and methods

2.1. Datasets

All data were de-identified. The UK National Health Research Authority and Research Ethics Committee approved this retrospective study (IRAS ID 235,658, REC ID 18/YH/0458).

2.1.1. Hospital head MRI dataset

All 81,936 adult (≥ 18 years old) head MRI examinations performed at King's College Hospital NHS Foundation Trust (KCH) and Guy's and St Thomas' NHS Foundation Trust (GSTT) between 2008 and 2019 were obtained for this study. The MRI scans were performed on Signa 1.5 T HDX (General Electric Healthcare, Chicago, US), Aera 1.5 T (Siemens, Erlangen, Germany), Ingenia 1.5 T (Philips Healthcare, Eindhoven, Netherlands) or Skyra 3 T (Siemens, Erlangen, Germany) scanners. The text of the corresponding radiology reports produced by expert neuroradiologists (UK consultant grade; US attending equivalent) were extracted from the Computerised Radiology Information System (CRIS) (Healthcare Software Systems, Mansfield, UK). These reports were largely unstructured and typically comprised 5–10 sentences of image interpretation, along with comments regarding the patient's clinical history, and recommended actions for the referring doctor. A subset of these examinations was identified as 'radiologically normal for age' (Section 2.1.1.2) and included for model training and testing. A separate subset of examinations that were reported as having atrophy 'excessive for age' was also identified (section 2.1.1.3) and included for additional model testing.

2.1.1.1. 'Normal for age' cohort identification for brain-age model development. 'Radiologically normal for age' cohort identification was performed using a transformer-based neuroradiology report classifier (Wood et al., 2020a)(Wood et al., 2021b). This model was trained and tested using a dataset of 5000 neuroradiology reports from KCH which had been manually labelled by a team of 5 expert neuroradiologists (UK consultant grade; US attending equivalent) as either 'radiologically normal for age' or 'radiologically abnormal for age', following comprehensive pre-determined criteria (Wood et al., 2020b)(Wood et al., 2021b). The model achieved near-perfect classification performance on a hold-out set of 500 manually-annotated KCH radiology reports (AUC = 0.991) and generalised to an external hold-out test set of 500 reports from GSTT (AUC = 0.990, ΔAUC = 0.001). For further information about the development of this model, see (Wood et al., 2020a)(Wood et al., 2020b)(Wood et al., 2021b).

Once validated, the model was used to classify all 75,778 examinations from KCH and GSTT which included an axial T2-weighted scan (Fig 2); in total, 23,302 examinations were identified as 'radiologically normal for age' and included for brain-age model training and testing (male/female = 10,711/ 12,591, mean age = 43.5 ± 15.3, age range = 18–95) (Table 1). Further dataset information is provided in Appendix A.

2.1.1.2. Atrophy 'excessive for age' dataset identification. A subset of examinations reported as having atrophy 'excessive for age' was identified using the specialised radiology report classifiers described in (Wood et al., 2021b) (Fig. B1, Table B1 in Appendix B). Candidate examinations from the larger set of 52,476 'radiologically abnormal for age' examinations were identified using the 'atrophy excessive for age' versus 'no atrophy excessive for age' classifier; these examinations were then passed to six additional classifiers ('mass' versus 'no mass', 'moderate or severe small vessel disease' versus 'no or mild small vessel disease', 'vascular abnormality' versus 'no vascular abnormality', 'stroke' versus 'no stroke', 'white matter inflammation' versus 'no white matter inflammation', 'previous brain damage' versus 'no previous brain damage') to exclude patients with these common pathologies. Two neuroradiologists then interrogated the radiology reports of the resulting 281 examinations to exclude rare abnormalities, and determined that 228 described examinations where excessive atrophy was the only abnormal finding (interrater agreement = 100%; Fleiss' kappa = 1) (male/female = 127/ 101, mean age = 53.1 ± 14.9, age range = 19 - 88). The axial T2-weighted scans from these examinations were then included for model testing (Fig. 2). Importantly, this dataset was clinically-representative, comprising patients with subtle (e.g., described by the reporting radiologist as 'slightly excessive for age' or similar) as well as more conspicuous (e.g., described by the reporting radiologist as 'markedly excessive for age' or similar) neurodegeneration.

2.1.2. External test dataset

To facilitate further testing of our model, as well as enable direct comparison of model performance with previous and future studies, all axial T2-weighted scans from the Information eXtraction from Images (IXI) healthy subject dataset were obtained (n = 563, male/female = 250/ 313, mean age = 48.6 ± 16.5 years, age range = 20 – 86 years). The scans were acquired at three different London institutions between 2005 – 2008 (Hammersmith Hospital, using a Phillips 3T system; Guy's Hospital, using a Phillips 1.5T system; Institute of Psychiatry, using a GE 1.5T system), and can be downloaded from https://brain-development.org/ixi-dataset/.

2.1.3. Research examination dataset for T1-weighted and T2-weighted brain-age comparison

Because our hospital datasets contained few 'radiologically normal for age' examinations where volumetric T1-weighted scans were acquired, a separate dataset was required to facilitate a fair comparison between volumetric T1-weighted and axial T2-weighted brain-age prediction. To this end, all healthy research volunteer examinations performed at the Institute of Psychiatry, Psychology & Neuroscience, King's College London, between 2013 and 2019 were obtained. We identified a 'normal for age' subset of 2387 examinations where both axial T2-weighted scans and volumetric T1-weighted scans were available (mean age = 32.8 years ± 12.3 years, age range = 18 – 87 years), and included these for additional brain-age model training and testing. These MRI scans were performed on a Signa 3T Discovery MR750 (General Electric Healthcare).

2.2. Neuroimaging processing

Axial T2-weighted or axial diffusion-weighted scans of arbitrary resolution and dimensions, stored as Digital Imaging and Communications

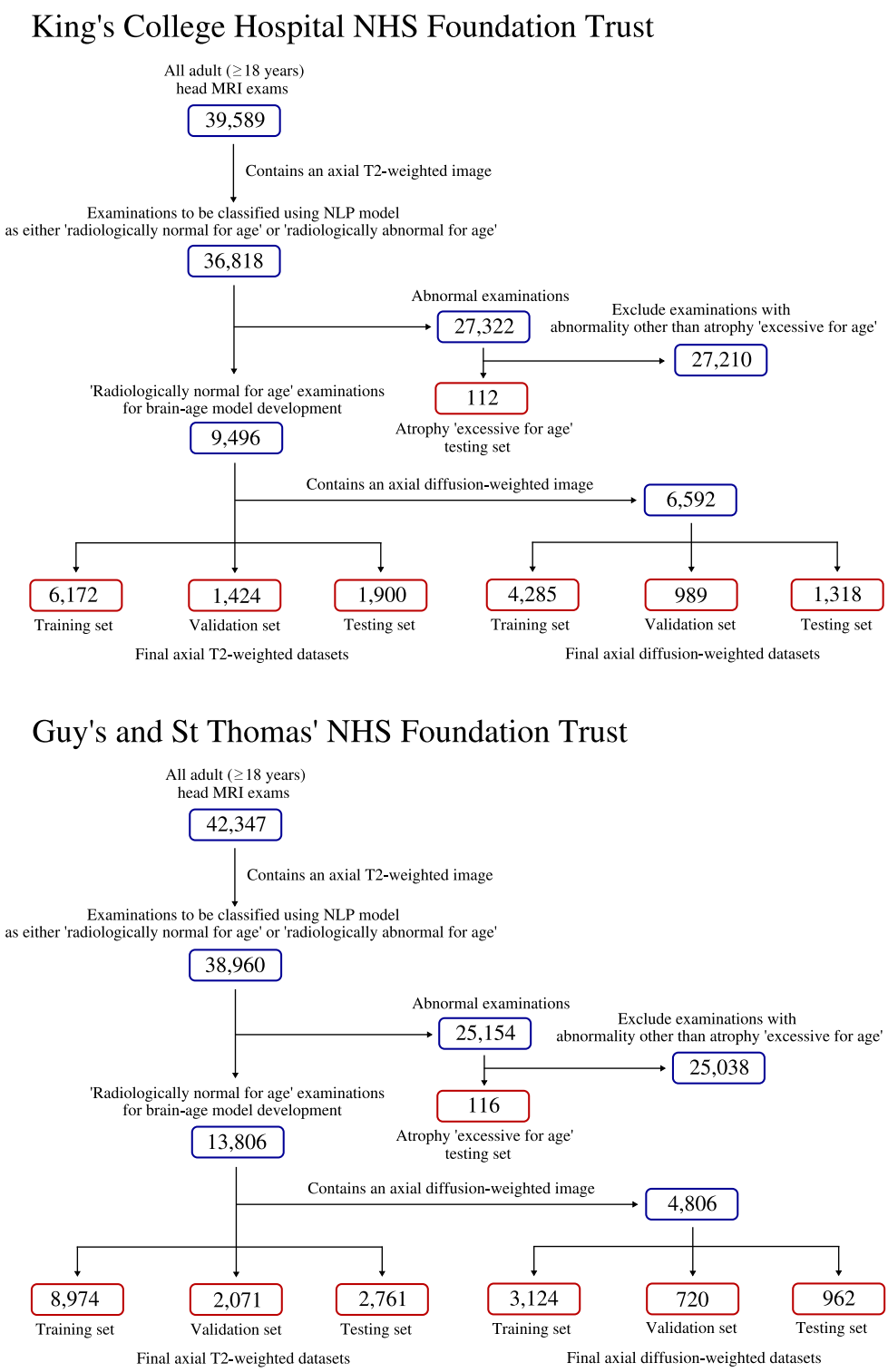

Fig. 2. Flow chart showing data sets used for training, validating, and testing our brain-age models. KCH (top), GSTT (bottom). To ensure that the training and test sets reflected the heterogeneity of examinations seen in routine clinical practice, no reported examinations were excluded on the basis of image quality.

in Medicine (DICOM) files, were converted into NIfTI format, resampled to common voxel sizes and dimensions (1mm^3), and then cropped or padded to a common array size (180 mm x 180 mm x 180 mm); this final step was necessary since CNNs require fixed-size images as input. The intensity of each image was normalised by subtracting the image mean and dividing by the image standard deviation. No spatial registration, bias field correction or skull-stripping was performed. All pre-processing was carried out using open-access python-based libraries: pydicom (Mason et al., 2020) was used to load DICOM files; dcm2niix (Li et al., 2016) was used to convert DICOM files to NifTI format; NiBabel (Brett et al., 2020) and numpy (Harris et al., 2020) were used to load and manipulate NifTI files; Project MONAI (MONAI, 2020) was used to resample, resize and normalise each image.

To compare brain-age prediction with and without skull-stripping, a separate dataset of skull-stripped scans was generated using HD-BET (Isensee et al., 2019), a deep learning-based PyTorch package, to re-

move non-brain tissue from all axial T2-weighted scans from GSTT (n scans = 13,806).[2] Further information about this open-access tool is available at https://github.com/MIC-DKFZ/HD-BET.

2.3. Brain-age modelling methods

Our brain-age models were based on the 'DenseNet121' architecture (Huang et al., 2017). DenseNet is a generalisation of the popular residual network ('ResNet') (He et al., 2016) which includes shortcut or 'skip' connections between internal neuron layers to overcome the 'vanishing gradient' problem (Pascanu et al., 2013) whereby earlier layers in a deep network fail to 'learn'. We elected to use an 'out-of-the-box' standard DenseNet121 configuration, rather than design a custom architecture, to ensure reproducibility and transparency of our framework. Specifically, our network consisted of an initial block of 64 convolutional filters (kernel size = [7 × 7 × 7], stride = 2) and a 'max-pooling' layer (kernel size = [3 × 3 × 3], stride = 3), followed by four 'densely connected' convolutional blocks. Each dense block consists of alternating point-wise (kernel size = [1 × 1 × 1]) and volumetric (kernel size = [3 × 3 × 3]) convolutions which are repeated 6, 12, 24 and 16 times in the four blocks, respectively. Between each dense block are 'transition layers' which consist of a point convolution (kernel size = [1 × 1 × 1]) and an average pooling layer (kernel size = [2 × 2 × 2], stride = 2). Global average pooling is applied to the output of the 4th dense block, resulting in a 1024-dimension feature vector which is converted by a fully-connected layer into a prediction for the patient's age (Fig. C1 in Appendix C).

Our DenseNet brain-age models were adapted from the implementation available at Project MONAI (https://docs.monai.io/en/latest/_modules/monai/networks/nets/densenet.html), and all modelling was performed with PyTorch 1.7.1 (Paszke et al., 2019) using two NVIDIA RTX 2080 11 GB graphics processing units (GPU). The Adam optimizer (Kingma and Ba, 2015) was used to update model weights during training, with the learning rate initially set to 1e-4 and reduced by a factor of 2 after every 5 epochs without validation loss improvement (i.e., learning rate scheduling). For all experiments, data were split into training (65%), validation (15%), and testing (20%) sets. This split was done at the patient level to prevent 'data leakage'. For each data split, model checkpoints were saved after each epoch, and the model with the lowest validation loss was used for testing. Mean absolute error (MAE) and Pearson's correlation were used to quantify model performance. Confidence intervals were generated by repeating this procedure 5 times for each model using different, randomly generated training/validation/testing splits. Paired *t*-testing was used to test the statistical significance of differences in computation time for the raw and skull-stripped models. Independent-sample *t*-testing was used to test the statistical significance of differences in brain-PAD between 'radiologically normal for age' and 'atrophy excessive for age' patients. 'Corrected paired t-testing' (Nadeau and Bengio, 2003) was used to test the statistical significance of differences in brain-age prediction between the raw and skull-stripped models, and between models using different MRI sequences. Scripts to enable readers to run our trained brain-age models using their own scans are available at https://github.com/MIDIconsortium/BrainAge .

2.4. Model interpretability

To scrutinise model predictions, guided backpropagation (Springenberg et al., 2015) and occlusion sensitivity analysis (Zeiler and Fergus, 2014) were performed. Briefly, guided backpropagation works by computing the derivative of the model predictions and 'back-propagating' this signal to the input image. In this way guided backpropagation highlights image regions which, if changed slightly, would alter the model's predictions. In contrast, occlusion sensitivity analysis involves 'masking out' a (e.g., cubic) region in an image and passing the 'occluded' image to a trained model. If the masked region contains features relevant for brain-age prediction, then the model's output is likely to differ from that generated for an unmasked image. By repeating this masking procedure at different locations in the image, a 'heatmap' of image regions which most influenced the model's predictions can be generated. For our experiments, we set the mask value to 0 (which corresponds to the mean image value after intensity normalisation), the mask size to 5 × 5 × 5, and the 'stride' (i.e., step size used to generate subsequent masks) to 3; these values represented a compromise between heatmap resolution and computational time.

[2] Experiments comparing brain-age prediction with and without skull-stripping were performed using data from a single hospital (GSTT) to limit computation time (skull-stripping required > 1 week per hospital dataset).

3. Results

3.1. Axial T2-weighted brain-age prediction

Accurate brain-age prediction (MAE = 2.97 years, 95% CI [2.94, 3.0], Pearson's correlation, $r = 0.972$ [0.970, 0.974]) was achieved using raw, clinical-grade axial T2-weighted scans pooled from both hospitals (n training = 15,146, n test = 4661) (Fig. 3a). Additional T2-weighted models generalised well between sites (Fig. 3b-d) and between scanner vendors (Fig. 4) (Δ MAE < 1.0 years).

By training additional models using different training data sample sizes, we observed that our brain-age framework is operating in an asymptotically optimal data regime (Fig. 5a-b); in other words, only minimal improvement can be expected by further increasing the training dataset size. We observed only minimal 'bias' (de Lange and Cole, 2020) in brain-age predictions (i.e., systematic overestimation and underestimation of age in younger and older subjects, respectively) (Pearson's correlation between brain-PAD and chronological age = −0.18) (Fig. 5c).

3.2. Skull-stripped versus 'raw' brain age prediction

Accurate brain-age prediction with and without skull-stripping was observed (raw MAE = 3.05 years [3.01, 3.09], skull-strip MAE = 3.65 years [3.60, 3.70], n training = 8974, n test = 2761). The difference in performance was significant ($p = 0.0002$). Guided backpropagation demonstrated that both models focus on similar regions for brain-age prediction (Pearson's correlation between raw and skull-stripped saliency maps aggregated across the entire test set ≥ 0.7); these primarily appear to be related to the cerebrospinal fluid spaces, such as the lateral ventricles (Fig. 6), in agreement with results derived using occlusion sensitivity analysis (Fig. E1, Appendix E). Computation time was significantly faster using raw scans (pre-processing + prediction time = 4.6 ± 0.8 s) compared with skull-stripped scans (pre-processing + prediction time = 48.9 ± 3.2 s) ($p < 0.0001$).

All results from Section 3.1 and 3.2 are summarised in Table 2:

3.3. Brain-age prediction with scans reported as having atrophy 'excessive for age'

Next, we tested the 228 patients reported as having atrophy 'excessive for age', using the axial T2-weighted model trained on scans pooled from both sites (n training = 15,146). These patients had systematically higher brain-predicted age than chronological age (atrophy 'excessive for age' mean brain-PAD = +5.89 years [5.21, 6.57], 'radiologically normal for age' mean brain-PAD = +0.05 years [−0.04, 0.14]) ($p < 0.0001$) (Fig. 7). Visualisations of subjects from the 'atrophy excessive for age' test set, including model predictions and saliency maps, are provided in Appendix F.

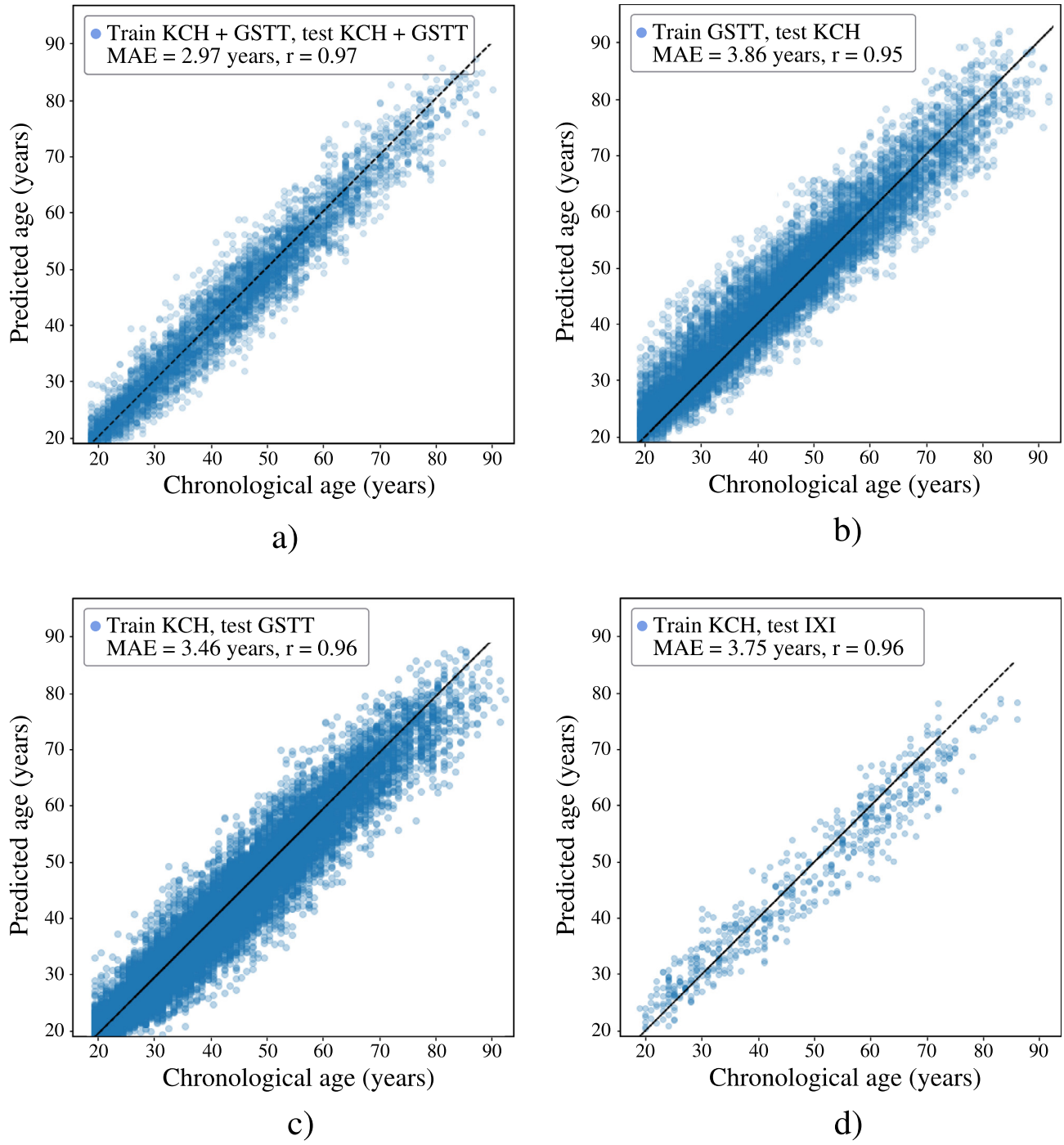

Fig. 3. Brain-age prediction using clinical-grade axial T2-weighted MRI scans. a) Highest accuracy was observed when training and testing using scans pooled from both KCH and GSTT (n training = 15,146, n test = 4661, MAE = 2.97 years 95% CI [2.94, 3.0], r = 0.972 [0.970, 0.974]). b) When trained on scans from GSTT, the model generalised to scans from KCH (MAE = 3.86 years [3.82, 3.90], r = 0.954 [0.951, 0.957], Δ MAE = 0.81 years). c) When trained on scans from KCH, the model generalised to scans from GSTT (MAE = 3.46 years [3.41, 3.51], r = 0.962 [0.959, 0.965], Δ MAE = 0.34 years). d) Generalisability to scans from the IXI dataset was also observed (MAE = 3.75 years [3.70, 3.80], r = 0.961 [0.958, 0.964], Δ MAE = 0.63 years).

3.4. Performance comparison with other MRI sequences

3.4.1. Axial diffusion-weighted scans

Using the subset of examinations from KCH and GSTT which included both an axial T2-weighted scan and an axial diffusion-weighted scan, additional models were trained (Table 3). Accurate brain-age prediction was achieved when training and testing using raw, clinical-grade axial diffusion-weighted (DWI) scans pooled from both sites (MAE = 3.98 years [3.93, 4.03], r = 0.944 [0.938, 0.950], n training = 7409, n test = 2280) (Fig. 8), although this was less accurate than for a model trained and tested on axial T2-weighted images alone from the same subset of examinations (MAE = 3.32 years [3.28, 3.36], r = 0.964 [0.961, 0.967]) ($p < 0.0001$). Averaging the predictions of the T2-weighted and diffusion-weighted models led to no statistically significant improvement over the axial T2-weighted predictions alone (ensemble MAE = 3.31 years [3.27, 3.35], r = 0.964 [0.960, 0.968], Δ MAE = −0.01 years, p = 0.41). We observed poor generalisability between these MRI modalities; a large drop in performance (Δ MAE = 6.51

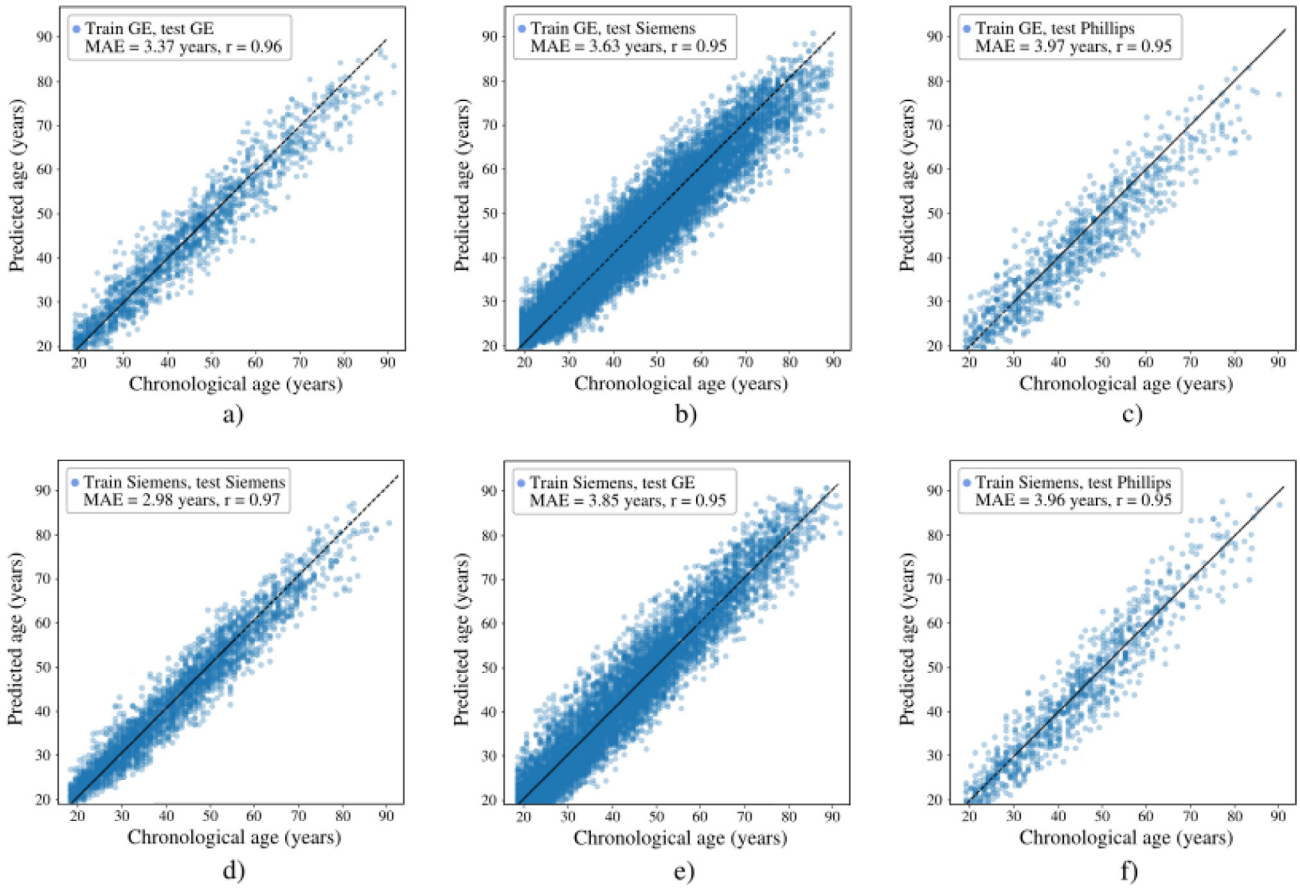

Fig. 4. Generalisability of brain-age predictions between scanner vendors. Top – when trained on scans obtained on GE scanners, the model generalised to scans obtained on Siemens scanners (MAE = 3.63 years 95% CI [3.58, 3.68], r = 0.953 [0.948, 0.958]) and Phillips scanners (MAE = 3.97 years [3.93, 4.01], r = 0.949 [0.945, 0.953]). Bottom - when trained on scans obtained on Siemens scanners, the model generalised to scans obtained on GE scanners (MAE = 3.85 years [3.81, 3.89], r = 0.954 [0.951, 0.957]) and Phillips scanners (MAE = 3.96 years [3.92, 4.00], r = 0.952 [0.948, 0.956]). Training on Phillips scanners was not performed due to the small number of these scans (n = 923) (see Fig. 5a-b).

Table 2
Brain-age results using clinical-grade axial T2-weighted MRI scans, including generalisability between hospital networks and scanner vendors, and a comparison of model performance with and without skull-stripping.

Training dataset	Test dataset	MAE (years) [95% CI]	Pearson's correlation [95% CI]
KCH + GSTT (n = 15,146)	KCH + GSTT (n = 4661)	2.97 [2.94, 3.0]	0.972 [0.970, 0.974]
KCH (n = 6172)	KCH (n = 1900)	3.12 [3.08, 3.16]	0.964 [0.961, 0.967]
KCH (n = 6172)	GSTT (n = 13,806)	3.46 [3.41, 3.51]	0.962 [0.959, 0.965]
KCH (n = 6172)	IXI (n = 563)	3.75 [3.70, 3.80]	0.961 [0.958, 0.964]
GSTT (n = 8974)	GSTT (n = 2761)	3.05 [3.01, 3.09]	0.964 [0.961, 0.967]
GSTT (n = 8974)	KCH (n = 9496)	3.86 [3.82, 3.90]	0.954 [0.951, 0.957]
GSTT (skull-stripped) (n = 8974)	GSTT (skull-stripped) (n = 2761)	3.65 [3.60, 3.70]	0.961 [0.957, 0.965]
GE scanner (n = 5445)	GE scanner (n = 1675)	3.37 [3.32, 3.42]	0.960 [0.956, 0.964]
GE scanner (n = 5445)	Siemens scanner (n = 14,002)	3.63 [3.58, 3.68]	0.953 [0.948, 0.958]
GE scanner (n = 5445)	Phillips scanner (n = 923)	3.97 [3.93, 4.01]	0.949 [0.945, 0.953]
Siemens scanner (n = 9101)	Siemens scanner (n = 2800)	2.98 [2.95, 3.01]	0.968 [0.965, 0.971]
Siemens scanner (n = 9101)	GE scanner (n = 8377)	3.85 [3.81, 3.89]	0.954 [0.951, 0.957]
Siemens scanner (n = 9101)	Phillips scanner (n = 923)	3.96 [3.92, 4.00]	0.952 [0.948, 0.956]

Note that training on GSTT and testing on IXI wasn't performed since IXI contains some data obtained from Guy's hospital. This ensured that the IXI dataset was a truly external dataset.

Table 3
Comparison of brain-age modelling using axial T2-weighted and axial diffusion-weighted scans, including generalisability between these modalities.

Modality	Training dataset	Test dataset	MAE (years) [95% CI]	Pearson's correlation (r) [95% CI]
Axial diffusion-weighted	KCH + GSTT (n = 7409)	KCH + GSTT (n = 2280)	3.98 [3.93, 4.03]	0.944 [0.938, 0.950]
	KCH + GSTT (n = 7409)	KCH + GSTT axial T2-weighted (n = 2280)	10.49 [9.25, 11.73]	0.608 [0.558, 0.658]
Axial T2-weighted	KCH + GSTT (n = 7409)	KCH + GSTT (n = 2280)	3.32 [3.28, 3.36]	0.964 [0.961, 0.967]
	KCH + GSTT (n = 7409)	KCH + GSTT axial diffusion-weighted (n = 2280)	14.84 [12.49,17.19]	0.176 [0.106, 0.246]
Ensemble model (axial T2-weighted + axial diffusion-weighted)	KCH + GSTT (n = 7409)	KCH + GSTT (n = 2280)	3.31 [3.27, 3.35]	0.964 [0.960, 0.968]

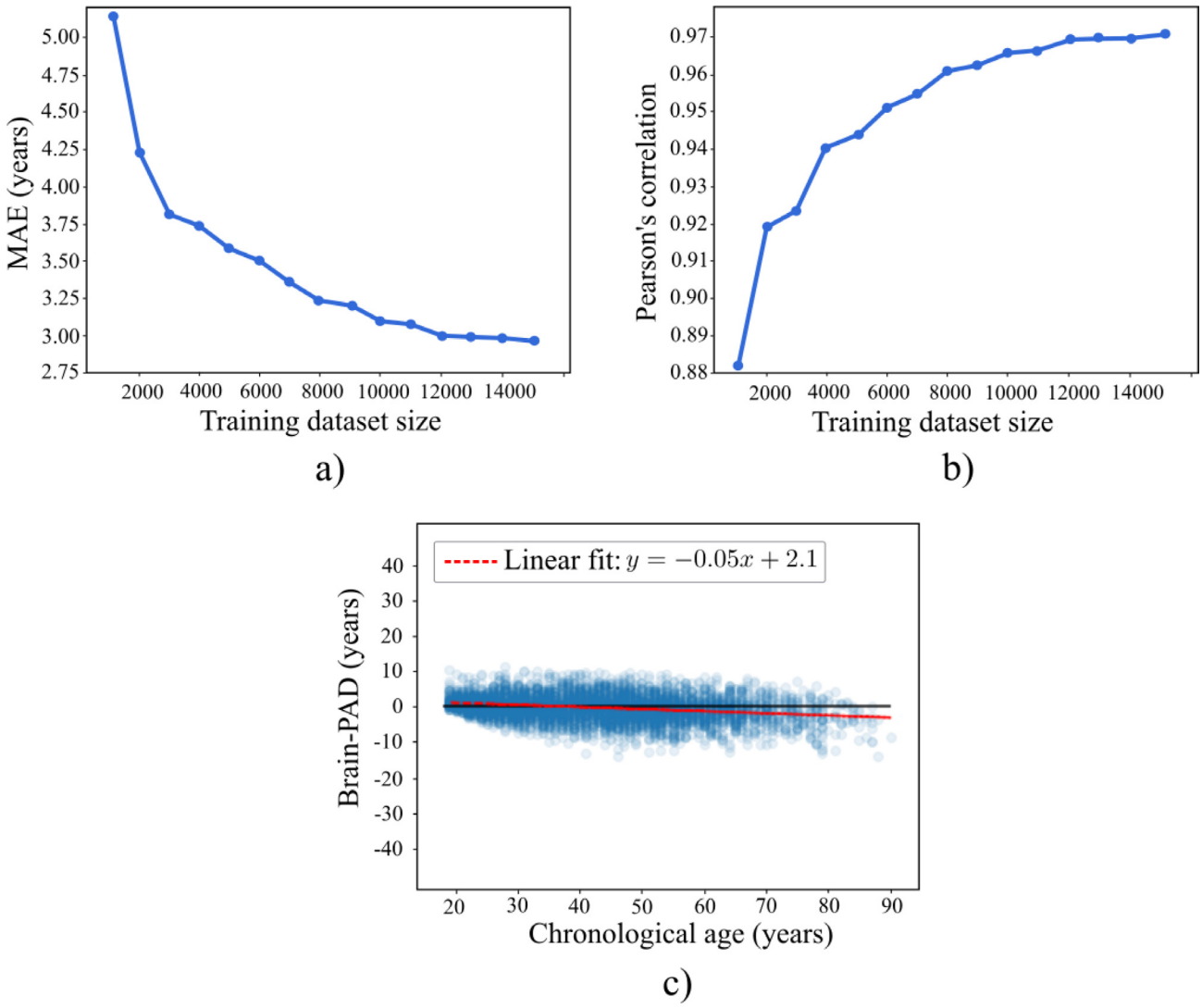

Fig. 5. Dataset size ablation study and brain-age 'prediction bias' analysis. a) Test set MAE as a function of training dataset size; rapid improvement (decreased MAE) with increasing dataset size is observed between 2000 - 12,000 scans, beyond which only modest improvement is seen. b) Test set Pearson's correlation as a function of training dataset size; a similar relationship to that seen for MAE is observed. c) Minimal 'bias' in brain-age prediction is seen (Pearson's correlation between brain-PAD and chronological age = −0.18).

years) was observed when applying the diffusion-weighted model to T2-weighted scans (Fig. 8a). Likewise, a large drop (Δ MAE = 11.52 years) was observed when applying the T2-weighted model to diffusion-weighted-weighted scans (Fig. 8b).

3.4.2. *Volumetric T1-weighted scans*

To facilitate direct comparison of axial T2-weighted and volumetric T1-weighted brain-age prediction, additional models were trained and validated (Fig. 9, Table 4) on a separate dataset of 2387 research examinations which included both sequences (mean age = 32.8 years ± 12.3 years, age range = 18 – 87 years). A separate dataset was required because at our institutions these sequences are rarely acquired together during routine examinations. Axial T2-weighted model performance on this dataset (MAE = 3.83 years [3.69, 3.97], $r = 0.950$ [0.943, 0.957]) was comparable to that of a pre-processed (skull-stripped, registered to MNI152 template) volumetric T1-weighted model (MAE = 3.86 years [3.67, 4.05], $r = 0.949$ [0.940, 0.958], $p = 0.43$), and better than a raw volumetric T1-weighted model (MAE = 4.86 years [4.64, 5.08], $r = 0.908$ [0.900, 0.916]) ($p = 0.002$). Poor generalisability between these MRI modalities was observed (Δ MAE > 7 years). Notably, an 'ensemble' model which averages the predictions of the pre-processed volumetric T1-weighted and axial T2-weighted models outperformed each individual single-sequence model (MAE = 3.35 years [3.20, 3.50], $r = 0.960$ [0.952, 0.968], $p = 0.02$), suggesting that these sequences provide complimentary information relevant to brain-age prediction.

4. Discussion

In this study we have demonstrated accurate brain-age prediction from clinical-grade, minimally processed, axial T2-weighted and axial diffusion-weighted scans. Our models generalise well between hospital trusts and scanner vendors, and show sensitivity to atrophy 'excessive for age'. Taken together, our brain-age framework shows feasibility for use as a screening tool during routine hospital examinations to automatically detect potentially pathological brain atrophy, with important implications for patient care, drug development, and adaptive MRI sequence acquisition.

To the best of our knowledge, this is the first study to present an accurate, generalizable 3D brain-age framework for use with axial T2-weighted scans. This is important because axial T2-weighted scans are typically the most commonly acquired sequence in clinical settings. At

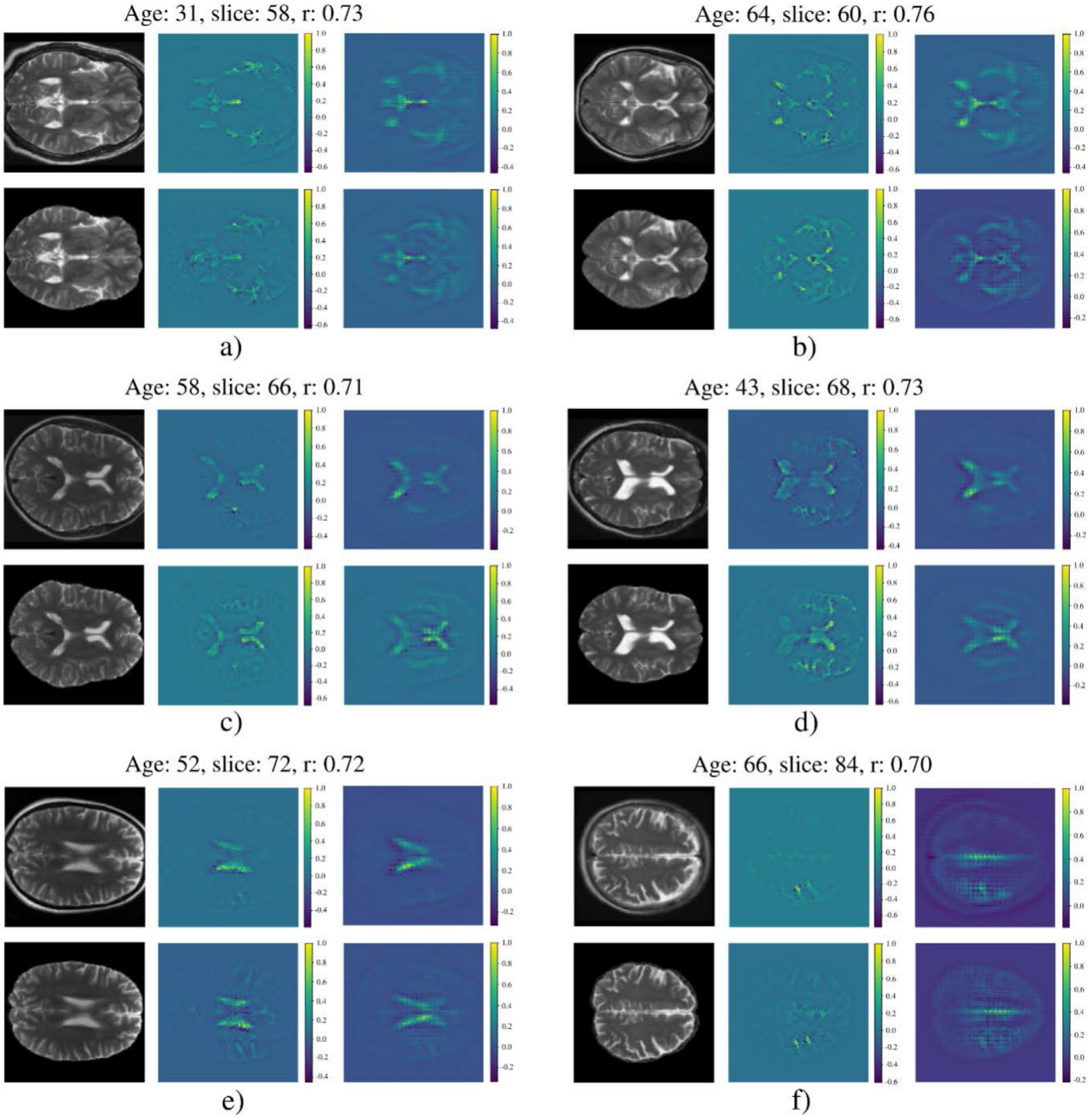

Fig. 6. Guided backpropagation analysis to interrogate brain regions influencing model predictions. Six representative subjects (a-f) from the test set are shown; in each panel the axial T2-weighted scan (left) and corresponding saliency map (middle) for that subject, along with the saliency map aggregated across the entire test set ($n = 2761$) for the same slice (right), are shown for the raw (top) and skull-stripped (bottom) models. Visually, the raw and skull-stripped models seem to focus on similar regions when predicting age, which primarily appear to be related to cerebrospinal fluid spaces. This is confirmed quantitatively by the high Pearson's correlation ($r \geq 0.7$) between aggregated saliency maps.

KCH and GSTT (two large and representative NHS hospital trusts), for example, axial T2-weighted scans are acquired during almost all head MRI examinations, both routine examinations and more targeted protocols, such as those for tumours, stroke or epilepsy. Axial diffusion-weighted scans are also commonly acquired during routine examinations. To the best of our knowledge, this is also first study to demonstrate accurate brain-age prediction using these scans. In contrast, previous brain-age studies have overwhelmingly used volumetric T1-weighted scans which, despite being commonly performed in research studies, are not part of most clinical head MRI examinations.

A notable exception is the recent study of (Hwang et al., 2021), which presented a 2D CNN-based brain-age model for use with axial T2-weighted scans. In that study, image-level brain-age prediction was achieved by averaging the predictions across all slices. However, two key limitations to this approach can be identified, and these likely contributed to the poorer generalisability of this framework (Δ MAE > 5

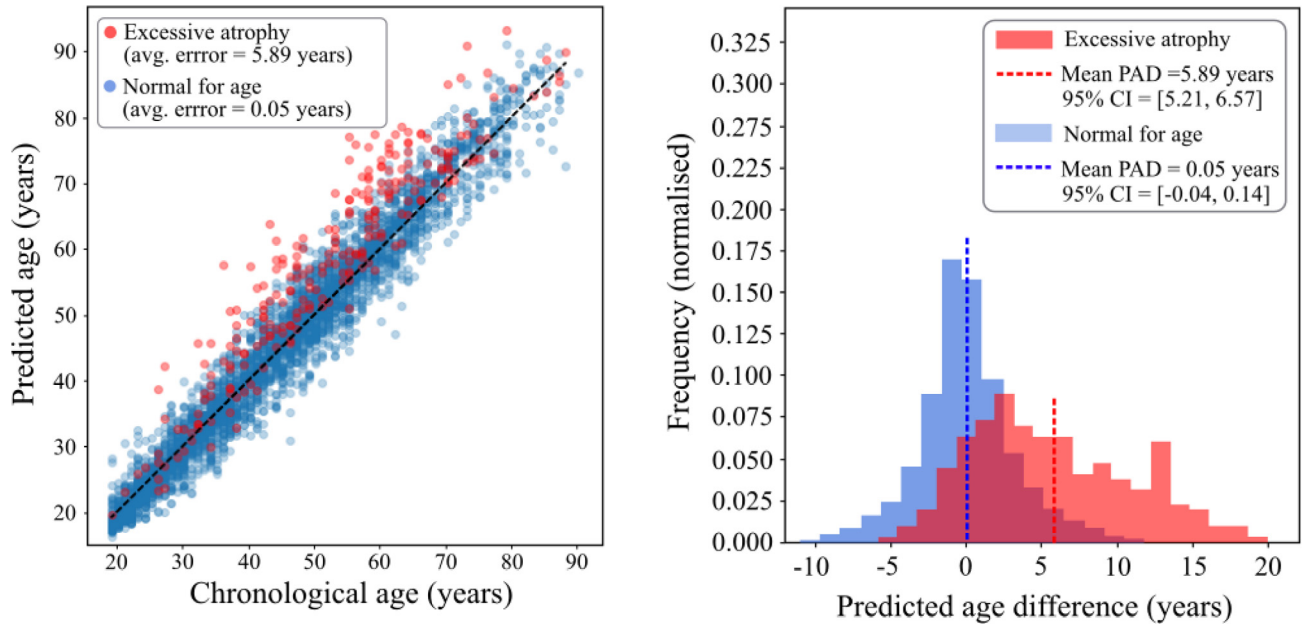

Fig. 7. Our axial T2-weighted model trained on scans from KCH and GSTT systematically predicted higher than chronological age for patients reported as having atrophy 'excessive for age' at these two hospital networks. Left: scatter plot of brain-age predictions for 'radiologically normal for age' (blue) and atrophy 'excessive for age' (red) patients. Right: Histogram of 'brain predicted age difference' (brain-PAD), generated by subtracting chronological age from predicted age, for 'radiologically normal for age' (blue) and atrophy 'excessive for age' (red) patients. Dotted lines represent the mean brain-PAD; 95% confidence intervals are also provided.

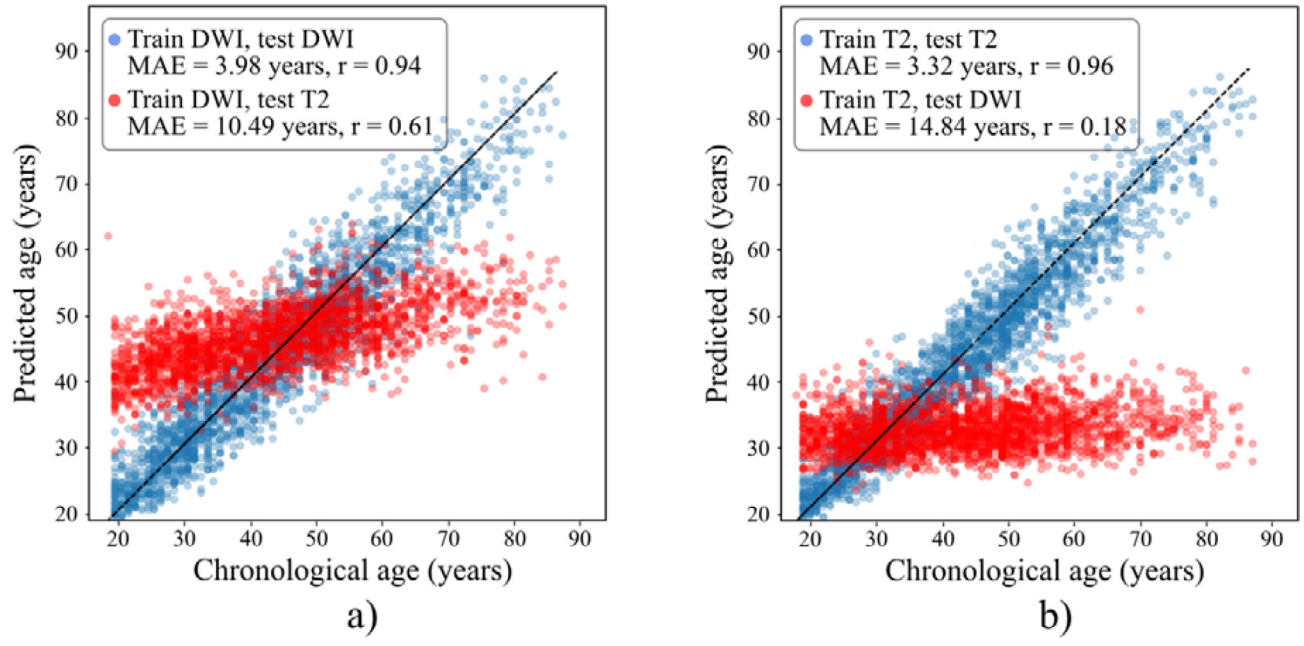

Fig. 8. (a) Accurate brain-age prediction was achieved using clinical-grade axial diffusion-weighted scans (MAE = 3.98 years, $r = 0.94$), although this was less accurate than an axial T2-weighted model (b) trained and tested on the same subset of examinations ($p < 0.0001$). Poor generalisability between modalities was observed (Δ MAE > 6 years).

years when tested on external data). First, those slices which do not contain any brain tissue - and therefore should *not* influence brain-age (e.g., the most inferior slice in an image which often shows the jaw and neck, or the most superior slice which is often outside the head) - contribute just as much to the final predicted age as do slices through the centre of the brain. Second, each slice is treated independently of all others; this precludes modelling non-linear interactions between axially-separated features (e.g., atrophic changes which have a significant axial extent, but are subtle within any given slice). In contrast, our 3D models are naturally able to model interactions between slices through the use of 'depth-wise convolutions' and have the flexibility to ignore irrelevant features for brain-age prediction (e.g., jaw, neck etc.). A quantitative comparison of our model with that of Hwang et al. is provided in Appendix D.

Our study has a number of additional strengths. Firstly, using a state-of-the-art neuroradiology report classifier, we have been able to generate a large, clinically-representative dataset for model training, overcoming a critical bottleneck. This is important as it ensured that our deep-leaning models could be trained on data from a range of scanner vendors and acquisition protocols, drawn from a clinically-

Table 4
Comparison of volumetric T1-weighted and axial T2-weighted brain-age prediction using a dataset of research examinations which included both sequences.

Modality	Training dataset	Test dataset	MAE (years) [95% CI]	Pearson's correlation (r) [95% CI]
Volumetric T1-weighted	IoPPN (n = 1551)	IoPPN (n = 477)	4.86 [4.64,5.08]	0.908 [0.900, 0.916]
	IoPPN (n = 1551)	IoPPN (Axial T2-weighted) (n = 477)	12.40 [9.6, 15.2]	0.266 [0.257, 0.275]
	IoPPN (skull-stripped and registered to MNI152) (n = 1551)	IoPPN (skull-stripped + registered to MNI152) (n = 477)	3.86 [3.67, 4.05]	0.949 [0.940, 0.958]
Axial T2-weighted	IoPPN (n = 1551)	IoPPN (n = 477)	3.83 [3.69,3.97]	0.950 [0.943, 0.957]
	IoPPN (n = 1551)	IoPPN (Volumetric T1-weighted) (n = 477)	13.90 [11.77, 16.03]	0.224 [0.084, 0.364]
Ensemble model (pre-processed volumetric T1-weighted + axial T2-weighted)	IoPPN (n = 1551)	IoPPN (n = 477)	3.35 [3.20, 3.50]	0.960 [0.952, 0.968]

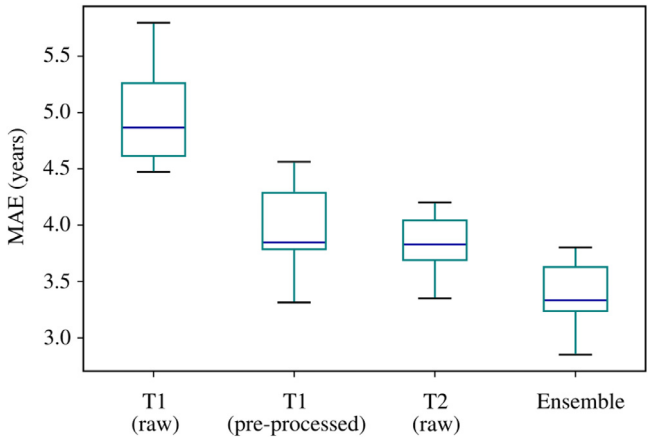

Fig. 9. Comparison of volumetric T1-weighted and axial T2-weighted brain-age prediction using a dataset of research examinations which included both sequences. Axial T2-weighted brain-age prediction (MAE = 3.83 years 95% CI [3.69, 3.97], r = 0.950 [0.943, 0.957]) was comparable to that using pre-processed (skull-stripped, registered to MNI 152 template) volumetric T1-weighted scans (MAE = 3.86 years [3.67, 4.05], r = 0.949 [0.940, 0.958], p = 0.43). An ensemble model which averaged the predictions of the raw T2-weighted and pre-processed volumetric T1-weighted models outperformed either model individually (MAE = 3.35 years [3.20, 3.50], r = 0.960 [0.952, 0.968], p = 0.02). Box and whisker distributions were generated by repeating the training/testing procedure for 10 different training/validation/testing splits.

representative population covering the full adult lifespan (18–95 years). This in turn enabled our models to generalise to out-of-sample data (e.g., scans from external hospitals and different scanner vendors). Generalisability beyond the training dataset is particularly important when using brain-predicted age for screening purposes, because in this case model 'errors' (i.e., disagreement between predicted and chronological age) are important. A brain-age model which has been trained on unrepresentative data may make an erroneously high prediction of age, not because the patient has an older-appearing brain, but because the scan is an outlier relative to the training distribution (e.g., due to differences in magnetic field strength and homogeneity, image resolution, or presence of motion or other artefacts).

Training at scale also enabled our models to *automatically* learn to focus on features relevant for brain-age prediction (i.e., brain parenchyma) and ignore irrelevant features such as extra-cranial tissue and absolute image position. We have therefore been able to avoid those additional pre-processing steps (e.g., stripping non-brain tissue and registering images to a common space) which are typically performed to mitigate overfitting to small training datasets. An important consequence of avoiding image pre-processing is that faster brain-age prediction can be achieved. Pre-processing can be computationally expensive, and due to the possibility of errors, requires manual quality assurance. Our framework is able to load axial T2-weighted or diffusion-weighted scans of arbitrary resolution and dimensions, stored as DICOM files, and ultimately return a brain-age prediction in under 5 s. This opens up a range of clinical applications, including real-time adaptive sequence acquisition, whereby patients with older-appearing brains are automatically transferred to a more targeted imaging protocol (e.g., 'dementia protocol') whilst still in the scanner.

Importantly, by comparing volumetric T1-weighted and axial T2-weighted brain-age prediction using a dataset of healthy participants for which both scans were performed during the same imaging session, we have demonstrated that brain-age prediction using raw, clinical-grade axial T2-weighted scans is comparable to that using volumetric T1-weighted scans. Notably, however, an 'ensemble model' which averages the predictions of the two models significantly outperformed both single-sequence models; this suggests a possible approach to improve brain-age prediction during 'dementia protocol' imaging sessions, since these examinations include both axial T2-weighted and volumetric T1-weighted scans.

Our study builds on recent transformative developments in the field of NLP. Until recently, automatic text-classification-based generation of a training cohort was infeasible, due in large part to the lexical complexity of neuroradiology reports. In the last two years, however, the development and open-source release of state-of-the-art 'transformer'-based language models - which have been pre-trained on huge collections of unlabelled text (e.g., all of English Wikipedia, all PubMed Central abstracts and articles etc.) - has made it feasible to derive accurate categorical radiological labels from radiology reports (e.g., 'radiologically normal for age', 'radiologically abnormal for age' etc.). In this way, large hospital databases of head MRI examinations can be automatically grouped according to these labels, facilitating downstream computer vision model development at scale (Wood et al., 2021a). To enable readers to generate large training datasets for brain-age model development using data from their own institution, we have made our neuroradiology report classifier training scripts, as well as a dedicated labelling 'app', available at https://github.com/MIDIconsortium/RadReports.

There are some limitations of our study to consider. First, although our results support the use of brain-age in clinical contexts to detect people with excessive atrophy who might be at risk of neurodegenerative disease and poor cognitive ageing, it is currently unclear how the model would perform in individuals with gross abnormalities, since the model was trained on radiologically 'normal for age' brains. Potentially, tumours and large strokes, for example, may be too far from the learned manifold/latent space to make the brain-age outputs meaningful. If excessive atrophy occurs alongside other abnormalities, however, the latter are typically the focus of clinical management and excessive atrophy is typically of minimal clinical relevance; therefore, this limitation may not be an issue in practice. Second, although we have used interpretability methods to confirm that our models focus on brain parenchyma for brain-age prediction and not, say, skull or other non-brain tissue, we have not systematically analysed which brain features are important for brain-age prediction. Such analyses could provide important informa-

tion about the ageing process, and as further work we plan to investigate this further.

Conclusions

In conclusion, we have demonstrated a framework that combines large hospital databases, NLP, and 3D CNNs, to deliver fast, accurate and radiologically-relevant brain-age predictions from minimally processed, clinical-grade structural MRI scans. This demonstrates the feasibility of using the brain-age paradigm for automatically detecting neurodegeneration during clinical examinations. Moreover, our framework could be used to leverage the wealth of existing large hospital databases to provide powerful new resources for the training, testing, and clinical validation of medical image analysis tools beyond brain-age.

Declaration of Competing Interest

Co-author Sebastien Ourselin is the co-founder of Brainminer; however, he did not control or analyse the data. The other authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Acknowledgements

We thank Joe Harper, Justin Sutton, Mark Allin, and Sean Hannah at KCH for their informatics and IT support, Ann-Marie Murtagh at KHP for research process support and KCL administrative support, particularly from Alima Rahman, Denise Barton, John Bingham, and Patrick Wong.

Funding

This work was supported by the Royal College of Radiologists, King's College Hospital Research and Innovation, King's Health Partners Challenge Fund, NVIDIA (through the unrestricted use of a GPU obtained in a competition), and the Wellcome/Engineering and Physical Sciences Research Council Centre for Medical Engineering (WT 203148/Z/16/Z).

Data and code availability

Scripts to enable readers to run our trained brain-age models using their own scans are available at https://github.com/MIDIconsortium/BrainAge. A dedicated 'labelling 'app' is made available at https://github.com/MIDIconsortium/RadReports to enable readers to label their own neuroradiology report datasets for natural language processing model development.

The datasets used in this study are not publicly available because the IRB of the study limits access to the data. Derived and supporting data are available from the corresponding author upon reasonable request.

Author contribution

David A. Wood: Methodology, Software, Formal analysis, Writing – original draft preparation, Sina Kafiabadi: Data curation, Validation, Ayisha Al Busaidi: Data curation, Validation, Emily Guilhem: Data curation, Validation, Antanas Montvila: Data curation, Validation, Jeremy Lynch: Data curation, Validation. Conceptualization, Matthew Townend: Software, Siddharth Agarwal: Validation, Writing – review and editing, Asif Mazumder: Data curation, Validation Gareth J. Barker: Writing – review and editing, Sebastian Ourselin: project administration, Funding acquisition, James H. Cole: Conceptualization, Supervision, Funding acquisition, Writing – review and editing, Thomas C. Booth: Conceptualization, Supervision, Funding acquisition, Writing – review and editing

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.118871.

References

ACR. ACR–ASNR–SPR practice parameter for the performance and interpretation of magnetic resonance imaging (MRI) of the brain. https://www.acr.org/-/media/ACR/Files/Practice-Parameters/MR-Brain.pdf. Published 2019. 2020-02-01.

Brett, M., Markiewicz, C., Hanke, M., Cote, M., Cipollini, B., (2020). Nipy/nibabel: 3.2.1 (Version 3.2.1). Zenodo. http://doi.org/10.5281/zenodo.4295521

Cole, J.H., Leech, R., Sharp, D.J.Alzheimer's Disease Neuroimaging Initiative, 2015. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. Ann. Neurol. 77 (4), 571–581.

Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., 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.

Biondo, F., Jewell, A., Pritchard, M., et al., 2020. Brain-age predicts subsequent dementia in memory clinic patients. Alzheimer's & Dementia 16. doi:10.1002/alz.037378.

Cole, J.H., Franke, K., 2017b. Predicting age using neuroimaging: innovative brain ageing biomarkers. Trends Neurosci. 40 (12), 681–690.

Cole, J.H., Ritchie, S.J., Bastin, M.E., Hernández, M.V., Maniega, S.M., Royle, N., Deary, I.J., 2018a. Brain age predicts mortality. Mol. Psychiatry 23 (5), 1385–1392.

Cole, J.H., Lorenz, R., Geranmayeh, F., Wood, T., Hellyer, P., Williams, S., & Leech, R. (2018b). Active Acquisition for multimodal neuroimaging. https://wellcomeopenresearch.org/articles/3-145.

de Lange, A.M.G., Cole, J.H., 2020. Commentary: correction procedures in brain-age prediction. NeuroImage: Clin. 26.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C.Alzheimer's Disease Neuroimaging Initiative, 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.

Franke, K., Gaser, C., 2012. Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. GeroPsych: J. Gerontopsychol. Geriatr. Psychiatry 25 (4), 235.

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., Celi, L.A., 2020. The myth of generalisability in clinical research and machine learning in health care. Lancet Digital Health 2 (9), e489–e492.

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.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585 (7825), 357–362.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J., 2018. Artificial intelligence in radiology. Nat. Rev. Cancer 18 (8), 500–510.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Hwang, I., Yeon, E.K., Lee, J.Y., Yoo, R.E., Kang, K.M., Yun, T.J., Kim, J.H., 2021. Prediction of brain age from routine T2-weighted spin-echo brain magnetic resonance images with a deep convolutional neural network. Neurobiol. Aging 105, 78–85.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Kickingereder, P., 2019. Automated brain extraction of multisequence MRI using artificial neural networks. Hum. Brain Mapp. 40 (17), 4952–4964.

Jónsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., Ellingsen, L.M., Walters, G.B., Gudbjartsson, D.F., Ulfarsson, M.O., 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nat Commun 10 (1), 1–10.

Kingma, D.P., & Ba, J. (2015, January). Adam: a method for stochastic optimization. In *ICLR (Poster)*.

Kocak, B., Kus, E.A., Kilickesmez, O., 2020. How to read and review papers on machine learning and artificial intelligence in radiology: a survival guide to key methodological concepts. Eur. Radiol. 1–12.

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Meisenzahl, E., 2014. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophr. Bull. 40 (5), 1140–1153.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lemaitre, H., Goldman, A.L., Sambataro, F., Verchinski, B.A., Meyer-Lindenberg, A., Weinberger, D.R., Mattay, V.S., 2012. Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume? Neurobiol. Aging 33 (3), 617 -e1.

Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C., 2016. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J. Neurosci. Methods 264, 47–56. doi:10.1016/j.jneumeth.2016.03.001, PMID: 26945974.

Mason, D., Scaramallion, R., mrbean-bremen, Suever, J., et al., 2020. Pydicom/pydicom: pydicom 2.1.2 (Version v2.1.2). Zenodo http://doi.org/10/5281/zenodo.4313150.

MONAI. Project monai, 2020. URL http://doi.org/10.5281/zenodo.4323059.

Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Mach. Learn. 52 (3), 239–281.

Pardoe, H.R., Cole, J.H., Blackmon, K., Thesen, T., Kuzniecky, RHuman Epilepsy Project Investigators, 2017. Structural brain changes in medically refractory focal epilepsy resemble premature brain aging. Epilepsy Res. 133, 28–32.

Pascanu, R., Mikolov, T., Bengio, Y., 2013. On the difficulty of training recurrent neural networks. In: International conference on machine learning. PMLR, pp. 1310–1318.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S., 2019. PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037.

Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Weiner, M.W., 2010. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. Neurology 74 (3), 201–209.

Springenberg, J., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for simplicity: the all convolutional net. *arXiv preprint arXiv:1412.6806*.

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010.

Wood, D.A., Lynch, J., Kafiabadi, S., et al. , 2020a. Automated labelling using an attention model for radiology reports of MRI scans (ALARM). In: Proceedings of the Third Conference on Medical Imaging with Deep Learning, in PMLR, 121 arXiv preprint arXiv:2106.08176. url: https://arxiv.org/abs/2106.08176.

Wood, D.A., Kafiabadi, S., Al Busaidi, A., Guilhem, E., Lynch, J., Townend, M., Booth, T.C, 2020b. Labelling imaging datasets on the basis of neuroradiology reports: a validation study. In: Interpretable and Annotation-Efficient Learning for Medical Image Computing. Springer, Cham, pp. 254–265.

Wood, D.A., Kafiabadi, S., Al Busaidi, A., Guilhem, E.L., Lynch, J., Townend, M.K., Booth, T.C, 2021b. Deep learning to automate the labelling of head MRI datasets for computer vision applications. Eur. Radiol. 1–12.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer, Cham, pp. 818–833.

Wood, D.A., Kafiabadi, S., Busaidi, A.A., Guilhem, E., Montvila, A., Agarwal, S., & Booth, T.C. (2021a). Automated triaging of head MRI examinations using convolutional neural networks. *arXiv preprint arXiv:2106.08176*