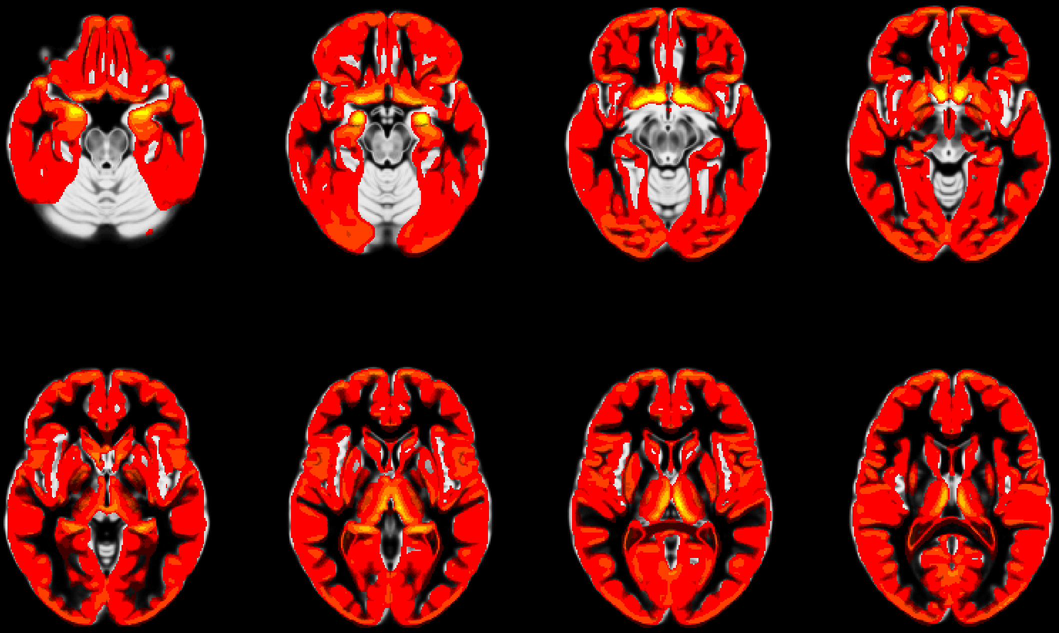
Grey Matter Age Prediction as a Biomarker for Risk of Dementia

A Population-based Study

Johnny Wang



TU Delft

Grey Matter Age Prediction as a Biomarker for Risk of Dementia

A Population-based Study

By

Johnny Wang

in partial fulfilment of the requirements for the degree of

Master of Science
in Mechanical Engineering

at the Delft University of Technology,
to be defended publicly on Monday January 7, 2019 at 10:30 AM.

Thesis committee:

Chairman:	Prof. dr. Wiro J. Niessen	TU Delft
Department representative:	Assistent prof. dr. Julian F.P. Kooij	TU Delft
External member:	Associate prof. dr. Frans M. Vos	TU Delft
Project supervisor:	Dr. Gennady V. Roshchupkin	Erasmus MC

Preface

Basis for the chosen topic of this Master thesis stems from my passion for programming and me wanting to expand my knowledge into the field of artificial intelligence. An opportunity was presented to work with interesting data at the Erasmus Medical Center (EMC) in Rotterdam, researching the possibilities of deep learning in the field of medical imaging. Thus, this graduation project was done in conjunction with the Biomedical Imaging Group Rotterdam (BIGR), part of the medical informatics department at the EMC in Rotterdam, The Netherlands. After more than one and a half years, what started as a project taking a deep learning approach towards investigating imaging genetics and brain MRIs, ended in a paper on deep learning age prediction based on MRIs for risk analysis of incident dementia. In this we found that the gap between a person's predicted brain age and actual chronological age shows potential as a risk biomarker for future dementia, i.e. a higher gap can indicate a higher risk for future dementia.

Part of this work has been submitted to and presented at the European Society of Medical Imaging Informatics (EuSoMII) annual meeting 2018, titled 'Advances in Medical Imaging with Informatics and Artificial Intelligence'. At the time of delivery of this thesis, an abstract has been submitted to the Organization of Human Brain Mapping (OHBM) annual meeting 2019 and a poster presentation is being prepared for the European Congress of Radiology (ECR) 2019, organized by the European Society of Radiology (ESR). Lastly, a paper is being prepared for submission to the Journal of American Medical Association (JAMA). Hence that this thesis has been written in the format of an article, extended from the original paper.

I hope you enjoy reading my thesis.

Johnny Wang, 28-12-2018.

Acknowledgements

So okay. I took a “little” longer than expected, but I reached the finish-line! I would first like to thank Professor Wiro Niessen for giving me the opportunity to work at his research group. Ofcourse my daily supervisor, Dr. Gennady Roshchupkin. Gena, many many thanks for ALL the guidance you have given me. I started with zero knowledge on deep learning, but now I was allowed to present our work at a conference and we are even about to submit a paper. I want to thank Professor Julian Kooij and Professor Frans Vos for being part of my Master exam committee, despite your (probably) busy schedules.

My thanks all my colleagues. To Maria for writing the paper with me. All other co-authors of the original paper: Aleksei, Florian, Professor De Bruijne, Professor Vernooij, Dr Adams and Professor Ikram. I thank everyone at BIGR and the medical informatics department. I really had a great time with all of you! (Honestly, the reason why I stuck around so long might just have been because I didn’t feel like leaving.)

My thanks to my friends from uni. Arno, we were a great team throughout the Master. It was a pleasure to have been stuck with you. Tito, (wherever you are) I’m going to finish the Master first, see you on the other side. Kenny, brother, thank you for being there for me at all times.

My thanks to my friends and relatives, for their continued support and cheering.

My thanks to my family. Sisters, Marleen and Ling. Mom and Dad. Even my nephew, Oscar. For their patience and continued support, and cheering me up whenever I was feeling down.

To all of you. Thank you!

Johnny Wang, 28-12-2018.

Contents

List of figures

List of tables

List of abbreviations

AD	Alzheimer’s disease
Batchnorm	Batch normalization
CI	Confidence interval
CNN	Convolutional neural network
CONV	Convolutional layer
CSF	Cerebral spinal fluid
DL	Deep learning
GM	Grey matter
HR	Hazards ratio
ICC	Intraclass correlation coefficient
ICV	Intracranial volume
MAE	Mean absolute error
MRI	Magnetic resonance images
MSE	Mean squared error
OR	Odds ratio
PCC	Pearson correlation coefficient
PH models	Proportional hazards models
ReLU	Rectified linear unit
VBM	Voxel-based morphometry
WM	White matter

Grey Matter Age Prediction as a Biomarker for Risk of Dementia: A Population-based Study

Abstract

The gap between predicted brain age and chronological age could serve as biomarker for early-stage neurodegeneration and as potentially as a risk indicator for dementia. We assess the utility of this age gap as a risk biomarker for incident dementia in a general elderly population. The brain age is estimated from longitudinal brain magnetic resonance imaging (MRI) data using deep learning models. From the population-based Rotterdam Study, 5656 dementia-free and stroke-free participants (mean age 64.67±9.82, 54.73% women) underwent brain MRI at 1.5T, including three-dimensional (3D) T1-weighted sequence, between 2006 and 2015. All participants were followed for incident dementia until 2016. During 6.66±2.46 years of follow-up, 159 subjects developed dementia. The entire dataset was split into control (N=5497) and incident dementia (N=159) groups. We then built a convolutional neural network (CNN) model trained on the control group to predict brain age based on brain MRI. Model prediction performance was measured in mean absolute error MAE=4.45±3.59 years of brain age prediction. Reproducibility of prediction was tested using the intra-class correlation coefficient ICC=0.97 (95% confidence interval CI=0.96-0.98), computed on a subset of 80 subjects. Hereafter, we investigated the gap between model predicted age and chronological age of the incident dementia group data, compared to control group. Logistic regressions and Cox proportional hazards models were used to assess the association of the age gap with incident dementia, adjusted for years of education, ApoEε4 allele carriership, GM and intracranial volume. These models showed that the age gap was significantly related to incident dementia (odds ratio OR=1.11 and 95% confidence intervals CI=1.05-1.16; hazard ratio HR=1.11 and 95% CI=1.06-1.15, respectively). Additionally, we computed the attention maps of CNN, which shows the importance of brain regions for age prediction. These were particularly focused on the amygdalae and hippocampi. We show that the gap between predicted and chronological brain age is a biomarker, associated with a risk of dementia development. This suggests that it can potentially be used as a complimentary biomarker for early-stage dementia risk screening.

Keywords: Deep Learning; age prediction; biomarker; dementia; magnetic resonance imaging; brain; voxel-based morphometry; survival analysis.

1. Introduction

The human brain continuously changes throughout the lifespan, including middle to old age. These changes reflect the normal aging process and are not necessarily pathological[1]. However, neurodegenerative diseases and dementia also affect brain structure and function[2,3]. A better understanding and modeling of normal brain aging can help to disentangle these two processes and improve the detection of early-stage neurodegeneration.

Magnetic resonance imaging (MRI) has widely been used to assist the diagnosis of brain diseases or find an association with epidemiological outcomes. Age prediction models based on brain MRIs are a popular trend in neuroscience[4–7]. The difference between predicted and chronological age is thought to serve as an important biomarker reflecting pathological processes in the brain. Several recent studies showed the relation between accelerated brain aging and various disorders, such as Alzheimer’s disease (AD), schizophrenia, epilepsy or diabetes[7–9]. In recent years, convolutional neural networks (CNN) have become the methodology of choice for analyzing medical images. These models are able to learn complex relations between input data and desired outcomes. Recent studies were able to demonstrate that CNN models can outperform complex machine learning models in brain MRI-based age prediction[5,6].

Although cross-sectional studies have suggested that the gap between predicted and chronological age may serve as a biomarker for dementia diagnosis, it remains unclear whether this is also the case for the years preceding dementia diagnosis[5,7]. Longitudinal studies examining the link between such a gap and future (incident) dementia are lacking and are crucial for validation of this biomarker for early-stage neurodegeneration detection.

One of the classical neuroimaging analysis approaches is the voxel-based morphometry analysis (VBM). Originally proposed as a hypothesis-free method, it has been widely used in brain imaging research field and demonstrated its effectiveness[4,10,11]. It allows analyzing the entire brain without any a priori defined brain regions. However, VBM analyzes images voxel-by-voxel and thereby do not take into account the spatial connectivity and more complex relations.

Therefore, using a deep learning (DL) model, we investigated the association of the grey matter (GM) age gap with incident dementia in a large population-based sample of middle-aged and elderly subjects.

2. Methods

2.1 Study Population

Data was acquired from the Rotterdam Study, an ongoing population-based cohort study among the inhabitants of Ommoord, a suburb of Rotterdam, the Netherlands[12]. 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 (**Figure 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.

2.2 Image processing

A 1.5 tesla GE Signa Excite MRI scanner was used to acquire multi-parametric MRI brain data, as previously reported[12]. Voxel-based morphometry (VBM) was performed according to an optimized VBM[10,11]. An overview of the VBM pipeline is shown in **Figure 2**[13]. First, all T1-weighted images were segmented into supratentorial GM, white matter (WM), and cerebrospinal fluid (CSF) using a previously described *k*-nearest neighbor algorithm, which was trained on six manually labeled atlases[14]. FMRIB's Software Library (FSL) software was used for VBM data processing[15]. All GM density maps were non-linearly registered to the standard

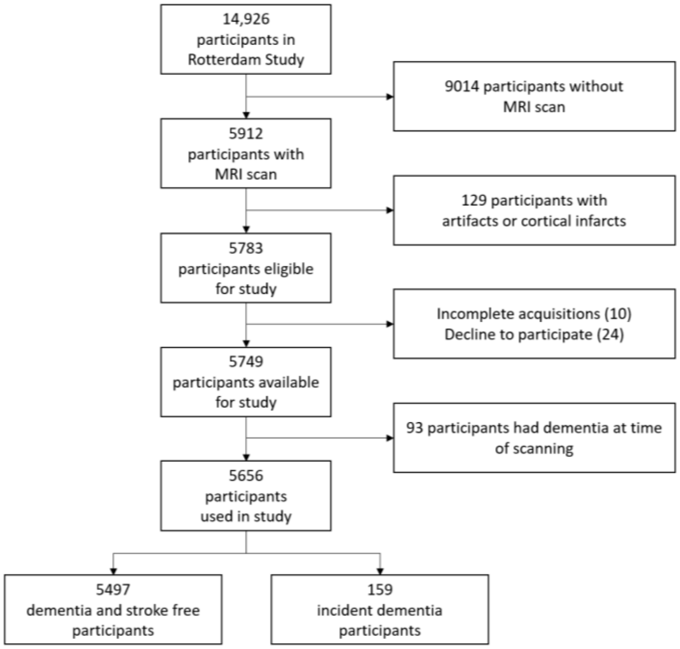

14,926 participants in Rotterdam Study
9014 participants without MRI scan
5912 participants with MRI scan
129 participants with artifacts or cortical infarcts
5783 participants eligible for study
Incomplete acquisitions (10)
Decline to participate (24)
5749 participants available for study
93 participants had dementia at time of scanning
5656 participants used in study
5497 dementia and stroke free participants
159 incident dementia participants

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

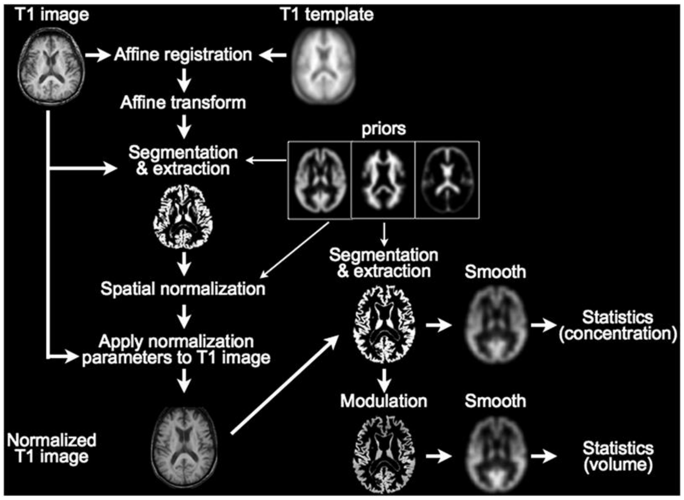


Figure 2. Graphical representation of the voxel-based morphometry pipeline[13]. T1-weighted magnetic resonance images are registered to a template, normalized and segmented according to priors. In this study, the final smoothing step is omitted.

Montreal Neurological Institute (MNI) GM probability template, with a 1x1x1 mm^3 voxel resolution.

A spatial modulation procedure was used to avoid differences in absolute GM volume due to the registration. This involved multiplying voxel density values by the Jacobian determinants estimated during spatial normalization. We did not apply smoothing. While VBM smoothing procedures increase the signal to noise ratio, they can affect the features which the network uses to learns from GM.

Intracranial volume (ICV) estimates were obtained by summing total GM, WM and CSF volumes.

Dementia assessment

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[16]. 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)) and AD (using the National Institute of Neurological and Communicative Diseases and Stroke–Alzheimer's Disease and Related Disorders Association (NINCDS–ADRDA) criteria)[17,18]. Follow-up until January 1st 2016 was virtually complete (95% of potential person-years). 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).

Other measurements

ApoEε4 carriership was determined using a polymerase chain reaction (PCR) on coded deoxyribonucleic acid (DNA) samples. If these values were missing, Haplotype Reference Consortium (HRC) imputed genotype values for rs7412 and rs429358 were used to define the ApoEε4 carrier status[19].

2.3 Deep Learning model

The concept of DL and its techniques are explained in **Appendix A.1-Appendix A.2**. Briefly, a DL model takes a set of inputs and respective outputs from a training set and finds an optimal non-linear relation between the two. A CNN is a class of DL techniques, which takes in multi-dimensional data as model input. These networks are generally used with a variety of different techniques and algorithms, which together define how the model optimizes the input-output relationship[20,21]. This is described in the model architecture.

Our 3-dimensional (3D) regression CNN model is designed to predict brain age using 3D GM density maps from VBM as input. It is inspired by ConvNet[22] and Deep CNN[21], as shown in **Figure 3** and detailed in **Appendix A.3**. Besides GM brain images, we provide information about the sex of the subject. This allows the network to adjust for GM differences between male and female subjects.

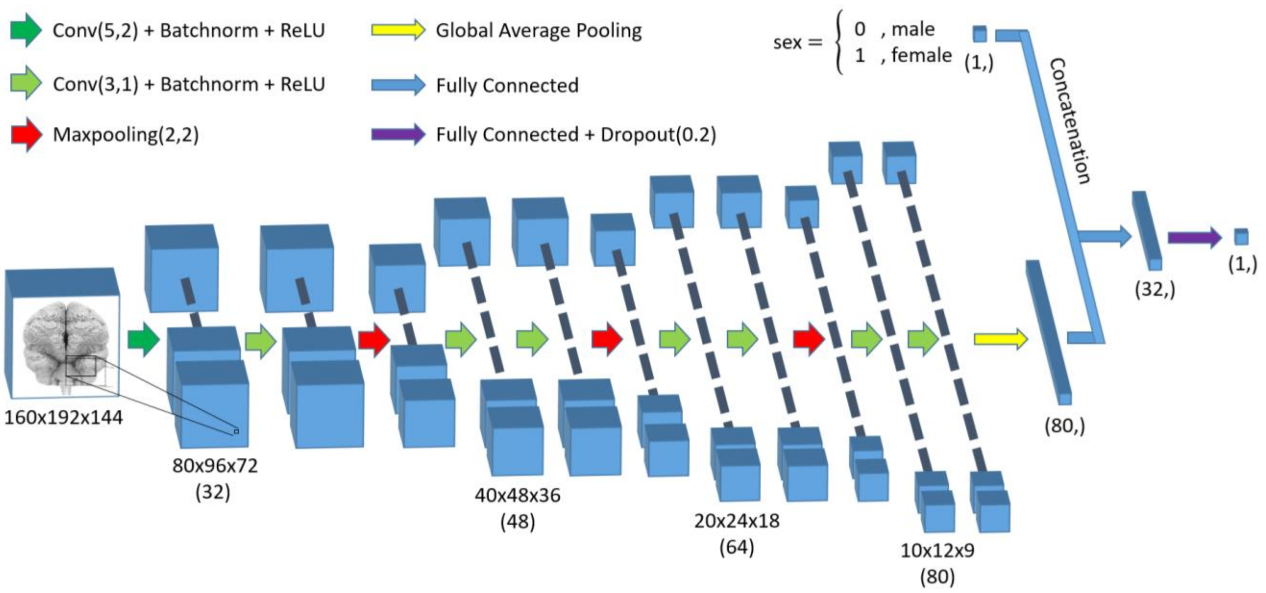


Figure 3. Graphical representation of the network architecture. The overall approach can be seen as four convolutional blocks ending on a pooling layer, which halves feature map dimensions. Hereafter, global average pooling extracts the final feature maps to a one-dimensional array of a single value per feature map. Fully connected layers are used to propagate to a single regression output. *Abbreviations:* kxkxk convolutional layer, with strides of s (CONV(k,s)); kxkxk max-pooling layer, with strides of s (Maxpooling(k,s)); batch normalization (Batchnorm); rectified linear unit (ReLU); dropout with probability p (Dropout(p)).

The dataset, excluding subjects with incident dementia, was randomly split into three independent sets: training (3688 subjects), validation (1099 subjects) and test (550 subjects). Subjects with incident dementia (159 subjects) were put in a fourth independent dataset. The CNN was trained using the training set, i.e. 3688 undersampled out of 3848 available subjects for training as described in **Appendix A.4**. For training we used all available scans for each subjects. Prediction accuracy was assessed on the test set. Model accuracy was measured based on the absolute gap, or mean absolute error (MAE) of prediction for all subjects $S = \{1,2,\dots,N\}$. This is equivalent to

$$\mathrm{MAE} = \frac{1}{N}\sum_{S} |\mathrm{gap}_S|$$
$$\mathrm{gap} = \mathrm{age}_{\mathrm{brain,predicted}} - \mathrm{age}_{\mathrm{chronological}} \qquad (1)$$

, where the gap of a subject gap_S is the difference between model output and real chronological age. Additionally, Pearson correlation coefficient (PCC) is reported as a measure of linear correlation between predicted model output and real chronological age[23].

Attention mapping

We retrieved attention maps from the trained networks using Gradient-weighted Class Activation Mapping (Grad-CAM)[24]. Attention maps show which areas on subject GM image are more important for age prediction. We expanded the Grad-CAM visualization technique to a 3D space and adopted it to a single regression output problem to obtain,

$$a_m = \frac{1}{Z}\sum_i\sum_j\sum_k \frac{\partial y}{\partial A^m_{ijk}}$$
$$L_{\mathrm{raw}} = \mathrm{ReLU}\left(\sum_m a_m A^m\right)$$
$$L = \mathrm{clip}\left(0.5 + \frac{1}{10}\cdot\frac{L_{\mathrm{raw}} - \mu(L_{\mathrm{raw}})}{\sigma(L_{\mathrm{raw}})}, 0, 1\right) \qquad (2)$$

. The weight a_m represents a partial linearization of network from activation maps A onwards based on gradient $\frac{\partial y}{\partial A^m}$, and captures the 'importance' of activation map m for the output y, summed for every pixel ijk and divided by map size Z. The weighted combination of the forward activation maps A results in localization Grad-

CAM map (attention map) L_{raw}, with intensity values >0 after the rectified linear unit (ReLU) function[21]. Finally, values of L_{raw} were normalized to range 0-1, where 1 indicates the most important area for the decision-making process, with $\mu(L_{\text{raw}})$ mean and $\sigma(L_{\text{raw}})$ standard deviation of attention map values, and clip thresholding the outcome to 0 minimum and 1 maximum to obtain the final normalized attention map L.

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 as

$$L_{\text{global},ijk} = \frac{1}{N}\sum_{S} L^{S}_{ijk} \tag{3}$$

, with L_{global} as the mean of L^S over all N subjects.

We computed the increase 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 L_{raw} value was performed, according to

$$\begin{aligned}
&\mathit{Assume\ linear\ regression\ model}:\\
&\qquad y_S = a + bx_S + \varepsilon_S\\
&\mathit{with}:\\
&\qquad y_S = L^{S}_{\text{raw},ijk}\ \mathit{and}\ x_S = \text{age}^{S}_{\text{chronological}}\\
&\mathit{then}:\\
&\qquad L_{\text{increase},ijk} = \frac{\partial L_{\text{raw},ijk}}{\partial \text{age}_{\text{chronological}}} = b_{ijk}
\end{aligned} \tag{4}$$

, where the slope b represents the increase in attention map value with 1 year age for voxel ijk[25]. Following is the resulting increase in attention map L_{increase}.

2.4 Statistical analysis

Reproducibility of the CNN age prediction was quantified using the intraclass correlation coefficient (ICC(3,1)), computed on a subset of 80 persons out of the test set who were scanned twice with a time interval of one to nine weeks[26]. Corresponding confidence interval was found by means of bootstrapping[27].

In order to be able to compare our findings with previous studies, logistic regression models and Cox proportional hazards models were used to assess the association between the age gap and the incidence of dementia. Mentioned analysis methods are briefly explained in **Appendix B**. We adjusted the regression models for biomarkers, which are known for their relation with dementia: age and sex (model I); additionally overall GM volume and ICV (model II); and years of education and APOEε4 carriership (model III)[19,28–30]. The logistic regression model used the occurrence of dementia-development during follow-up as output. The proportional hazards and linearity assumption were met for the Cox proportional hazards models. Python and R were used to perform the statistical analyses[31–34].

3. Results

The study population characteristics are described in **Table 1**. The algorithm was trained and validated on random subsets of subjects with mean age 66.09±10.76 years and 55% females; and mean age 64.84±9.69 years and 54% females, respectively. The following results are reported for the test set (mean age 64.85±10.82 years and 55% females).

3.1 Network performance

The overall performance measured on the test set was MAE=4.45±3.59 years (**Figure 4**), with a Pearson correlation between chronological and predicted brain age of PCC=0.85 (p-value=4.76x10^{-156}). A reproducibility score of ICC=0.97 (95% confidence interval CI 0.96-0.98) was achieved.

A split evaluation can be considered between male and female subjects. **Figure 5** shows the network found no significant difference between the two groups (p-value=0.34).

Attention map

Figure 6 shows the global attention map of the test set, indicating the areas contributing to age prediction in bright color, as well as the increase of attention map values over age. A quantitative analysis per brain region is presented in **Table 2** and **Figure C-1**, which show that highest mean

Table 1. Quantitative description of the data used from the population-based Rotterdam Study.

	Train	Validation	Test**	Incident dementia**
N_{subj}	3688	1099	550	159
N_{img}	5865	2353	550	159
Mean age* (years±sd)	66.09±10.76	64.84±9.69	64.85±10.82	77.33±7.15
Sex proportion* (female/male)	0.55/0.45	0.54/0.46	0.55/0.45	0.58/0.42
Education* (years±sd)	12.64±3.89	12.63±3.81	12.58±4.00	11.43±3.57
GM volume* (liters±sd)	0.60±0.06	0.60±0.06	0.60±0.06	0.55±0.05
ICV* (liters±sd)	1.48±0.16	1.47±0.16	1.48±0.16	1.45±0.17
ApoE4 carriership* (0/1/2)	0.72/0.26/0.02	0.72/0.25/0.02	0.74/0.23/0.03	0.57/0.36/0.06
Follow-up time* (years±sd)	5.42 ±2.81	4.93±2.80	6.68±2.29	4.29±2.26

* Values are based on N_{img}.
** Selection only includes baseline image of subjects.
Abreviations: number of subjects (N_{subj}); number of images (N_{img}); grey matter (GM); intracranial volume (ICV)

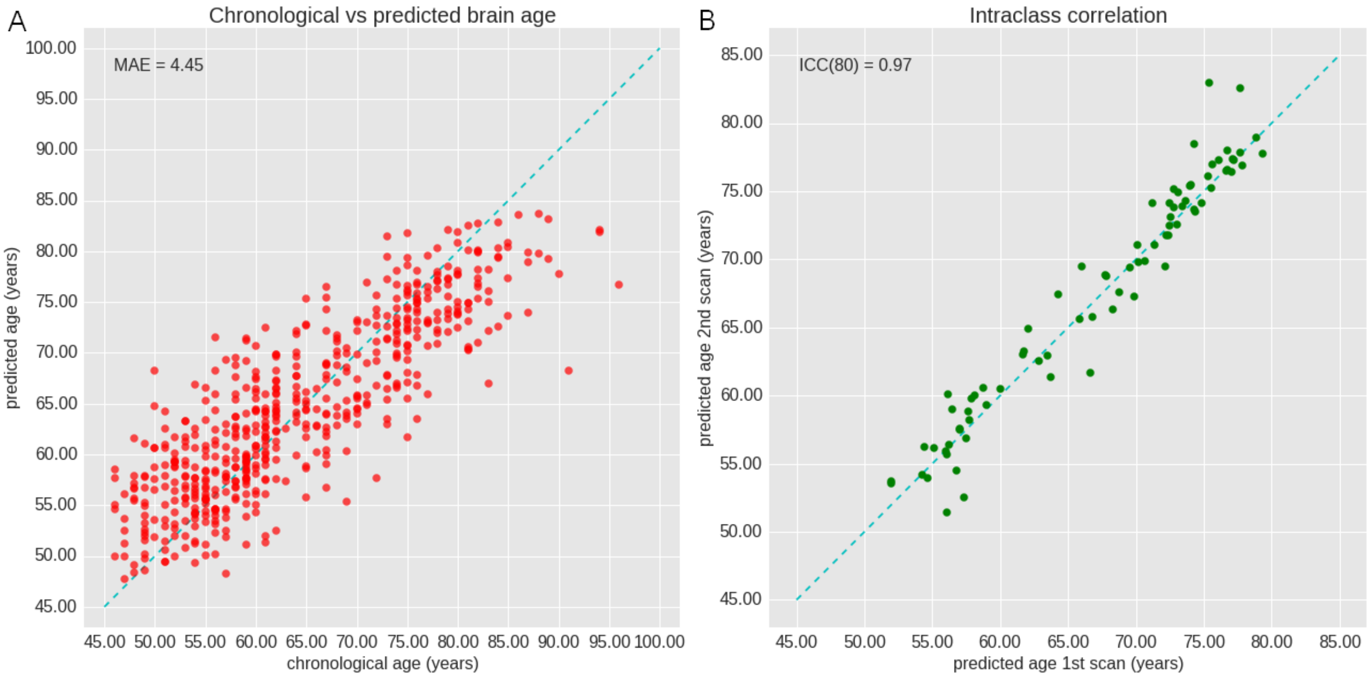


Figure 4. Performance of convolutional neural network on test dataset. **(A)** The plot depicts chronological age (x-axis) and brain-predicted age (y-axis) with mean absolute error (MAE) and Pearson correlation coefficient (PCC). The dashed line indicates perfect prediction x=y. **(B)** The figure shows reproducibility of the CNN performance. Scan 1 and 2 are taken with one to nine weeks interval. The dashed line indicates a perfect reproducibility.

intensities were computed for the nucleus accumbens (0.89) and amygdala (0.71). Highest intensity quintiles were computed for the nucleus accumbens (0.99), amygdala (0.98) and subcallosal area (0.98). Amongst the higher intensity regions, we found that brain regions such as amygdalae and hippocampi are not only important for predicting brain age, but that they also become more important with increasing chronological age, which is shown in **Figure 6B** and **Table D-1**.

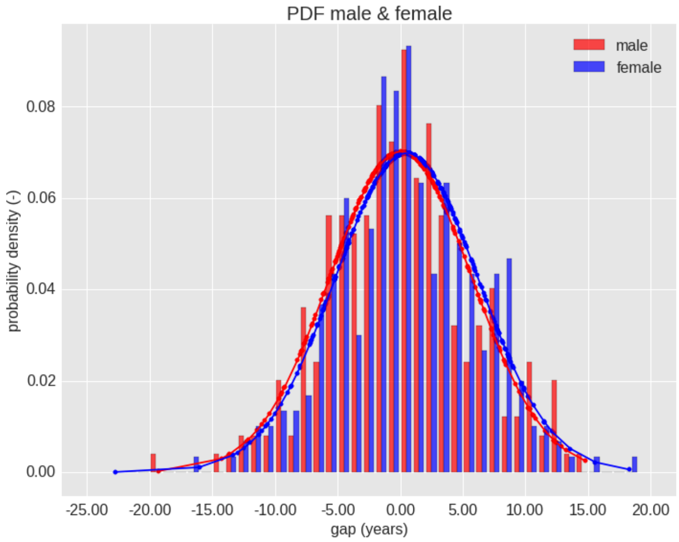


Figure 5. The probability density of the gap value (PAD) for male and female subjects. The distribution shows the difference in prediction for these two groups. Distributions are similar as η_{female}=0.51 and σ^2_{female}=5.72 for female, whereas η_{male}=0.04 and σ^2_{male}=5.69 for male. Resulting t-test showed no significant difference between the two groups as t(550)=-0.96 and p=0.34.

3.2 Logistic regression

The regression analyses are performed on baseline MRI data of validation, test and incident dementia dataset subjects. We computed a logistic regression for the three models, as shown in **Table 3**. The age gap was significantly associated with dementia incidence while age, sex, education years, GM and ICV volume and the ApoEε4 allele carriership were included in the model, with model III: odds ratio OR=1.11 (95% CI 1.05-1.16) per year age gap. These associations were similar in a subsample with a follow-up time for incident dementia of more than 5 years, model III OR=1.09 (95% CI 1.01-1.16) per year age gap.

3.3 Survival analysis

Table 3 shows the age gap was significantly associated with the incidence of dementia, with model III hazard ratio HR=1.11 (95% CI 1.06-1.15) per year age gap. As in the

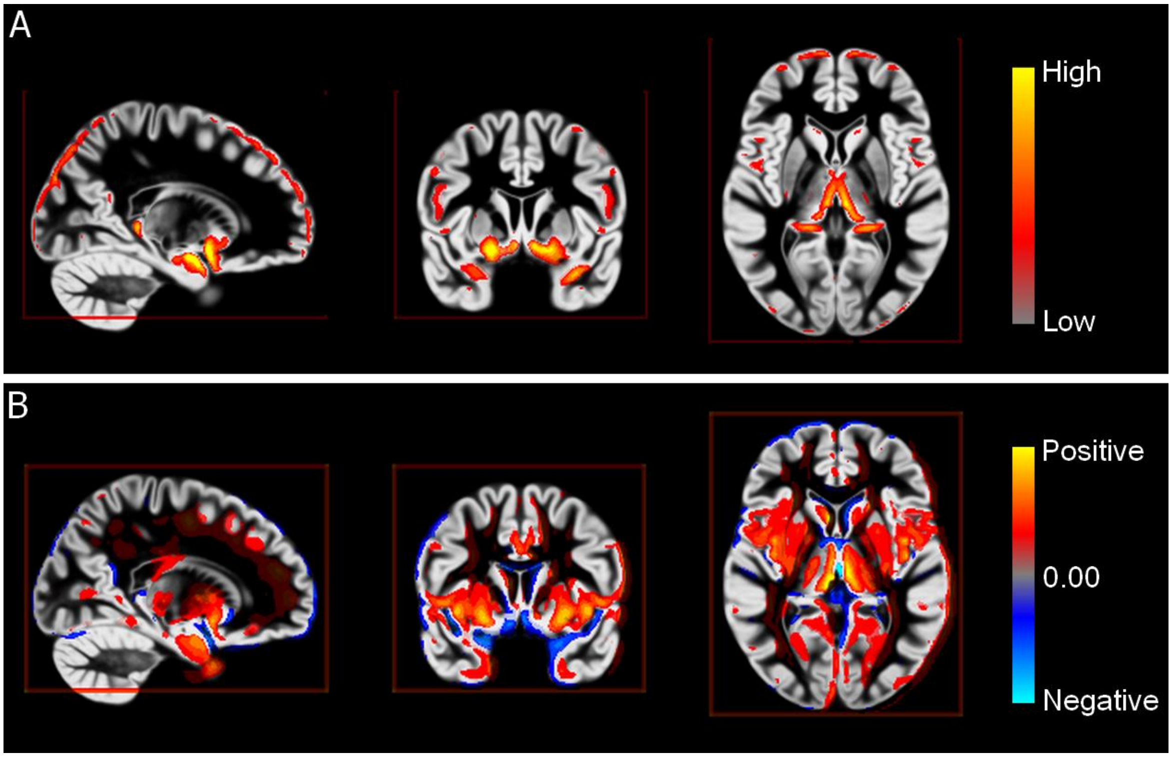


Figure 6. Grad-CAM attention map and increase in attention map overlaid on a brain template. **(A)** Grad-Cam attention map intensity per voxel. Voxel values in the attention map have been set at 0.65 minimum threshold and capped at 0.95 maximum to exclude background values and focus on more important 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).

Table 2. Quantitative description of the attention map. Mean and (lower boundary of) fifth quintile of attention map intensity are shown per brain region. Brain regions are grouped by lobes.

Brain region	Size (voxels)	Attention map Intensity	
		Mean	5th quartile
Temporal Lobe			
Amygdala	4,398	0.71	0.98
Hippocampus	6,687	0.61	0.80
Anterior temporal lobe medial part	22,842	0.54	0.78
Superior temporal gyrus, anterior part	14,369	0.54	0.74
Lateral occipitotemporal gyrus (gyrus fusiformis)	12,908	0.53	0.62
Posterior temporal lobe	143,237	0.52	0.68
Superior temporal gyrus, central part	42,794	0.52	0.68
Gyri parahippocampalis et ambiens	13,767	0.51	0.63
Medial and inferior temporal gyri	55,102	0.50	0.68
Anterior temporal lobe lateral part	11,999	0.49	0.65
Insula and Cingulate gyri			
Cingulate gyrus anterior part (supragenual)	24,751	0.53	0.63
Cingulate gyrus posterior part	24,235	0.52	0.64
Insula	44,328	0.51	0.64
Frontal Lobe			
Subcallosal area	788	0.70	0.98
Posterior orbital gyrus	15,061	0.54	0.72
Straight gyrus (gyrus rectus)	11,826	0.54	0.67
Inferior frontal gyrus	55,754	0.53	0.72
Superior frontal gyrus	166,766	0.52	0.77
Precentral gyrus	106,145	0.52	0.77
Medial orbital gyrus	18,554	0.52	0.77
Pre-subgenual anterior cingulate gyrus	2,451	0.52	0.61
Middle frontal gyrus	161,999	0.51	0.74
Anterior orbital gyrus	19,514	0.51	0.73
Lateral orbital gyrus	11,112	0.51	0.77
Subgenual anterior cingulate gyrus	4,287	0.50	0.71
Occipital Lobe			
Cuneus	28,209	0.57	0.67
Lingual gyrus	36,627	0.55	0.65
Lateral remainder of occipital lobe	131,852	0.54	0.73
Parietal Lobe			
Superior parietal gyrus	130,908	0.54	0.74
Remainder of parietal lobe (including supramarginal and angular gyrus)	131,972	0.52	0.75
Postcentral gyrus	89,087	0.52	0.74
Central Structures			
Nucleus accumbens	888	0.89	0.99
Thalamus	20,953	0.61	0.79
Putamen	14,502	0.60	0.74
Pallidum (globus pallidus)	3,835	0.58	0.69
Caudate nucleus	12,229	0.56	0.67

Table 3. Logistic regression and Cox regression analysis for age gap and incident dementia. Association of gap between predicted brain age and chronological age with incident dementia assessed by logistic regression and Cox proportional hazards models, both in the total study sample and in a subsample with a minimum follow-up time of 5 years.

	Logistic Regression			Cox Regression		
Model	n/N	OR (95% CI)	p-value	n/N	HR (95% CI)	p-value
			Total sample			
Model I	159/1808	1.15 (1.10-1.20)	2.67×10^{-10}	159/1808	1.15 (1.11-1.20)	1.0×10^{-12}
Model II	154/1790	1.11 (1.06-1.16)	2.57×10^{-5}	154/1790	1.11 (1.07-1.16)	4.6×10^{-7}
Model III	150/1714	1.11 (1.05-1.16)	4.80×10^{-5}	150/1714	1.11 (1.06-1.15)	1.2×10^{-6}
			Sample follow-up time > 5 years			
Model I	62/1366	1.11 (1.04-1.18)	1.26×10^{-3}	62/1366	1.13 (1.06-1.20)	1.4×10^{-4}
Model II	60/1352	1.09 (1.02-1.16)	1.43×10^{-2}	60/1352	1.10 (1.03-1.17)	3.2×10^{-3}
Model III	58/1305	1.09 (1.01-1.16)	2.08×10^{-2}	58/1305	1.09 (1.02-1.17)	7.2×10^{-3}

Model I: age + sex.
Model II: model I + grey matter volume + intracranial volume.
Model III: model II + years of education + APOEε4 carrier status.
Abbreviations: confidence interval (CI); odds ratio (OR); hazard ratio (HR); number of cases (n); total number of participants (N).

logistic regression model, associations were similar in the subsample with a follow-up time for incident dementia of more than 5 years, model III HR=1.09 (95% CI 1.02-1.17) per year age gap. Additionally, **Figure 7** shows the Kaplan-Meier curves for the test set separated by age gap[35]. A clear transition can be seen from the higher gap group to lower gap groups, coinciding with an increasing probability of being free of dementia in follow-up.

3.4 Gap-associated features

Table 4 shows a list of features that can affect the brain pathology and may be associated with the gap[9]. Significantly lower values were found for GM volume in the highest quintile. However, systolic blood pressure and mild cognitive impairment were already only nominally significant, after Bonferroni correction[36].

4. Discussion

In a large sample of community-dwelling middle-aged and older adults, using a DL model for brain age prediction on MRI-derived grey matter tissue density, we found that the gap between predicted brain age and chronological age was related to an increased risk of dementia, independent of other known risk factors for dementia.

Our trained CNN model showed a similar performance in age prediction compared to previous studies that use a multimodal data model[5] and DL model[6], which achieved

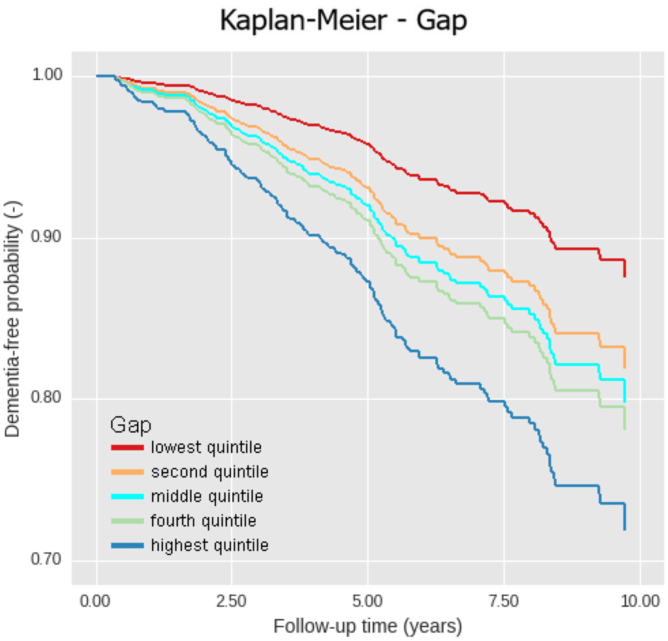


Figure 7. Kaplan-Meier curves presenting the dementia-free probability over time for participants with different age gap values, divided into quintiles. Low gap values correspond to chronological ages surpassing brain age, whereas high gap values correspond to chronological ages that are lower than the brain age.

Table 4. Characteristics comparison of subjects grouped by age gap. Groups consist 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.

Characteristic	Value lowest quintile (n=340)	Value highest quintile (n=350)	p-value
Age gap (years)	-5.7 ± 3.9	6.9 ± 4.5	<0.001
Grey matter volume (mL)	605 ± 56.9	577.6 ± 56.2	<0.001
Systolic blood pressure (mmHg)	138.9 ± 21.6	143.1 ± 21.0	0.009
Mild cognitive impairment, n (%)	15 (4.4)	31 (8.9)	0.013
Diastolic blood pressure (mmHg)	82.1 ± 10.8	84.1 ± 11.1	0.014
Fasting glucose level (mmol/L)	5.5 ± 1.2	5.7 ± 1.1	0.021
Current or past smoker, n (%)	102 (30.0)	130 (37.1)	0.027
Body mass index (kg/m^2)	27.2 ± 3.9	27.8 ± 4.5	0.043
Mini-Mental State Examination score	28.0 ± 1.8	27.7 ± 2.1	0.095
Total cholesterol (mmol/L)	5.6 ± 1.0	5.5 ± 1.1	0.323
APOEε4 carrier, n (%)	92 (27.1)	103 (29.4)	0.418
Female, n (%)	187 (55.0)	203 (58.0)	0.428
HDL cholesterol (mmol/L)	1.4 ± 0.4	1.5 ± +- 0.4	0.549
Age (years)	65.5 ± 10.8	65.3 ± 11.0	0.771
Years of education	12.4 ± 3.8	12.3 ± 4	0.829
Intracranial volume (mL)	1465.8 ± 163.2	1466.3 ± 164.1	0.971

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

performances of MAE=4.29 and MAE=4.16, respectively. Previous studies looked cross-sectionally[5,6] at the association of the age gap and dementia occurrence, but in this study we were also able to look at this association longitudinally. As non-reversible pathological changes already occur years prior to diagnosis, identifying early-stage biomarkers for dementia is of importance. The gap has potential to be utilized alongside other clinical risk factors and biomarkers to separate the population into categories with sufficiently distinct degrees of risk to drive clinical or personal decision-making, e.g. dementia screening and informed life planning.

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 **Figure C-2**. Here we present the performance in gap on male and female subjects, of one model including sex covariate and one without. 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). Also bringing the mean gap for males and females closer to zero (one-sample t-test: p_{male}=0.88 and p_{female}=0.05).

Moreover, we retrieved attention maps from the model. These maps show which brain regions are most important for age prediction, which also provides insights into processes in aging and neurodegeneration. In that regard, literature[2,29] has reported that aging affects the entire GM volume in the brain, as is also confirmed by the attention maps retrieved from the model shown in **Figure C-3**, but more 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. From this we know that the gap can hold information on differences in specific brain regions linked to the age prediction, compared to exclusively using GM volume for predicting. According to literature[2,29] the insula, superior temporal areas and multiple gyri have shown significant age-related GM volume difference. However, amongst these, most of the larger regions were often only partially highlighted by the network. On the other hand, brain structures less significantly affected by age in literature[2,29], were more highlighted by the network, e.g.: caudate nuclei, amygdalae, hippocampi and thalami.

The amygdalae and (parts of the) hippocampi in particular proved to be more associated to age prediction and also increased in attention map intensity in older subjects. This is in accordance to literature where significant negative associations between GM volume and age have been reported for these regions[2,29]. Atrophy of these two structures also has also shown to be more prevalent in dementia patients, including years before diagnosis[37,38].

Limitations

We were not able to perfectly predict the age based only on MRI for healthy subjects. We assume that due to biological similarity of the brain within a range of several years, there will always be an according level of uncertainty in the age prediction.

Furthermore, although we excluded subjects with dementia and stroke while training the model, there are a number of other factors that can affect overall or local GM volume, in turn affecting the age prediction and gap (**Table 4**). These additional features can introduce bias, which may be solved by adding the information as a covariate to the model. This however requires the respective information on the subjects, which can make the method less accessible for general use.

We were unable to utilize the full scale of the input data. In terms of DL implementation, the model uses larger receptive fields and strides in the first layer compared to following layers. This was due to the restricted computational power (GPU memory) that was available to train the network. Thus, the model might have excluded valuable finer details in the input data.

Lastly, the current CNN model is incapable of handling unfamiliar datasets, limiting its practical use. A drawback of CNN's is that the training data should be representative for the data for which the trained network is used. Thus limiting the generalizability of our method. However, this can be addressed by training models on more diverse or new datasets.

Future recommendations

The CNN built for this study, uses a fairly standard architecture. It may be valuable to investigate more complex CNN architectures for this application, as smaller details might not have been noticed by our model. In that regard, our current approach uses minimally pre-processed GM density maps as input. It would be interesting to test models for raw T1-weighted brain MRIs, for the purpose of using non-preprocessed MRIs for gap estimation. Note that this however may complicate visualization analysis, as images are no longer segmented and registered.

This study has investigated the association between the age gap and incident dementia, whilst adjusting for five known biomarkers, i.e. age, sex, GM volume, ICV, years of education and APOEε4 carriership. However, it is still required to investigate the correlation of age gap to other biomarkers to prove whether it is an independent risk biomarker for incident dementia. Additionally, it would be interesting to compare the power of the age gap as a risk biomarker to other known biomarkers.

As mentioned in Limitations, several cofactors also influence brain pathology and can affect age estimation and gap. Further research is needed into these gap-associated features, which may explain gap differences. Investigation of the exact association between these features and age gap might also be interesting when regarding human lifestyle. On the other hand, to try and get a more accurate gap correlation to dementia, more of these features can be introduced as covariates in the architecture. Although, this requires the study to consider the practical application of acquiring and using such additional features.

Lastly, this study has investigated association between age gap and incidence of dementia, but that is not to say that dementia is the only neurodegenerative disease that can be traced by this variable. Follow-up studies can include early-stages of other neurodegenerative diseases or even mental diseases for investigation after their association with age gap, to improve our understanding of the human brain pathology.

5. Conclusion

We showed that the gap between predicted and chronological brain age is a biomarker associated with a risk of dementia development. DL visualization allows further investigation of the gap and neurodegeneration with respect to the human brain. This suggests that the age gap may be applicable for dementia risk screening, but there is still room for improvement of the model and for further research into the association between gap and brain diseases.

Acknowledgments

This research was made possible by The Erasmus Medical Center, Rotterdam. I thank my direct supervisors (Gennady V. Roshchupkin and Wiro J. Niessen) for their continued guidance during this project. I thank the other co-authors of the original paper (Maria J. Knol, Aleksei Tiulpin, Florian Dubost, Marleen de Bruijne, Meike W. Vernooij, Hieab H.H. Adams and M. Arfan Ikram) for their insights, expertise and contributions that greatly assisted in this research. Lastly, I thank the many colleagues from the department of medical informatics who provided their insights on this research.

References

1. Vinke, E. J. *et al.* Trajectories of imaging markers in brain aging: The Rotterdam Study. *Neurobiol. Aging* (2018). doi:10.1016/j.neurobiolaging.2018.07.001
2. Manard, M., Bahri, M. A., Salmon, E. & Collette, F. Relationship between grey matter integrity and executive abilities in aging. *Brain Res.* **1642,** 562–580 (2016).
3. Abbott, A. Dementia: A problem for our age. *Nature* **475,** (2011).
4. Franke, K., Luders, E., May, A., Wilke, M. & Gaser, C. Brain maturation: Predicting individual BrainAGE in children and adolescents using structural MRI. *Neuroimage* **63,** 1305–1312 (2012).
5. Liem, F. *et al.* Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage* **148,** 179–188 (2017).
6. Cole, J. H. *et al.* Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage* **163,** 115–124 (2017).
7. Kaufmann, T. *et al.* Genetics of brain age suggest an overlap with common brain disorders. *bioRxiv* (2018). doi:10.1101/303164
8. Holmes, G. L., Milh, M. D. M. & Dulac, O. *Maturation of the human brain and epilepsy*. *Handbook of Clinical Neurology* **107,** (2012).
9. Franke, K., Gaser, C., Manor, B. & Novak, V. Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Front. Aging Neurosci.* **5,** (2013).
10. Good, C. D. *et al.* A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* **14,** 21–36 (2001).
11. Roshchupkin, G. V. *et al.* Fine-mapping the effects of Alzheimer's disease risk loci on brain morphology. *Neurobiol. Aging* **48,** 204–211 (2016).
12. Ikram, M. A. *et al.* The Rotterdam Scan Study: design update 2016 and main findings. *Eur. J. Epidemiol.* **30,** 1299–1315 (2015).
13. Matsunari, I. *et al.* Comparison of 18F-FDG PET and Optimized Voxel-Based Morphometry for Detection of Alzheimer's Disease: Aging Effect on Diagnostic Performance. *J. Nucl. Med.* **48,** 1961–1970 (2007).
14. Vrooman, H. A. *et al.* Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. *Neuroimage* **37,** 71–81 (2007).
15. Smith, S. M. & Nichols, T. E. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* **44,** 83–98 (2009).
16. Mutlu, U. *et al.* Association of Retinal Neurodegeneration on Optical Coherence Tomography With Dementia. *JAMA Neurol.* 1–8 (2018). doi:10.1001/jamaneurol.2018.1563
17. McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease. *Neurology* **34,** 939 (1984).
18. Román, G. *et al.* Vascular dementia: diagnostic

criteria for research studies. *Neurology* **43,** 250–260 (1993).

19. Seripa, D. *et al.* TOMM40, APOE, and APOC1 in primary progressive aphasia and frontotemporal dementia. *J. Alzheimer's Dis.* (2012). doi:10.3233/JAD-2012-120403
20. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86,** 2278–2323 (1998).
21. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1–9 (2012). doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007
22. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv Prepr.* 1–10 (2014). doi:10.1016/j.infsof.2008.09.005
23. Williams, S. Pearson's correlation coefficient. *The New Zealand medical journal* (1996). doi:10.1136/bmj.e4483
24. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proc. IEEE Int. Conf. Comput. Vis.* **2017–Octob,** 618–626 (2017).
25. Preacher, K. J., Curran, P. J. & Bauer, D. J. Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis. *J. Educ. Behav. Stat.* **31,** 437–448 (2006).
26. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86,** 420–428 (1979).
27. DiCiccio, T. J. & Efron, B. Bootstrap confidence intervals. *Stat. Sci.* (1996). doi:10.1214/ss/1032280214
28. Ruitenberg, A., Ott, A., Van Swieten, J. C., Hofman, A. & Breteler, M. M. B. Incidence of dementia: Does gender make a difference? *Neurobiol. Aging* (2001). doi:10.1016/S0197-4580(01)00231-7
29. Matsuda, H. Voxel-based Morphometry of Brain MRI in Normal Aging and Alzheimer's Disease. *Aging Dis.* **4,** 29–37 (2013).
30. Roses, A. D. & Saunders, A. M. APOE is a major susceptibility gene for Alzheimer's disease. *Curr. Opin. Biotechnol.* **5,** 663–667 (1994).
31. Rossum, G. Van & Drake, F. L. Python Reference Manual. *Python Software Foundation* (2001). Available at: http://www.python.org.
32. Ascher, D., Dubois, P., Hinsen, K., Hugunin, J. & Oliphant, T. Numerical Python. *Lawrence Livermore National Laboratory* (2001). Available at: http://www.pfdubois.com/numpy/.
33. Chollet, F. Keras. *Github repository* (2015). Available at: https://github.com/fchollet/keras.
34. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* Available at: http://www.r-project.org/.
35. Rich, J. T. *et al.* A practical guide to understanding Kaplan-Meier curves. *Otolaryngol. - Head Neck Surg.* **143,** 331–336 (2010).
36. Abdi, H. The Bonferonni and Šidák Corrections for Multiple Comparisons. *Encycl. Meas. Stat.* 103–107 (2007). doi:10.4135/9781412952644
37. Aylward, E. H. *et al.* MRI volumes of the hippocampus and amygdala in adults with Down's syndrome with and without dementia. *Am. J. Psychiatry* (1999). doi:10.1176/ajp.156.4.564
38. Wachinger, C., Salat, D. H., Weiner, M. & Reuter, M. Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala. *Brain* (2016). doi:10.1093/brain/aww243
39. Sathya, R. & Abraham, A. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *Int. J. Adv. Res. Artif. Intell.* **2,** 34–38 (2013).
40. Xinghuo Yu, M. Onder Efe, and O. K. A General Backpropagation Algorithm for Feedforward Neural Networks Learning. *IEEE Trans. Neural Networks* **13,** 251–254 (2002).
41. Gill, J. K. Automatic Log Analysis using Deep learning and AI - XenonStack. (2017). Available at: https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/. (Accessed: 19th December 2018)
42. Xue-Wen Chen & Xiaotong Lin. Big Data Deep Learning: Challenges and Perspectives. *IEEE Access* **2,** 514–525 (2014).
43. Işin, A., Direkoğlu, C. & Şah, M. Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods. *Procedia Comput. Sci.* **102,** 317–324 (2016).
44. Ker, J., Wang, L., Rao, J. & Lim, T. Deep Learning Applications in Medical Image Analysis. *IEEE Access* 1–1 (2018). doi:10.1109/ACCESS.2017.2788044
45. Szegedy, C. *et al.* Going Deeper with Convolutions. 1–9 (2014). doi:10.1109/CVPR.2015.7298594

46. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Miccai* 234–241 (2015). doi:10.1007/978-3-319-24574-4_28
47. Lin, M., Chen, Q. & Yan, S. Network In Network. *arXiv Prepr.* 10 (2014). doi:10.1109/ASRU.2015.7404828
48. Perez, L. & Wang, J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. (2017).
49. Kingma, D. P. & Ba, J. L. Adam: A Method for Stochastic Optimization. *ICLR Conf. Proc.* **1631,** 13–15 (2015).
50. Ruder, S. An overview of gradient descent optimization algorithms. 1–14 (2016). doi:10.1111/j.0006-341X.1999.00591.x
51. Cox, D. R. The Regression Analysis of Binary Sequences. *J. R. Stat. Soc. Ser. B* **20,** 215–242 (1958).
52. Fox, J. Cox Proportional-Hazards Regression for Survival Data. *Append. to An R S-PLUS Companion to Appl. Regres.* (2002). doi:10.1016/j.carbon.2010.02.029

Appendix A. Deep learning and convolutional neural networks

Appendix A.1. Deep learning

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 each iteration's trained model performance, to create an optimized model using backpropagation by supervised or unsupervised learning. 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.

Deep Learning is a subset of machine learning, which is a form of artificial intelligence often used to develop models. Compared to classical model building, machine learning techniques require a set of input and respective output to find and optimize a non-linear relation between the two. By providing this data to a set of algorithms, the method is able to train a by the user designed model, as illustrated in **Figure A-1**. Generally, the user designs the model architecture by selecting the model components. After which the machine learning method optimizes the model, i.e. model parameters are iteratively adjusted using backpropagation according to a loss function, to create an optimal model by supervised or unsupervised learning[39,40].

In machine learning, the input is usually made up of a set of user-defined features that are correlated to the output. However, deep learning allows the model itself to choose which relevant features to extract from the input, shown in **Figure A-2**[41]. Although this method requires a lot of data for training, it facilitates the model to freely search the input-space and find the most important, possibly new, input features.

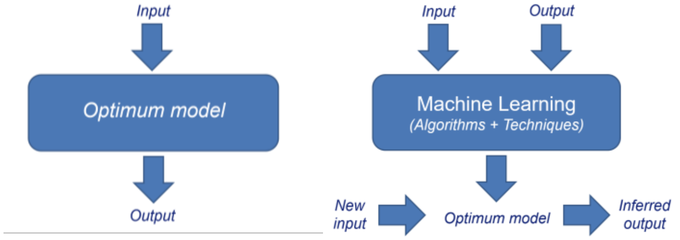


Figure A-1. Graphic representation of a classical model (left) and machine learning model (right). Classical model requires the user to optimize the model, whereas machine learning optimizes the model for the user.

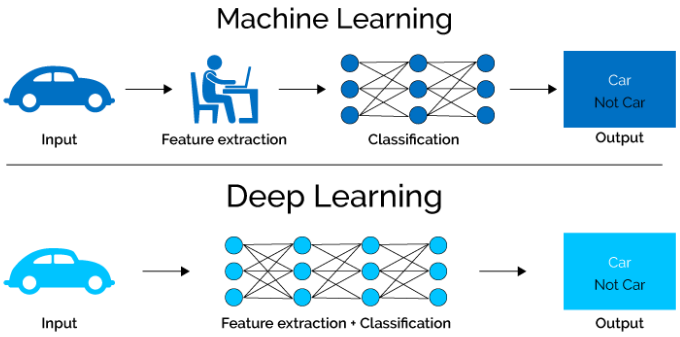


Figure A-2. Graphic representation of feature extraction in machine learning and deep learning[41]. Machine learning techniques require user specified features (e.g. wheels, windows, etc.) extracted from the input as input for the classification method. Deep learning techniques exclude the user from the process, take in entire inputs and perform feature extraction alongside classification in its method.

Appendix A.2. Convolutional Neural Networks

Convolutional neural networks (CNNs) are a class amongst deep learning techniques. They allow multi-dimensional input images and inspect these inputs by scanning them for relevant information[20,21]. 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[42–44].

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), as illustrated in **Figure A-3**. 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 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, fully connected layers, batch normalization and dropout[21,44].

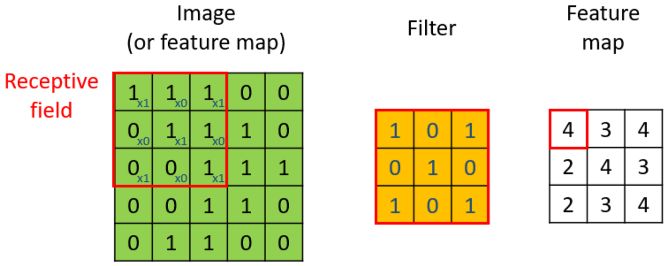


Figure A-3. Principle of convolutional layers. An image can be seen a field of numbers. A CNN uses small loops called filters to calculate and produce new images, called feature maps, which hold information on the image.

Appendix A.3. Selecting network architecture

Several different CNN model architectures were built, trained and tested in this study. Examples amongst the models were very deep CNNs, CNNs with many feature maps and U-Nets[22,45,46]. All models achieved similar performances in MAE between 4.20-4.80 years and ICC of around 0.90. As such we decided on our current CNN model based on the good overall performance and simplicity of the architecture, which allows visualization by Grad-CAM and a better interpretation of the results.

As mentioned in Section 2.3, we built a 3D regression CNN model to predict brain age using brain MRI of GM voxels and the sex covariate as input. Thus, our architecture consists of two branches. The first branch can be described as four convolutional blocks ending on a pooling layer. The first layer takes the input image through a 5x5x5 convolutional layer with strides of 2. This is done to effectively decrease feature map sizes, allowing the network to accept larger resolution input images allowed by GPU memory space. Hereafter, convolutional blocks use 3x3x3 convolution with strides of 1, as recommended by literature[22]. Each block ends on a MP layer, which sequentially halves feature map dimensions but increases their number from 32 to 48, 64 and 80, respectively. The last convolutional block applies global average pooling to extract the final feature maps to a one-dimensional array of a single variable per feature map[47].

After, the first branch is merged with the second branch, which consists of the binary input sex. Merging is performed by concatenation between the array of 80 single variable feature map representations and sex input, finally followed by one more fully connected layers of 32 channels to propagate to a single regression output.

Appendix A.4. Network training

The CNN has been trained using the data from the training set of 3848 subjects. Here, over- and undersampling had been applied to the training set. Thus effectively using data of 3688 subjects to distribute the samples more evenly over the age range of the population ($N_{\text{img,train\_balanced}}$=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[48]. 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.

For optimization the performance is measured in model accuracy based on the mean squared error (MSE) of the prediction

$$\mathrm{MSE} = \frac{1}{N}\sum_{S}(\mathrm{gap}_{S})^{2} \tag{5}$$

, as MSE penalizes outliers more than MAE. The model is optimized by the iterative process of supervised learning by backpropagation, using an Adaptive Moment Estimation (Adam) optimizer as its loss function[49]. Adam handles an adaptive learning rate whilst carrying momentum. The resulting optimization function is known to be robust whilst the hyper-parameters typically require little tuning, making it easy to implement[50]. Hyper-parameter tuning was done empirically and the best model was selected based on its performance on the validation set.

Appendix B. Analysis methods

Appendix B.1. Logistic regression analysis

Logistic regression is a machine learning and statistical model approach, which is able to explain binary dependent variables[51]. It is used to relate a binary outcome to one or more variables, by fitting a logistic function (a variant of the sigmoid function) on the data samples as shown in **Figure B-1**. The logistic function for a single variable x is written as

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{6}$$

, with probability p and model coefficients β_0 and β_1. This is expanded for multiple m variables to

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}} \tag{7}$$

, with additional model coefficients β_i for every explanatory variable x_i $(i = \{1,2,\dots,m\})$.

The relationship between probability and variable can be derived from **Equation 7**. For each variable the odds ratio (OR) can be defined as

$$\begin{gathered}\text{odds}_i = e^{\beta_0 + \beta_i x_i} \\ \text{OR}_i = \frac{\text{odds}(x_i + 1)}{\text{odds}(x_i)} = e^{\beta_i}\end{gathered} \tag{8}$$

. The OR_i provides an interpretation for β_i, as e^{β_i} indicate the incease in probability by multiplication for every 1-unit increase in x.

For each variable, the p-value can be computed in logistic regression, indicating whether changes in the variable are associated with a significant increase or decrease in probability. Note that for multiple variables, this means high p-values may indicate that the difference is already explained by another variable in the model.

Thus, logistic regression is utilized here as a tool to compare the gap with other known biomarkers in terms of their correlation to follow-up for dementia.

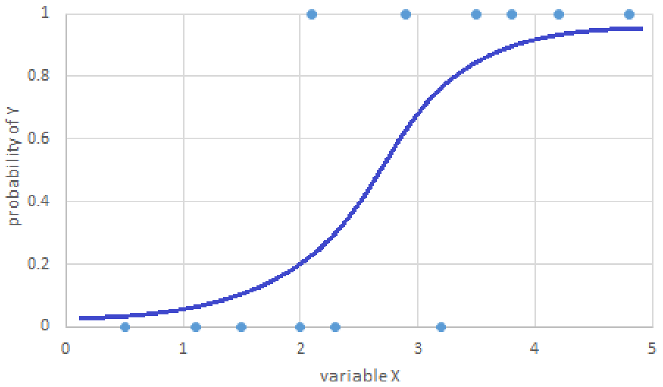


Figure B-1. Graph of a logistic regression curve showing probability of outcome Y versus input variable X.

Appendix B.2. Cox proportional hazards regression analysis

Proportional hazards models (a.k.a. Cox models or PH models) are a class of survival analysis models in statistics, which examines the time it takes for events to occur[52]. Such models are able to relate the time passed till the occurrence of an event to one or more variables, which may be associated with that quantity of time. Similar to logistic regression (Appendix B.1), it fits a so called hazard function to the data that has the form

$$\begin{gathered}h_i(t) = e^{\alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{im}} \\ h_i(t) = h_0(t) e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{im}}\end{gathered} \tag{9}$$

, which includes hazard h_i at time t for observation i and constant α as a baseline hazard $h_0(t)$. β_i are model coefficients for every explanatory variable x_{ij} $(j = \{1,2,\dots,m\})$. Following, similar to the OR, we find the hazard ratio (HR) as

$$\begin{gathered}\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{im} \\ \text{HR}_i = \frac{h_i(t)}{h_{i'}(t)} = \frac{e^{\eta_i}}{e^{\eta_{i'}}}\end{gathered} \tag{10}$$

, which considers two observations i and i' that differ in x_i. The HR_i represents the risk of even in observation group i compared to group i'.

If the HR is close to 1, then that respective variable does not affect survival; less than 1, then that variable is protective and associated with improved survival; or greater than 1, then that variable is associated with increased risk or decreased survival. Also in accordance to the logistic regression analysis, p-values are computed for each variable.

Appendix C. Additional figures

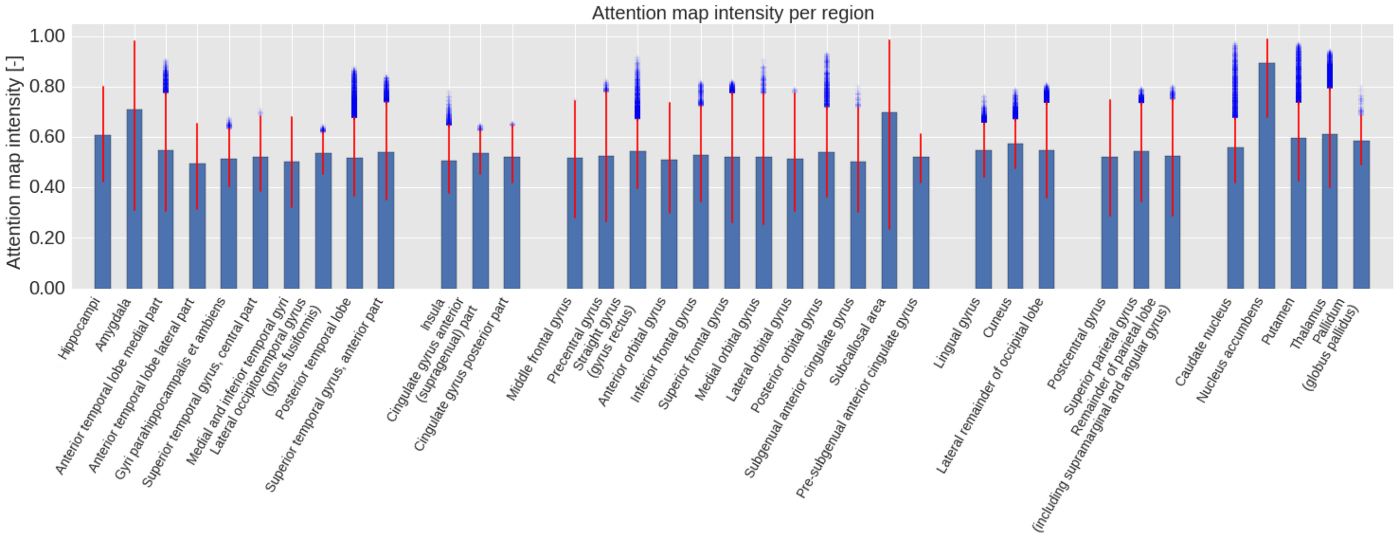


Figure C-1. Bar plot of attention map values per region. Bars show mean of each region. Variance from first to fifth quintile (upper and lower boundary, respectively) are indicated in red. Outliers are indicated in dark blue..

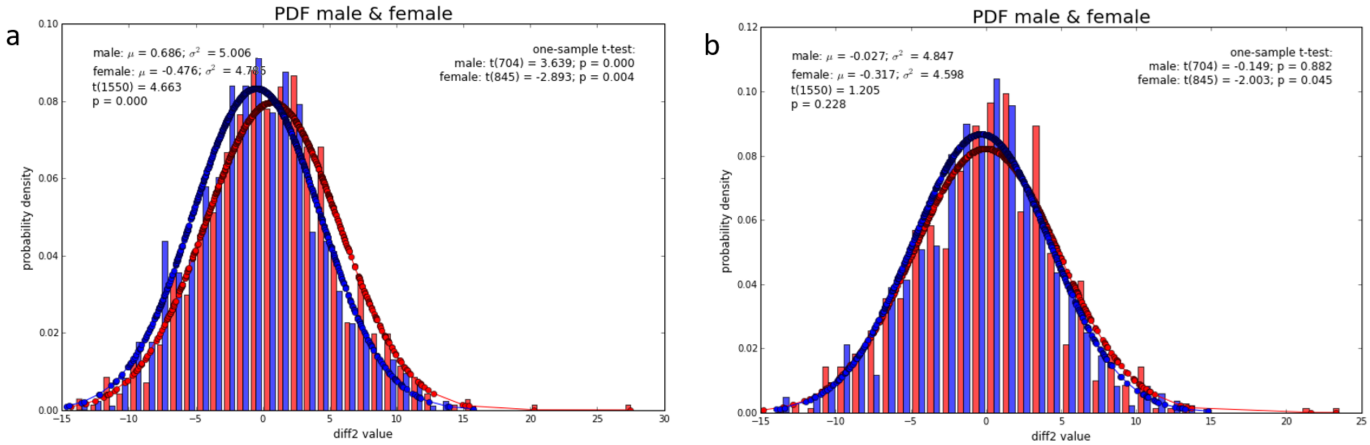


Figure C-2. Effect on the gap value distribution (red=male; blue=female) of adding sex as a covariate to the model. A comparison of the probability density functions for gap of two early trained models along with their respective t-test results. Both models have the exact same architecture with one the 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.

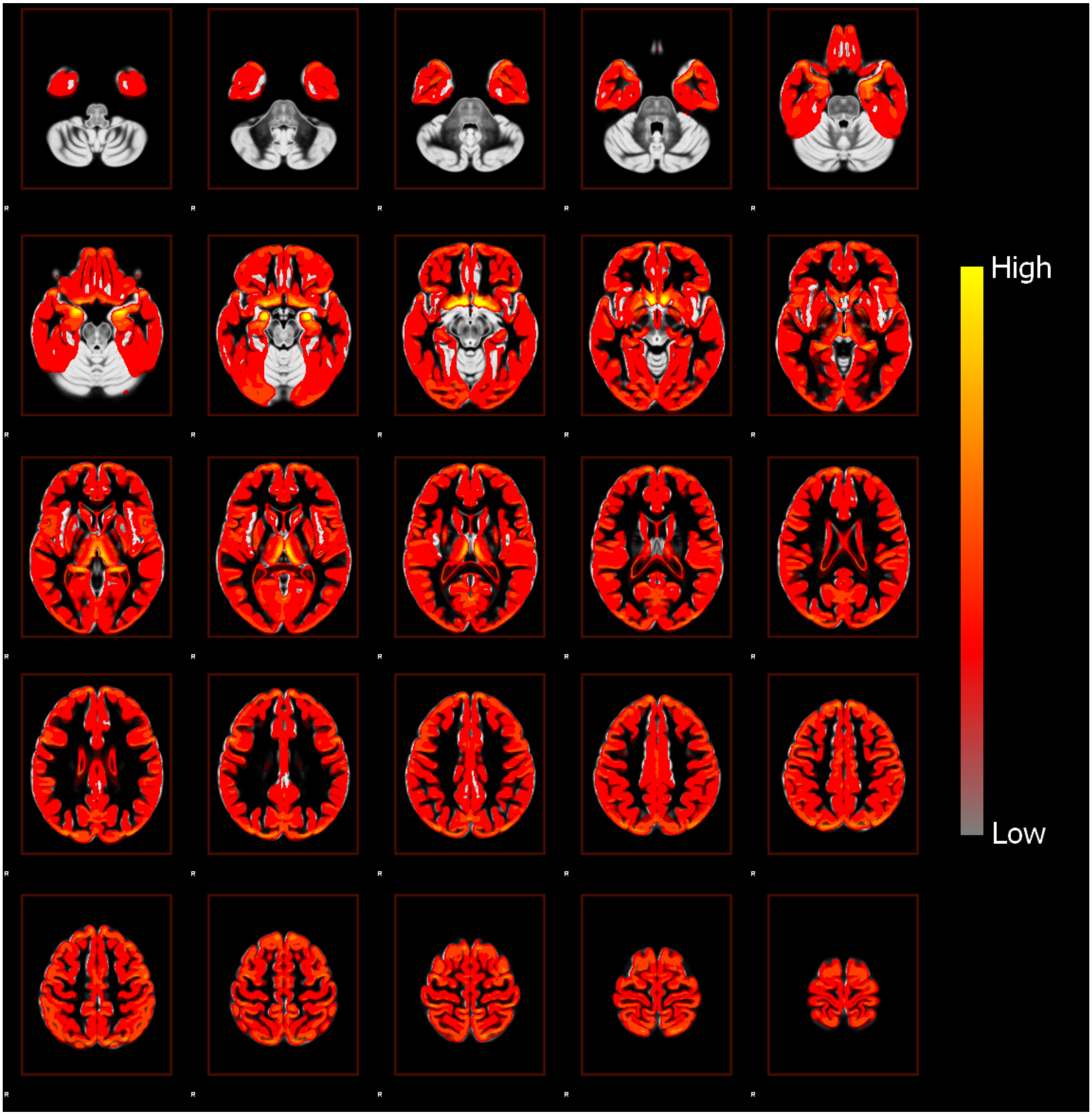


Figure C-3. Grad-CAM attention map intensity per voxel overlaid on a brain template. Voxel values in the attention map have been set at 0.50 minimum and 1.00 maximum threshold to exclude background values and focus on more highlighted regions, according to normalization around 0.50 in the Grad-CAM implementation.

Appendix D. Additional tables

Table D-1. Quantitative description of the increase in attention map. Number of positive voxels (intensity increase) and negative voxels (intensity decrease), and their respective fraction of region size are shown per brain region. Brain regions are grouped by lobes.

Brain region	Attention map intensity increase		Attention map intensity decrease	
	fraction	(n/N)	fraction	(n/N)
Temporal Lobe				
Amygdala	0.48	(2,093/4,398)	0.01	(65/4,398)
Hippocampus	0.31	(2,056/6,687)	0.00	(0/6,687)
Gyri parahippocampalis et ambiens	0.24	(3,244/13,767)	0.01	(98/13,767)
Lateral occipitotemporal gyrus (gyrus fusiformis)	0.22	(2,798/12,908)	0.00	(0/12,908)
Posterior temporal lobe	0.20	(28,128/143,237)	0.00	(113/143,237)
Anterior temporal lobe medial part	0.18	(4,190/22,842)	0.02	(418/22,842)
Superior temporal gyrus, anterior part	0.17	(2,435/14,369)	0.03	(420/14,369)
Superior temporal gyrus, central part	0.12	(5,321/42,794)	0.00	(65/42,794)
Medial and inferior temporal gyri	0.12	(6,529/55,102)	0.00	(0/55,102)
Anterior temporal lobe lateral part	0.04	(467/11,999)	0.00	(2/11,999)
Insula and Cingulate gyri				
Insula	0.55	(24,188/44,328)	0.03	(1,199/44,328)
Cingulate gyrus posterior part	0.43	(10,368/24,235)	0.01	(245/24,235)
Cingulate gyrus anterior (supragenual) part	0.31	(7,648/24,751)	0.00	(53/24,751)
Frontal Lobe				
Subgenual anterior cingulate gyrus	0.53	(2,271/4,287)	0.00	(2/4,287)
Straight gyrus (gyrus rectus)	0.43	(5,033/11,826)	0.00	(27/11,826)
Subcallosal area	0.43	(335/788)	0.13	(106/788)
Pre-subgenual anterior cingulate gyrus	0.24	(597/2,451)	0.00	(0/2,451)
Superior frontal gyrus	0.19	(31,890/166,766)	0.01	(1,407/166,766)
Inferior frontal gyrus	0.16	(8,875/55,754)	0.01	(560/55,754)
Medial orbital gyrus	0.14	(2,603/18,554)	0.02	(410/18,554)
Middle frontal gyrus	0.12	(19,296/161,999)	0.00	(464/161,999)
Precentral gyrus	0.12	(12,646/106,145)	0.00	(426/106,145)
Posterior orbital gyrus	0.08	(1,273/15,061)	0.01	(205/15,061)
Anterior orbital gyrus	0.01	(181/19,514)	0.02	(389/19,514)
Lateral orbital gyrus	0.01	(165/11,112)	0.00	(10/11,112)
Occipital Lobe				
Lingual gyrus	0.15	(5,618/36,627)	0.00	(86/36,627)
Cuneus	0.13	(3,645/28,209)	0.00	(29/28,209)
Lateral remainder of occipital lobe	0.13	(16,571/131,852)	0.00	(27/131,852)
Parietal Lobe				
Superior parietal gyrus	0.17	(22,515/130,908)	0.00	(145/130,908)
Postcentral gyrus	0.12	(10,703/89,087)	0.00	(163/89,087)
Remainder of parietal lobe (including supramarginal and angular gyrus)	0.10	(13,458/131,972)	0.00	(129/131,972)
Central Structures				
Pallidum (globus pallidus)	0.81	(3,113/3,835)	0.00	(0/3,835)
Putamen	0.72	(10,508/14,502)	0.01	(90/14,502)
Thalamus	0.52	(10,988/20,953)	0.02	(421/20,953)
Nucleus accumbens	0.51	(451/888)	0.02	(18/888)
Caudate nucleus	0.30	(3,645/12,229)	0.01	(62/12,229)

Abbreviations: number of positive/negative voxels in region (n); region size in number of voxels (N).