SUPPLEMENTARY INFORMATION

Grey Matter Age Prediction as a Biomarker for Risk of Dementia

Content:

www.pnas.org/cgi/doi/10.1073/pnas.1902376116

- **Supplementary Figure 6.** Grad-CAM attention map intensity per voxel overlaid on a brain template.
- **Supplementary Figure 7.** Correlation between brain pathology associated features and the age gap.
-

***CORRESPONDENCE:**

G.V. Roshchupkin PhD, Department of Epidemiology, Department of Medical Informatics, Department of Radiology and Nuclear Medicine, Erasmus Medical Center

Dr. Molewaterplein 50, 3015 GE, Rotterdam, the Netherlands.

E-mail: g.roshchupkin@erasmusmc.nl

Supplementary methods 1: Study Population

The cohort started in January 1990 (n=7983) and was extended in February 2000 (n=3011) and February 2006 (n=3932). Follow-up examinations take place every 3 to 4 years. MRI was implemented in 2005, and 5912 persons scanned until 2015 were eligible for this study. We excluded individuals with incomplete acquisitions, scans with artifacts hampering automated processing, participants with MRI-defined cortical infarcts and participants with dementia or stroke at the time of scanning (**SI Appendix Fig. 1**). This resulted in 5656 subjects to be included in this study. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the Wet Bevolkingsonderzoek ERGO (Population Studies Act: Rotterdam Study). All participants provided written informed consent to participate in the study and to obtain information from their treating physicians. The scan protocol of the Rotterdam Study is carefully balanced between the restrictions of time, costs and inconvenience for the participants and the relevance and quality of the acquired imaging data, to ensure participant compliance and reproducible image quality (to reduce motion artefacts). Each MRI scan that is acquired is visually examined by a research physician in the Rotterdam Scan Study.

Supplementary methods 2: Measurements of characteristics

All participants were monitored for dementia at baseline and following visits to the study center using the Mini-Mental State Examination (MMSE) and the Geriatric Mental State (GMS) organic level. Further investigation was initiated for participants who scored lower than 26 for their MMSE or above 0 for their GMS[1]. Additionally, the entire cohort was continuously checked for dementia through electronic linkage between the study center and medical records from general practitioners and the regional institute for outpatient mental health care. Available information on cognitive testing and clinical neuroimaging was used when required for diagnosis of dementia subtype. Final diagnosis was established by a consensus panel led by a consultant neurologist, according to a standard criteria for dementia (using the Third Revised Edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III-R))[2,3]. Until January 1st 2016, 92% of the potential person-time follow-up was complete. Participants were censored at date of dementia diagnosis, death or loss to follow-up, or at January 1st 2016, whichever came first. Of 5496 subjects included in this analysis, 159 developed dementia within 10 years of follow-up (mean follow-up time 4.34±2.25 years).

Mild cognitive impairment (MCI) was assessed in individuals over the age of 60 years, for which both subjective and objective cognitive deficits were required. An objective cognitive deficit was based on a cut-off of 1.5 standard deviations below the Rotterdam Study age- and education-specific means in three cognitive domains, i.e. the memory, information processing speed and executive functioning domain. Subjective cognitive deficits were defined as having answered yes to any of six questions regarding difficulties in memory (difficulties finding words, or remembering plans) or daily functioning (difficulties managing finances, getting dressed, or using the phone).

71 Systolic and diastolic blood pressure was measured twice in the right arm in sitting position after
72 five minutes of rest, of which the average was used. Body mass index (BMI) was defined as
73 weight in kilograms (kg) divided by height in meters squared (m^2). Participants were asked by
74 interview whether they were a current or past smoker, which was used to define their smoking
75 status. Glucose, total cholesterol and HDL cholesterol were measured in blood of the fasting
76 state.

77

Supplementary methods 3: Deep learning and convolutional neural networks

Deep learning techniques require a set of input and respective output to find and optimize a non-linear relation between the two. By providing data to a set of algorithms, the method is able to train a by the user designed model. Generally, the user designs the model architecture by selecting the model components. Subsequently, the machine learning method iteratively adjusts the model parameters according to that iteration's trained model performance, to create an optimized model using backpropagation by supervised or unsupervised learning[4,5]. By letting the model itself choose which relevant features to extract from the input, deep learning facilitates the model to freely search the input-space and find the most important, possibly new, input features.

Convolutional neural networks (CNNs) are a subset amongst deep learning techniques. They allow multi-dimensional input images and inspect these inputs by scanning them for relevant information[6,7]. Deep learning and CNN models have been rising in popularity and have been actively studied in recent years, reaching state-of-the-art performances in many applications amongst which medical imaging[8–10].

CNNs regard an image as a field of numerical values, view small portions of this image (receptive field) and perform multiplications with a weight-matrix (filter) to extract certain information (feature) from this portion. By inspecting the entire image using this filter in a grid-wise manner, the filter extracts specific information which is then saved to a new matrix or image (feature map). Repeating this process for the resulting feature maps, the network iteratively refines or searches for more information inside of the image that is relevant to the output.

These convolutional layers (CONV layers) are then typically combined with a variety of different techniques and algorithms that allow the network to appropriately extract the information from the input. Commonly used techniques are rectified linear units activation (ReLU), max-pooling layers (MP), fully connected layers (FC), batch normalization and dropout[7,10].

Supplementary methods 4: Network training

The CNN has been trained using the data from the training set of 3688 subjects. Here, over- and undersampling had been applied to the training set. Thus, effectively data of 3688 subjects was used out of 3848 available subjects available for the training set, to distribute the samples more evenly over the age range of the population ($N_{img,train\_balanced}$=8060 images, mean age 68.52±13.71sd). To avoid overfitting on the training set and to improve overall model performance, data augmentation was also applied during training[11]. Data augmentation included random small translations and mirroring in planes. We also used follow-up MRI scans of each subject as a 'natural data augmentation' technique.

The best model was selected based on its performance on the validation set. Here the performance is measured as the model accuracy based on the root mean squared error (RMSE) of the gap, as RMSE penalizes outliers more than MAE.

Supplementary methods 5: Attention mapping

Attention map intensity values were normalized to range 0-1, where 1 indicates the value for areas most associated with the network's decision. We expanded the Grad-CAM visualization technique to a 3-dimensional space.

Attention maps were computed for every individual. Since all brain images were registered to the same template space, a global average voxel-wise attention map could be made over attention maps of all subjects to obtain a global attention map for the age prediction network.

We computed the change in attention map over age per voxel, to investigate the change in regions predictive for brain age between age groups. To this end, for each voxel, a linear regression from age to attention map value was performed, resulting in a line of which the slope represents the increase in attention map value with age for the given voxel

Supplementary text 1: Sex covariate effect on CNN model performance

We can consider a split evaluation between male and female subjects. **Supplementary Figure 3** shows the network found no significant difference between the two groups (p=0.34). By including sex as a covariate, the covariate can reduce the difference in resulting age predictions between male and female subjects.

The trained model was able to reduce prediction error and correct for male and female biases observed in the image by the model. By including the additional input of sex, the model is able to prevent over- and under prediction for male and female ages, respectively, as shown in **Supplementary Figure 4**. Here we present the performance in gap on male and female subjects, both early adapted models were trained under the same training settings and used the exact same training and validation sample sets. The model that includes the additional input of the subject's respective sex, was able to reduce the overall gap between male and female subjects to be insignificant (p-value=0.23). This also brought the mean gap for males and females closer to zero (one-sample t-test: p_{male}=0.88 and p_{female}=0.05).

Given that the sex as covariates improved the model performance, we hypothesised that brain age prediction gap might have different predictive value for males and females. We did stratified analysis and have not found any differences between males (HR=1.16, 95% CI 1.09-1.24) and females (HR=1.14, 95% CI 1.09-1.20)

Supplementary text 2: Important regions attention map

Although aging affects the entire GM volume in the brain, as shown in **Supplementary Figure 5**, significant negative association between GM volume and age have been reported for several specific brain regions, i.e. a reduction in GM volume with age[12,13]. According to literature[12,13] the insula, superior temporal areas and multiple gyri have shown significant age-related GM volume differences. However, due to the large size of most of these regions often only parts of these region were highlighted by the network. Interestingly, brain structures affected by age with higher p-value in literature[12,13], were also more highlighted by the network, e.g.: caudate nucleus, amygdala, hippocampus and thalamus.

Supplementary references:

1. Mutlu U, Colijn JM, Ikram MA, et al. Association of Retinal Neurodegeneration on Optical Coherence Tomography With Dementia. *JAMA Neurol*. 2018:1-8. doi:10.1001/jamaneurol.2018.1563

2. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease. *Neurology*. 1984;34(7):939.

3. Román G, Tatemichi T, Erkinjuntti T, et al. Vascular dementia: diagnostic criteria for research studies. *Neurology*. 1993;43(2):250-260. doi:http://dx.doi.org/10.1212/WNL.43.2.250

4. Sathya R, Abraham A. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *Int J Adv Res Artif Intell*. 2013;2(2):34-38. doi:10.14569/IJARAI.2013.020206

5. Xinghuo Yu, M. Onder Efe and OK. A General Backpropagation Algorithm for Feedforward Neural Networks Learning. *IEEE Trans Neural Networks*. 2002;13(1):251-254.

6. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278-2323. doi:10.1109/5.726791

7. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst*. 2012:1-9. doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007

8. Xue-Wen Chen, Xiaotong Lin. Big Data Deep Learning: Challenges and Perspectives.

IEEE Access. 2014;2:514-525. doi:10.1109/ACCESS.2014.2325029

9. Işin A, Direkoğlu C, Şah M. Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods. *Procedia Comput Sci*. 2016;102(August):317-324. doi:10.1016/j.procs.2016.09.407

10. Ker J, Wang L, Rao J, Lim T. Deep Learning Applications in Medical Image Analysis. *IEEE Access*. 2018:1-1. doi:10.1109/ACCESS.2017.2788044

11. Perez L, Wang J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. 2017.

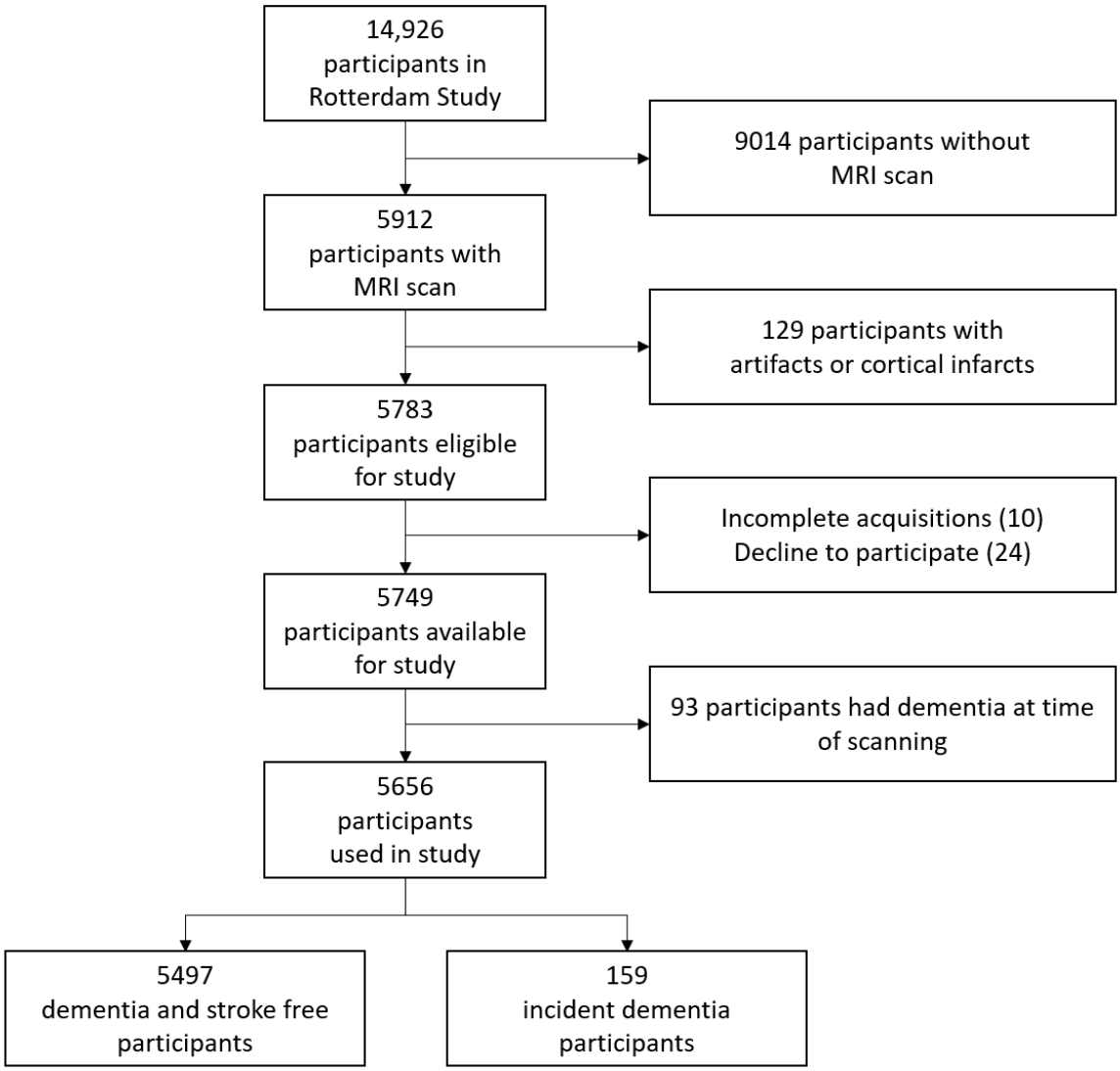
12. Manard M, Bahri MA, Salmon E, Collette F. Relationship between grey matter integrity and executive abilities in aging. *Brain Res*. 2016;1642:562-580. doi:10.1016/j.brainres.2016.04.045

13. Matsuda H. Voxel-based Morphometry of Brain MRI in Normal Aging and Alzheimer's Disease. *Aging Dis*. 2013;4(1):29-37.

Supplementary Table 1. Characteristics of subjects with the 5-year age-stratified lowest quintile age gap values compared to the 5-year age-stratified highest quintile age gap values, and correlation estimates between these characteristics and the age gap in the full sample.

Characteristic	Value lowest quintile (n=340)*	Value highest quintile (n=350)*	p-value	Correlation estimates (95% CI)
Age gap (years)	-5.7 ± 3.9	6.9 ± 4.5	<0.001	-
Grey matter volume (mL)	605.0 ± 56.9	577.6 ± 56.2	<0.001	0.11 (0.07;0.16)
Systolic blood pressure (mmHg)	138.9 ± 21.6	143.1 ± 21.0	0.009	-0.19 (-0.24;-0.15)
Mild cognitive impairment, n (%)	15 (4.4)	31 (8.9)	0.013	0.07 (0.01;0.13)
Diastolic blood pressure (mmHg)	82.1 ± 10.8	84.1 ± 11.1	0.014	0.06 (0.01;0.11)
Fasting glucose level (mmol/L)	5.5 ± 1.2	5.7 ± 1.1	0.021	-0.01 (-0.06;0.03)
Current or past smoker, n (%)	102 (30.0)	130 (37.1)	0.027	0.34 (0.30;0.38)
Body mass index (kg/m^2)	27.2 ± 3.9	27.8 ± 4.5	0.043	0.08 (0.04;0.13)
Mini-Mental State Examination score	28.0 ± 1.8	27.7 ± 2.1	0.095	0.07 (0.02;0.12)
Total cholesterol (mmol/L)	5.6 ± 1.0	5.5 ± 1.1	0.323	0.00 (-0.05;0.05)
APOEε4 carrier, n (%)	92 (27.1)	103 (29.4)	0.418	0.01 (-0.04;0.06)
Female, n (%)	187 (55.0)	203 (58.0)	0.428	0.02 (-0.03;0.07)
HDL cholesterol (mmol/L)	1.4 ± 0.4	1.5 ± 0.4	0.549	0.00 (-0.04;0.05)
Age (years)	65.5 ± 10.8	65.3 ± 11.0	0.771	-0.58 (-0.61;-0.55)
Years of education	12.4 ± 3.8	12.3 ± 4.0	0.829	0.10 (0.05;0.14)
Intracranial volume (mL)	1465.8 ± 163.2	1466.3 ± 164.1	0.971	0.10 (0.06;0.15)

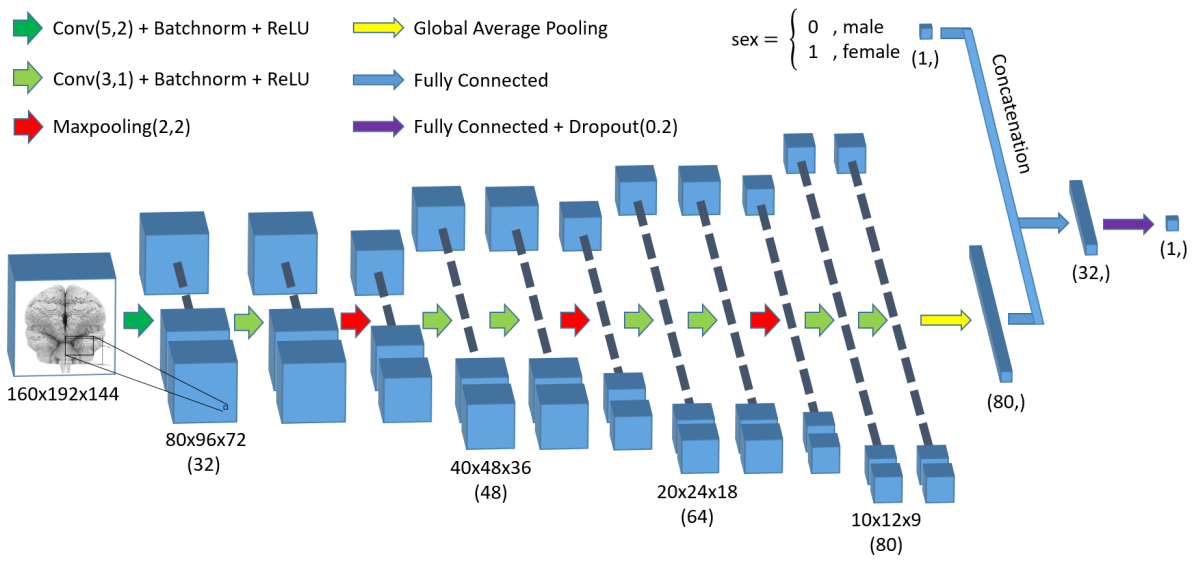
*Values are presented in mean ± SD unless stated otherwise.
Abbreviations: confidence interval (CI); number of participants (n); standard deviation (SD).



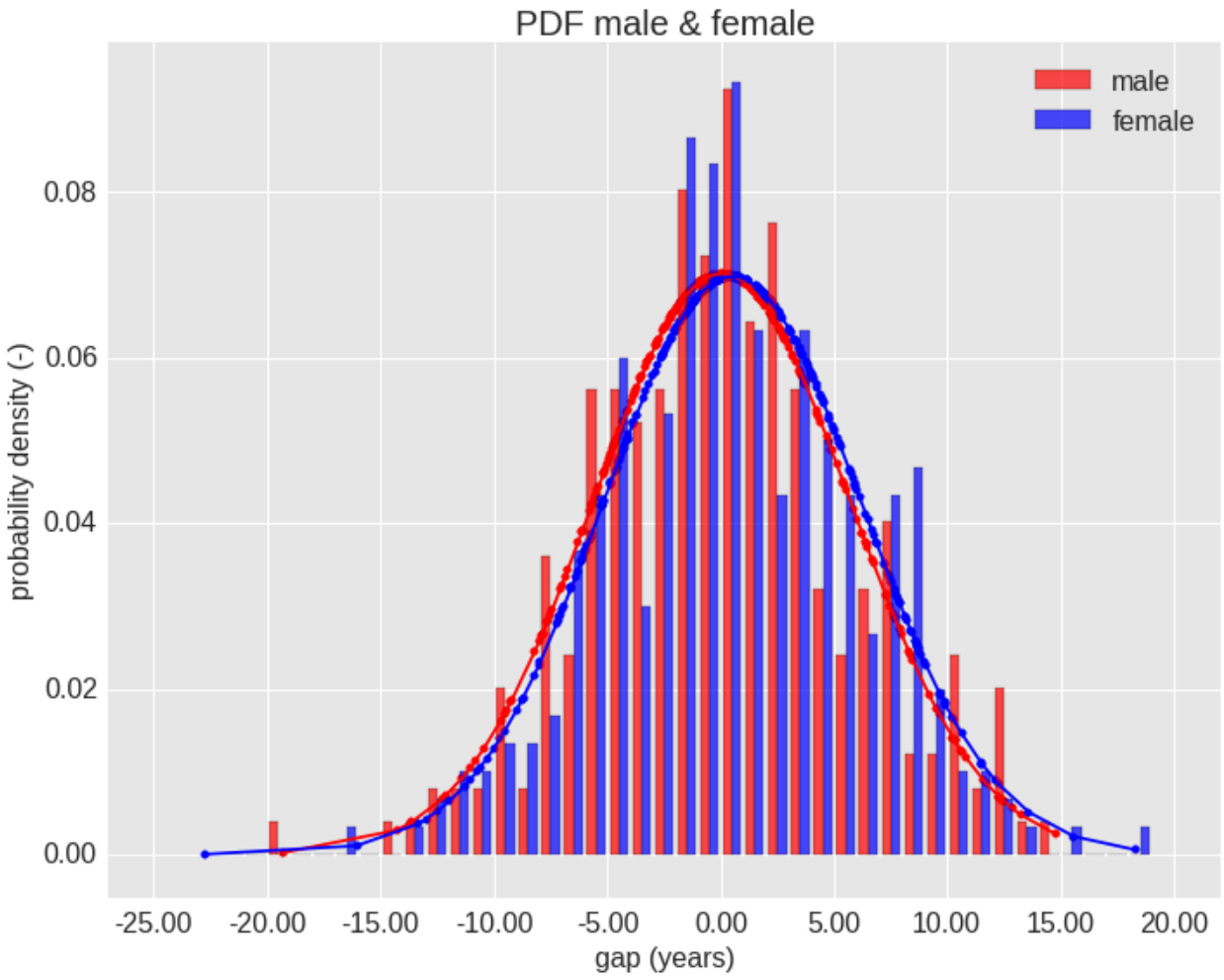
198

199 **Supplementary Figure 1.** Flowchart showing the number of excluded participants per category.

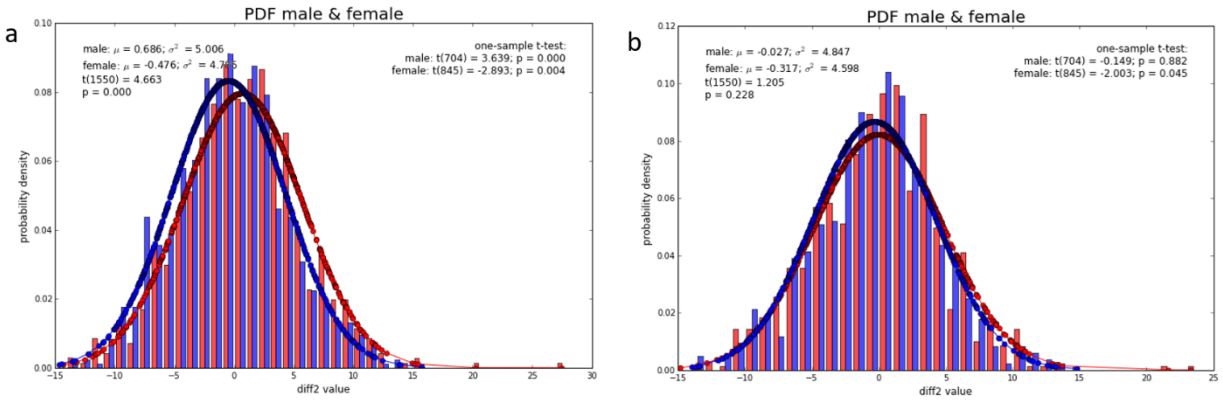
200



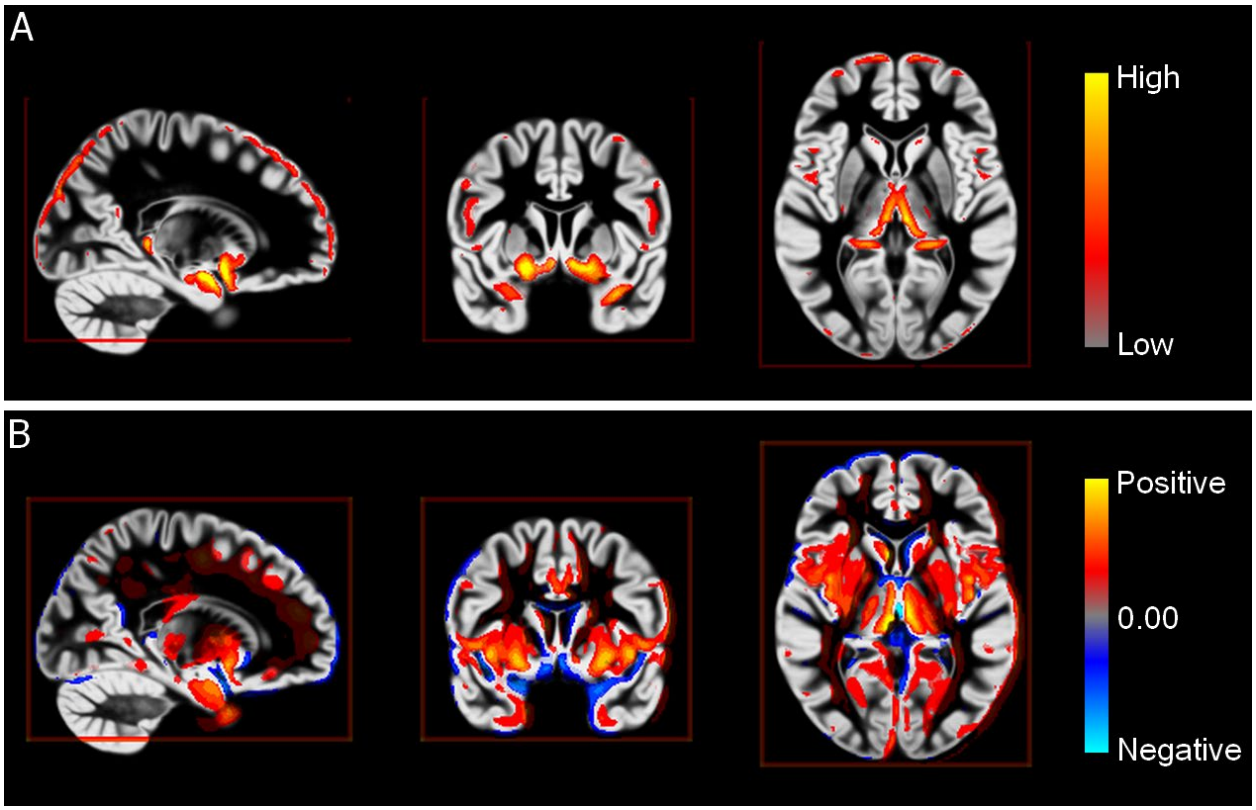
Supplementary Figure 2. Graphical representation of the network architecture. The overall approach can be seen as four convolutional blocks ending on a pooling layer, which halves feature map dimensions. Hereafter, global average pooling extracts the final feature maps to a one-dimensional array of a single value per feature map. Fully connected layers are used to propagate to a single regression output. *Abbreviations:* kxkxk convolutional layer, with strides of sxsxs (CONV(k,s)); kxkxk max-pooling layer, with strides of sxsxs (Maxpooling(k,s)); batch normalization (Batchnorm); rectified linear unit (ReLU); dropout with probability p (Dropout(p)).



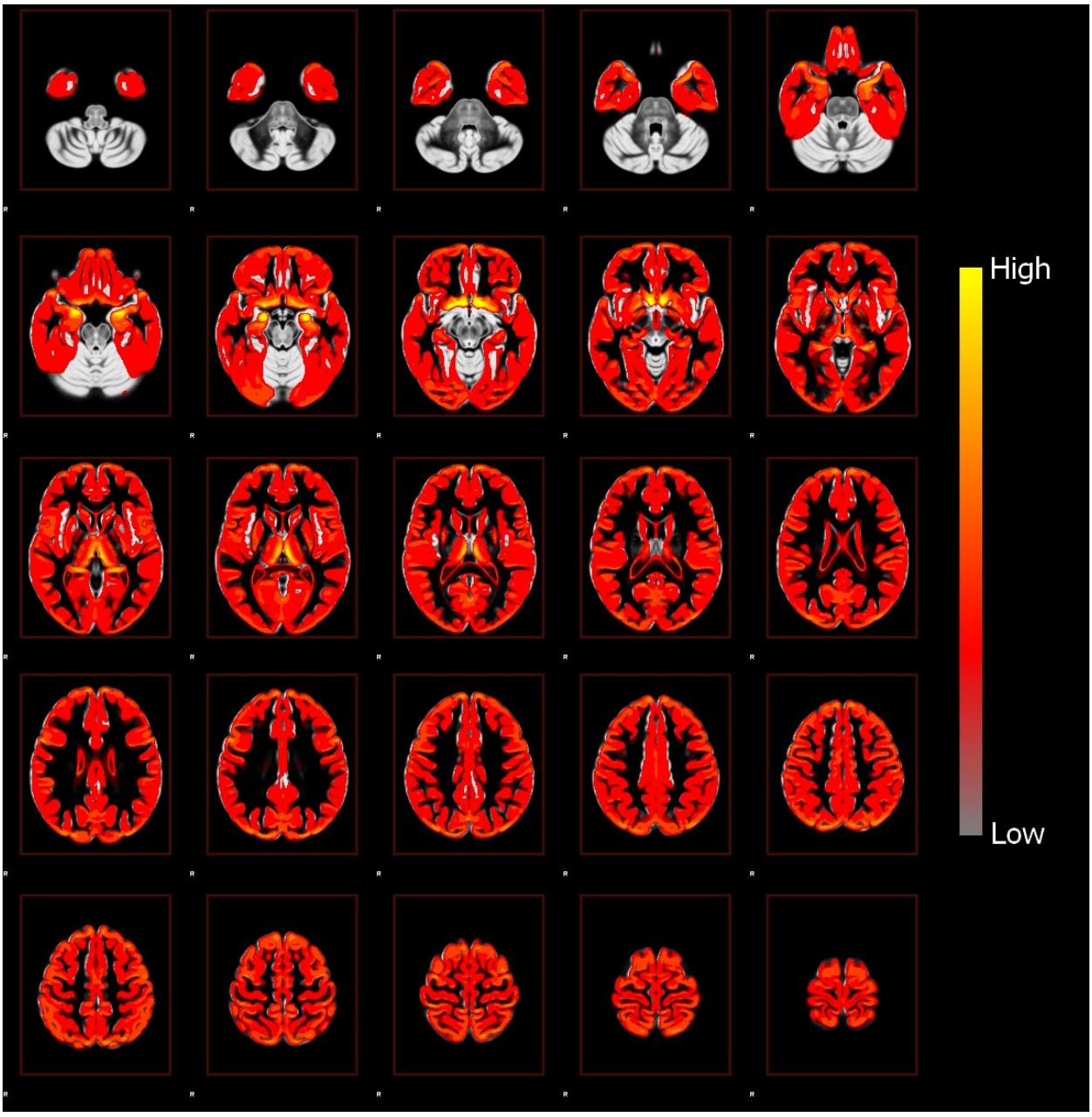
Supplementary Figure 3. The probability density of the gap value (PAD) for male and female subjects. The distribution shows the difference in prediction for these two groups. Distributions are similar as mean $\eta_{female} = 0.51$ and variance $\sigma^2_{female} = 5.72$ for female, whereas $\eta_{male} = 0.04$ and $\sigma^2_{male} = 5.69$ for male. Resulting t-test showed no significant difference between the two groups as $t(550) = -0.96$ and $p = 0.34$.



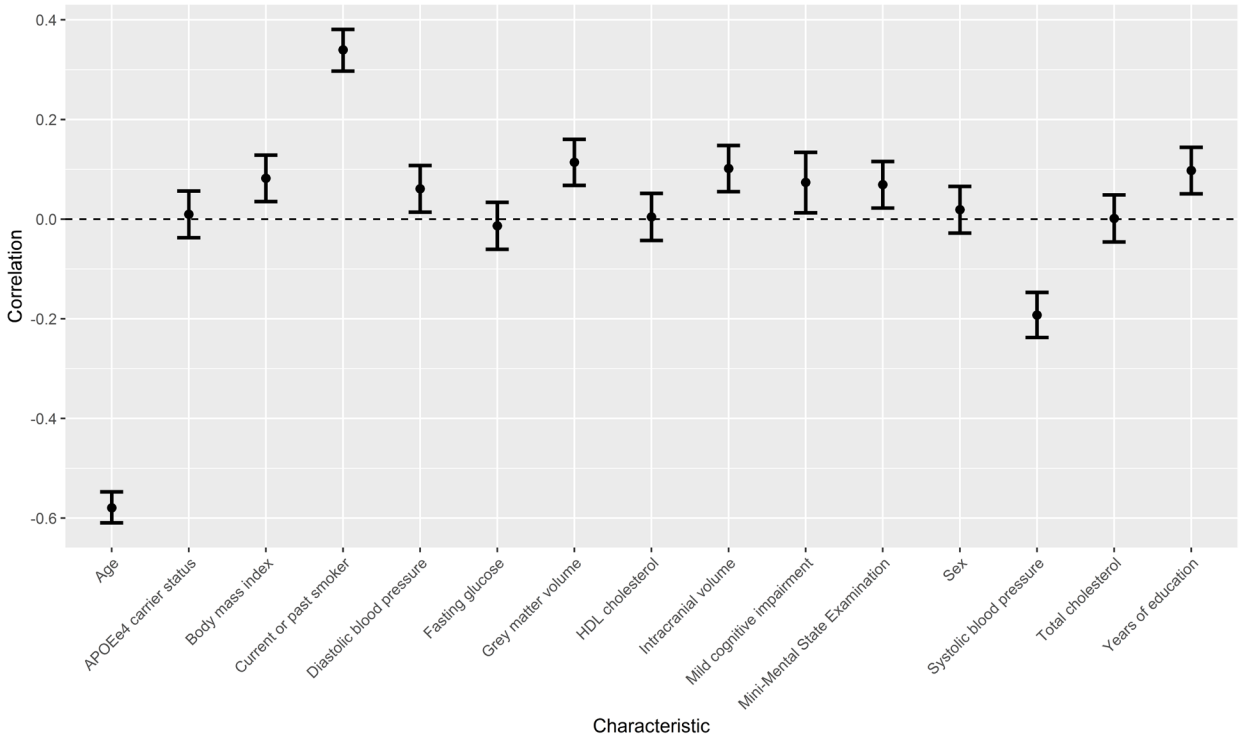
Supplementary Figure 4. Effect of adding sex as a covariate to the model on the gap value distribution (red=male; blue=female). A comparison of the probability density functions for gap of two early trained models along with their respective t-test results. Both models have the exact same architecture with one the exception. a) Model uses only a single brain-MRI voxels input. b) Model uses two inputs, i.e. brain-MRI voxels and respective sex. Models were trained under the exact same settings.



Supplementary Figure 5. Grad-CAM attention map and attention map change overlaid on a brain template. **(A)** Grad-Cam attention map intensity per voxel. Voxel values in the attention map have been set at 0.65 minimum threshold and capped at 0.95 maximum to exclude background values and focus on more important regions. **(B)** Increase in attention map intensity over chronological age per voxel. Map include only voxels with a significant increase in voxel values ($p<3e^{-7}$ after Bonferroni correction by number of GM voxels).



Supplementary Figure 6. Grad-CAM attention map intensity per voxel overlaid on a brain template.



Supplementary Figure 7. Correlation between brain pathology associated features and the age gap.