Cross-sectional brain age assessments are limited in predicting future brain change

Max Korbmacher[1,2*], Didac Vidal-Pineiro[3], Meng-Yun Wang[4], Dennis van der Meer[5], Thomas Wolfers[6], Hajer Nakua[7], Eli Eikefjord[1,2], Ole A. Andreassen[5,9], Lars T. Westlye[5,8,9], Ivan I. Maximov[1]

[1]Department of Health and Functioning, Western Norway University of Applied Sciences, Bergen, Norway.
[2]Mohn Medical Imaging and Visualization Centre (MMIV), Bergen, Norway.
[3]Center for Lifespan Changes in Brain and Cognition, Department of Psychology, University of Oslo, Oslo, Norway.
[4]Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands.
[5]Centre for Precision Psychiatry, Division of Mental Health and Addiction, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway.
[6]H, Tübingen, Germany.
[7]Columbia University Irving Medical Centre, Columbia University, New York City, USA.
[8]Department of Psychology, University of Oslo, Oslo, Norway.
[9]KG Jebsen Centre for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway.

Draft

Corresponding author: max.korbmacher@hvl.no;

ABSTRACT

The concept of brain age (BA) describes an integrative imaging marker of brain health, often suggested to reflect ageing processes. However, the degree to which cross-sectional MRI features, including BA, reflect past, ongoing and future brain changes across different tissue types from macro- to microstructure remains controversial. Multimodal imaging data of $\mathbf{39,325}$ UK Biobank participants, aged $\mathbf{44-82}$ years at baseline and $\mathbf{2,520}$ follow-ups within $\mathbf{1.12-6.90}$ years, provide insufficient evidence that BA reflects the rate of brain ageing.

Biomarkers which successfully characterise ageing still need to be established. An emerging candidate for such a marker is the concept of biological brain age (BA). Algorithms that predict BA provide insight into the differences between imaging metrics of healthy populations and independent target populations, for example, presenting a certain pathology. BA can be predicted from different types of imaging data, such as different modalities or brain regions[1]. The difference between BA and chronological age, called the brain age gap (BAG), has been used as a proxy for brain health. Previous studies identified the largest group-level differences in BAG between healthy controls and individuals with neurodegenerative disorders[2, 3] which makes BAG particularly interesting in the context of ageing; both healthy and pathological. To increase the clinical utility of BAG metrics, it is necessary to understand the degree to which cross-sectional BAG can predict brain ageing later in life [4, 5].

Here, we capitalised on the largest accessible multimodal magnetic resonance imaging (MRI) dataset featuring T_1-weighted and diffusion MRI from the UK Biobank including thousands of healthily ageing participants. These two modalities have previously been shown to be accurate BA predictors[1, 2, 6]. After exclusions based on poor MRI data, participant withdrawal from the study, and presence of a psychiatric or neurological disorder based on the ICD-10 (see Online Methods), we retained baseline brain scans of a total of N_{TP1}=39,325 individuals. BA prediction models were trained on data from participants without available follow-up scans (N = 36,805, aged 64.63 ± 7.70 years, range: $44.57 - 82.75$ years).

For model training, different machine learning algorithms were implemented, using k-fold cross-validation with 5 outer and 10 inner folds, including hyperparameter tuning, to determine the best performing model (see Online Methods for details on algorithms probed). Among the various approaches, the best performing algorithm was linear regression which was ultimately used for testing, predicting individual BA from T_1-weighted and diffusion MRI-extracted brain features individually as well as their combination (multimodal MRI) on two data points from a subset of participants with follow-up scans (N=2,520), aged 62.22 ± 7.23 years (range: $46.63 - 80.30$ years) at baseline. The follow-up scan was obtained within 2.45 ± 0.75 years from the baseline scan (range: $1.12 - 6.90$ years). To account for age bias introduced by the training sample's age-distribution, we used a linear age correction (Online Methods). The brain features used in the three models were region-averaged cortical surface area, volume, and thickness measures extracted from FreeSurfer[7] recon-all pipeline (208 total features), and various diffusion measures from conventional and advanced diffusion approaches (1,794 total features, see Online Methods).

Despite the short inter-scan interval (ISI), we could observe tissue maturation indicated by significant time-point differences in MRI-derived regional brain features (cortical thickness, surface area, cortical volume and diffusion metrics across brain-regions, Fig. 1d). More than 90% of the T_1-weighted, and more than 78% of the diffusion-derived features changed significantly ($|\bar{d}_{T_1w}| = 0.26$, $|\bar{d}_{dMRI}| = 0.15$; Fig.1d, Suppl. Data 1), with larger magnitude of these changes observed for T_1-weighted ($|d| = 0.20 - 0.30$; Suppl. Fig. 5) compared to diffusion metrics ($|d| = 0.12 - 0.16$; see Suppl. Fig. 6 metric-level changes). The features that showed significant change (p <0.05, N_{T_1w}=202,N_{dMRI}=1618) between baseline and follow-up were then used to

compute principal components of the centercepts/averages ($\bar{PC}$) and the annual rate of change of these features (ΔPC; Suppl. Figs. 1, 3).

Fig. 1 demonstrates that BA trained on cross-sectional data can be applied in longitudinal data. Training and test sample characteristics were similar (Fig. 1a). BA predictions in the test sample were strongly associated with chronological age ($r_{uncorrectedBA} > 0.68$, $r_{correctedBA} > 0.97$; Fig. 1b, Suppl. Table 1) and correlated between time points ($r > 0.91$; Suppl. Table 2). The resulting BAGs were also strongly correlated between time points ($r > 0.80$), reflecting the feature correlations between time points ($r_{\bar{T_1}w} > 0.85, r_{d\bar{M}RI} > 0.76$, Suppl. Fig 19), and BAGs increased significantly over time($\beta_{std} > 0.08, \beta_{years} = 0.08 - 2.32, p < 2{\times}10^{-10}$; Fig. 1c, Suppl. Tables 3,12), indicating accelerated brain ageing.

Fig. 2 illustrates that baseline (cross-sectional) BAG, represented by centercept/average of the BAGs ($B\bar{A}G$), based on T_1-weighted features is limited in predicting longitudinal brain changes. Within modalities, only the T_1-weighted based $B\bar{A}G$ was significantly and positively associated with the annual rate of BAG change (ΔBAG; $\beta_{std} = 0.028 \pm 0.148$; Fig. 2a) and the principal component of longitudinal feature changes ΔPC ($\beta_{std} = 0.054 \pm 0.015$; Fig. 2b, Suppl. Table 6). Moreover, $B\bar{A}G$ from dMRI ($\beta_{std} = 0.046 \pm 0.018$) and multimodal MRI ($\beta_{std} = 0.096 \pm 0.037$) were associated with the T_1-weighted ΔPC. Other BAGs associated negatively and/or non-significantly with ΔPCs across modalities. Cross-sectional principal components ($\bar{PC}$) were limited in predicting principal components of longitudinal feature changes ΔPC (all associations were negative: $|\beta_{std}| < 0.063$; Suppl. Table 4). Finally, Fig. 2 highlights the importance of age-bias correction: corrected $B\bar{A}Gs$ were stronger related to ΔPCs and $\Delta BAGs$ than uncorrected $B\bar{A}Gs$.

The stronger association between $B\bar{A}G$ from T_1-weighted features and both the respective principal component of change and ΔBAG were not necessarily reflected when predicting the regional ageing patterns across metrics within modalities, outlining significant associations in only 9% of change in specific brain regions. While cross-sectional T_1-weighted BAG was limited in predicting future change in T_1-weighted features, longitudinal T_1-weighted BAG significantly correlated with 75% of these changes (Fig. 2c).

DMRI-based cross-sectional BAG reflected a higher proportion (38%) of regional brain change, also indicated by larger significant effect sizes ($|\bar{\beta}| = 0.140$), compared to T_1-weighted ($|\bar{\beta}| = 0.121$), but not multimodal BAG ($|\bar{\beta}| = 0.190$). The strongest regional dMRI-based BAG associations were found for BRIA's microscopic fractional anisotropy, in the anterior corona radiata ($\beta = -0.263$), superior longitudinal fasciculus ($\beta = -0.262$), and intra-axonal water fraction in the corpus callosum ($\beta = -0.262$) and superior longitudinal fasciculus ($\beta = -0.238$). However, the overall strongest associations with $B\bar{A}G$ were found for skeleton averages of microscopic fractional anisotropy ($\beta = -0.280$) and the intra-axonal water fraction ($\beta = -0.277$).

As a higher BAG can be expected at higher ages and potentially also the rate of change in BAG to accelerate, we show that our analyses are independent of both, by correcting for the age bias (see Online Methods), and by showing that baseline BAG can predict future changes in BAG, independent of ISI, between baseline and follow-up. This was indicated by the effect of the interaction between the ISI and $B\bar{A}G$

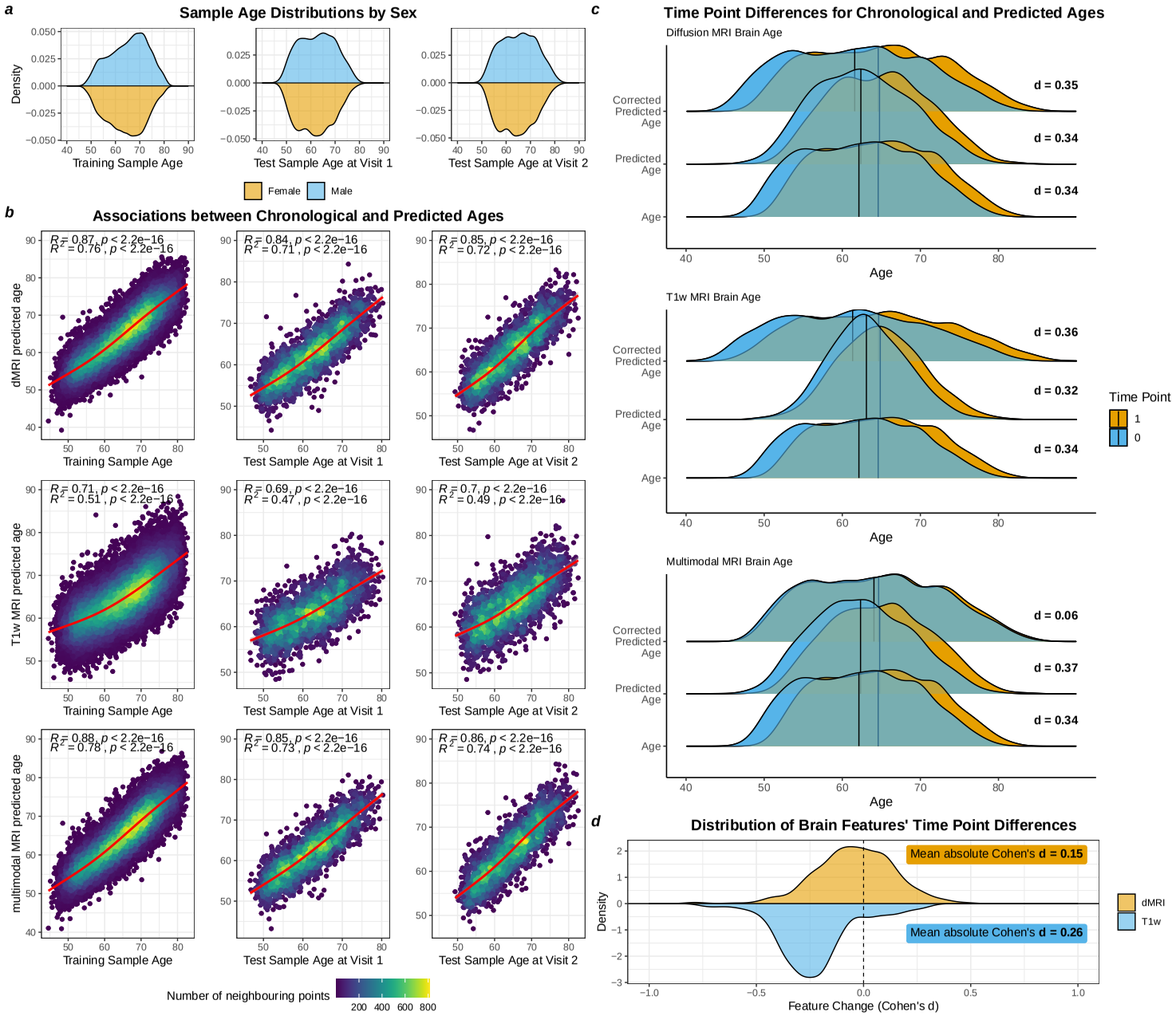

Fig. 1 Training and test sample had similar characteristics and brain age was predicted with high accuracy in training and each test data time point individually. Moreover, we denote brain changes indicated by regional a) Sample age distribution at each visit, separating the cross-sectional training data from the longitudinal test data. b) Model Performance for the training set, and the two test points for each MRI modality. Uncorrected estimates are presented, which were overlaid with a cubic spline with k = 4 knots. c) Time point differences for age and both crude and age-bias corrected BAs for each MRI modality indicated by Cohen's d. d) Distribution of effect sizes indicating the change in anatomical features of diffusion MRI (dMRI) and T_1-weighted MRI (T1w).

on ΔBAG being non-significant ($p > 0.05$; Suppl. Table 5, Suppl. Fig. 2) when using either a linear or cubic interaction term. This indicates that the observed associations between cross-sectional $B\bar{A}G$ and ΔBAG were independent of the ISI in the current study, and hence not just an artefact of study design, age or ageing.

BAG was limited in reflecting sub-clinical health characteristics. Our sample was selected to not contain neurological or psychiatric disorders and showed relatively stable health based on various health indicators. Yet, these health indicators were limited in reflecting BAG. Health characteristics were evaluated by examining different risk factors for age-related diseases and mortality, including cardiometabolics, depression, neuroticism, and polygenic risk scores (PGRS) of different disorders. Small associations were found between both BAGs and ΔBAG and different cross-sectional health indicators (Suppl. Fig. 4), including PGRS of common psychiatric disorders and Alzheimer's disease ($|\beta_{std}| < 0.06$,$p_{Bonferroni} > 0.05$), clinically relevant state (depression rating) and trait (neuroticism) assessment scores ($\beta_{std} < 0.08$,$p_{Bonferroni} > 0.05$), with larger group-level differences for cardiometabolic factors hypertension and diabetes ($\beta_{std} < 0.35$,$p_{Bonferroni} < 0.047$). Among the longitudinally available phenotypes, only waist-to-hip ratio (WHR), previously shown to be related to BAG[8], changed significantly between time points at the group level ($t = 10.36, d = 0.15, p < 2.2 \times 10^{-16}, p_{Bonferroni} < 2.2 \times 10^{-16}$). However, while WHR showed a small, significant association with BAG_{T_1w} at baseline ($\beta_{std} < 0.08$), WHR changes were not predicted by BAGs or PCs ($p > .05$; Suppl. Fig. 4). Neuroticism ($t = 2.83, d = 0.04, p = 0.005, p_{Bonferroni} = 0.030$) and depression ($t = 2.13, d = 0.04, p = 0.033, p_{Bonferroni} = 0.198$) scores decreased over time, however, changes in these scores were not found to be predicted by BAGs or PCs of brain feature change or centercepts ($p > .05$).

Taken together, our findings indicate that BAG is limited in reflecting longitudinal brain changes. Overall, a) cross-sectional BAGs presented small associations with longitudinal brain ages across modalities, b) only BAGs from T_1-weighted MRI features showed significant *but small* positive association with the respective longitudinal principal components, and c) BAGs explained less than 1% of the variance of the mentioned principal components and BAG change.

Yet, dMRI-based cross-sectional BAG correlated significantly with the annual change in around 38% of the region-level features (at a relatively small average effect of $|\bar{\beta}| = 0.140$). Assessing the rate of BAG change, T_1-weighted BAG correlated significantly with the largest portion of regional brain change (75%). Hence, despite BAG correlating weakly with future change in BAG and future change principal components, a single-time-point dMRI BAG might allow to capture a portion of future changes in brain morphometry on the region level, whereas T_1-weighted BAG is most reflective of the brain state. Future investigations might focus on the constructing explainable brain age models which leverage region-level data, and further investigate the potential of brain age in datasets with multiple follow-ups. Alternatively, other markers which reflect a person's deviation from a norm defined by the characteristics of a training dataset, might be of interest.

Brain age allows to reduce large amounts of information into a single personalised health score. Although brain age can be vulnerable to individual differences[5], such a

single number is intuitive when set in contrast to a person's chronological age, and does not require expert knowledge to be interpreted: a very high brain age in contrast to the chronological age might be alarming. Hence, brain age predictions hold the promise to provide additional information on routine clinical scans, and for example support incidental findings. Closer examinations of how BAG reflects developmental trajectories under different conditions and on different samples offer room for future research.

The identified increase in BAG over time indicates an acceleration of brain ageing during ageing without neurological or neuropsychiatric diagnosis, which has also previously been highlighted in white matter microstructure[9]. In contrast, BAG during adulthood without pathology can be expected to be stable, since tissue changes remain small[5]. However, here, we show that during pathology-free ageing, which is accompanied by regional brain changes, also the BAG will change over time. Hence, BAG provides an indicator of the brain's morphometric state. Such state has been shown to be influential for the future development of disorders, as indicated by disability accumulation in multiple sclerosis[10] or changes in dementia ratings[11]. However, the biological underpinnings of these associations remain unclear.

We observed considerable modality-dependent differences in brain ages. $B\bar{A}G_{dMRI}$ was most predictive of brain change. Modality-dependent differences might originate from the attempt to reduce a more complex feature space into single scores, such as brain age or principal components. Future modelling might focus on different spatial scales, such as voxel-level analysis, and simultaneously on different biophysical modelling approaches to extract meaningful brain metrics.

In conclusion, we find that cross-sectional BAG estimates are limited in reflecting future brain changes. This limits the potential of BAG for longitudinal inference and establishing BAG as a biomarker. During generally pathology-free ageing, BAG is not stable but increases, together with morphometric changes, potentially due to accelerated ageing. Yet, only dMRI-based BAG also reflected regional morphometric changes. These findings provide new and more pronounced insights into the mechanism of BAG. For example, a higher BAG does not automatically indicate the presence of a disorder, which would however be crucial for diagnostics. Instead, the observed modality dependencies suggest that dMRI and multimodal BAGs reflect the morphometric state, which is influenced by early life factors[4]. DMRI BAG might reflect regional brain changes better than the other approaches. This more nuanced understanding of BAG underscores the need for closer examinations of the biological underpinnings of BAG to aid the general interpretation of the marker and to increase clarity around BAG's clinical utility.

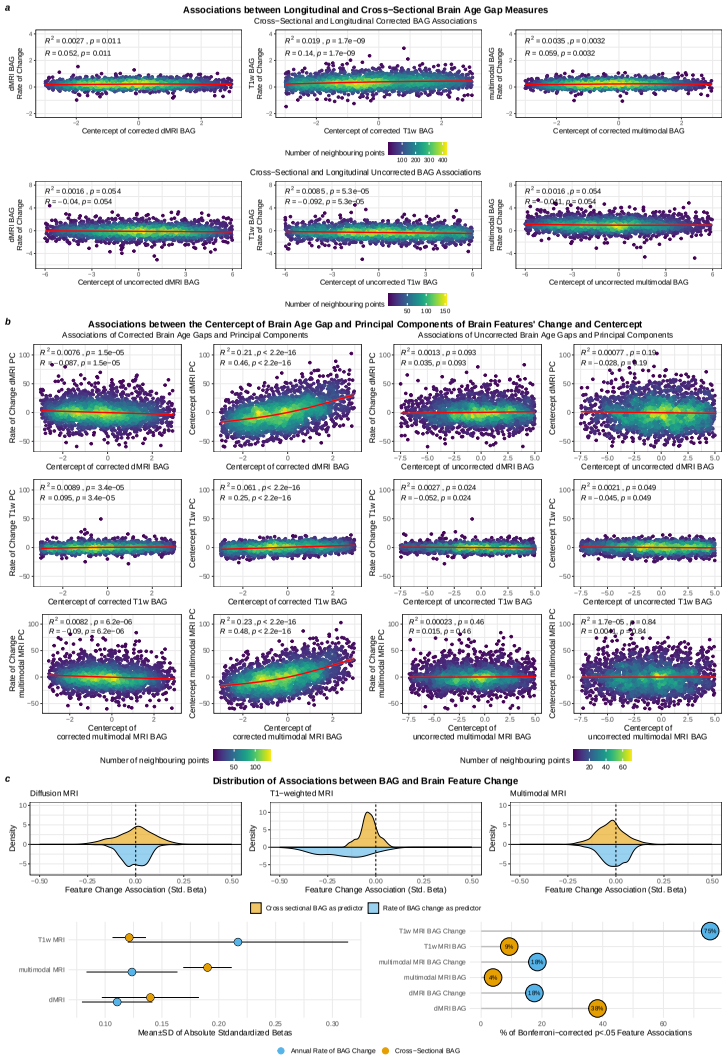

Fig. 2 BAG is overall limited in reflecting brain change, yet, T_1-weighted brain age reflects the strongest regional brain changes. a) Associations between uncorrected $B\bar{A}G$ and ΔBAG in the top row, and corrected associations in the bottom row. Associations were obtained specific to each modality: T_1-weighted (T1w), diffusion (dMRI), and multimodal MRI. The displayed line fits were cubic splines with k = 4 knots. b) Associations between the centercepts (proxy for cross-sectional BA measures) of each modality-specific BAG and PCs of both the centercepts and the annual rate of change in brain features. The left two columns present associations of uncorrected BAG estimates, and the right two columns of training-sample age-corrected BAG estimates, respectively. The displayed line fits were cubic splines with k = 4 knots. c) Top row: Distribution of associations between corrected BAG and brain features and annual change of brain features (including associations with $p_{Bonferroni} < .05$). Bottom left: Absolute mean and standard deviation of the associations between centercept and rate of change in corrected BAG and annual change of brain features. Bottom right: Percentage of significant associations between centercept and rate of change in corrected BAG and the annual change of brain features after Bonferroni-correction.

Online Methods

Sample characteristics

We obtained UKB data[12] containing dMRI data of $N = 46{,}637$ cross-sectional datasets, of which $N = 4{,}871$ entailed data available at two time points, and $N = 48{,}044$ T_1-weighted MRI datasets of which $N = 4{,}960$ were followed up. Participant data were excluded when consent had been withdrawn, or data quality deemed to be insufficient based on the YTTRIUM method[13] applied to dMRI data, and for T_1-weighted data based on Euler numbers[14], leading to exclusions when three standard deviations from the mean were exceeded. Additionally, we excluded participants which were diagnosed with any mental and behavioural disorder (ICD-10 category F), disease of the nervous system (ICD-10 category G), and disease of the circulatory system (ICD-10 category I). The remaining datasets, after the exclusions were applied, entailed $N = 36{,}805$ purely cross-sectional participants (52.17% females), which were used as a training set. The participants in the training sample were on average 64.79 ± 7.70 years old (range: $44.57 - 82.75$ years), with MRI scans obtained at four sites: (1) Cheadle (57.81%), (2) Newcastle (25.34%), (3) Reading (16.70%), and Bristol (0.15%). The independent testing set, not being a subset of the above mentioned training data, consisted of $N = 2{,}520$ participants (52.17% females) aged 62.26 ± 7.19 years at baseline (range: $46.12 - 80.30$ years), and at time-point two, the mean age was 64.67 ± 7.11 years (range: $49.33 - 82.59$ years), indicating an average age difference of $\Delta Age = 2.45 \pm 0.75$ years (range: $1.12 - 6.90$ years). The test data were collected at three sites: (1) in Cheadle (57.36%), (2) Newcastle (37.04%), and (3) Reading (5.60%).

MRI acquisition and post-processing

UKB MRI data acquisition procedures and protocols are described elsewhere[12, 15, 16]. After access to the raw dMRI data was obtained, we processed the data using an optimised pipeline[17]. The pipeline includes corrections for noise[18], Gibbs ringing[19], susceptibility-induced and motion distortions, and eddy current artifacts[20]. Isotropic $1\,\mathrm{mm}^3$ Gaussian smoothing was carried out using FSL's[21, 22] *fslmaths*. Employing the multi-shell data, Diffusion Tensor Imaging (DTI), Diffusion Kurtosis Imaging (DKI)[23] and White Matter Tract Integrity (WMTI)[24] metrics were estimated using Matlab 2017b code (https://gitgub.com/NYU-DiffusionMRI/DESIGNER). Spherical mean technique SMT[25], and multi-compartment spherical mean technique (mcSMT)[26] metrics were estimated using original code (https://github.com/ekaden/smt)[25, 26]. Estimates from the Bayesian Rotational Invariant Approach (BRIA) were evaluated by the original Matlab code (https://bitbucket.org/reisert/baydiff/src/master/)[27].

T_1-weighted images were processed using FreeSurfer (version 5.3)[7] automatic *recon-all* pipeline for a cortical reconstruction and subcortical segmentation of the T_1-weighted images (http://surfer.nmr.mgh.harvard.edu/fswiki)[28]. Notably, the influence of the FreeSurfer version on the brain age predictions was estimated as well and assumed to be small in this case[29].

In total, we obtained 26 WM metrics from six diffusion approaches (DTI, DKI, WMTI, SMT, mcSMT, BRIA; see for overview Suppl. Table 8). In order to normalise all metrics, we used Tract-based Spatial Statistics (TBSS)[30], as part of FSL[21, 22]. In brief, initially all brain-extracted[31] fractional anisotropy (FA) images were aligned to MNI space using non-linear transformation (FNIRT)[22]. Following, the mean FA image and related mean FA skeleton were derived. Each diffusion scalar map was projected onto the mean FA skeleton using TBSS. To provide a quantitative description of diffusion metrics at a region level, we used the John Hopkins University (JHU) atlas[32], and obtained 48 white matter regions of interest (ROIs) and 20 tract averages based on a probabilistic white matter atlas (JHU)[33] for each of the 26 metrics. Altogether, $1,794$ diffusion features were derived per individual [26 metrics $\times$ (48 ROIs + 20 tracts + 1 global skeleton mean value)]. For T_1-weighted data, we applied the Desikan-Killiany Atlas[34] to obtain regional estimates of thickness, area, and volume, leading to 208 features [3 metrics $\times$ (34 ROIs + 2 global mean values (left and right hemispheres)]. This results in a total of $2,002$ multimodal MRI features per individual.

Cardiometabolic risk factors

We used a selection of cardiometabolic risk factors, which have association with BAG and relevant to brain ageing[1]. Smoking, hypertension, and diabetes were binary and the waist-hip ratio (WHR) is a scalar value.

Depression and neuroticism scores

Depression scores were computed using the Recent Depressive Symptoms (RDS-4) score (fields 2050, 2060, 2070, 2080), which was suggested in a previous investigation using UKB imaging data.[35] Neuroticism scores (UKB data-field 20127) were derived as a summary score from the Eysenck Neuroticism (N-12) inventory which includes items describing neuroticism traits.

Polygenic risk scores (PGRS)

We estimated PGRS for each participant with available genomic data, using PRSice2[36] with default settings. As input for the PGRS, we used summary statistics from recent genome-wide association studies of Autism Spectrum Disorder (ASD)[37], Major Depressive Disorder (MDD)[38], Schizophrenia (SCZ)[39], Attention Deficit Hyperactivity Disorder (ADHD)[40], Bipolar Disorder (BIP)[41], Obsessive Compulsive Disorder (OCD)[42], Anxiety Disorder (ANX)[43], and Alzheimer's Disease (AD)[44]. We used a minor allele frequency of 0.05, as the threshold most commonly used in PGRS studies of psychiatric disorders.

While psychiatric disorders were p-values thresholded at $\alpha = 0.05$[37–45], recommendations for AD ($\alpha = 1.07^{-4}$)[46] lead to the application of a lower threshold of $\alpha = 0.0001$, with the goal of optimising signal to noise in comparison to previously used $\alpha = 0.001$[47]. The goal of the PGRS estimation was to relate the PGRS to cross-sectional and longitudinal BAG and the principal components. PGRS data were available for $N = 2{,}166$ of the longitudinal datasets (after exclusions).

Principal components (PC) of brain physiology

We conducted six principal component analyses, one for the centercepts of the features and one for the annual rate of change of features for each modality: 1) T_1-weighted, 2) diffusion, and 3) multimodal MRI features (as described in the MRI acquisition and processing section). The first component of each of the PC analyses was selected (rate of change components (ΔPC): T_1-weighted R^2=18.5%, dMRI R^2=23.5%, multimodal R^2=20.8; centercept components ($\bar{PC}$): T_1-weighted R^2=30.5%, dMRI R^2=39.1%, multimodal R^2=34.6; see Suppl. Fig. 1). For categorical feature contributions of the different approaches to the first two components per modality see Suppl. Fig. 3, and for the relative contribution by feature type see Suppl. Fig. 7. In brief, contributions of feature classes across regions (e.g., forceps fractional anisotropy) to PCs were relatively evenly distributed. This suggests that there is not a single dominantly contributing feature for dMRI and multimodal components. Yet, thickness contributed more than volume and surface area for the first T_1-weighted MRI component (Suppl. Figs. 3, 7).

Statistical analyses

All statistical analyses were carried out using *R* version (v4.2.0, www.r-project.org), Python (v3.7.1) and FSL (v6.0.2)[22]. *P*-values which adjusted for multiple comparison using FDR-correction[48] are marked with p_{FDR}, and *p*-values corrected using the the Bonferroni method are marked with $p_{Bonferroni}$. Standardised regression coefficients are labelled as β_{std}, unstandardized coefficients as β. For example, in a regression function $\hat{y} = \beta_0 + \beta_1 \times x + \epsilon$, β_1 is the slope, with β_0 the intercept, and ϵ the modelling error. When scaling variables to a standardized value (M = 0, SD = 1), β_{std} can be obtained. The proportion of true null effects/*p*-values was estimated using the Storey-Tibshirani method [49].

Brain age prediction

Brain age predictions refer to age predictions from a trained model in unseen data based on a set MRI features, here, regional brain metrics representing different grey and white matter microstructure characteristics. A higher predicted age indicates that the model assumes this person to be older, based on the presented brain features.

As the goal is to train models which are generalisable, we estimated the power of our model to do so under varying assumptions of parameter shrinkage, which quantifies the extend to which the model is generalisable. A larger shrinkage indicates lower generalisability. With the conservative assumption of a large to extremely large parameter shrinkage, for example, of 30%, 40%, 50%, and 60%, we estimated required training sample sizes of 14,826, 19,522, 25,848, and 35,117 participants to train a BA model on the selected maximum of 2,002 features (the multimodal MRI brain age model), respectively, using the *pmsampsize* package[50].

We tested several algorithms, including eXtreme Gradient Boosting (XGBoost)[51], the least absolute shrinkage and selection operator (LASSO)[52], and simple linear regression models using *k*-fold nested cross-validation with included hyperparamenter tuning on five inner and ten outer folds. Across all algorithms tested, linear regression models performed best in terms of variance explained and correlations of brain age

predictions and chronological age and commonly used error metrics, including Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) on the training sample and were, therefore, used to predict BA (Suppl. Table 7). The superiority of linear models was also underscored by a lower parameter shrinkage when predicting in the test data. These predictions are presented in the main text, whereas the results from the other algorithms are presented in the Supplement. Altogether, 2,002 features (i.e., brain regional metrics based on T_1-weighted or dMRI measures) were used per individual. After the training procedure was completed on the participants for which only a baseline scan was available, we predicted BA in the remaining participants ($N =$ 2,678) with tow available data points for each of these two study time points (baseline and follow-up).

We calculated corrected BA estimates by first calculating the intercept (α) and slope (β) of the linear associations between predicted BA (γ_{train}) and chronological age (Ω_{train}) in the *training* (baseline) sample (Eq. 1):

$$\hat{\Omega}_{train} = \alpha + \beta \times \gamma_{train}. \tag{1}$$

The calculated intercept (α) and slope (β) from the training sample were then used to estimate a corrected BAG ($BAGc$), as previously suggested[53], from the predicted age (γ_{test}) and chronological age (Ω_{test}) separately in each of data points of the *testing* (longitudinal) sample:

$$B\hat{A}Gc = (\gamma_{test} + (\Omega_{test} - (\alpha + \beta \times \gamma_{test}))) - \Omega_{test}. \tag{2}$$

We present the results for both corrected and uncorrected BAG.

As a control, we randomly split the longitudinal data into equal parts, trained models and predicted in $N_{TP1-2} = 1{,}339$ at each time point within the same individuals (due to the high dimensionality of the dMRI and multimodal data in contrast to the degrees of freedom, only T_1-weighted data were considered; Suppl. Tables 9−11). These predictions were used to repeat the analyses presented in the main text (Suppl. Note 1).

Rate of change and centercepts

In order to investigate how single time point BA predictions relate to longitudinal changes in BA and features, we estimated the annual rate of change and centercepts in both features and BAs. Centercepts were used to establish cross-sectional proxies (of the BAG, PCs, and brain features) which are statistically independent from the annual rate of change. Centercepts are the average of two measures without considering the inter-scan interval (ISI). The annual rate of change, on the other hand, has the ISI as denominator

Exploratory analyses

Time point correlations between brain ages at each time point were assessed using uncorrected Pearson's correlations. To assess time point (TP) difference in the a) corrected and b) uncorrected brain age gaps (BAG), we used mixed linear models

(MLMs) with ID, $Site$, Age, Sex, and the $Age * Sex$ interaction as fixed effects, the subject/ID as random effect (u), and the subject residuals (e).

$$\begin{aligned} B\hat{A}G = \beta_0 + \beta_1 \times TP + \beta_2 \times Age + \beta_3 \times Age * Sex + \\ \beta_4 \times Sex + \beta_5 \times Site + u_{ID} + e \end{aligned} \tag{3}$$

We used paired samples t-tests to assess features changes over time. Changing features were included in the the principal components analyses.

We evaluated the association of BA centercept or the annual rate of change in BA with a) the first cross-sectional principal component (of the centercepts of features) and b) the principal component of features' annual rate of change (PC), correcting for the ISI.

$$\begin{aligned} B\hat{A}G = \beta_0 + \beta_1 \times PC + \beta_2 \times ISI + \beta_3 \times Age + \\ \beta_4 \times Sex + \beta_5 \times Age * Sex + u_{Site} + e \end{aligned} \tag{4}$$

When assessing how a) the centercept of the BAG and b) the annual rate of change in BAG (BAG) reflect brain features and change in brain features (F), to ensure model convergence, we used simple linear models.

$$\begin{aligned} \hat{F} = \beta_0 + \beta_1 \times BAG + \beta_2 \times ISI + \beta_3 \times Age + \\ \beta_4 \times Sex + \beta_5 \times Age * Sex + \beta_6 \times Site \end{aligned} \tag{5}$$

Finally, we explored the associations between PC, BAG, and their annual rates of change $B\bar{A}G, \Delta BAG, \bar{PC}, \Delta PC$ (four outcome variables), and time-point specific principal components PC_{TP1} and PC_{TP2} and BAGs BAG_{TP1} and PC_{TP2} (another four outcome variables; all summarized in the formula as PC/BAG) with pheno- and genotypes (P/G), including PGRS of psychiatric disorders and Alzheimer's, depression and neuroticism scores, and cardiometabolic risk factors.

$$\begin{aligned} \hat{PC}/B\hat{A}G = \beta_0 + \beta_1 \times P/G + \beta_2 \times Age + \\ \beta_3 \times Sex + \beta_4 \times Age * Sex + \beta_5 \times Site \end{aligned} \tag{6}$$

The associations between PGRS and rate of change of PC and BAG had the lowest statistical power due to the limited availability of participant's genetic data ($N = 2,160$). We conducted a power analysis to ensure we would be able to detect meaningful effect sizes. We aimed for a power of 80%, and an α-level of 0.05 in simple linear regression models, as described above, indicating that effect sizes as small as Cohen's $f^2 = 0.006$ can be detected, corresponding to a Pearson's correlation coefficient of $r = 0.006$ or Cohen's $d = 0.012$.

Confirmatory analyses

We tested several preregistered hypotheses (see https://aspredicted.org/7bd4e.pdf)[1]. First, to test whether there were relationships between cross-sectional and longitudinal measures of the obtained BAGs and PCs, we associated their centercepts[54] and annual rates of change. We ran first MLMs predicting BAG changes (BAG_Δ) from cross sectional BAG ($B\hat{A}G$).We also predicted the longitudinal principal component of brain feature changes (PC_Δ) from the cross-sectional principal component ($\hat{PC}$) controlling for the inter-scan interval (ISI), age, sex, and the age-sex interaction as fixed effects, and scanning site as random effect (Eq. 7, 8).

$$\begin{aligned} H1a : B\hat{A}G_\Delta = \beta_0 + \beta_1 \times BAG_c + \beta_2 \times ISI + \beta_3 \times Age + \\ \beta_4 \times Sex + \beta_5 \times Age * Sex + u_{ID} + e \end{aligned} \tag{7}$$

$$\begin{aligned} H1b : P\hat{C}_\Delta = \beta_0 + \beta_1 \times PC_c + \beta_2 \times ISI + \beta_3 \times Age + \\ \beta_4 \times Sex + \beta_5 \times Age * Sex + u_{ID} + e \end{aligned} \tag{8}$$

To test the interaction effect of ISI and the respective cross-sectional measures (BAG_c and PC_c) we used generalised additive models (GAM) taking the same form of Eq. 8 and 8, yet introducing a spline chaining $k = 4$ cubic functions s on the interaction of interest:

$$\begin{aligned} H2_a : B\hat{A}G_\Delta = \beta_0 + \beta_1 \times BAG_c + \beta_2 \times ISI + \beta_3 \times Sex + \\ \beta_4 \times Age + \beta_5 \times Age * Sex + s(ISI * BAG_c) + (1|Site) \end{aligned} \tag{9}$$

$$\begin{aligned} H2_b : P\hat{C}_\Delta = \beta_0 + \beta_1 \times PC_c + \beta_2 \times ISI + \beta_3 \times Sex + \\ \beta_4 \times Age + \beta_5 \times Age * Sex + s(ISI * PC_c) + (1|Site) \end{aligned} \tag{10}$$

DATA AVAILABILITY

This study has been conducted using UKB data under Application 27412. All raw data are available from the UKB (www.ukbiobank.ac.uk). UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval.

The raw and processed UK Biobank MRI data are protected and are not openly available due to data privacy laws. However, access can be obtained by applying for access and paying an access fee (see https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access).

CODE AVAILABILITY

Analysis code is available at https://github.com/MaxKorbmacher/longitudinalBAG.git.

[1]Note: Instead of estimating brain ages for dMRI only, we also used T_1-weighted and multimodal data as well. For more meaningful results, we did not conduct a genome-wide association study with brain age and used instead polygenic risk scores of psychiatric disorders and Alzheimer's disease, as well as cardiometabolics and depression and neuroticism scores.

REFERENCES

[1] Korbmacher, M. *et al.* Brain asymmetries from mid-to late life and hemispheric brain age. *Nature Communications* **15**, 956 (2024).

[2] Franke, K. & Gaser, C. Ten years of brainage as a neuroimaging biomarker of brain aging: what insights have we gained? *Front. Neur.* 789 (2019).

[3] Kaufmann, T. *et al.* Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat. Neur.* **22**, 1617–1623 (2019).

[4] Vidal-Pineiro, D. *et al.* Individual variations in 'brain age'relate to early-life factors more than to longitudinal brain change. *elife* **10**, e69995 (2021).

[5] Korbmacher, M. *et al.* Considerations on brain age predictions from repeatedly sampled data across time. *Brain & Beh.* **13**, 1–8 (2023).

[6] De Lange, A.-M. G. *et al.* Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study. *NeuroImage* **222**, 117292 (2020).

[7] Fischl, B. Freesurfer. *NeuroImage* **62**, 774–781 (2012).

[8] Korbmacher, M. *et al.* Bio-psycho-social factors' associations with brain age: a large-scale uk biobank diffusion study of 35,749 participants. *Front. Psych.* **14**, 1117732 (2023).

[9] Korbmacher, M. *et al.* Distinct longitudinal brain white matter microstructure changes and associated polygenic risk of common psychiatric disorders and alzheimer's disease in the uk biobank. *medRxiv (Preprint)* (2023).

[10] Brier, M. R. *et al.* "brain age" predicts disability accumulation in multiple sclerosis. *Annals of clinical and translational neurology* **10**, 990–1001 (2023).

[11] Tseng, W.-Y. I., Hsu, Y.-C. & Kao, T.-W. Brain age difference at baseline predicts clinical dementia rating change in approximately two years. *Journal of Alzheimer's Disease* **86**, 613–627 (2022).

[12] Alfaro-Almagro, F. *et al.* Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *NeuroImage* **166**, 400–424 (2018).

[13] Maximov, I. I. *et al.* Fast qualitY conTrol meThod foR derIved diffUsion Metrics (YTTRIUM) in big data analysis: UK Biobank 18,608 example. *HBM* **42**, 3141–3155 (2021).

[14] Rosen, A. F. *et al.* Quantitative assessment of structural image quality. *NeuroImage* **169**, 407–418 (2018).

[15] Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neur.* **19**, 1523–1536 (2016).

[16] Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

[17] Maximov, I. I., Alnæs, D. & Westlye, L. T. Towards an optimised processing pipeline for diffusion magnetic resonance imaging data: Effects of artefact corrections on diffusion metrics and their age associations in uk biobank. *Human brain mapping* **40**, 4146–4162 (2019).

[18] Veraart, J., Fieremans, E. & Novikov, D. S. Diffusion MRI noise mapping using random matrix theory. *MR in Med.* **76**, 1582–1593 (2016).

[19] Kellner, E., Dhital, B., Kiselev, V. G. & Reisert, M. Gibbs-ringing artifact removal based on local subvoxel-shifts. *MR in Med.* **76**, 1574–1581 (2016).

[20] Andersson, J. L. & Sotiropoulos, S. N. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage* **125**, 1063–1078 (2016).

[21] Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23**, S208–S219 (2004).

[22] Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. FSL. *NeuroImage* **62**, 782–790 (2012).

[23] Jensen, J. H., Helpern, J. A., Ramani, A., Lu, H. & Kaczynski, K. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **53**, 1432–1440 (2005).

[24] Fieremans, E., Jensen, J. H. & Helpern, J. A. White matter characterization with diffusional kurtosis imaging. *Neuroimage* **58**, 177–188 (2011).

[25] Kaden, E., Kruggel, F. & Alexander, D. C. Quantitative mapping of the per-axon diffusion coefficients in brain white matter. *Magnetic resonance in medicine* **75**, 1752–1763 (2016).

[26] Kaden, E., Kelm, N. D., Carson, R. P., Does, M. D. & Alexander, D. C. Multi-compartment microscopic diffusion imaging. *NeuroImage* **139**, 346–359 (2016).

[27] Reisert, M., Kellner, E., Dhital, B., Hennig, J. & Kiselev, V. G. Disentangling micro from mesostructure by diffusion MRI: a Bayesian approach. *Neuroimage* **147**, 964–975 (2017).

[28] Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage* **9**, 179–194 (1999).

[29] Korbmacher, M., Westlye, L. T. & Maximov, I. I. Freesurfer version-shuffling can enhance brain age predictions. *Neuroimage: Reports* **4**, 100214 (2024).

[30] Smith, S. M. *et al.* Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage* **31**, 1487–1505 (2006).

[31] Smith, S. M. Fast robust automated brain extraction. *HBM* **17**, 143–155 (2002).

[32] Mori, S., Wakana, S., Nagae-Poetscher, L. & Van Zijl, P. MRI atlas of human white matter. *Am. J. Neurorad.* **27**, 1384 (2006).

[33] Hua, K. *et al.* Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *NeuroImage* **39**, 336–347 (2008).

[34] Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage* **31**, 968–980 (2006).

[35] Dutt, R. K. *et al.* Mental health in the uk biobank: A roadmap to self-report measures and neuroimaging correlates. *Human Brain Mapping* **43**, 816–832 (2022).

[36] Choi, S. W. & O'Reilly, P. F. Prsice-2: Polygenic risk score software for biobank-scale data. *Gigasci* **8**, giz082 (2019).

[37] Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of gwas of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Mol Autism* **8**, 1–17 (2017).

[38] Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Gen* **50**, 668–681 (2018).

[39] Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nat* **604**, 502–508 (2022).

[40] Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Gen* **51**, 63–75 (2019).

[41] Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Gen* **53**, 817–829 (2021).

[42] Arnold, P. D. *et al.* Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol Psych* **23**, 1181–1181 (2018).

[43] Otowa, T. *et al.* Meta-analysis of genome-wide association studies of anxiety disorders. *Mol Psych* **21**, 1391–1399 (2016).

[44] Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for alzheimer's disease. *Nat Gen* **53**, 1276–1282 (2021).

[45] Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat Gen* **45**, 1452–1458 (2013).

[46] Clark, K., Leung, Y. Y., Lee, W.-P., Voight, B. & Wang, L.-S. Polygenic risk scores in alzheimer's disease genetics: methodology, applications, inclusion, and diversity. *J Alz Dis* **89**, 1–12 (2022).

[47] Euesden, J., Lewis, C. M. & O'reilly, P. F. Prsice: polygenic risk score software. *Bioinf* **31**, 1466–1468 (2015).

[48] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

[49] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).

[50] Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: Part ii-binary and time-to-event outcomes. *Statistics in medicine* **38**, 1276–1296 (2019).

[51] Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system*, 785–794 (2016).

[52] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**, 267–288 (1996).

[53] de Lange, A.-M. G. *et al.* Mind the gap: Performance metric evaluation in brain-age prediction. *HBM* **43**, 3113–3129 (2022).

[54] Wainer, H. The centercept: An estimable and meaningful regression parameter. *Psychological Science* **11**, 434–436 (2000).

ACKNOWLEDGEMENTS

This study has been conducted using UKB data under Application 27412. UKB has received ethics approval from the National Health Service National Research Ethics Service (ref 11/NW/0382). The work was performed on the Service for Sensitive Data (TSD) platform, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo IT-Department (USIT). Computations were performed using resources provided by UNINETT Sigma2 – the National Infrastructure for High Performance Computing and Data Storage in Norway.

This research was funded by the Research Council of Norway (#223273, L.T.W.; #324252, O.A.A.); the South-Eastern Norway Regional Health Authority (#2022080, O.A.A.); and the European Union's Horizon2020 Research and Innovation Programme (#847776, O.A.A.; #802998 L.T.W.).

AUTHOR CONTRIBUTIONS

M.K.: Study design, Software, Formal analysis, Visualisations, Project administration, Writing—original draft, Writing—review & editing. D.V.-P.: Study design, Software, Writing—review & editing. M.-Y.W.: Writing—review & editing, D.v.d.M.: Software, Writing – review & editing. T.W.: Writing – review & editing. H.N.: Writing – review & editing. E.E.: Writing—review & editing, Funding acquisition. O.A.A.: Writing—review & editing, Funding acquisition. L.T.W.: Writing—review & editing, Funding acquisition. I.I.M.: Study design, Data preprocessing and quality control, Writing—review & editing, Funding acquisition.

COMPETING INTERESTS

The authors declare the following competing interests: OAA has received a speaker's honorarium from Lundbeck and is a consultant to Coretechs.ai. The remaining authors declare no other competing interests.