Deep neural networks learn general and clinically relevant representations of the ageing brain

Esten H. Leonardsen [a,b,*], Han Peng [c], Tobias Kaufmann [b,d], Ingrid Agartz [b,e,f], Ole A. Andreassen [b], Elisabeth Gulowsen Celius [g,h], Thomas Espeseth [a,i], Hanne F. Harbo [g,h], Einar A. Høgestøl [a,b,g], Ann-Marie de Lange [a,j,k], Andre F. Marquand [l], Didac Vidal-Piñeiro [a], James M. Roe [a], Geir Selbæk [m,n], Øystein Sørensen [a], Stephen M. Smith [c], Lars T. Westlye [a,b,o,1], Thomas Wolfers [a,b,1], Yunpeng Wang [a,1]

[a] Department of Psychology, University of Oslo, Oslo, Norway
[b] Norwegian Centre for Mental Disorders Research (NORMENT), Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[c] Wellcome Centre for Integrative Neuroimaging (WIN FMRIB), University of Oxford, Oxford, OX3 9DU, United Kingdom
[d] Department of Psychiatry and Psychotherapy, Tübingen Center for Mental Health, University of Tübingen, Germany
[e] Department of Psychiatric Research, Diakonhjemmet Hospital, Oslo, Norway
[f] Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet & Stockholm Health Care Services, Stockholm County Council, Stockholm, Sweden
[g] Department of Neurology, Oslo University Hospital, Norway
[h] Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[i] Department of Psychology, Bjørknes University College, Oslo, Norway
[j] LREN, Centre for Research in Neurosciences-Department of Clinical Neurosciences, CHUV and University of Lausanne, Lausanne, Switzerland
[k] Department of Psychiatry, University of Oxford, Oxford, UK
[l] Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, Netherlands
[m] Norwegian National Advisory Unit on Aging and Health, Vestfold Hospital Trust, Tønsberg, Norway
[n] Department of Geriatric Medicine, Oslo University Hospital, Oslo, Norway
[o] KG Jebsen Center for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway

ABSTRACT

The discrepancy between chronological age and the apparent age of the brain based on neuroimaging data — the brain age delta — has emerged as a reliable marker of brain health. With an increasing wealth of data, approaches to tackle heterogeneity in data acquisition are vital. To this end, we compiled raw structural magnetic resonance images into one of the largest and most diverse datasets assembled (n=53542), and trained convolutional neural networks (CNNs) to predict age. We achieved state-of-the-art performance on unseen data from unknown scanners (n=2553), and showed that higher brain age delta is associated with diabetes, alcohol intake and smoking. Using transfer learning, the intermediate representations learned by our model complemented and partly outperformed brain age delta in predicting common brain disorders. Our work shows we can achieve generalizable and biologically plausible brain age predictions using CNNs trained on heterogeneous datasets, and transfer them to clinical use cases.

Introduction

Neurodevelopmental and age-related changes in the brain play a crucial role in the etiology of complex neurological Stephan and Brayne (2009) and mental disorders (Thapar and Riglin, 2020). Predictive models for an individual's age based on magnetic resonance imaging (MRI) have been used to estimate normative trajectories across the lifespan (Cole and Franke, 2017; Cole et al., 2017; Franke et al., 2012; 2010). Individual deviations from these trajectories, often called the brain age delta, have been linked to brain health (Franke and Gaser,

2019; Kaufmann et al., 2019), and more extreme deviations observed in patients with schizophrenia (SCZ) (Nenadi et al., 2017; Rokicki et al., 2021; Schnack et al., 2016), depression (Han et al., 2020), cognitive impairment (Elliott et al., 2019; Liem et al., 2017), dementia (Wang et al., 2019), Alzheimer's disease (AD) (Gaser et al., 2013) and multiple sclerosis (MS) (Høgestøl et al., 2019), implying that such deviations could be a feasible biological marker for various brain disorders.

The brain age of an individual is typically estimated from brain images using statistical learning techniques. The first-generation models were relatively simple, typically based on independent voxels

---

Franke et al. (2010) or a limited number of imaging-derived phenotypes (IDPs) reflecting brain properties such as volumetric measures of different regions (Smith et al., 2019). These models generally estimated linear relationships, were restricted in scale, and were trained on datasets with tens, hundreds or a few thousand participants (Franke and Gaser, 2019). In parallel with continuous computational advances the exponential growth of MRI data has enabled deep learning models of scale for accurate brain age estimation (Abrol et al., 2021; Cole et al., 2017). Deep learning models can take minimally- or non-preprocessed 3-D images as input - avoiding computationally demanding and hypothesis-driven (Oquendo et al., 2012) choices during image processing - and model complex non-linear relations between voxels. One such model is the Simple Fully Convolutional Network (SFCN), a novel deep Convolutional Neural Network (CNN) that won the Predictive Analysis Challenge for brain age prediction in 2019 (PAC2019) (Gong et al., 2021; Peng et al., 2021). While such deep learning approaches allow for the prediction of brain age with unprecedented accuracy, and can potentially help us identify idiosyncratic regional patterns of neurodevelopment and ageing at the individual level (Dinsdale et al., 2021; Hofmann et al., 2021), their complexity also comes with a risk of overfitting, namely finding patterns in the training data which do not generalize well to new, previously unseen, participants (Arbabshirani et al., 2017).

Transfer learning, a deep learning technique widely used in other areas of applied machine learning research, has recently gained momentum in neuroimaging (Valverde et al., 2021). Here, learned intermediate representations can be shared between tasks, allowing a model to be transferred to a problem or a dataset it was initially not trained for (Bengio, 2012). This approach has arguably been one of the core developments underlying the practical success of deep learning in a range of computer vision problems (SharifRazavian et al., 2014) by using models trained on general purpose datasets, typically ImageNet (Russakovsky et al., 2015), and fine-tuning them towards a wide array of tasks. Recent studies have shown that transfer learning yields promising results also for brain age predictions (Jonsson et al., 2019) and clinical classifications based on MRI data (Bashyam et al., 2020; Lu et al., 2020). This exemplifies the need for robust, pretrained models on massive multisite datasets, that can be translated to smaller clinical samples, and, ultimately, to individual cases in a clinical setting. Modelling brain age as a pre-training step has obvious advantages as age is a variable which is available in most current MRI datasets. Additionally, the representations learned by the brain age models, representing partly independent dimensions of age-related variance, could be of direct importance in individual level brain phenotyping.

In the present study we trained deep neural network models for brain age prediction on structural MRI data from 53,542 healthy individuals between 3 and 95 years of age to test their ability to generalize, and demonstrate the downstream applicability and biological relevance of a properly generalizing model. We used the Simple Fully Convolutional Network with a softmax output (SFCN-sm) which predicts age as a discrete probability distribution, and compared its accuracy and generalizability with two proposed variants of the architecture; a regression variant (SFCN-reg), directly predicting continuous age, and a ranking variant (SFCN-rank) encoding age as an ordinal vector. Brain age delta was computed as the difference between the predicted and chronological age. To estimate the sensitivity and clinical relevance of our best model, we tested for associations between brain age delta and a range of clinical and biological phenotypes, and sociodemographic and lifestyle variables, in unseen data from a population sample. To further demonstrate the applicability of the model, we employed the pretrained SFCN-reg in a transfer learning context to predict case-control status for AD, MS, mild cognitive impairment (MCI), SCZ, mood disorders and psychotic disorders on datasets obtained from a range of MRI scanners. To promote transparency and reproducibility we have implemented an easy-to-use Keras interface for all the trained models, both the brain age and clinical predictors, and a pipeline for preprocessing images, available on our GitHub http://www.github.com/estenhl/pyment-public.

Results

We compiled 21 publicly available datasets with T1-weighted MRI scans into a large and diverse imaging dataset (total N = 53542; female N = 27715; age range = 3-95), and trained a Simple Fully Convolutional 3-dimensional CNN with 6 convolutional blocks and a softmax output layer (SFCN-sm, Fig. 1b), as introduced in Peng et al (Peng et al., 2021). We then proposed two alternatives for the prediction layer of this architecture, the first based on regression (SFCN-reg) and the second based on ranking (Chen et al., 2017) (SFCN-rank) (see Materials and Methods for details). Due to the time-consuming process of model evaluation, we restricted our search to these three variants of the given architecture, and trained a handful of versions of each variant with different hyperparameter settings. To evaluate the generalization performance of the models, we divided our data at two levels; A reference dataset (Fig. 1 and Supplementary Table 1) and an external dataset (Fig. 2a and Supplementary Table 2). In the reference dataset, we evaluated the performance of the trained models on an independent test split from known scanners with an age distribution resembling the training split. We then tested model performance on the external dataset compiled from different sources, originating from scanners unseen by the models during training with a divergent age distribution.

Superior generalization performance of SFCN-reg

We used the training and validation sets to optimize and tune the models, and a conjunction of the test set and the external dataset in a final model comparison. For each model variant we trained three versions, with different hyperparameter settings, on the training data, and selected the best model based on the mean absolute error (MAE) on the validation set (Fig. 1c and Supplementary Table 8). We then compared the performances of the best version for each model variant on the test set. In this comparison the results mirrored those of the validation set, with SCFN-sm achieving the best result with an MAE of 2.23 years, followed by SCFN-reg with 2.47 years and SCFN-rank with 2.55 years (Fig. 2b). Given the added complexity of including multiple datasets from a large range of scanners, we consider these results to be approximately on par with the MAE of 2.14 reported in the original SFCN paper (Peng et al., 2021), and thus among the best performing models in the field He et al. (2021). Additionally, the heterogeneous origins of the dataset facilitate cross-site generalization, an essential property when training large multisite models (Dockès et al., 2021).

As a conclusive test of model generalization, we performed the same comparison on the external dataset, containing unseen data from different MRI scanners with an age distribution diverging from that of the reference dataset. SCFN-reg substantially outperformed the two alternatives, with an MAE of 3.90 compared to 5.04 and 5.82 for SFCN-sm and SFCN-rank respectively (Fig. 2). While the performance of all models were lower on the external dataset, the extent of the generalization error was considerably different. When compared with MAEs from the test set, the average error of SFCN-reg increased by approximately half, while it more than doubled in SFCN-sm and SFCN-rank (Supplementary Table 9). This difference coincides with the architectural differences between the models: Where both the SFCN-sm and SFCN-rank used age-bins, with an output node for each age in its prediction range, SFCN-reg had a single output predicting a single continuous number. Therefore, the predictions of the SFCN-reg reflect a simpler combination of the learned representations in the preceding layer of the model, which we hypothesize may be the reason for the improved generalization performance. In the subsequent applications we use the SFCN-reg version that achieved the best MAE on the external dataset. Additionally, to facilitate cross-study comparisons, we have compiled a range of performance metrics for our three models on the external dataset in Table 1.

To better understand the visual patterns underlying the predictions of the SFCN-reg we conducted two post-hoc correlational analyses. First,
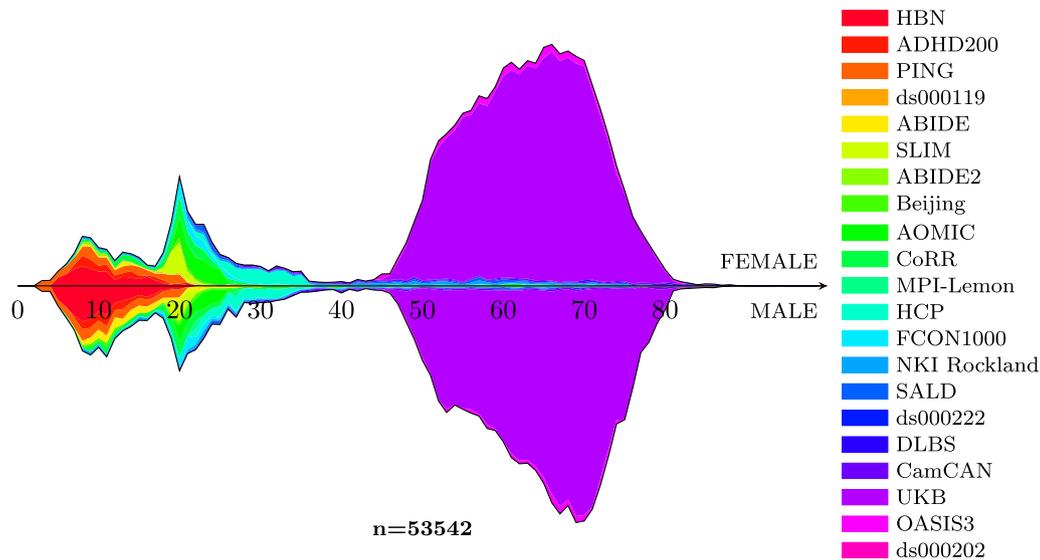
(a) Age distribution of the reference dataset
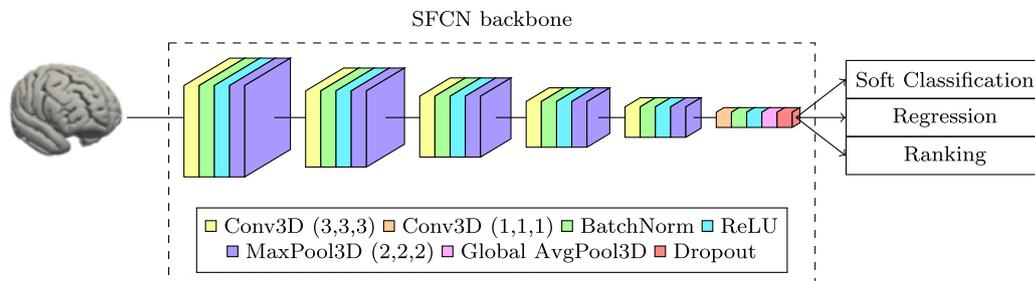
(b) Brain age model architecture
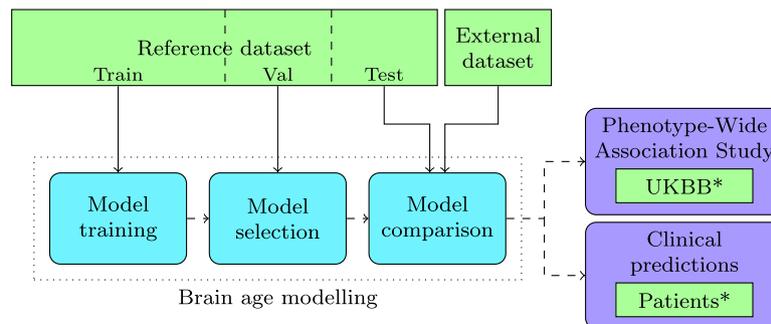
(c) Flowchart of the modelling process and applications

Fig. 1. An overview of the dataset and models used for brain age modelling. (a) The reference dataset contains 53,542 healthy participants from 21 datasets, with ages ranging from 3 to 95 years. (b) We implemented three model variants for predicting brain age, all based on the contest-winning SFCN architecture (Peng et al., 2021). All models take minimally preprocessed T1-weighted MRI images as input. (c) The modelling process consisted of three steps, utilizing different parts of the reference data and the external dataset from previously unseen scanners. The best brain age model was applied in a phenome-wide association study, and a case-control comparison including several clinical conditions.

we examined brains in groups of participants with extreme brain age deltas. Second, we correlated the delta with FreeSurfer-extracted imaging measures. High deltas were broadly associated with a general pattern of reduced cortical thickness (Fig. 3a and Supplementary Figure 8), reduced volumes of several subcortical areas, and increases in ventricle size, cerebrospinal fluid, and both white matter and non-white matter hypointensities (Fig. 3b). We also correlated the delta with voxel-wise volume and area, seeing a less pronounced pattern (Supplementary Figure 9). A full overview of the subcortical correlations can be seen in Supplementary Table 11.

Observing that all the models performed worse in the external dataset than in the test set, we performed post-hoc analyses to further understand the causes underlying the generalization problems. First, we investigated whether the inflated error could be explained by an additive offset, a linear multiplicative bias, or a general increase in variance. We observed that all three measures were amplified in the exter-
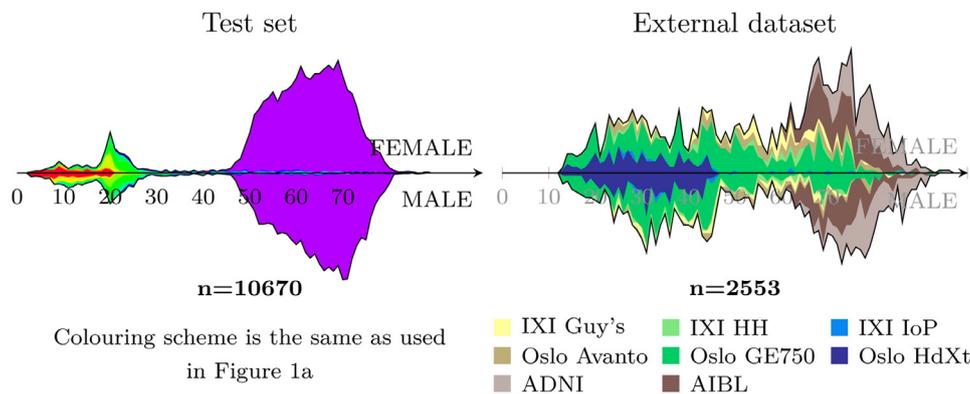
(a) Age distribution of the two datasets used in the comparison

(b) Brain age predictions

Fig. 2. The two datasets used in the model comparison, and the predictive performance of the three model variants. (a) For comparing the models we employed two distinct datasets: A test set sampled from the reference dataset, and an external dataset. The former was drawn from the reference dataset using stratification, and as a consequence has a similar age and sex-distribution. The latter was compiled from a different subset of datasets, and thus was acquired from different scanners. The age range of the external dataset is somewhat narrower, spanning a region of 13 to 95 years, and has a more uniform distribution. (b) The six scatter plots display the predictions of a given model on the x-axis, against the ground truth age of the participants on the y-axis. The top row contains the predictions of the three model variants on the test set, and the bottom row the external dataset.

Table 1

Predictive performance of the models. We compared the soft classification model (SFCN-sm), the regression model (SFCN-reg) and the ranking model (SFCN-rank) on the external dataset, originating from scanners which has not been seen by the models during training. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R and R2 are sensitive to the age range of the dataset, while normalized RMSE (nRMSE) and Relative Absolute Error (RAE) are not, facilitating comparisons across datasets.

Model	MAE	RMSE	R	R^2	mRMSE	RAE
SFCN-sm	5.04	6.51	0.961	0.903	0.078	0.26
SFCN-reg	**3.90**	**5.11**	**0.975**	**0.940**	**0.061**	**0.20**
SFCN-rank	5.92	7.54	0.959	0.870	0.090	0.31

nal dataset when compared to the test set, and thus jointly contributed to the increase (Supplementary Table 10). Secondly, we tried isolating two sources of generalization error: Differences in population, represented here by age and sex distributions, and differences in scanners and

acquisition protocols. We approached this by resampling two artificial datasets, both with participants previously unseen by the models. First, we sampled a dataset with an "Unknown population", with participants from the test set following the empirical age and sex distribution of the external dataset. Secondly, we created a dataset with "Unknown scanners", sampling participants from the external dataset while following the distributions of the test set (see Materials and Methods for further details). Due to the stratification used in the initial train/validation/test split the latter set also directly matches the distributions of the training set. For each of these two new datasets we computed an MAE per model, which naturally fell between the MAEs on the test set and on the external dataset. While this approach is exploratory and inherently limited to the characteristics and actual data points making up our datasets, the results clearly indicate that the main driver of generalization error is the unknown scanners (Table 2 and Supplementary Figure 1) which had higher errors (MAE_{US}) than the unknown population (MAE_{UP}) for all the models. Additionally, these two sources of generalization error seem to work additively, with their sum closely matching the full gen-
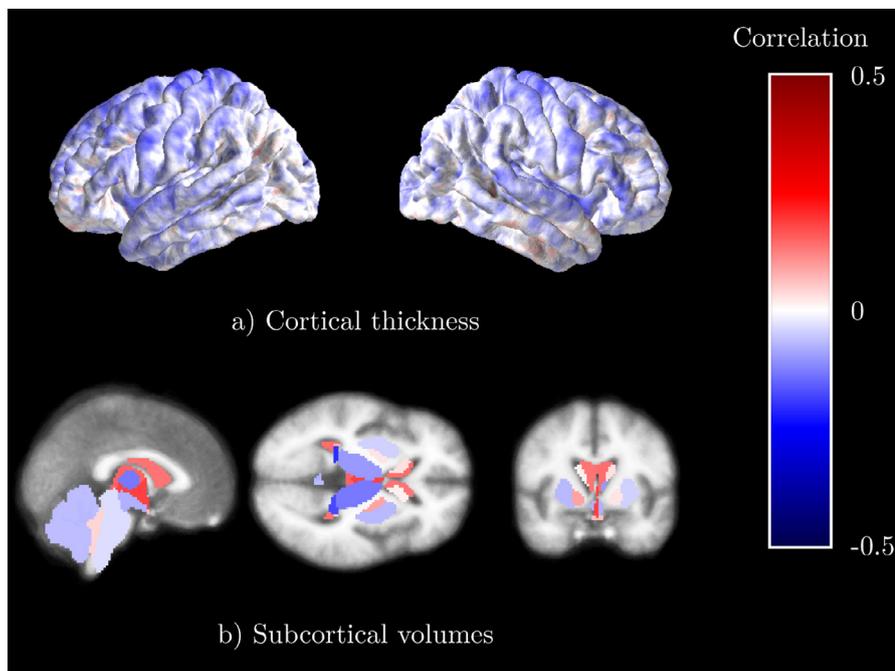
Fig. 3. Correlation between brain age delta and vertex-wise and subcortical measures in imaging space. (a) Correlations with cortical thickness at the vertex-level computed by FreeSurfer, plotted on two generic surface hemispheres. (b) Correlations with subcortical volumes using the standard FreeSurfer atlas, overlayed over an average brain. The reported associations were computed as Pearson correlations in the test portion of UKBB, with the imaging measures derived using FreeSurfer's recon-all pipeline.

Table 2

Results of the post-hoc generalization source analysis. We measured Mean Absolute Errors for the three model variants on the test set, drawn from the same distributions of scanners and ages as the training dataset, the "Unknown population" dataset (MAE_{UP}) and the "Unknown scanners" dataset (MAE_{US}), both representing a single source of generalization error, and the external dataset, different from the training set in both regards.

Model	MAE_{TEST}	MAE_{UP}	MAE_{US}	$MAE_{EXTERNAL}$
SFCN-sm	**2.23**	3.48	3.84	5.04
SFCN-reg	2.47	**2.89**	**3.42**	**3.90**
SFCN-rank	2.55	3.95	4.32	5.92

eralization error observed in the external dataset. A final observation was that SFCN-reg handled both sources of error best, which is further evidence of its superior ability to generalize.

Brain age predictions associate with biological phenotypes and lifestyle factors

Next, we examined the biological relevance of the model predictions by correlating their deviations from chronological age with an array of phenotypes (n=394) in a phenome-wide association study. We performed this analysis in the subset of the UK Biobank (UKBB) data that was not used for brain age modelling or validation (n=8066), and tested associations with all the biological phenotypes and lifestyle variables accessible, manually divided into thirteen thematic categories for interpretability (Supplementary Table 7). For each phenotype we computed a univariate correlation while correcting for age and sex (Alfaro-Almagro et al., 2021; Smith and Nichols, 2018), and assessed its significance using a Bonferroni-corrected p-value threshold of $p < 1.26 \times 10^{-4}$ (see Materials and Methods). All continuous variables were standardized, such that their effect sizes denote the impact a one standard deviation increase has on the brain age delta. In general, our results corroborated several findings derived from previous studies using smaller samples (Fig. 4a). We observed significantly higher delta in participants with high blood pressure ($\beta = 0.41$, $p = 1.86 \times 10^{-7}$), those currently on blood pressure medication ($\beta = 0.54$, $p = 1.17 \times 10^{-10}$), and a positive correlation with blood pressure readings (diastolic (DBP):

$\beta = 0.15$, $p = 2.53 \times 10^{-6}$, systolic (SBP): $\beta = 0.16$, $p = 3.30 \times 10^{-6}$). The associations with the largest effects indicated higher delta in patients with a diabetes diagnosis ($\beta = 0.74$, $p = 2.25 \times 10^{-7}$) or diabetes-related eye problems ($\beta = 1.78$, $p = 4.59 \times 10^{-7}$). Among the biochemical measurements, significant associations with brain age delta were found for blood glucose levels ($\beta = 0.23$, $p = 3.18 \times 10^{-11}$), Insulin-Like Growth Factor-1 levels ($\beta = -0.22$, $p = 5.81 \times 10^{-11}$), glycated haemoglobin levels ($\beta = 0.16$, $p = 7.97 \times 10^{-7}$) and mean corpuscular volume ($\beta = 0.13$, $p = 3.33 \times 10^{-5}$). Associations with variables we categorized as related to diet and lifestyle were dominated by previous smoking, with a positive correlation with number of cigarettes per day (absolute pack years: $\beta = 0.24$, $p = 8.55 \times 10^{-5}$, pack years as proportion of age: $\beta = 0.26$, $p = 2.41 \times 10^{-5}$) and age stopped smoking ($\beta = 0.25$, $p = 7.59 \times 10^{-5}$). We also observed significant associations with average weekly beer and cider intake ($\beta = 0.21$, $p = 1.13 \times 10^{-7}$) and alcohol intake frequency ($\beta = 0.12$, $p = 1.09 \times 10^{-4}$), cereal intake ($\beta = -0.16$, $p = 7.61 \times 10^{-7}$) and participation in "Other group activity" (e.g. social activities not related to a sports or social club, religious group or adult education, $\beta = -0.27$, $p = 5.91 \times 10^{-5}$). Further, we observed a significant correlation with the number of people living in the participants household ($\beta = -0.13$, $p = 8.20 \times 10^{-5}$) and higher deltas in those born outside the United Kingdom and the Republic of Ireland (compared to the baseline group born in England, $\beta = 0.62$, $p = 4.62 \times 10^{-6}$). An overview of all the 394 associations can be found in Supplementary Table 14.

Transferring brain age predictions to developmental and degenerative brain disorders

For six different disorders we compiled a patient cohort and a matched control group, and calculated a brain age delta per participant based on the prediction from SFCN-reg (Supplementary Figure 4). In all control groups, the brain age prediction accuracy was approximately the same as for the full test set (MAEs=2.91-4.05, Supplementary Figure 3). Patients with MS showed significantly higher brain age estimates than their matched healthy controls (brain age group mean difference $\Delta = 4.42$ years, $p = 1.71 \times 10^{-22}$, Cohen's $d = 0.87$). A similar pattern was also observed for patients with AD ($\Delta = 2.81$, $p = 4.20 \times 10^{-20}$, $d = 0.59$), MCI ($\Delta = 2.13$, $p = 1.25 \times 10^{-15}$, $d = 0.46$) and SCZ ($\Delta = 1.40$, $p = 4.29 \times 10^{-5}$, $d = 0.34$). For the individuals with mood disorders (MOOD, see Mate-
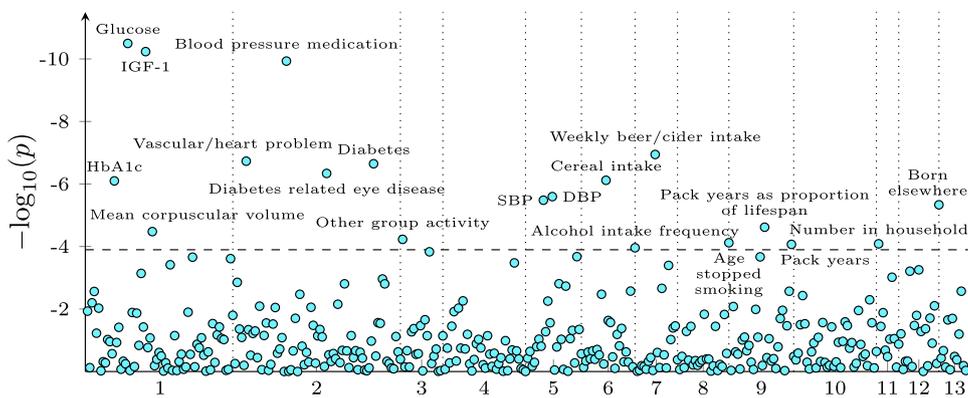
(a) **Significances of the PheWAS correlations**
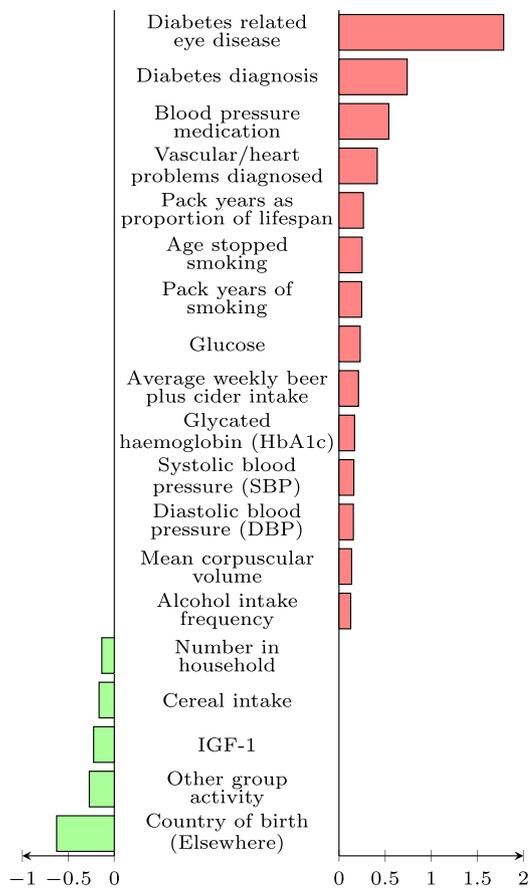
(b) **Effect sizes of significant hits**
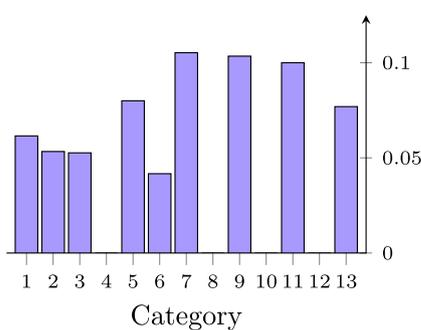
(c) **Proportion of significant hits per category**

	Category
1	Biochemical measures
2	Health related
3	Lifestyle
4	Cognitive measures
5	Physical measures
6	Diet
7	Alcohol
8	Microbial infection
9	Smoking
10	Mental health
11	Socioeconomic
12	Family
13	Early life factors

Fig. 4. Associations between the brain age delta and a wide range of phenotypes. We correlated the brain age delta originating from the SFCN-reg with 394 phenotypic variables categorized into thirteen thematic categories. (a) A Manhattan plot visualizing the significances of the 394 associations. A Bonferroni-corrected threshold revealed 19 significant associations with the delta. (b) Effect sizes of the significant associations. For binary variables the effect size express the mean difference between the groups, while for continuous variables it denotes the change in brain age delta associated with a one standard deviation increase. (c) For each of the thirteen categories we calculated the proportion of significant hits by dividing the number of significant hits within that category with the total number of variables in the same category.

rials and Methods) this difference was the smallest ($\Delta = 0.64$, $p = .04$, $d = 0.17$), while the difference was not significant for patients with a mix of psychotic diagnoses (PSY) ($\Delta = 0.74$, $p = .15$, $d = 0.20$). Both the relative ordering of the disorders in terms of group difference, the magnitude of the disparities, and the observed significance resemble a previous study Kaufmann et al. (2019) using a different model based on a smaller dataset.

To demonstrate the predictive power of our best performing pre-trained brain age model for clinical conditions, we trained multiple binary classifiers to predict whether a participant had a diagnosis or not (Fig. 1c, Fig. 5 and Materials and Methods). We used logistic regression models with an l_1-penalization for this purpose (LASSO models),

optimized via a nested cross validation procedure (Supplementary Figure 5), and started with a baseline model classifying participants based only on age and sex. The second model included the brain age delta originating from the brain age prediction of SFCN-reg, and the third model replaced the brain age delta by 64 features encoded in the second to last layer of the same model (Fig. 5b and Materials and Methods). Across the six disorders, the baseline models achieved an area under the receiver operating curve (AUCs) ranging from 0.47 to 0.54, indicating that our matching procedure was satisfactory. Using the second set of models, quantifying the predictive power of the brain age deltas in Supplementary Figure 4, greatly improved the prediction performance when compared to the corresponding baseline models for MS (AUC=0.71 vs.

(a) Patient and control group age distributions

(b) Conceptual model compositions
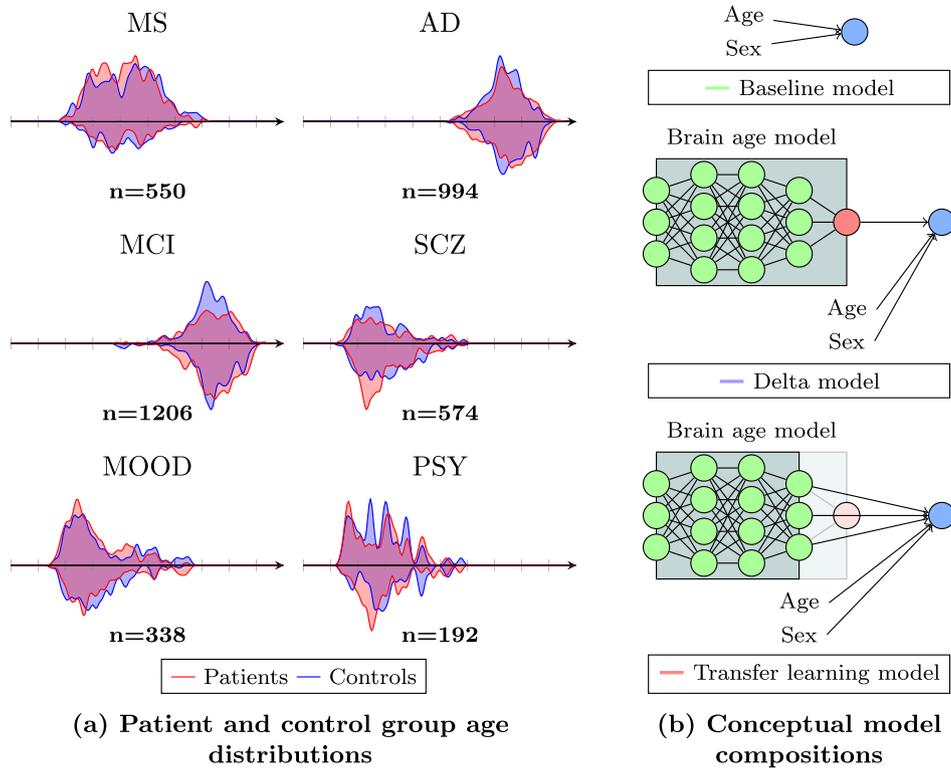
Fig. 5. The datasets and models used for clinical predictions, and an overview over their performance. We trained binary classifiers to separate cases and controls for multiple common brain disorders, using different levels of information from the brain age model. (a) We used a strict matching procedure, drawing a set of controls for each scanner-specific patient dataset matching its empirical age and sex distribution. (b) For each of the six disorders we trained three logistic regression models. The first used age and sex as predictors, the second included the brain age delta, and the third used internal features from SFCN-reg in a transfer learning setting. (c) For each disorder we compared the classifiers using AUCs.

(c) ROC curves for the predictors
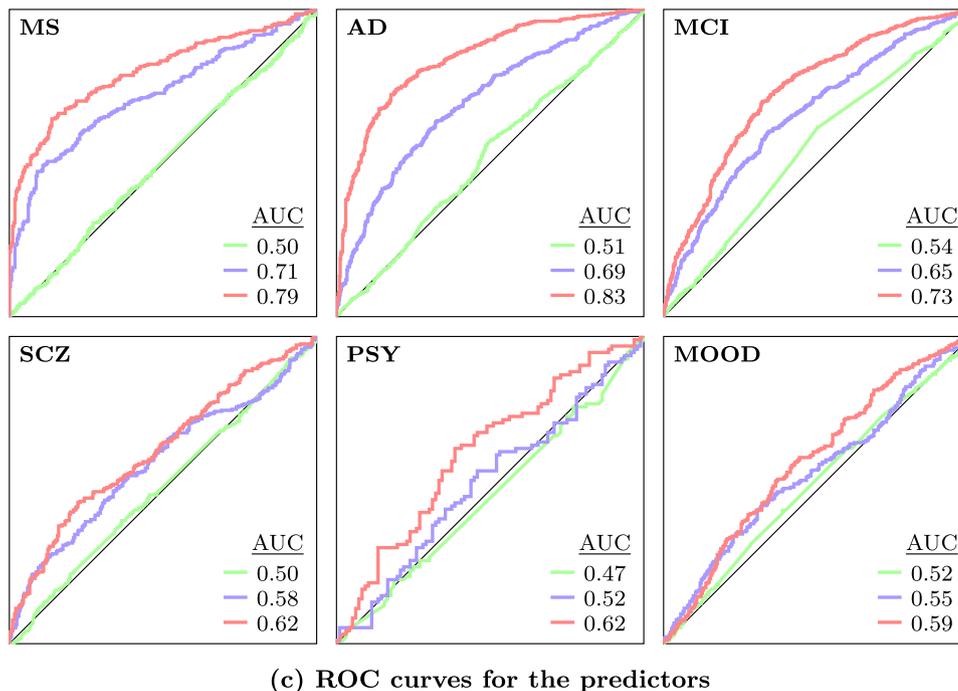
0.50), AD (0.69 vs. 0.51) and MCI (0.65 vs. 0.54); but to a minimal extent for SCZ (0.58 vs. 0.50), MOOD (0.55 vs. 0.52), and PSY (0.52 vs. 0.47). A third set of classifiers were implemented in a strict transfer learning context, utilizing the first part of the SFCN-reg as an encoder. Here, we ran all images in the case-control dataset through the model up until the second-to-last layer, encoding them as 64-dimensional feature vectors. We then used these vectors as predictors in a secondary modelling step, performing the regular cross-validation scheme and training LASSO models with the vectors from the training folds as inputs (Materials and Methods). These features represent high-level, data-driven ab-

stractions of the brain imaging data, and underlie the singular brain age prediction. We refer to this variant of transfer learning as strict because we kept the weights of the initial brain age model locked while optimizing for the new binary objective, which in turn allow us to keep treating these as ageing features and thus promote interpretability. While this complicates contextualizing the performance of our models in terms of existing case-control classifiers, it gives us an indication of the information content of these learned features. This third set of models improved AD prediction substantially (AUC=0.83), and also were notably better for MS (0.79), MCI (0.73) and PSY (0.62), while only a marginal im-

provement was observed in SCZ (0.62) and MOOD (0.59). Overall, our results show that our brain age model can be transferred to make case-control predictions of these common clinical brain disorders.

Discussion

Brain maturation and ageing, and its interactions with clinical brain disorders and conditions, are complex processes with pivotal environmental and genetic contributions (Fjell and Walhovd, 2010; Johnson, 2001). Brain age prediction and the accompanying brain age delta has the potential to provide intuitive and useful measures for summarizing individual brain aberrations. However, technical differences between studies, e.g. the use of different scanners and MRI scan parameters, have represented challenges for the direct generalization and applicability of brain age models based on large training sets. We assembled a large and diverse neuroimaging dataset to train multiple state-of-the-art deep learning models for brain age prediction, and extensively tested their ability to generalize, and their sensitivity to various common brain disorders. Our best model, the SFCN-reg, showed superior performance on an external dataset with differing scanners and age distributions compared to that of the training set, enabling it for applications in other datasets. We then demonstrated the relevance of the model predictions by showing associations between the brain age delta and a range of complex human traits and health outcomes in a population sample. Lastly, we transferred the trained SFCN-reg to clinical data in a transfer learning setting, showing that both the brain age delta and the internal features learned by the model have predictive value when differentiating between controls and patients with common brain disorders.

In our experiment the SFCN-reg outperformed the other models in terms of generalization performance. The age prediction accuracy on the test data (MAE=2.47) is among the best in the field (He et al., 2021). Crucially, the prediction accuracy on data coming from unknown scanners (MAE=3.90) fares very favourably when compared to other studies attempting to transfer between datasets of similar size and complexity (Bashyam et al., 2020; Boyle et al., 2021; He et al., 2021; Jonsson et al., 2019; Ren et al., 2019). Using data from different scanners, protocols and populations in neuroimaging comes with the problem of modelling the effects of these appropriately (Butler et al., 2021; Smith et al., 2019). Previous brain age studies have often explicitly included a scanner term in the modelling or corrected the computed brain age for various biases (Lange A. Marie and Cole, 2020). Recent approaches have tried to address this problem directly via specific deep learning paradigms (Dinsdale et al., 2020). Our results show that given sufficiently large and heterogeneous training data, deep CNNs achieve state-of-the-art performance for brain age predictions even when scanner effects are not explicitly modelled, and more importantly that this performance translates to scanners and protocols that are unknown to the model. This suggests that the representations learned by the model are dominated by age-related variance, not scanner-dependent artefacts, an extension of the model robustness shown in earlier studies (Abrol et al., 2021). From a practical point of view this suggests that our trained model may be employed in other applications in new datasets, without the need for retraining or applying corrective procedures.

Evaluating the biological relevance of the brain age predictions is essential to further understand and trust these models. Therefore, we correlated brain age delta obtained from the SFCN-reg with a wide range of phenotypes in a subset of the UKBB not used for model training. When applying a Bonferroni-corrected threshold, we found nineteen significant hits spread across seven of our thirteen categories. Almost all of these have in earlier studies been found to have a relationship with ageing, either generally or specifically in the brain, or with brain health and/or cognitive function: Glucose level (Karasik et al., 2005), Insulin-like Growth factor-1 (Ashpole et al., 2015), and glycated haemoglobin (Roth et al., 2016) are known to change with age, and corpuscular volume has been associated with cognitive functioning (Gamaldo et al., 2013). Lifestyle factors involving alcohol and smoking impact various

biological and bodily ages (Karasik et al., 2005), including that of the brain (Cole, 2020; Lange et al., 2020; Wrigglesworth et al., 2021). Elevated blood pressure and other cardiovascular risk factors have established associations with increased brain age (Lange et al., 2020; Lange A. Marie et al., 2021), and an increase in predicted age has been observed in patients with diabetes (Franke et al., 2013). These results need to be assessed carefully due to the large sample size and relatively small effect sizes (Sullivan and Feinn, 2012), but in sum we believe they indicate that our model makes biologically meaningful predictions. Further, transferring the model to unseen datasets comprised of patients with different clinical conditions and matched healthy peers revealed both high accuracy in terms of age prediction and higher deltas among patients with brain disorders, in line with previous studies (Kaufmann et al., 2019). Importantly, the data in the case-control datasets were obtained from scanners not included in the training set, supporting that the model generalizes to previously unseen scanners; a highly valuable asset.

As a further attempt to understand the SFCN-reg model we correlated variance in the prediction-space with variation in imaging-space after standard preprocessing. Explaining the predictions of deep neural networks is inherently hard, and the plethora of methods for interpretation that exist are largely tailored to classification models. Additionally, they often rely on humanly discernable features for validations, such as faces or clearly identifiable objects. Furthermore, in our analyses we were not interested in interpreting the predicted age per se, but rather its deviation from the chronological age of the participant, the brain age delta. Given these factors, we resolved to correlating the brain age delta with standard measures derived from MRI images, such as cortical measures and subcortical volumes, using linear methods. The delta was associated with surface-wide patterns of cortical thinning, as well as volumetric decreases of multiple subcortical areas. Such atrophic trajectories have been linked to the ageing brain (Fjell et al., 2015; Walhovd et al., 2005) both in healthy populations (Fjell et al., 2009) and those with brain disorders (Jacobsen et al., 2014; Pini et al., 2016). While these overall patterns are plausible, and provide further confidence in our model, methodological advances are required to precisely describe its inner workings.

Compared to the all-in-one brain age delta, a single number describing the difference between apparent and chronological age, our results showed increased predictive value for MS, AD and MCI when using the internal representations of the SFCN-reg underlying the brain age predictions. This supports the view that for some applications the constituent components of the singular brain age delta are relevant beyond the age prediction alone (Smith et al., 2020). The innate ability of deep neural networks to form abstractions of the brain at different spatial resolutions throughout the layers of the model may help disentangle individual differences in neurodevelopmental and age-related processes related to complex disorders and traits. Furthermore, we see this result as evidence that deep learning models trained to predict age in large multisite datasets constitute excellent starting points for transfer learning, which can subsequently be fine-tuned to a variety of tasks.

There is an ongoing discussion in the field on whether brain age models that are precise, or those that allow for sufficient variance in their single-subject predictions, are the most useful in a downstream analysis of behavioural and clinical traits (Bashyam et al., 2021; 2020; Hahn et al., 2021). An argument for a model which allows for more variation (a 'looser' fit) is that this would more accurately depict brain age as a complex process which appears differently in different individuals. One challenge with this approach is that the brain age delta is a residual, and recognizing what portion of this error comes from biological variation and what is modelling imprecision is practically unfeasible. As more complex models such as deep CNNs become competitive for brain age modelling, it becomes possible to minimize the overall model error, including the methodological portion, while still allowing the model to accurately represent the necessary biological variability internally.

There are some limitations of the present study which we acknowledge here. Given the computational cost of training complex deep learn-

ing models on such a large dataset we restrained our study to a limited number of possible models, both in terms of model architectures and hyperparameter settings. With further increases in sample sizes and diversity, deeper architectures may be sensible, as the risk of overfitting is directly alleviated by the larger datasets. While we refer to our dataset as diverse, most participants included in the study are of white European background. However, this limitation is not specific to the present study, and global collaborations are needed to build models that generalize across cultural and genetic backgrounds (Thompson et al., 2020). Relatedly, the current investigation included T1-weighted MRI data only. Future integration of information spanning various imaging modalities may increase both age prediction accuracy and the sensitivity and specificity of various biological and clinical traits and conditions (Lange A. Marie et al., 2021; Richard et al., 2018; Rokicki et al., 2021; Smith et al., 2020). To evaluate the clinical relevance of the brain age model in the context of disorders we employed transfer learning in a strict way, by training LASSO models on top of the SFCN architecture. This setup increases the interpretability of the transfer process as the lower-level representations of our model used as predictors capture age-related features of the brain. In turn, this allows for the interpretation that illnesses that are predictable by the model must also rely on these representations, and thus implicitly relate to age. There are multiple steps which could have been taken instead to maximize predictability, a natural starting point is to fine-tune the entire model (Bashyam et al., 2020). Lastly, in order to reach broad adoption of these models and, ultimately, approach clinical usability, a better understanding of the regional patterns driving the prediction, their specific biological significance and how it changes across time and contexts is needed (Vidal-Piñeiro et al., 2021).

In conclusion, we have trained multiple variants of a deep neural network to predict brain age on a large and heterogeneous sample of raw structural MRI data, and observed distinct differences in their ability to generalize to unseen samples and scanners. The predictions of our best model were linked to biochemical biomarkers, cardiovascular risk factors, smoking and alcohol intake, among others. Using transfer learning, we demonstrated that clinical conditions with a neurodevelopmental or neurodegenerative aetiology were predictable by our model, initially trained to predict age. Jointly, these findings add to the growing literature documenting the tremendous potential of advanced techniques for statistical learning to decode biologically and clinically relevant information from brain MRI scans.

Materials and Methods

Data

All data sets used in the present study have been obtained from previously published studies which have been approved by their respective institutional review board or relevant research ethics committee.

Reference dataset

The reference dataset used for training the brain age models was T1-weighted MRI scans derived from 21 non-overlapping and publicly available datasets (total n=53542; female n=27715) of healthy individuals, with ages ranging from 3 to 95 years (Fig. 1 and Supplementary Table 1). The age distribution can be seen in Fig. 1a. The younger age-range (3–30 years) was mainly composed of participants from multiple different datasets. Though the older age-range (40–80) also included multiple datasets, UKBB (obtained from the data repository under accession number 27412) accounted for most of these participants. The most sparsely populated age-ranges were in the very young (147 participants with age $\leq$ 5), very old (17 participants with age $\geq$ 85), and in midlife (42 participants with $35 \leq$ age < 45). For each of the datasets, participants that had one or more psychiatric, neurological and/or other relevant diagnoses (Supplementary Table 4), and those withdrawn from the respective study were excluded before model training. In addition, for

participants having multiple brain scans, the baseline data were used, such that in the final dataset each data point represents a unique participant drawn from a normative population.

External dataset

To evaluate the generalizability of our trained brain age prediction model, an external dataset was collected (Supplementary Table 2). This dataset included the IXI project (Supplementary Table 5) and healthy controls from the datasets underlying the clinical data described below (total n=2553). Importantly, the external dataset contained images generated by scanners not used in the reference dataset, and subsequently unknown to the models during training and validation. This dataset can be seen as having a more uniform age distribution (Fig. 2a), meaning that our test would capture whether any given model relies too heavily on information observed in the training data.

Clinical data

The clinical data consisted of six patient cohorts diagnosed with MCI, AD, MS, SCZ, a mix of psychotic diagnoses and mood disorders, where the latter was a combination of two cohorts with depression and bipolar disorder (Fig. 2a and Supplementary Table 3). The individual cohorts were compiled from ADNI, AIBL, and multiple scanners at the Oslo University Hospital (Supplementary Tables 5 and 6). In addition to the patients, we used healthy controls from the same scanners in the external dataset to create matched control groups for the clinical predictions.

Quality control

To ensure data quality, we executed a quality control (QC) pipeline, consisting of checking whether any of the image preprocessing steps failed, and a manual control via visual inspection. To take advantage of as much data as possible this manual control was lenient, removing samples where either a significant portion of the brain was missing, or where the orientation of the head was dramatically off, and resulted in dropping only 39/53581 participants in the reference dataset (0.07%) and 1/2554 participants in the external dataset (0.03%).

Image preprocessing

We first performed skull-stripping with the FreeSurfer 5.3 auto-recon pipeline (Sgonne et al., 2004) to produce a brainmask, minimizing the amount of non-brain information in the data, then reoriented the images to the standard FSL (Jenkinson et al., 2012) orientation using fslreorient2std. The resulting images were linearly registered to the MNI152 space using FLIRT (Jenkinson, 2001) with linear interpolation and the default 1 mm FSL template (version 6.0). We cropped away borders of [6:173, 2:214, 0:160] voxels, in the sagittal, coronal and axial dimensions respectively. This cropping yielded the smallest cuboid with marginal loss of brain-related information across the dataset, minimizing the memory footprint of the models during training. As a last preprocessing step the voxel intensity values of all brain images were normalized to the range [0, 1].

Brain age models

The state-of-the-art network architecture, the Simple Fully Convolutional Network (SFCN) (Peng et al., 2021), was implemented as the backbone in all our brain age models. The SFCN architecture consists of a VGG (Simonyan and Zisserman, 2015)-like structure, with five repeated convolutional blocks, each with a three-dimensional convolutional layer with a filter size of (3, 3, 3), a batch normalization layer, rectified linear activation function (ReLU) activation, and a max-pooling layer with a pooling size of (2, 2, 2) (Fig. 1b). The model then has a channel-wise convolutional layer, a last batch normalization layer and a global average pooling layer. From this backbone we defined three end-to-end variants for brain age prediction: The original soft classification model (SFCN-sm), a regression variant with a single output

node with a linear activation (SFCN-reg) and a ranking model from the general age-regression literature (SFCN-rank) (Chen et al., 2017), an approach which has also been successful for brain age predictions (Xia et al., 2021). All the model definitions rely on a matrix X of dimensions $[N, h, w, d]$ containing MRI images as input, an N-dimensional vector age=$[\text{age}_0, \text{age}_1, \ldots, \text{age}_{N-1}]$ containing the ground truth ages of the participants to compute its loss, and are ultimately able to produce an N-dimensional vector aĝe = $[\text{aĝe}_0, \text{aĝe}_1, \ldots, \text{aĝe}_{N-1}]$ with a single brain age prediction per participant (although this is not necessarily the direct output of the model).

SFCN-sm

The soft classification variant formulates the age regression problem as a multiclass classification problem, by having $m = \lceil \max(\text{age}) \rceil - \lfloor \min(\text{age}) \rfloor$ output neurons, where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote the ceiling and floor operators respectively. It is denoted as soft because it uses a target vector per participant generated by a normal distribution centered around the ground truth age, instead of the one-hot encoding used in regular classification

$$y_i = \mathcal{N}(\text{age}_i, 1).$$

The predictions of the model are similarly a vector of length m with a softmax activation

$$\hat{y}_i = [\hat{y}_{i,0}, \hat{y}_{i,1}, \ldots, \hat{y}_{i,(m-1)}], \sum_{j=0}^{m-1} \hat{y}_{i,j} = 1.$$

The loss for a single datapoint is the Kullback-Leibler divergence between the two vectors

$$\text{loss}_i = \text{KL}(y_i || \hat{y}_i).$$

The final age-prediction of a participant is calculated as a weighted sum of the prediction vector

$$\text{aĝe}_i = \lfloor \min(\text{age}) \rfloor + \sum_{j=0}^{m-1} j \hat{y}_{i,j}.$$

SFCN-reg

The regression variant has a single output neuron predicting a single value $\hat{y}_i$ per participant, limited to the range (min(age), max(age)) with a bounded ReLU activation. The value of $\hat{y}_i$ can be used directly as the predicted age for a participant, $\text{aĝe}_i = \hat{y}_i$. During training, the model optimizes the mean squared error.

SFCN-rank

The ranking model formulates the age regression problem as a set of binary "*Is participant X older than age y?*"-questions. Like the soft classification model, the model has $m = \lceil \max(\text{age}) \rceil - \lfloor \min(\text{age}) \rfloor$ output neurons, each representing one such binary question. The target vector for a participant is a binary vector of length m, with a 1 in bins corresponding to ages younger than the participants age and 0 in the rest.

$$y_i = [y_{i,0}, y_{i,1}, \ldots, y_{i,(m-1)}], y_{i,x} = \begin{cases} 1 & \lfloor \min(\text{age}) \rfloor + x \leq \text{age}_i \\ 0 & else \end{cases}.$$

For each participant the model predicts a vector of length m, where each output neuron is limited to the range [0, 1] by a sigmoid activation

$$\hat{y}_i = [\hat{y}_{i,0}, \hat{y}_{i,1}, \ldots, \hat{y}_{i,(m-1)}], \hat{y}_{i,x} \in [0, 1].$$

The model optimizes the mean binary cross entropy across all the output neurons

$$\text{loss}_i = -\frac{1}{m} \sum_{j=0}^{m} y_{i,j} \log(\hat{y}_{i,j}) + (1 - y_{i,j}) log(1 - \hat{y}_{i,j}).$$

To calculate a predicted age for the model we sum up the number of age bins for which the model predicts that a participant is older than

the given age (it is worth noting that this limits the model to integer predictions)

$$\text{aĝe}_i = \lfloor \min(\text{age}) \rfloor + \sum_{j=0}^{m} \mathbb{1}_{x \geq 0.5}(\hat{y}_{i,j}).$$

Brain age model training and comparison

All brain age models were trained on 2 NVIDIA V100 GPUs with 32GB memory, using the Keras Chollet (2015) interface of Tensorflow 2.3 Abadi et al. (0000)) on top of cuda 10.0. Using a batch size of 14 the models took approximately 1 second per step, translating into roughly 45 minutes per epoch or about 2.5 days per full training session. To train the brain age models, 80% (n=42829) of the reference dataset was used for model building (training and validation) and 20% (n=10713) for testing. Among the data for model building 80% (n=34285) and 20% (n=8544) were used for training and validation of the models, respectively (Fig. 1c and Supplementary Table 1). Before these splits, the data was stratified by age and original study to ensure that all subsets had resembling age distributions and came from multiple scanners. Given the great computational cost of model training, determining optimal hyperparameter values by searching over the full configuration space for each model is impractical. Instead, we employed post-hoc heuristics, i.e., tweaking the models based on previous runs. For each training run we trained the model from scratch (with randomly initialized parameters) for 80 epochs, optimized by vanilla stochastic gradient descent, employing an annealing, step-wise, learning rate schedule. This schedule had three steps, reducing the learning rate by a factor of 3 after epochs 20, 40 and 60. The initial learning rate was found independently for each model variant using a learning rate sweep (Smith, 2017) (Supplementary Figure 7). We used mean absolute error (MAE) on the validation split to determine the best epoch for each run. We also report RMSE, R, R^2 for all models in Table 1, and to enable comparisons with other studies with possibly different age ranges the normalized measures normalized RMSE (nRMSE)

$$\text{nRMSE} = \frac{\text{RMSE}}{\max(\text{age}) - \min(\text{age})}$$

and Relative Absolute Error (RAE)

$$\text{RAE} = \frac{\sum_{i=0}^{N} |\hat{y}_i - y_i|}{\sum_{i=0}^{N} |y_i - \bar{y}|}$$

The first model we trained was SFCN-sm with the hyperparameters specified in the original SFCN paper (Peng et al., 2021). Seeing that this model was underfitting we relaxed the regularization for a second run of the same model, and subsequently a third. The two hyperparameters we tuned in this process were the weight decay, and the dropout rate between the two final layers of the model. Having trained three soft classification models, we moved on to train three regression models and three ranking models using these same heuristics (Supplementary Table 8). To select a candidate model for each variant we compared the MAEs on the validation split. In the final model selection, we compared the MAEs of the candidate models for each variant on the test set and the external dataset.

Correlating the brain age delta with imaging measures

To investigate the influence of different brain regions and voxel-wise patterns on our predictions, we performed two post-hoc analyses. First, we carried out a qualitative comparison of groups with unusually low and high brain age deltas, stratified by age. To minimize the impact of scanner effects, both in prediction- and voxel-space, while retaining large enough groups to enable meaningful comparisons, we performed this analysis in the test-portion of UKBB. We also executed the analysis

independently for each sex, to ensure similarities within the groups. For each age bin $b \in [50, 55, 60, 65, 70, 75]$ we selected 10 participants with $age = b \pm 0.5$ with the lowest and highest brain age delta, and compared them with a group of participants the same age with the smallest absolute delta. For each of these groups we created an average brain using FreeSurfer's make_average_brain. The average brains are seen in Supplementary Figure 8.

Additionally, we performed a quantitative analysis correlating variation in brain age delta with structural imaging measures derived by FreeSurfer at the vertex (for cortical features) and atlas level (for subcortical features). This analysis also used subjects from a single scanner to minimize potential biases in the results. Being less reliant on a very large dataset, we were able to use data from the largest scanner in the external dataset, the Oslo GE 750 scanner (n=876). For each subject we computed cortical thickness for each vertex, and subcortical volumes, using FreeSurfer's recon-all pipeline and the default FreeSurfer atlas. For each of these measures we computed the Pearson correlation with the brain age delta across all healthy subjects from the given scanner. The vertex-wise and subcortical correlations are shown in Fig. 3, Supplementary Figure 9 and Supplementary Table 11.

Post-hoc generalization analysis

To study the causes of the differences in generalization, we designed an experiment to isolate the underlying sources of this error. Based on previous knowledge of the problems of new scanners, and the predictions of the models at different ages (Fig. 4) we specifically targeted two possible sources: Differences in population, represented by different distributions of age and sex, and data coming from unknown scanners. For each source we sampled an artificial, bootstrapped dataset based on our existing data. For the "Unknown population" dataset we sampled participants from the test set (originating from the reference dataset), to match the empirical age and sex distribution of the external dataset. Similarly, for the "Unknown scanners" dataset we sampled participants from the external dataset (coming from scanners unknown to the model) to match the age and sex distribution of the test set (and thus also the training set). The idea behind both datasets is to isolate a single source of generalization error. For robustness, we bootstrapped each of these two artificial datasets 100 times and reported the mean MAE achieved by the different models. Each sample was drawn probabilistically, with replacement, with the probability of drawing participant x of age x_a and sex x_s from dataset source based on the age and sex distribution of dataset target given by

$$p(x) = \frac{\hat{P}_{target}(x_a, x_s)}{\hat{P}_{source}(x_a, x_s)}$$

where $\hat{P}_{dataset}(x_a, x_s)$ denotes the proportion of participants in dataset *dataset* with age x_a and sex x_s. In "Unknown population", the test set plays the role of source and the external dataset is the target, while this is switched for "Unknown scanners".

Phenome-wide association study

In the PheWAS we calculated the univariate correlation between the brain age delta and 402 phenotypic variables from the UKBB, manually divided into thirteen thematically defined categories for interpretability (Supplementary Table 7). We performed this analysis in the UKBB portion of the test split (n=8066), and used all the variables available to us at the time. We encoded all phenotypic variables according to the PHESANT (Millard et al., 2018) datatypes, and removed non-informative levels (Supplementary Table 12) based on the UKBB coding schemes. Additionally, we re-coded the ordinal variables as categorical or continuous by hand (Supplementary Table 13). Variables which were impossible to model (i.e. singular or all missing values) were discarded. We then fitted a linear model per variable, modelling the delta as a function of the given covariate, age and sex, using the Python statsmodels

API (Seabold and Perktold, 2010). All continuous variables were standardized using a z-score normalization pre-modelling, such that the reported effect sizes refer to the change in brain age delta associated with a one standard deviation increase in the given variable. For assessing the significance of the associations, we computed a Bonferroni-corrected p-value threshold $p_{bonferroni} = 0.05/394 = 1.12 \times 10^{-4}$.

Transfer learning to clinical samples

In the transfer learning analysis we trained multiple binary models to predict whether a participant had a given diagnosis or belonged to the control group, based on various levels of information from the brain age model. For this purpose, we used the clinical dataset which was previously unseen by the model, and matched controls groups drawn from the external dataset used in the generalization test.

Control matching

To avoid biases in the case-control datasets we drew the subsets of controls independently for each scanner in the patient dataset, matching empirical distributions of age and sex in the corresponding case subset. For each disorder, for each scanner, the control group was created by drawing n_{case} controls, without replacement, using the sampling procedure described for the post-hoc generalization analysis.

Feature extraction

To generate the feature vectors for each participant we used the trained SFCN-reg model as an encoder, up until and including the global average-pooling layer. Running a single MRI through the model up until this point results in a 64-dimensional vector representing the original image. Each of the dimensions $n = [0, \ldots, 63]$ in this space represent a high level feature of the brain, and each participant $X_i = [X_{i,0}, \ldots, X_{i,63}]$ is encoded as a point in this space. When used as a predictor in the subsequent modelling phase, each of these 64 dimensions were treated as an independent variable.

Modelling

For each disorder we compared three different LASSO models, all trained and evaluated using the following general procedure, but on different covariate sets. We first stratified the given dataset on disorder, age and sex, respectively, and split it into 5 folds. We performed an outer cross-validation over these splits to allow us to have an out-of-sample prediction for each participant. When training a model on the training folds we performed an inner cross-validation to find the optimal value of the penalty parameter λ. The nested cross-validation procedure is illustrated in Supplementary Figure 5. Having found λ, we retrained the model on all the data from the four training folds. The models were implemented using sklearn's LogisticRegression Pedregosa et al. (0000)) with an l_1-penalty. Having the out-of-sample predictions for all the participants allowed us to calculate and compare AUCs based on the entire case-control dataset for the given disorder.

In addition to training the LASSO-model based on the brain age features, we trained an MLP using Keras with the same inputs. We did not optimize hyperparameters for this model, but observed similar results as the best LASSO models with the initial configuration (Supplementary Figure 6). The main benefit of the MLP is that it does not require a two-step process for the clinical prediction models, first processing the images with the encoder and then doing a prediction via a separate API, but can be implemented as an end-to-end binary classifier in Keras taking MRIs as inputs, and thus are more accessible for use by others.

Data availability

The raw data incorporated in this work were gathered from various resources. Material requests will need to be placed with individual principal investigators. A detailed overview of the independent datasets, and their origins, is provided in Supplementary Table 5.

Code availability

All of the trained brain age models and a pipeline for preprocessing images is released in our GitHub repo at http://www.github.com/estenhl/pyment-public

Funding sources

Declaration of Competing Interest

Declarations of interest: None

Credit authorship contribution statement

Esten H. Leonardsen: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Han Peng:** Conceptualization, Methodology, Software, Validation, Visualization, Writing – review & editing. **Tobias Kaufmann:** Conceptualization, Methodology, Software, Validation, Visualization, Writing – review & editing. **Ingrid Agartz:** Data curation, Validation, Writing – review & editing. **Ole A. Andreassen:** Data curation, Funding acquisition, Validation, Writing – review & editing. **Elisabeth Gulowsen Celius:** Data curation, Validation, Writing – review & editing. **Thomas Espeseth:** Data curation, Validation, Writing – review & editing. **Hanne F. Harbo:** Data curation, Validation, Writing – review & editing. **Einar A. Høgestøl:** Data curation, Investigation, Validation, Writing – review & editing. **Ann-Marie de Lange:** Data curation, Validation, Writing – review & editing. **Andre F. Marquand:** Investigation, Methodology, Validation, Writing – review & editing. **Didac Vidal-Piñeiro:** Investigation, Methodology, Validation, Writing – review & editing. **James M. Roe:** Investigation, Methodology, Validation, Writing – review & editing. **Geir Selbæk:** Data curation, Validation, Writing – review & editing. **Øystein Sørensen:** Investigation, Methodology, Validation, Writing – review & editing. **Stephen M. Smith:** Investigation, Methodology, Software, Validation, Writing – review & editing. **Lars T.**

Westlye: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft. **Thomas Wolfers:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft. **Yunpeng Wang:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft.

Supplementary material

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 19

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., Calhoun, V., 2021. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nat Commun 12 (1), 353.

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J.L.R., Bastiani, M., Miller, K.L., Nichols, T.E., Smith, S.M., 2021. Confound modelling in UK Biobank brain imaging. Neuroimage 224, 117002.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage 145, 137–165.

Ashpole, N.M., Sanders, J.E., Hodges, E.L., Yan, H., Sonntag, W.E., 2015. Growth hormone, insulin-like growth factor-1 and the aging brain. Exp. Gerontol. 68, 76–81.

Bashyam, V., Shou, H., Davatzikos, C., 2021. Reply: from 'loose fitting' to high-performance, uncertainty-aware brain-age modelling. Brain 144 (3), 32.

Bashyam, V.M., Erus, G., Doshi, J., Habes, M., Nasrallah, I.M., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., Fan, Y., Launer, L.J., Masters, C.L., Maruff, P., Zhuo, C., Vlzke, H., Johnson, S.C., Fripp, J., Koutsouleris, N., Satterthwaite, T.D., Wolf, D., Gur, R.E., Gur, R.C., Morris, J., Albert, M.S., Grabe, H.J., Resnick, S., Nick, B.R., Wolk, D.A., Shou, H., Davatzikos, D., 2020. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. Brain 143 (7), 2312–2324.

Bengio, Y., 2012. Deep Learning of Representations for Unsupervised and Transfer Learning. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp. 17–36.

Boyle, R., Jollans, L., Rueda-Delgado, L.M., Rizzo, R., Yener, G.P., McMorrow, J.P., Knight, S.P., Carey, D., Robertson, I.H., Emek-Sava, D.D., Stern, Y., Kenny, R.A., Whelan, R., 2021. Brain-predicted age difference score is related to specific cognitive functions: a multi-site replication analysis. Brain Imaging Behav 15 (1), 327–345.

Butler, E.R., Chen, A., Ramadan, R., Le, T.T., Ruparel, K., Moore, T.M., Satterthwaite, T.D., Zhang, F., Shou, H., Gur, R.C., Nichols, T.E., Shinohara, R.T., 2021. Pitfalls in brain age analyses. Hum Brain Mapp 42 (13), 4092–4101.

Chen, S., Zhang, C., Dong, M., Le, J., Rao, M., 2017. Using ranking-CNN for age estimation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 742–751.

Chollet, F., 2015. Keras.

Cole, J.H., 2020. Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. Neurobiol. Aging 92, 34–42.

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

Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., Montana, G., 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. Neuroimage 163, 115–124.

Dinsdale, N.K., Bluemke, E., Smith, S.M., Arya, Z., Vidaurre, D., Jenkinson, M., Namburete, A.I.L., 2021. Learning patterns of the ageing brain in MRI using deep convolutional networks. Neuroimage 224, 117401.

Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L., Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Kevin, Z.S., Racoceanu, D., Joskowicz, L., 2020. Unlearning Scanner Bias for MRI Harmonisation. In: Medical Image Computing and Computer Assisted Intervention 'MICCAI' 2020, pp. 369–378.

Dockès, J.m., Varoquaux, G., Poline, J.B., 2021. Preventing dataset shift from breaking machine-learning biomarkers. Gigascience 10 (9).

Elliott, M.L., Belsky, D.W., Knodt, A.R., Ireland, D., Melzer, T.R., Poulton, R., Ramrakha, S., Caspi, A., Moffitt, T.E., Hariri, A.R., 2019. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. Mol Psychiatry 1–10.

Fjell, A.M., Grydeland, H., Krogsrud, S.K., Amlien, I., Rohani, D.A., Ferschmann, L., Storsve, A.B., Tamnes, C.K., Sala-Llonch, R., Due-Tønnessen, P., Bjørnerud, A., Sølsnes, A.E., Håberg, A.K., Skranes, J., Bartsch, H., Chen, C.H., Thompson, W.K., Panizzon, M.S., Kremen, W.S., Dale, A.M., Walhovd, K.B., 2015. Development and aging of cortical thickness correspond to genetic organization patterns. Proc. Natl. Acad. Sci. U.S.A. 112 (50), 15462–15467.

Fjell, A.M., Walhovd, K.B., 2010. Structural brain changes in aging: courses, causes and cognitive consequences. Rev Neurosci 21 (3), 187–222.

Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M., 2009. One-year brain atrophy evident in healthy aging. J. Neurosci. 29 (48), 15223–15231.

Franke, K., Gaser, C., 2019. Ten years of brainAGE as a neuroimaging biomarker of brain aging: what insights have we gained? Front. Neurol. 10.

Franke, K., Gaser, C., Manor, B., Novak, V., 2013. Advanced brainAGE in older adults with type 2 diabetes mellitus. Front. Aging Neurosci. 5.

Franke, K., Luders, E., May, A., Wilke, M., Gaser, C., 2012. Brain maturation: predicting individual brainAGE in children and adolescents using structural MRI. Neuroimage 63 (3), 1305–1312.

Franke, K., Ziegler, G., Klppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50 (3), 883–892.

Gamaldo, A.A., Ferrucci, L., Rifkind, J., Longo, D.L., Zonderman, A.B., 2013. Relationship between mean corpuscular volume and cognitive performance in older adults. J Am Geriatr Soc 61, 84–89.

Gaser, C., Franke, K., Klåppel, S., Koutsouleris, N., Sauer, H., Initiative, A.s., 2013. BrainAGE in mild cognitive impaired patients: predicting the conversion to alzheimers disease. PLoS ONE 8 (6), e67346.

Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., Peng, H., 2021. Optimising a simple fully convolutional network for accurate brain age prediction in the PAC 2019 challenge. Front Psychiatry 12, 658.

Hahn, T., Fisch, L., Ernsting, J., Winter, N.R., Leenings, R., Sarink, K., Emden, D., Kircher, T., Berger, K., Dannlowski, U., 2021. From 'loose fitting' to high-performance, uncertainty-aware brain-age modelling. Brain 144 (3), 31.

Han, L.K.M., Dinga, R., Hahn, T., Ching, C.R.K., Eyler, L.T., Aftanas, L., Aghajani, M., Aleman, A., Baune, B.T., Berger, K., Brak, I., Filho, G.B., Carballedo, A., Connolly, C.G., Couvy-Duchesne, B., Cullen, K.R., Dannlowski, U., Davey, C.G., Dima, D., Duran, F.L.S., Enneking, V., Filimonova, E., Frenzel, S., Frodl, T., Fu, C.H.Y., Godlewska, B.R., Gotlib, I.H., Grabe, H.J., Groenewold, N.A., Grotegerd, D., Gruber, O., Hall, G.B., Harrison, B.J., Hatton, S.N., Hermesdorf, M., Hickie, I.B., Ho, T.C., Hosten, N., Jansen, A., Khler, C., Kircher, T., Klimes-Dougan, B., Krmer, B., Krug, A., Lagopoulos, J., Leenings, R., MacMaster, F.P., MacQueen, G., McIntosh, A., McLellan, Q., McMahon, K.L., Medland, S.E., Mueller, B.A., Mwangi, B., Osipov, E., Portella, M.J., Pozzi, E., Reneman, L., Repple, J., Rosa, P.G.P., Sacchet, M.D., Smann, P.G., Schnell, K., Schrantee, A., Simulionyte, E., Soares, J.C., Sommer, J., Stein, D.J., Steinstrter, O., Strike, L.T., Thomopoulos, S.I., Tol, M.J., Veer, I.M., Vermeiren, R.R.J.M., Walter, H., Wee, N.J.A., Werff, S.J.A., Whalley, H., Winter, N.R., Wittfeld, K., Wright, M.J., Wu, M.J., Vlzke, H., Yang, T.T., Zannias, V., Zubicaray, G.I., Zunta-Soares, G.B., Ab, C., Alda, M., Andreassen, O.A., Bnm, E., Bonnin, C.M., Canales-Rodriguez, E.J., Cannon, D., Caseras, X., Chaim-Avancini, T.M., Elvsshagen, T.r., Favre, P., Foley, S.F., Fullerton, J.M., Goikolea, J.M., Haarman, B.C.M., Hajek, T., Henry, C., Houenou, J., Howells, F.M., Ingvar, M., Kuplicki, R., Lafer, B., Landn, M., Machado-Vieira, R., Malt, U.F., McDonald, C., Mitchell, P.B., Nabulsi, L., Otaduy, M.C., Overs, B.J., Polosan, M., Pomarol-Clotet, E., Radua, J., Rive, M.M., Roberts, G., Ruhe, H.G., Salvador, R., Sarr, S., Satterthwaite, T.D., Savitz, J., Schene, A.H., Schofield, P.R., Serpa, M.H., Sim, K., Soeiro-de Souza, M.G., Sutherland, A.N., Temmingh, H.S., Timmons, G.M., Uhlmann, A., Vieta, E., Wolf, D.H., Zanetti, M.V., Jahanshad, N., Thompson, P.M., Veltman, D.J., Penninx, B.W.J.H., Marquand, A.F., Cole, J.H., Schmaal, L., 2020. Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group. In: Mol Psychiatry, pp. 1–16. Publisher: Nature Publishing Group.

Høgestøl, E.A., Kaufmann, T., Nygaard, G.O., Beyer, M.K., Sowa, P., Nordvik, J.E., Kolskår, K., Richard, G., Andreassen, O.A., Harbo, H.F., Westlye, L.T., 2019. Cross-Sectional and Longitudinal MRI Brain Scans Reveal Accelerated Brain Aging in Multiple Sclerosis. Front. Neurol., Vol. 0. Publisher: Frontiers

He, S., Pereira, D., DavidPerez, J., Gollub, R.L., Murphy, S.N., Prabhu, S., Pienaar, R., Robertson, R.L., EllenGrant, P., Ou, Y., 2021. Multi-channel attention-fusion neural network for brain age estimation: accuracy, generality, and interpretation with 16,705 healthy MRIs across lifespan. Med Image Anal 72, 102091.

Hofmann, S.M., Beyer, F., Lapuschkin, S., Loeffler, M., Müller, K.R., Villringer, A., Samek, W., Veronica, W.A., 2021. Towards the Interpretability of Deep Learning Models for Human Neuroimaging. bioRxiv, 2021.06.25.449906. Publisher: Cold Spring Harbor Laboratory Section: New Results

Jacobsen, C., Hagemeier, J., Myhr, K.M., Nyland, H., Lode, K., Bergsland, N., Ramasamy, D.P., Dalaker, T.O., Larsen, J.P., Farbu, E., Zivadinov, R., 2014. Brain atrophy and disability progression in multiple sclerosis patients: a 10-year follow-up study. Journal of Neurology, Neurosurgery & Psychiatry 85 (10), 1109–1115. Publisher: BMJ Publishing Group Ltd Section: Multiple sclerosis

Jenkinson, M., 2001. A global optimisation method for robust affine registration of brain images. Med Image Anal 5 (2), 143–156.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage 62 (2), 782–790.

Johnson, M.H., 2001. Functional brain development in humans. Nat Rev Neurosci 2 (7), 475–483. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Reviews Publisher: Nature Publishing Group

Jonsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., Ellingsen, L.M., Bragi, W.G., Gudbjartsson, D.F., Stefansson, H.H., Stefansson, K.K., Ulfarsson, M.O., 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nat Commun 10 (1), 5409.

Karasik, D., Demissie, S., Adrienne, C.L., Kiel, D.P., 2005. Disentangling the genetic determinants of human aging: biological age as an alternative to the use of survival measures. The Journals of Gerontology: Series A 60 (5), 574–587.

Kaufmann, T., Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Aln, D., Barch, D.M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M.K., Bøem, E., Borgwardt, S., Brandt, C.L., Buitelaar, J., Celius, E.G., Cervenka, S., Conzelmann, A., Crdova-Palomera, A., Dale, A.M., Quervain, D.J.F., DiCarlo, P., Djurovic, S., Dnm, D.S., Eisenacher, S., Elvsshagen, T.r, Espeseth, T., Fatouros-Bergman, H., Flyckt, L., Franke, B., Frei, O., Haatveit, B., Hberg, A.K., Harbo, H.F., Hartman, C.A., Heslenfeld, D., Hoekstra, P.J., A, H.E., Jernigan, T.L., Jonassen, R., Jnsson, E.G., Kirsch, P., Knm, I., Kolskår, K., Landr, N.I., Hellard, S.L., Lesch, K.P., Lovestone, S., Lundervold, A., Lundervold, A.J., Maglanoc, L.A., Malt, U.F., Mecocci, P., Melle, I., Meyer-Lindenberg, A., Moberget, T., Norbom, L.B., Nordvik, J.E., Nyberg, L., Oosterlaan, J., Papalino, M., Papassotiropoulos, A., Pauli, P., Pergola, G., Persson, K., Richard, G.v., Rokicki, J., Sanders, A.M., Selb, G., Shadrin, A.A., Smeland, O.B., Soininen, H., Sowa, P., Steen, V.M., Tsolaki, M., Ulrichsen, K.M., Vellas, B., Wang, L., Westman, E., Ziegler, G.C., Zink, M., Andreassen, O.A., Westlye, L.T., 2019. Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nat Neurosci 22 (10), 1617–1623.

de Lange, A.-M.G., Anatürk, M., Suri, S., Kaufmann, T., Cole, J.H., Griffanti, L., Zsoldos, E., Jensen, D.E.A., Filippini, N., Singh-Manoux, A., Kivimäki, M., Westlye, L.T., Ebmeier, K.P., 2020. Multimodal brain-age prediction and cardiovascular risk: the Whitehall II MRI sub-study. Neuroimage 222, 117292.

de Lang, A.-M.G., Cole, J.H., 2020. Commentary: correction procedures in brain-age prediction. Neuroimage Clin 26, 102229.

Beck, D., de Lang, A.-M.G., Pedersen, M.L., Alnaes, D., Maximov, I.I., Voldsbekk, I., Richard, G., Sanders, A.M., Ulrichsen, K.M., Dørum, E.S., Kolskår, K.K., Høgestøl, E.A., Steen, N.E., Djurovic, S., Andreassen, O.A., Nordvik, J.E., Kaufmann, T., Westlye, L.T., 2021. Cardiometabolic Risk Factors Associated with Brain Age and Accelerate Brain Ageing. Hum Brain Mapp.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., KharabianMasouleh, S., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Cameron, C.R., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M.L., Witte, A.V., Villringer, A., Margulies, D.S., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. Neuroimage 148, 179–188.

Lu, B., Li, H. X., Chang, Z. K., Li, L., Chen, N. X., Zhu, Z. C., Zhou, H. X., Li, X. Y., Wang, Y. W., Cui, S. X., Deng, Z. Y., Fan, Z., Yang, H., Chen, X., Thompson, P. M., Castellanos, F. X., Yan, C. G., for the Alzheimer's Disease Neuroimaging Initiative, 2020. A Practical Alzheimer Disease Classifier via Brain Imaging-Based Deep Learning on 85,721 Samples.

Millard, M., Davies, N.M., Gaunt, T.R., DaveySmith, G., Tilling, K., 2018. Software application profile: PHESANT: a tool for performing automated phenome scans in UK biobank. Int J Epidemiol 47 (1), 29–35.

Nenadi, I., Dietzek, M., Langbein, K., Sauer, H., Gaser, C., 2017. BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. Psychiatry Research: Neuroimaging 266, 86–89.

Oquendo, M.A., Baca-Garcia, E., Artès-Rodrìguez, A., Perez-Cruz, F., Galfalvy, H.C., Blasco-Fontecilla, H., Madigan, D., Duan, N., 2012. Machine learning and data mining: strategies for hypothesis generation. Mol Psychiatry 17 (10), 956–959.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Scikit-learn: Machine Learning in Python. In: MACHINE LEARNING IN PYTHON, p. 6.

Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021. Accurate brain age prediction with lightweight deep neural networks. Med Image Anal 68, 101871.

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., Frisoni, G.B., 2016. Brain atrophy in Alzheimer's disease and aging. Ageing Res. Rev. 30, 25–48.

Ren, Y., Luo, Q., Gong, W., Lu, W., 2019. Transfer learning models on brain age prediction. In: Proceedings of the Third International Symposium on Image Computing and Digital Medicine - ISICDM 2019, pp. 278–282.

Richard, G.v., Kolskår, K., Sanders, A.M., Kaufmann, T., Petersen, A., Doan, N.T., Snchez, J.M., Aln, D., Ulrichsen, K.M., Dnm, E.S., Andreassen, O.A., Nordvik, J.E., Westlye, L.T., 2018. Assessing distinct patterns of cognitive aging using tissue-specific brain age prediction based on diffusion tensor imaging and brain morphometry. PeerJ 6, e5908.

Rokicki, J., Wolfers, T., Nordh, W., Tesli, N., Quintana, D.S., Aln, D., Richard, G., Lange, A.-M.d., Lund, M.J., Norbom, L., Agartz, I., Melle, I., Nnm, T., Selb, G., Persson, K., Nordvik, J.E., Schwarz, E., Andreassen, O.A., Kaufmann, T., Westlye, L.T., 2021. Multimodal imaging improves brain age prediction and reveals distinct abnormalities in patients with psychiatric and neurological disorders. Hum Brain Mapp 42 (6), 1714–1726.

Roth, J., Müller, N., Lehmann, T., Heinemann, L., Wolf, G., Müller, U.A., 2016. HbA1c and age in non-diabetic subjects: an ignored association? Exp Clin Endocrinol Diabetes 124 (10), 637–642.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. Int J Comput Vis 115 (3), 211–252.

Schnack, H.G., Haren, N.E.M., Nieuwenhuis, M., HulshoffPol, H.E., Cahn, W., Kahn, R.S., 2016. Accelerated brain aging in schizophrenia: alongitudinal pattern recognition study. AJP 173 (6), 607–616.

Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and Statistical Modeling with Python. 92–96

Sgonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22 (3), 1060–1075.

SharifRazavian, A., Azizpour, H., Sullivan, J., Carlsson, S., 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. 806–813

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv: 1409.1556.

Smith, L. N., 2017. Cyclical Learning Rates for Training Neural Networks. ArXiv: 1506.01186.

Smith, S.M., Elliott, L.T., Alfaro-Almagro, F., McCarthy, P., Nichols, T.E., Douaud, G., Miller, K.L., Peelle, J.E., Lange, F.P., Madan, C., Nyberg, L., 2020. Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. Elife 9, e52677.

Smith, S.M., Nichols, T.E., 2018. Statistical challenges in 'big data' human neuroimaging. Neuron 97 (2), 263–268.

Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E., Miller, K.L., 2019. Estimation of brain age delta from brain imaging. Neuroimage 200, 528–539.

Stephan, B.C.M., Brayne, C., 2009. Assessing the risk of dementia in the aging population. Nat Rev Neurol 5 (8), 417–418.

Sullivan, G.M., Feinn, R., 2012. Using effect size or why the p value is not enough. J Grad Med Educ 4 (3), 279–282.

Thapar, A., Riglin, L., 2020. The importance of a developmental perspective in psychiatry: what do recent genetic-epidemiological findings show? Mol Psychiatry 25 (8), 1631–1639.

Thompson, P.M., Jahanshad, N., Ching, C.R.K., Salminen, L.E., Thomopoulos, S.I., Bright, J., Baune, B.T., Bertoln, S., Bralten, J., Bruin, W.B., Blow, R., Chen, J., Chye, Y., Dannlowski, U., Kovel, C.G.F., Donohoe, G., Eyler, L.T., Faraone, S.V., Favre, P., Filippi, C.A., Frodl, T., Garijo, D., Gil, Y., Grabe, H.J., Grasby, K.L., Hajek, T., Han, L.K.M., Hatton, S.N., Hilbert, K., Ho, T.C., Holleran, L., Homuth, G., Hosten, N., Houenou, J., Ivanov, I., Jia, T., Kelly, S., Klein, M., Kwon, J.S., Laansma, M.A., Leerssen, J., Lueken, U., Nunes, A., Neill, J.O., Opel, N., Piras, F., Piras, F., Postema, M.C., Pozzi, E., Shatokhina, N., Soriano-Mas, C., Spalletta, G., Sun, D., Teumer, A., Tilot, A.K., Tozzi, L., Merwe, C., VanSomeren, E.J.W., Wingen, G.A., Vlzke, H., Walton, E., Wang, L., Winkler, A.M., Wittfeld, K., Wright, M.J., Yun, J.Y., Zhang, G., Zhang-James, Y., Adhikari, B.M., Agartz, I., Aghajani, M., Aleman, A., Althoff, R.R., Altmann, A., Andreassen, O.A., Baron, D.A., Bartnik-Olson, B.L., MarieBas-Hoogendam, J., Baskin-Sommers, A.R., Bearden, C.E., Berner, L.A., Boedhoe, P.S.W., Brouwer, R.M., Buitelaar, J.K., Caeyenberghs, K., Cecil, C.A.M., Cohen, R.A., Cole, J.H., Conrod, P.J., DeBrito, S.A., Zwarte, S.M.C., Dennis, E.L., Desrivieres, S., Dima, D., Ehrlich, S., Esopenko, C., Fairchild, G., Fisher, S.E., Fouche, J.P., Francks, C., Frangou, S., Franke, B., Garavan, H.P., Glahn, D.C., Groenewold, N.A., Gurholt, T.P., Gutman, B.A., Hahn, T., Harding, I.H., Hernaus, D., Hibar, D.P., Hillary, F.G., Hoogman, M., HulshoffPol, H.E., Jalbrzikowski, M., Karkashadze, G.A., Klapwijk, E.T., Knickmeyer, R.C., Kochunov, P., Koerte, I.K., Kong, X.Z., Liew, S.L., Lin, A.P., Logue, M.W., Luders, E., Macciardi, F., Mackey, S., Mayer, A.R., McDonald, C.R., McMahon, A.B., Medland, S.E., Modinos, G., Morey, R.A., Mueller, S.C., Mukherjee, P., Namazova-Baranova, L., Nir, T.M., Olsen, A., Paschou, P., Pine, D.S., Pizzagalli, F., Rentera, M.E., Rohrer, J.D., Smann, P.G., Schmaal, L., Schumann, G., Shiroishi, M.S., Sisodiya, S.M., Smit, D.J.A., Snm, I.E., Stein, D.J., Stein, J.L., Tahmasian, M., Tate, D.F., Turner, J.A., Heuvel, O.A., Wee, N.J.A., Werf, Y.D., Erp, T.G.M., Haren, N.E.M., Rooij, D., Velzen, L.S., Veer, I.M., Veltman, D.J., Villalon-Reina, J.E., Walter, H., Whelan, C.D., Wilde, E.A., Zarei, M., Zelman, V., 2020. ENIGMA and global neuroscience: a decade of large-scale studies of the brain in health and disease across more than 40 countries. Transl Psychiatry 10 (1), 1–28.

Valverde, J.M., Imani, V., Abdollahzadeh, A., DeFeo, R., Prakash, M., Ciszek, R., Tohka, J., 2021. Transfer learning in magnetic resonance brain imaging: a systematic review. Journal of Imaging 7 (4), 66.

Vidal-Piñeiro, D., Wang, Y., Krogsrud, S., Amlien, I., Baarè, W., Bartrès-Faz, D., Bertram, L., Brandmaier, A.M., Drevon, C., Düzel, S., Ebmeier, K., Henson, R.N., Junque, C., Kievit, R.A., Khn, S., Leonardsen, E., Lindenberger, U., Madsen, K., Magnussen, F., Mowinckel, A., Nyberg, L., Roe, J., Segura, B., Smith, S., Suri, S., Westerhausen, R., Zalesky, A., Zsoldos, E., Ageing, T., Walhovd, K., Fjell, A., 2021. Individual Variations in 'Brain Age' Relate to Early Life Factors More than to Longitudinal Brain Change. bioRxiv, 2021.02.08.428915.

Walhovd, K.B., Fjell, A.M., Reinvang, I., Lundervold, A., Dale, A.M., Eilertsen, D.E., Quinn, B.T., Salat, D., Makris, N., Fischl, B., 2005. Effects of age on volumes of cortex, white matter and subcortical structures. Neurobiol. Aging 26 (9), 1261–1270.

Wang, J., Knol, M.J., Tiulpin, A., Dubost, F., Bruijne, M.d., Vernooij, M.W., Adams, H.H.H., Arfan, I.M., Niessen, W.J., Roshchupkin, G.V., 2019. Gray matter age prediction as a biomarker for risk of dementia. PNAS 116 (42), 21213–21218.

Wrigglesworth, J., Ward, P., Harding, I.H., Nilaweera, D., Wu, Z., Woods, R.L., Ryan, J., 2021. Factors associated with brain ageing - a systematic review. BMC Neurol 21 (1), 312.

Xia, T., Chartsias, A., Wang, C., Tsaftaris, S.A., 2021. Learning to synthesise the ageing brain without longitudinal data. Med Image Anal 73, 102169.