Age as a moderator variable in the investigation of the relationship between brain structure and cognition

Ben Griffin[1,2,3], Chetan Gohil[3], Mark Woolrich M[3], Stephen Smith[1] and Diego Vidaurre[2,3]

[1]Oxford Centre for Functional MRI of the Brain (FMRIB), Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, United Kingdom

[2]Center of Functionally Integrative Neuroscience, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

[3]Oxford Centre for Human Brain Activity (OHBA), Wellcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford, Oxford, United Kingdom

1 Abstract

Understanding how differences in brain structure relate to differences in cognition across the lifespan is essential for addressing age-related cognitive decline. Since age is strongly associated with both brain structure and cognition, predictive models often risk simply capturing age effects. To mitigate this risk, deconfounding is typically applied to remove the linear effects of age. Here, we propose to treat age instead as a moderator variable, therefore capturing changes in how brain structure and cognitive abilities are statistically connected. For this view to hold, variations in brain structure linked to differences in cognitive performance in older subjects (e.g. related to disease) would differ from those in younger subjects. Using structural brain imaging data from the UK Biobank, we found evidence supporting this, and that the generalisability of prediction models depends on the age group used to train the model. These findings highlight the complex nature of age-related brain-cognition interactions, and suggest that the optimal modelling approach depends on the population under study.

2 Introduction

Identifying robust associations between brain structure and cognition is critical for both basic neuroscience and clinical applications. Predictive models that reveal relationships between brain data (e.g., structural MRI) and cognitive abilities (e.g., fluid intelligence) are important for uncovering these associations, improving our understanding of cognitive decline and neurological disorders such as Alzheimer's disease (Solé-Padullés et al., 2009) and other dementias (Suzuki et al., 2019).

Capturing these brain-cognition relationships is complicated by the fact that ageing affects both brain structure and cognition in complex ways that vary across individuals. For instance, different cognitive abilities show different relationships with age (Horn & Cattell, 1967); crystallised abilities generally improve until 60 before plateauing, whereas fluid abilities decline from age 20 (Murman, 2015; Salthouse, 2010). Similarly, different brain regions show distinct patterns of age-related atrophy (Bethlehem et al., 2022; Fjell & Walhovd, 2010). The complexity increases when considering structure and cognition together. For example, dividing adults aged 18-88 into younger and older groups revealed a weaker relationship between white matter and certain cognitive functions in older adults, suggesting that some brain-cognition associations become less stable with age (de Mooij et al., 2018).

These complexities make robust predictive modelling challenging, particularly when accounting for age. While age can be considered a key variable of interest in certain studies of age-related conditions (Davis et al., 2018; Jack et al., 2019), it is often treated as a confound, and typically linearly regressed out from both the brain data and cognitive variables (Azevedo et al., 2023; Bracher-Smith et al., 2022). This approach ensures that identified relationships are truly reflective of brain-behaviour associations independent of age, at least under assumptions of linearity. Despite this, predictive models may still show bias toward certain age groups. While predictive biases have been widely recognised with respect to sex, ethnicity, and socio-economic status (Li et al., 2022; Martin et al., 2019; Obermeyer et al., 2019), similar biases can also arise for age (Greene et al., 2022; Wu et al., 2023). These complexities highlight the importance of developing predictive approaches that perform robustly across different age groups.

Beyond the treatment of age as a linear confound, we address whether the nature of the brain-cognition relationship is consistent across age groups, or if, on the contrary, age *moderates* this association. While some studies test moderation by including interaction terms in regression models, these approaches may fail to distinguish between true

interaction effects and underlying nonlinear relationships (Rimpler et al., 2024). Given that ageing involves different underlying mechanisms such as neurodegeneration or disease-related structural changes, we specifically hypothesise that the structural brain features associated with cognitive performance in older adults may differ from those in younger populations.

Using data from UK Biobank (UKB) (Sudlow et al., 2015), we found that models predicting cognition from brain structure generalised differently across age groups. Specifically, models trained on younger subjects generalised well to older subjects, but models trained on older subjects did not generalise as effectively to younger individuals. This difference in generalisation suggests that while certain brain-cognition associations are stable across the lifespan, there may be increased noise, variability, or complexity in the relationships for older subjects that impact model generalisability. These results highlight a trade-off between specificity and generalisability, with important implications for both research and clinical applications.

3 Results

We investigated the role that age plays in moderating the relationship between brain structure and cognition by comparing these relationships across age groups and assessing how well predictive models trained on one group generalised to others. Specifically, we trained predictive models using features from 1,439 structural imaging-derived phenotypes (IDPs) to predict cognitive performance (measured by the first principal component of multiple cognitive tests) for 25,170 UKB subjects (see **Subsection 5.1** for dataset details).

First, we evaluated whether the association between individual structural IDPs and cognitive performance was stable across the lifespan by fitting linear regression models separately within non-overlapping age groups and examining changes in regression coefficients. Next, to explore how training models on different age groups affected model generalisation, we used elastic net regression to make out-of-sample predictions using all IDPs simultaneously, comparing models trained within specific age groups to those trained across groups or on a broader age range than the target test group. Finally, we explored how the age group of the training subjects impacted the importance given to different categories of structural brain measures for predicting cognition.

3.1 Age as a moderator between brain structure and cognition

We first examined how age moderates the relationship between brain structure and cognition by modelling the association between each individual IDP and cognitive performance across age groups (corresponding to age quartiles; see **Supplementary Figure SI-1**).

For each IDP, we fitted a linear regression model to predict cognitive performance, using a model that allows the regression coefficients to vary between age groups. To achieve this, we created four binary variables, each representing one age quartile, with a value of 1 for subjects within the respective quartile and 0 otherwise. These variables were multiplied by the selected IDP, resulting in four new features (i.e., the IDP split by age quartiles). All non-age-related confounds were regressed out using the procedure described in **Subsection 5.1.1.** Additionally, age was included as a separate regressor to control for its confounding effect, as is commonly done in neuroimaging studies. The target variable was a composite cognitive trait, constructed as described in **Subsection 5.1.2.** We then ran a linear regression model to predict this trait, before comparing the coefficients for each model across all 1,439 IDPs (see **Subsection 5.2.1** for full methodological details).

Figure 1a shows the coefficients for the five features in each regression model across all IDPs. As observed, certain IDPs show distinct patterns of regression coefficients across age quartiles. For example, cortical thickness IDPs tend to show stronger positive coefficients in the youngest quartile (indicated in red) and stronger negative coefficients in the oldest quartile (indicated by blue). **Figure 1b** presents scatterplots comparing model coefficients between quartiles, highlighting the particularly low Pearson correlation between the youngest (Q1) and oldest (Q4) quartiles (r = 0.40). These findings are supported by **Supplementary Figure SI-2a**, which shows a violin plot summarising the full coefficient distributions across quartiles, showing both the systematic shift in medians and the increased variability in Q4.

The distribution of scatter points across the different quartiles suggested that the magnitude of the coefficients varied between quartiles. To investigate how the importance of the IDPs in predicting cognition changed across age groups, we took the absolute value of the coefficients (to reflect the magnitude of the association) and calculated Spearman's correlation between quartiles. **Figure 1c** shows the resulting heatmap, indicating that the most predictive IDPs in Q1 (i.e., the youngest subjects) differ from those in older subjects. Specifically, the correlations between Q1 and the other quartiles were particularly low (Q1 vs. Q2: r_s = 0.30; Q1 vs. Q3: r_s = 0.26; Q1 vs. Q4: r_s = 0.21), further emphasising the weak resemblance between Q1 and the other age groups. A Pearson equivalent of this analysis is shown in **Supplementary Figure SI-2b**, which also highlights the differences between coefficients of Q1 and the remaining quartiles.

Finally, we examined whether the differences between the regression coefficients were statistically significant across the age groups. To test this, we performed permutation testing. For each IDP, we investigated whether a coefficient for a given age quartile was significantly different to the mean of the other three quartiles. This procedure resulted in 5,756 p-values, one for each age quartile and each IDP (see **Subsection 5.2.2** for methodological details).

Figure 1d presents histograms of the p-values (uncorrected for multiple comparisons) for each quartile from the permutation testing. Under the null hypothesis, that there are no differences between the coefficient for each age quartile and the other three quartiles, we would expect a uniform distribution of p-values. To assess this, we performed a Kolmogorov-Smirnov (KS) test to evaluate uniformity. The results show that the youngest and oldest quartiles did not follow a uniform distribution (Q1: p_{KS}<0.001; Q4: p_{KS}<0.001), suggesting that these quartiles have a different relationship between IDPs and cognition to the other age quartiles. However, the middle quartiles followed a uniform distribution (Q2: p_{KS}=0.20; Q3: p_{KS}=0.73), suggesting no significant differences from the null expectation.

Having established that age moderates the brain-cognition relationship for at least some of the IDPs (particularly in the youngest and oldest quartiles), we next aimed to identify which specific IDPs showed significant effects after correcting for multiple comparisons. To account for the large number of tests across 1,439 IDPs and four age quartiles, we applied Benjamini-Hochberg false discovery rate (FDR) correction (see **Supplementary Figure SI-2c** for histograms of corrected p-values). Although the number of significant IDPs was modest after correction, a subset remained significant in the youngest and oldest quartiles, with 26 IDPs in Q1 and 13 IDPs in Q4, with some IDPs appearing in both. In total, 35 unique IDPs were significant, falling into the following categories: Cortical Thickness (26), Regional and Tissue Volume (7), and Regional and Tissue Intensity (2). A full list of significant IDPs is provided in **Supplementary Table SI-1**. As expected under the null hypothesis, the middle quartiles showed no significant effects after correction, aside from one regional and tissue volume IDP in Q3. This supports the fact that age-moderated brain-cognition effects are most evident at the extremes of the age distribution.

In summary, these findings reveal that, beyond the general effects of age, the relationship between structural IDPs and cognitive performance varies across age groups. These results therefore support our hypothesis that between-subject variations in brain structure associated with variations in cognition in older people may be different to those in younger people. **Supplementary Figure SI-2d** further illustrates this by showing Bland-Altman plots of Fisher-transformed IDP-cognition correlations across age quartiles, revealing consistent shifts and widening differences in association strength between Q1 and Q4.

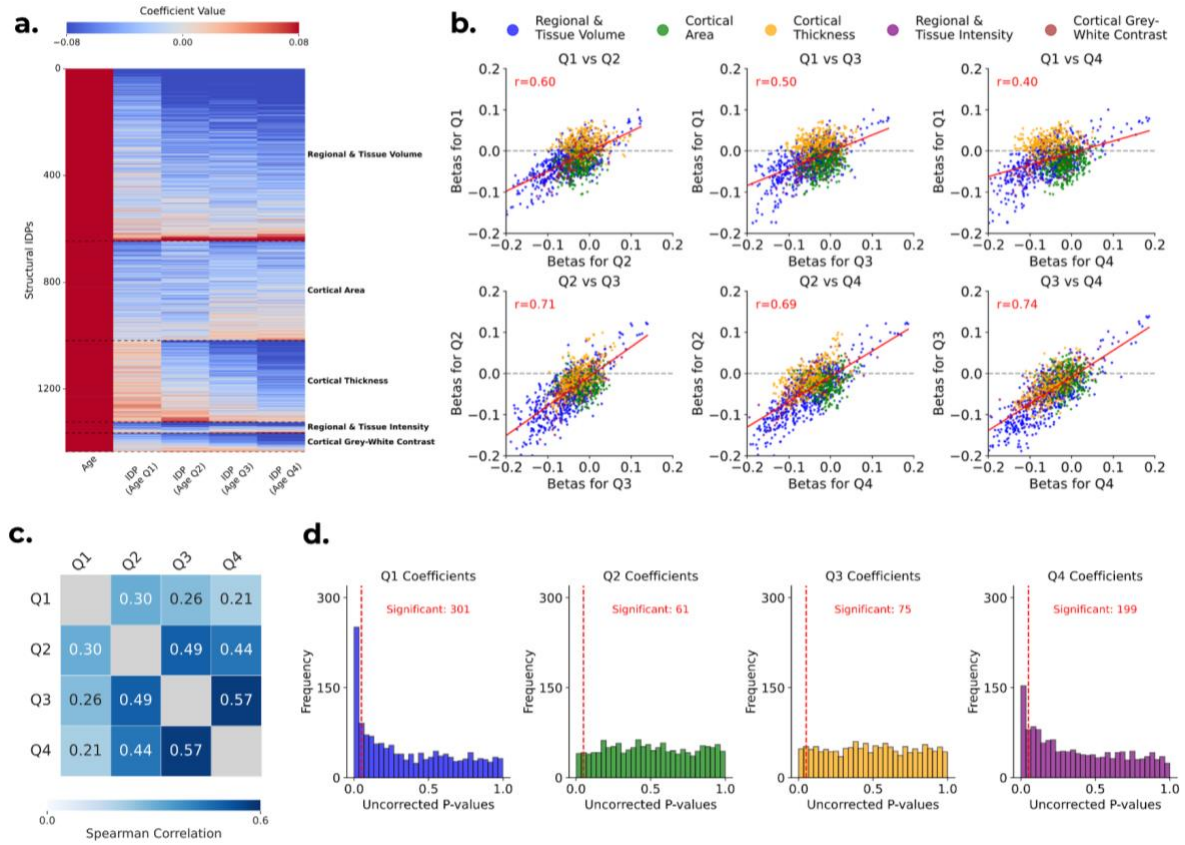


Figure 1 **(a)** Heatmap of coefficients from the linear regression model predicting cognitive performance using each of the 1,439 IDPs individually. The five features used in the model were age and each IDP split into four quartiles based on age (x-axis) across all 1,439 IDPs (y-axis). IDPs are categorised into five groups: Regional and Tissue Volume, Cortical Area, Cortical Thickness, Regional and Tissue Intensity, and Cortical Grey-White Contrast. **(b)** Scatterplots comparing regression coefficients for IDPs between age quartiles. Each scatterplot compares the coefficients between two quartiles: (i) Q1 vs. Q2. (ii) Q1 vs. Q3. (iii) Q1 vs. Q4. (iv) Q2 vs. Q3. (v) Q2 vs. Q4. (vi) Q3 vs. Q4. **(c)** Heatmap of Spearman's correlations between the absolute values of all IDP coefficients across age quartiles. Darker shades indicate stronger positive correlations, highlighting the particularly low correlation between the youngest quartile (Q1) and the others. **(d)** Histograms of uncorrected p-values from permutation testing across 1,439 IDPs, with each distribution representing one age quartile. The red dashed lines mark the p = 0.05 threshold. Under the null hypothesis of no coefficient differences across age quartiles, each distribution would be uniform, and approximately 72 p-values (5% of 1,439) would fall below this threshold by chance.

3.2 Prediction accuracy depends on training age group, with younger models generalising better

To expand on this result, we turned to out-of-sample prediction of cognitive performance using brain data. Here, using a fully cross-validated framework, we ran an elastic net regression using all 1,439 structural IDPs to predict cognition (see **Subsection 5.3** for full methodological details). Unlike the previous section, which examined IDPs one at a time, this multivariate approach uses all structural features simultaneously. This allows the model to detect brain-wide patterns associated with cognition, improving predictive performance and aligning more closely with real-world or clinical applications. Since we are now making full predictions, rather than simply fitting a model as in the previous section, we regressed

out all confounds (including age and age-related confounds such as age^2) from both the dependent and the independent variables (see **Subsection 5.1.1** for full details). This is considered the gold standard for deconfounding when analysing large-scale neuroimaging datasets.

The population was divided into two equally sized age groups, a younger and an older half. We chose two groups, rather than four as in the previous section, to simplify the analysis and obtain larger sample sizes per group. This grouping improved model stability and statistical power during cross-validation, while still capturing meaningful age differences and allowing us to assess generalisability across age ranges. After dividing the population, we trained one elastic net model for each age group. This yielded a vector of regression coefficients (with as many coefficients as IDPs) per group. We then assessed model generalisability by predicting cognition for subjects both within-age-group (e.g., Train: Young; Test: Young) and across-age-group (e.g., Train: Young; Test: Old). To improve robustness of predictions, this was carried out for 10 repeats of 10-fold cross-validation for all 25,170 subjects

Figure 2a compares within-age-group and across-age-group out-of-sample prediction accuracy. The figure shows that within-age-group prediction accuracy was higher than across-age-group accuracy (mean $r_{\text{WITHIN-AGE-GROUP}}$: 0.095 vs. mean $r_{\text{ACROSS-AGE-GROUP}}$: 0.085; $p<0.001$), suggesting that age-specific models are more effective at capturing the brain-cognition relationships unique to each age group.

To further investigate these results, we examined the individual models within each training and testing subgroup. **Figure 2b** illustrates how the prediction accuracy is impacted by training and testing on different age groups. The figure shows that models trained on the younger half consistently outperformed those trained on the older half, regardless of the testing group, although the extent of this difference varied depending on the test group. When testing on older subjects, models trained on younger subjects still outperformed those trained on older subjects (mean $r_{\text{TR: Y; TE: O}}$: 0.099 vs. mean $r_{\text{TR: O; TE: O}}$: 0.090; $p<0.001$). However, when testing on younger subjects, the performance gap was much larger, with models trained on older subjects performing significantly worse than those trained on younger subjects (mean $r_{\text{TR: Y; TE: Y}}$: 0.100 vs. mean $r_{\text{TR: O; TE: Y}}$: 0.070; $p<0.001$). This suggests that models trained on older subjects struggle to capture brain–cognition relationships relevant to younger individuals, whereas those trained on younger subjects remain predictive even for older individuals. Furthermore, while models trained on younger subjects performed similarly across test groups (mean $r_{\text{TR: Y; TE: Y}}$: 0.100 vs. mean $r_{\text{TR: Y; TE: O}}$: 0.099; $p=1.00$), performance

dropped significantly when models trained on older subjects were tested on younger subjects rather than older subjects (mean $r_{TR: O; TE: O}$: 0.090 vs. mean $r_{TR: O; TE: Y}$: 0.070; p<0.001).

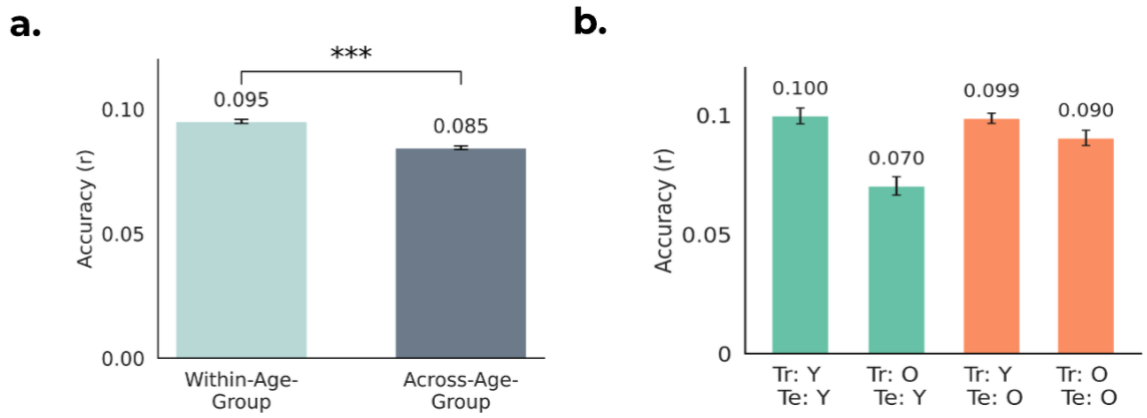


Figure 2 *Performance of different training and testing splits when fitting a multiple regression model with elastic net regularisation using 1,439 structural IDPs to predict cognition. Results are based on 10 repeats of 10-fold cross-validation, each using randomly chosen folds.* **(a)** Bar chart comparing mean prediction accuracy for within-age-group training and testing (Train: Young, Test: Young; and Train: Old, Test: Old) versus across-age-group training and testing (Train: Young, Test: Old; and Train: Old, Test: Young). Both the IDPs and cognitive trait were deconfounded for all confounds, including age-related confounds. Accuracy is measured as the correlation between the predicted and observed cognitive scores, with error bars showing the standard deviation across the 10 repeats for each set of accuracies. *** p < 0.001. **(b)** Comparison of prediction accuracy across four training-testing splits (Train: Young, Test: Young; Train: Old, Test: Old; Train: Young, Test: Old; Train: Old, Test: Young), highlighting that model generalisibility depends on age group.

3.3 Pooled training improves generalisation for older subjects but not younger subjects

Next, we repeated our multiple regression analysis using the same modelling and deconfounding approach, but this time comparing three model types:

(i) **Within-age-group models:** Training and testing within the same age group (e.g., Train: Young; Test: Young)

(ii) **Across-age-group models:** Training in one age group and testing in the other (e.g., Train: Young; Test: Old)

(iii) **Pooled models:** Training on the entire population and testing within a specific age group (e.g., Train: ALL; Test: Young).

This comparison aimed to assess whether training on a broader age range (i.e., the pooled model, which is the more conventional approach to prediction) could compensate for the age-specific variability observed in our previous analyses (see **Subsection 5.3.2** for full methodological details).

Figure 3a compares the mean prediction accuracy of the three model types across 10 repeats of 10-fold cross-validation. The figure highlights that within-age-group models

outperformed pooled models (mean $r_{WITHIN\text{-}AGE\text{-}GROUP}$: 0.095 vs. mean r_{POOLED}: 0.090; p=0.002) and across-age-group models (mean $r_{WITHIN\text{-}AGE\text{-}GROUP}$: 0.095 vs. mean $r_{ACROSS\text{-}AGE\text{-}GROUP}$: 0.085; p<0.001), while pooled models outperformed the across-age-group models (mean r_{POOLED}: 0.090 vs. mean $r_{ACROSS\text{-}AGE\text{-}GROUP}$: 0.085; p=0.02). These results are consistent with our expectations, since the models which are performing better are those trained on data that are more representative of the test population.

However, **Figure 3b**, which compares the performance of models trained and tested on different age groups, reveals that the impact on model performance of training across the full population depends on the test group. On the one hand, the trend seen in **Figure 3a** continues when the test group is the younger half of the subjects. Specifically, the within-age-group model achieved the highest accuracy compared to both the pooled model (mean $r_{Train: Y, Test: Y}$=0.100 vs. $r_{Train: ALL, Test: Y}$=0.089; p<0.001) and the across-age-group model (mean $r_{Train: Y, Test: Y}$=0.100 vs. $r_{Train: O, Test: Y}$=0.070; p<0.001). The pooled model, in turn, outperformed the across-age-group model ($r_{Train: ALL, Test: Y}$=0.089 vs. $r_{Train: O, Test: Y}$=0.070; p<0.001). These results support the previous findings that training on a dataset more representative of the test population improves prediction accuracy. On the other hand, when testing on older subjects, model performance was more similar across conditions. The within-age-group model performed very similarly to the pooled model ($r_{Train: O, Test: O}$=0.090 vs. $r_{Train: ALL, Test: O}$=0.091; p=1.00), suggesting that pooling training data is as effective as training within the same age group for predicting cognition in older individuals. However, the across-age-group model, which was trained on younger subjects, achieved significantly higher accuracy than the pooled model ($r_{Train: Y, Test: O}$=0.099 vs. $r_{Train: ALL, Test: O}$=0.091; p=0.006), indicating that models trained on younger subjects generalise more effectively to older subjects than the conventional approach of training across all subjects.

Overall, these results support our previous findings, highlighting a key distinction in model generalisation across age groups. We found that pooled training produced an intermediate performance level between within-age-group and across-age-group training and testing. While pooled models trained across a broader age range of subjects resulted in lower accuracy for younger subjects, they achieved comparable accuracy to age-specific models for older subjects. These analyses highlight a trade-off between model generalisability and age-group specificity.

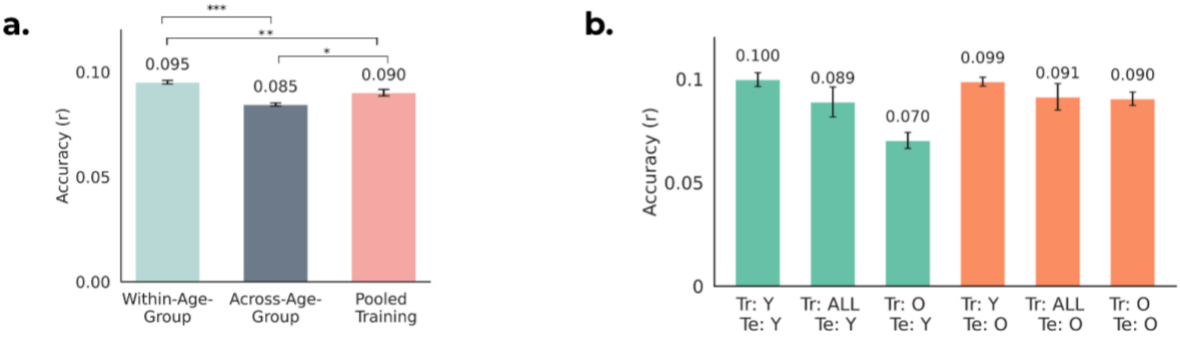


Figure 3 *Comparison of model performance, regularisation strength, and regression coefficients across different training strategies when using 1,439 structural IDPs to predict cognition using 10 repeats of 10-fold cross-validation*. **(a)** Bar chart comparing prediction accuracy for within-age-group training and testing (Train: Young, Test: Young; and Train: Old, Test: Old), across-age-group training and testing (Train: Young, Test: Old and Train: Old, Test: Young), and pooled training (Train: All, Test: Young; and Train: All, Test: Old). Both the IDPs and cognitive trait were deconfounded for all confounds (including age-related confounds). Accuracy is measured as the correlation between predicted and observed cognitive scores, with error bars showing the standard deviation across the 10 repeats. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. **(b)** Comparison of prediction accuracy across four model training/testing splits, using the deconfounding procedure described previously. Models were trained and tested on different age groups (Train: Young, Test: Young; Train: Old, Test: Old; Train: Young, Test: Old; Train: Old, Test: Young).

3.4 Models trained on older subjects require strong regularisation to prevent overfitting

To investigate the factors driving the observed differences in model prediction accuracy (e.g., overfitting), we examined the elastic net hyperparameter values across models (selected by an inner cross-validation loop). **Figure 4a** compares the strength of elastic net regularisation and model accuracy across 10 repeats of 10-fold cross-validation for each of the six models. The top panel presents bar charts ranking model accuracy across cross-validation repetitions. The bottom panel shows the regularisation strength (α) selected for each outer fold, represented by black crosses. For each model, these values were obtained using nested cross-validation, with the inner loop selecting the optimal α. The coloured circle indicates the mean of these α values across the 10 outer folds, summarising the average regularisation strength used during that cross-validation repetition.

The figure highlights a relationship between the age of subjects in the training group, the regularisation strength used to fit the model, and model accuracy. Models trained on older subjects show higher levels of regularisation than models trained across the entire population for both the younger test group (mean $\alpha_{\text{Train: Old, Test: Young}} = 7.87$ vs. mean $\alpha_{\text{Train: All, Test: Young}} = 4.97$) and the older test group (mean $\alpha_{\text{Train: Old, Test: Old}} = 8.40$ vs. mean $\alpha_{\text{Train: All, Test: Old}} = 5.17$). Similarly, models trained on younger subjects show lower regularisation

levels than those trained on older subjects, and are more comparable to those trained across the entire population (mean $\alpha_{\text{Train: Young, Test: Old}} = 5.00$; mean $\alpha_{\text{Train: Young, Test: Young}} = 6.20$).

These results suggest that models trained on older subject require stronger regularisation to shrink regression coefficients and prevent overfitting, potentially due to increased noise or greater inter-subject variability in older populations. Models trained on younger subjects, or across a broader age range, require less regularisation, which may explain why pooled models perform comparably to those trained only on older subjects. While including younger subjects may introduce a slight bias, the lower noise levels relative to training only on older subjects results in less regularisation, allowing the regression coefficients to be more predictive of cognition.

Figure 4b presents the mean absolute regression coefficients for each IDP category across different training groups. Across all IDP categories, models trained on older subjects showed consistently lower coefficient magnitudes than those trained on younger subjects, with the largest reductions seen in Regional and Tissue Volume (31.7% reduction across 646 IDPs), Cortical Area (26.5% reduction across 372 IDPs) and Cortical Thickness (22.9% reduction across 306 IDPs). These categories are likely to be particularly impactful on the final result, given the relatively large number of IDPs in each category. Although Regional and Tissue Intensity (19.8% reduction across 41 IDPs) and Cortical Grey-White Contrast (16.9% reduction across 70 IDPs) showed smaller reductions across fewer IDPs, there was still a consistent reduction compared to models trained in older populations. Pooled training resulted in mid-range values across all IDP categories, providing a balance between the younger and older subjects. This supports the idea that including a broader age range in training helps stabilise coefficient estimates by reducing overfitting to age-specific noise, rather than the age-specific models discovering unique features in each age group.

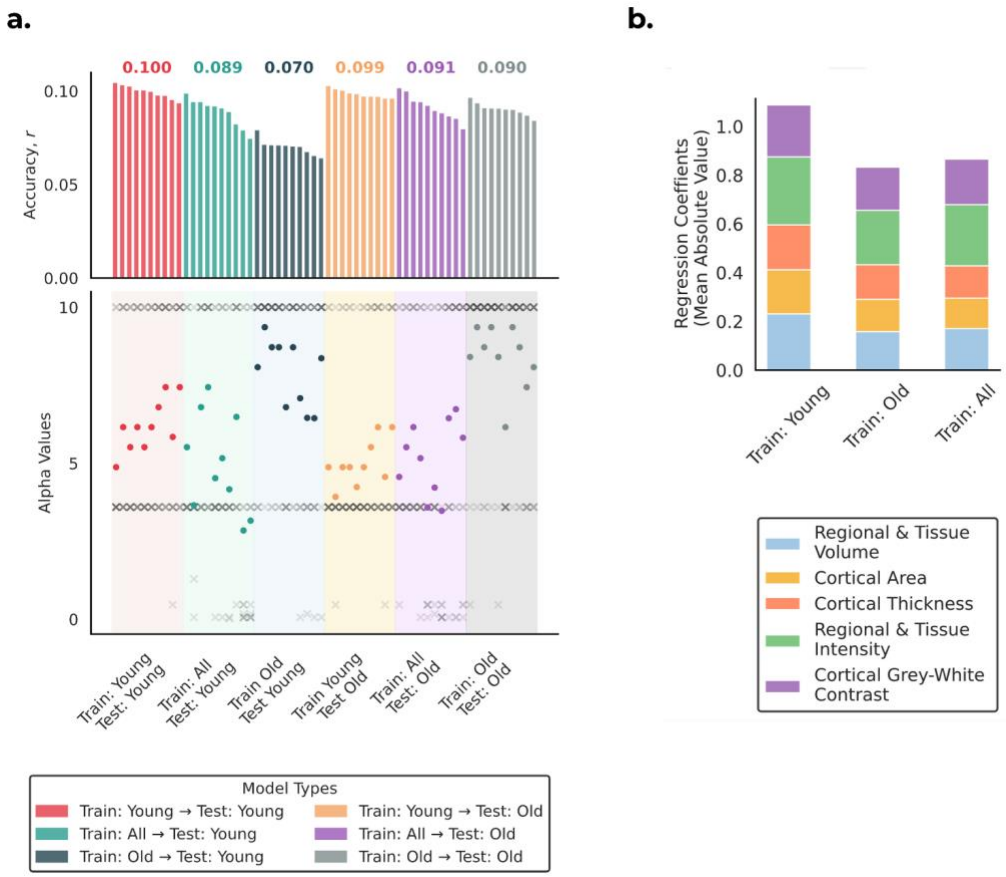


Figure 4 *Overview of model regularisation and feature importance across training strategies.* **(a)** Comparison of elastic net regularisation strength and model accuracy. The top panel ranks models by accuracy across cross-validation repetitions. The bottom panel shows regularisation strength per cross-validation repetition (black crosses), with the mean regularisation level across 10 folds indicated by a coloured circle. **(b)** Stacked bar plot showing mean absolute regression coefficients values across different structural IDP groups for the three training approaches.

4 Discussion

Age, brain structure, and cognition are intrinsically linked: aging drives changes in the brain, which in turn results in changes in cognition. Therefore, accounting for the influence of age is of importance when studying brain-cognition relationships. In this study, we explored whether the nature of the brain-cognition relationship is consistent across different age groups by analysing structural IDPs from UKB. Our analyses revealed statistically significant age-related differences in the relationship between cognition and brain structure.

First, we found that when fitting a regression model using a single IDP at a time, but split across age quartiles, the coefficients for each quartile were significantly different from each other for a subset of the IDPs. Furthermore, in out-of-sample predictions, within-age-group models significantly outperformed across-age-group models, highlighting the importance of training on data with a similar distribution to the test set. These results support the hypothesis that the differences in brain structure that are behind the differences in cognitive performance at old age may be different from those at young age.

However, when investigating the results from specific training and testing splits, our findings highlighted a key distinction in model generalisation across age groups, where prediction accuracy strongly depended on the specific age group of training subjects. We found that models trained on younger subjects generalised well to older subjects, likely reflecting certain brain-cognition relationships remaining stable across the lifespan. However, models trained on older subjects generalised poorly to younger subjects. This poorer generalisation might result from variations in signal-to-noise ratio across the lifespan (Nomi et al., 2024), leading to a higher chance of overfitting to the training set for older subjects. Additionally, the underlying relationships between brain structure and cognition may be more complex (or just more variable) in older subjects. Part of this complexity may arise because age-related changes accelerate in later life, or because additional sources of subject variability (e.g., neurodegeneration) complicate brain-cognition relationships in older subjects. In contrast, between-subject variability in younger subjects may primarily reflect natural baseline differences, independent of aging. Although this complexity may not result in improved prediction accuracy, it may impact the older model’s ability to generalise to younger subjects.

Previous studies have also reported age-related differences in prediction generalisability, including work based on smaller datasets using single cognitive outcomes (Yu & Fischer, 2022). Our findings replicate and extend these effects using the large sample size of UK Biobank to focus on age's role as a moderator of brain–cognition relationships, rather than treating age purely as a confound.

We then investigated how pooled training compared to age-specific models. While within-age-group models achieved the highest accuracy, pooled training offered improved generalisation compared to across-age-group models. Therefore, while age-specific models may be beneficial for certain subgroups (e.g., younger adults), our results indicate that their performance can be inconsistent, particularly when applied to other age groups. Models trained on older subjects, in particular, performed poorly when tested on younger subjects, highlighting a limitation of age-specific training when generalisability is required.

By contrast, pooled models provided more robust and consistent predictions, likely due to the inclusion of younger subjects with more stable brain-cognition relationships. These findings suggest that for certain subgroups (here, older subjects) having a more diverse training set can help improve model generalisation while maintaining accuracy for the age-specific test group. Therefore, our findings suggest that the conventional approach to model training (i.e., pooled training) may offer the most reliable approach for datasets with greater heterogeneity or limited sample sizes, where robust performance is important, even if age-specific models can sometimes achieve higher accuracy.

Our analysis of regression coefficients suggests that pooled models achieve better generalisation by reducing overfitting and stabilising predictions. Models trained only on older subjects required stronger regularisation, possibly due to higher noise or variability in brain structure and cognition within this group, resulting in overly aggressive coefficient shrinking. On the other hand, pooled models, which include younger subjects with more consistent brain-cognition relationships, required less regularisation and produced more robust predictive features, despite producing a similar level of prediction accuracy to the models trained only on older subjects. This suggests that pooling across age groups can help avoid overfitting to age-specific variability and therefore improve generalisation.

In any case, several limitations must be acknowledged. First, we focused on the first principal component of cognition throughout our analysis; however, cognition is formed of various intelligence subtypes (e.g., crystallised and fluid intelligence) which may be better explored separately. Second, although UKB offers a large and diverse dataset, there are still likely to be issues with generalising to underrepresented populations. Future studies should aim to validate our findings in more diverse cohorts, and investigate how additional factors (e.g., genes) interact with age and brain structure to influence cognition. Finally, dividing subjects into halves and quartiles based on age simplifies the continuous nature of age. A fully continuous, smoothly age-varying coefficient prediction model could provide further insights into the nonlinear nature of the role of age as a moderator. Although this approach

may improve predictions for specific age-groups, it may sacrifice some of the generalisability that we have been able to capture with our more simplistic approach.

Additionally, while we used linear regression to address the confounding effects of age and other potential confounds, this may fail to account for more complex, nonlinear relationships between brain structure and cognition across the lifespan. Although we included an expanded set of confounds which incorporate transformed versions of the original confounds (e.g., age^2, age x sex) to try to capture these effects (Alfaro-Almagro et al., 2021), alternative approaches such as propensity score matching (Rosenbaum & Rubin, 1983) and confound-isolating cross-validation (Chyzhyk et al., 2022) may offer more rigorous solutions. These methods have previously highlighted important limitations of traditional linear deconfounding (Benkarim et al., 2022). Incorporating these approaches into future analyses could improve robustness, and reduce the risk of residual confounding and making misleading conclusions (Power et al., 2012).

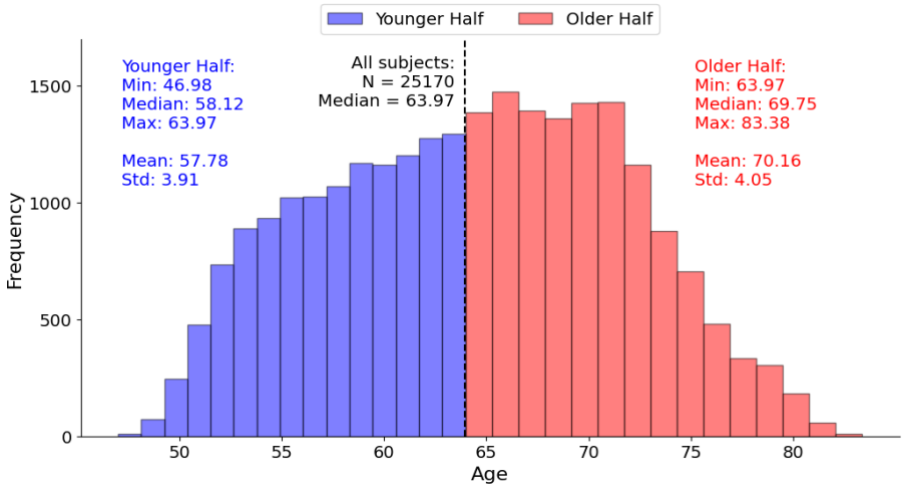
Overall, this work demonstrates that the nature of the brain-cognition relationship varies across the lifespan. Our findings highlight a trade-off between age-specific and pooled modelling approaches, with implications for both research and clinical applications, where the optimal approach may depend on the population under study. For situations requiring high accuracy within a specific target group (e.g., clinical diagnostics), age-specific models may be preferable. However, pooled models provide advantages for generalisation across diverse populations. Recognising this trade-off is essential for improving prediction models in both research and clinical settings, ensuring that predictions are optimal for the specific population under study.

5 Methods

5.1 UKB neuroimaging and non-imaging data

We analysed data from UKB, a large-scale neuroimaging dataset that includes subjects aged 47 to 83 years. Our study explored the relationship between a composite cognitive measure, specifically the first principal component of the top 30 cognitive traits selected from the 1,331 available traits in UKB (see details in **Subsection 5.1.2**), and 1,439 structural IDPs. These IDPs comprised 1,436 T1-weighted MRI measures, including regional and tissue volume, cortical area, cortical thickness, region and tissue intensity, cortical grey-white contrast, and white matter hyperintensity volumes, and three T2-weighted MRI measures. Our analysis included 25,170 UKB subjects, after excluding those with missing values for any of the 30 cognitive traits or for age. Missing values in structural IDPs were imputed using the mean to maintain a larger sample size. **Figure 5a** shows the age distribution, with subjects split into younger and older halves at the median age to ensure equally sized groups. The figure highlights the balanced age distribution of the dataset. To investigate the impact of age on the relationship between structural IDPs and cognition, we stratified the IDP and cognitive data according to these two age groups. **Supplementary Figure SI-3** shows the age distributions further split by sex, confirming similar male-female distributions within each age group.

a.



b.

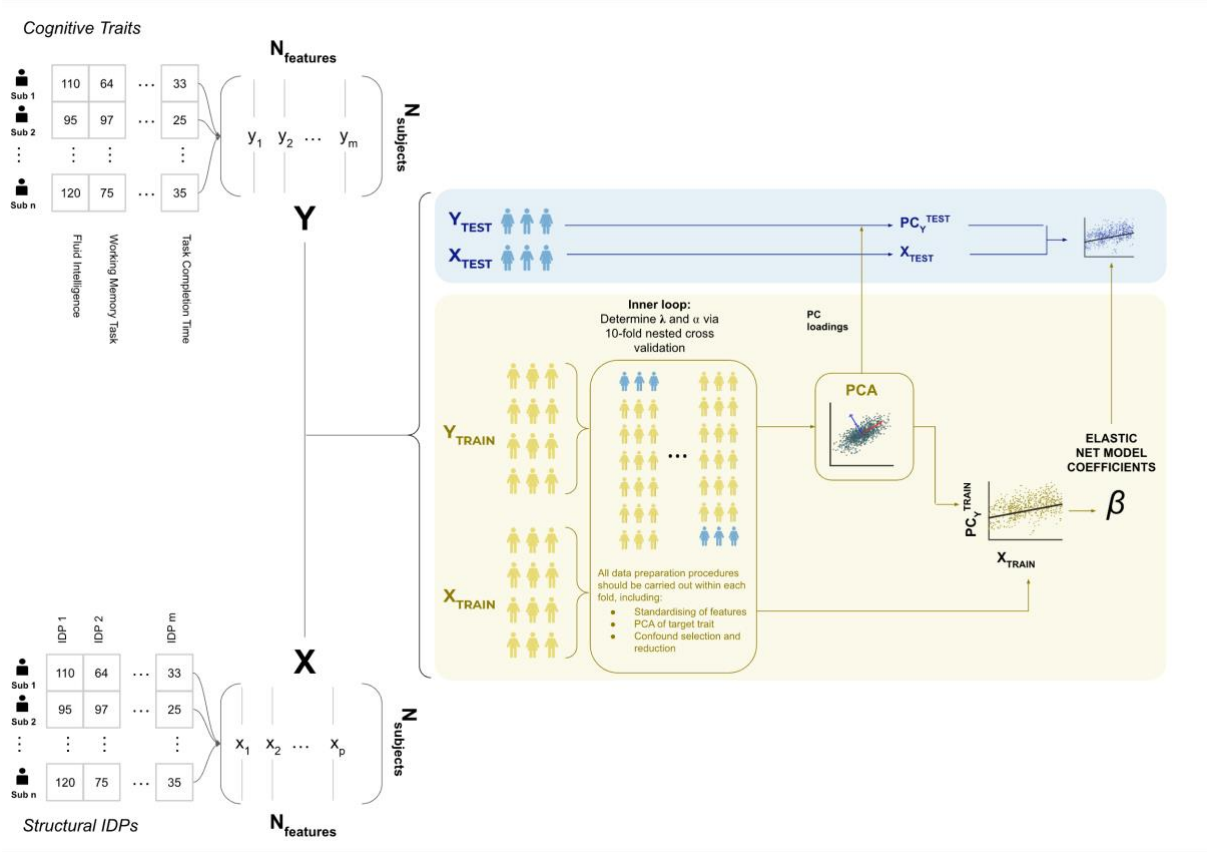


Figure 5 **(a)** Age distribution of UKB subjects, divided into younger (blue) and older (red) halves, each containing 12,585 subjects. Summary statistics indicate that both age groups have similar standard deviations and a median equidistant from the population mean, demonstrating a balanced age distribution. **(b)** Cross-validation framework for extracting a composite cognitive measure via PCA while avoiding data leakage. Each subject was represented by a set of structural IDPs (X) and corresponding cognitive trait scores (Y). A fully nested 10-fold cross-validation framework was used to ensure that all data processing and modelling steps were carried out in a strictly out-of-sample manner. Within each outer fold, the training data underwent an inner loop including standardisation, deconfounding, and elastic net hyperparameter tuning. PCA was applied to the training cognitive scores (Y_{TRAIN}), and the first principal component (PC_Y^{TRAIN}) was used as the regression target to determine elastic net coefficients. Separately, the learned PCA loadings were applied to the test data (Y_{TEST}) to generate a composite cognitive score (PC_Y^{TEST}), which was used as the out-of-sample target for evaluating model performance using test data IDPs (X_{TEST}) as features.

5.1.1 Confound removal

Large-scale neuroimaging datasets such as UKB have significantly increased the statistical power for detecting small brain-behaviour relationships but have also increased the risk of confounding biases (e.g., age, head motion) which can bias findings and complicate interpretation (Smith & Nichols, 2018). To control for confounding effects, we linearly regressed out the confounds from both imaging and non-imaging variables. To prevent data leakage, we applied the deconfounding within the cross-validation framework to avoid information from the test set leaking into the training data. We created a reduced set of confounds from a comprehensive set of 559 UKB imaging confounds provided by Alfaro-Almagro et al., (2021; see Supplementary Table SI-2). This was done by first selecting a subset of conventional confounds including age, sex, age*sex, scanning site, head size, and head motion (see **Supplementary Table SI-3** for full list), before reducing the remaining confounds, which includes other age-related confounds such as age^2, through Principal Component Analysis (PCA). We retained the principal components which explained >85% of the total variance, ensuring that the reduced confound set captured most of the variability of the full set, while minimising redundancy.

5.1.2 Composite cognitive measure

Prediction accuracy for individual cognitive traits is often limited (Pervaiz et al., 2020), which can limit our ability to distinguish age-specific models, as they require subsampling the data. Therefore, we developed a composite cognitive measure from UKB cognitive traits to improve prediction.

To create a robust cognitive target for prediction, we selected the top 30 cognitive traits based on a preliminary elastic net regression, using all 1,439 IDPs to predict each cognitive measure across all subjects (a list of these 30 cognitive traits can be found in **Supplementary Table SI-4**). Since generalisability can only be meaningfully assessed for traits that are at least moderately predictable within the same age agroup, focusing on well-predicted traits ensured that model comparisons reflected meaningful cognitive variation. This trait selection step was not cross-validated within the modelling framework, both to avoid the considerable computational cost of evaluating all 1,331 traits in a nested framework and to ensure we have a consistent subset of target cognitive traits across all model comparisons. While the trait selection was performed on the full dataset and could in principle favour pooled models (see **Subsection 5.3.2** for description of pooled models), the chosen traits were broadly representative across age groups.

To further boost the prediction accuracy, we aimed to create a copmisitve cognitive measure. To do so, we applied PCA to extract first principal component from the top 30 performing traits, capturing the greatest variance across them. At this modelling stage, all steps in developing this composite cognitive trait were performed using a fully cross-validated procedure to prevent data leakage (**Figure 5b**), which could otherwise lead to biased test set predictions. The composite measure was then used as the target variable to evaluate prediction performance across regression models.

5.2 Univariate modelling: age-moderated effects

To investigate how brain-cognition relationships vary across the lifespan, we first applied univariate linear models to test whether the effect of individual IDPs on cognition differs across age quartiles. This section outlines the modelling approach and describes the permutation testing framework used to assess the statistical significance of the effects.

5.2.1 Linear modelling of age-moderated brain-cognition relationships

To estimate age-specific associations between brain structure and cognitive performance, we fitted a linear regression model using as regressors age and a single IDP at a time, treating age as a moderator by stratifying subjects into age quartiles. We created four binary variables representing the age quartiles, assigning a value of 1 for subjects within the respective quartile and 0 otherwise. Each binary variable was multiplied by the selected IDP, creating four new features representing the IDP stratified by age quartile. Additionally, age was included as a regressor to account for its confounding effect, as is commonly done in neuroimaging studies (Jenkinson & Chappell, 2017; Rao et al., 2015).

5.2.2 Permutation testing of age-moderated effects

To assess statistical differences in the regression coefficients across age quartiles, we used permutation testing. Specifically, for each IDP, we applied a linear regression model across all subjects, resulting in five coefficients: one for age, and one for each of the four age quartile-IDP interaction features. We then computed the difference between the coefficient for each quartile and the mean of the other three quartiles.

This procedure was repeated 10,000 times, with age-quartile labels randomly permuted in each iteration to shuffle age-group assignments. This generated 10,000 sets of four coefficients per IDP. For each permutation and coefficient, we recalculated the difference between the permuted coefficient and the mean of the remaining three. After all permutations, we computed p-value by determining the proportion of times the permuted

difference exceeded the observed difference for each coefficient. This resulted in four p-values for each IDP, one for each quartile.

Under the null hypothesis of no significant coefficient differences across quartiles, the distribution of p-values should be uniform. To assess this hypothesis, we applied a Kolmogorov-Smirnov (KS) test. The KS statistic quantifies the deviation between the empirical cumulative distribution function (formed by the permutation-derived p-values), and the cumulative distribution of a uniform distribution. A significant p-value for an age quartile indicates systematic differences in the IDP regression coefficients, suggesting age-related variation in the relationship between brain structure and cognition.

5.3 Multivariate Modelling: Elastic Net Prediction

We then turned to multivariate prediction, using elastic net regression to predict cognition from all 1,439 structural IDPs simultaneously. Here, we describe the prediction framework used to assess model generalisation across different age groups.

5.3.1 Multivariate prediction using elastic net regression

Elastic net is a common approach for handling multicollinearity in large imaging datasets while avoiding overtfitting. Elastic net combines the L_1 penalty of lasso (Tibshirani, 1996) and the L_2 penalty of ridge regression (Hoerl & Kennard, 1970), enabling both coefficient shrinkage and sparsity in the model (Vidaurre et al., 2013; Zou & Hastie, 2005).

Let $X \in \mathbb{R}^{n \times p}$ be the design matrix with n subjects and p features, and let $y \in \mathbb{R}^n$ be the target vector, representing the composite cognitive measure. Following the implementation in `scikit-learn`, the elastic net regression coefficients, $\hat{\beta}$, are determined by:

$$\hat{\beta} = \underset{\alpha}{\operatorname{argmin}} \left\{ \frac{1}{2n} ||y - X\beta||_2^2 + \alpha \left[\rho \left| |\beta| \right|_1 + \frac{1-\rho}{2} ||\beta||_2^2 \right] \right\}.$$

Here, α is the overall regularisation strength, and ρ controls the mix between L1 and L2 penalties. Both hyperparameters were selected via nested cross-validation.

5.3.2 Age-split training and generalisation performance

Using a fully cross-validated framework, we applied elastic net regression with 1,439 IDPs to predict cognition. The population was divided into two equally sized age groups, a younger half and an older half. We examined six distinct training and testing approaches, as shown in **Figure 6**:

1. **Within-age-group models:** Training and testing within the same age group:

    a. Train: Young; Test: Young
    b. Train: Old; Test: Old
2. **Across-age-group models:** Training in one age group and testing in the other:
    a. Train: Young; Test: Old
    b. Train: Old; Test: Young
3. **Pooled models:** Training on the entire dataset and testing within each specific age group:
    a. Train: ALL; Test: Young
    b. Train: ALL; Test: Old

To ensure comparability across models, we matched the sample size of the "Train: ALL" dataset to each age-specific training set by randomly subsampling subjects. This provided a standardised framework for evaluating out-of-sample predictions across all approaches.

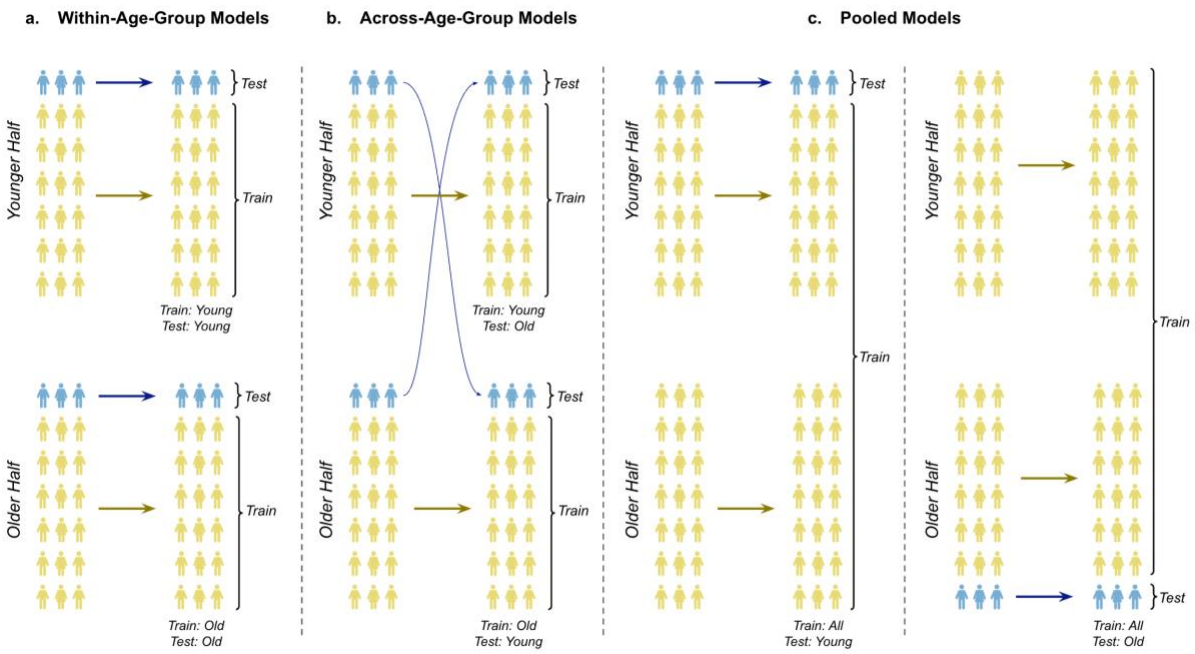


Figure 6 *Example fold within the cross-validation framework for predicting cognition from structural IDPs using different model approaches. Three approaches were implemented:* **(a) Within-age-group models:** Elastic net regression was applied within the same age group for both training and testing (e.g., Train: Young; Test: Young). **(b) Across-age-group models:** Models were trained on one age group and tested on the other (e.g., Train: Young; Test: Old). **(c) Pooled models:** Models were trained on the full population but subsampled to match the sample sizes of the other approaches, with testing restricted to a specific age group (e.g., Train: ALL; Test: Old). This schematic presents a single fold of the 10-fold cross-validation procedure. Each model was trained and tested on an equal number of subjects per fold across all approaches.

Acknowledgements

D. Vidaurre is supported by a Novo Nordisk Foundation Emerging Investigator Fellowship (NNF19OC-0054895), an ERC Starting Grant (ERC-StG-2019-850404), and a DFF Project from the Independent Research Fund of Denmark (2034-00054B).

This research has been conducted in part using the UK Biobank Resource under Application Number 8107. We are grateful to UK Biobank for making the data available, and to all UK Biobank study participants, who generously donated their time to make this resource possible.

Analysis was carried out on the clusters at the Oxford Biomedical Research Computing (BMRC) facility. BMRC is a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute, supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre.

This research was funded in part by the Wellcome Trust (215573/Z/19/Z). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

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*. https://doi.org/10.1016/j.neuroimage.2020.117002

Azevedo, T., Bethlehem, R. A. I., Whiteside, D. J., Swaddiwudhipong, N., Rowe, J. B., Lió, P., Rittman, T., Li, G., Petrie, E. C., Peskind, E. R., Villena, T., Martinez, W., Sadowsky, C., Lee, A., Ravdin, L., Lin, M., Chiang, G., Relkin, N., Santulli, R. B., … Silbert, L. C. (2023). Identifying healthy individuals with Alzheimer's disease neuroimaging phenotypes in the UK Biobank. *Communications Medicine*, *3*(1). https://doi.org/10.1038/s43856-023-00313-w

Benkarim, O., Paquola, C., Park, B. Y., Kebets, V., Hong, S. J., Vos de Wael, R., Zhang, S., Yeo, B. T. T., Eickenberg, M., Ge, T., Poline, J. B., Bernhardt, B. C., & Bzdok, D. (2022). Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. *PLoS Biology*, *20*(4). https://doi.org/10.1371/journal.pbio.3001627

Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., … Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*(7906). https://doi.org/10.1038/s41586-022-04554-y

Bracher-Smith, M., Rees, E., Menzies, G., Walters, J. T. R., O'Donovan, M. C., Owen, M. J., Kirov, G., & Escott-Price, V. (2022). Machine learning for prediction of schizophrenia using genetic and demographic factors in the UK biobank. *Schizophrenia Research*, *246*. https://doi.org/10.1016/j.schres.2022.06.006

Chyzhyk, D., Varoquaux, G., Milham, M., & Thirion, B. (2022). How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience*, *11*. https://doi.org/10.1093/gigascience/giac014

de Mooij, S. M. M., Henson, R. N. A., Waldorp, L. J., & Kievit, R. A. (2018). Age differentiation within gray matter, white matter, and between memory and white matter in an adult life span cohort. *Journal of Neuroscience*, *38*(25). https://doi.org/10.1523/JNEUROSCI.1627-17.2018

Fjell, A. M., & Walhovd, K. B. (2010). Structural brain changes in aging: Courses, causes and cognitive consequences. In *Reviews in the Neurosciences* (Vol. 21, Issue 3). https://doi.org/10.1515/REVNEURO.2010.21.3.187

Greene, A. S., Shen, X., Noble, S., Horien, C., Hahn, C. A., Arora, J., Tokoglu, F., Spann, M. N., Carrión, C. I., Barron, D. S., Sanacora, G., Srihari, V. H., Woods, S. W., Scheinost, D., & Constable, R. T. (2022). Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature*, *609*(7925). https://doi.org/10.1038/s41586-022-05118-w

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1). https://doi.org/10.1080/00401706.1970.10488634

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, *26*(C). https://doi.org/10.1016/0001-6918(67)90011-x

Jenkinson, M., & Chappell, M. (2017). Short introduction to the General Linear Model for Neuroimaging. In *Oxford Neuroimaging Primers*.

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., Ge, T., Patil, K. R., Jabbi, M., Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, *8*(11). https://doi.org/10.1126/sciadv.abj1812

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *51*(4). https://doi.org/10.1038/s41588-019-0379-x

Murman, D. L. (2015). The Impact of Age on Cognition. *Seminars in Hearing*, *36*(3). https://doi.org/10.1055/s-0035-1555115

Nomi, J. S., Bzdok, D., Li, J., Bolt, T., Chang, C., Kornfeld, S., Goodman, Z. T., Yeo, B. T. T., Spreng, R. N., & Uddin, L. Q. (2024). Systematic cross-sectional age-associations in global fMRI signal topography. *Imaging Neuroscience*, *2*. https://doi.org/10.1162/imag_a_00101

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464). https://doi.org/10.1126/science.aax2342

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage*, *211*. https://doi.org/10.1016/j.neuroimage.2020.116604

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3). https://doi.org/10.1016/j.neuroimage.2011.10.018

Rao, A., Monteiro, J. M., Ashburner, J., Portugal, L., Fernandes, O., De Oliveira, L., Pereira, M., & Mourao-Miranda, J. (2015). A Comparison of Strategies for Incorporating Nuisance Variables into Predictive Neuroimaging Models. *Proceedings - 2015 International Workshop on Pattern Recognition in NeuroImaging, PRNI 2015*. https://doi.org/10.1109/PRNI.2015.28

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1). https://doi.org/10.1093/biomet/70.1.41

Salthouse, T. A. (2010). Selective review of cognitive aging. In *Journal of the International Neuropsychological Society* (Vol. 16, Issue 5). https://doi.org/10.1017/S1355617710000706

Solé-Padullés, C., Bartrés-Faz, D., Junqué, C., Vendrell, P., Rami, L., Clemente, I. C., Bosch, B., Villar, A., Bargalló, N., Jurado, M. A., Barrios, M., & Molinuevo, J. L. (2009). Brain structure and function related to cognitive reserve variables in normal aging, mild cognitive impairment and Alzheimer's disease. *Neurobiology of Aging*, *30*(7). https://doi.org/10.1016/j.neurobiolaging.2007.10.008

Suzuki, H., Venkataraman, A. V., Bai, W., Guitton, F., Guo, Y., Dehghan, A., & Matthews, P. M. (2019). Associations of Regional Brain Structural Differences with Aging, Modifiable Risk Factors for Dementia, and Cognitive Performance. *JAMA Network Open*, *2*(12). https://doi.org/10.1001/jamanetworkopen.2019.17257

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B: Methodological*, *58*(1). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vidaurre, D., Bielza, C., & Larrañaga, P. (2013). A Survey of L1 Regression. *International Statistical Review*, *81*(3). https://doi.org/10.1111/insr.12023

Wu, J., Li, J., Eickhoff, S. B., Scheinost, D., & Genon, S. (2023). The challenges and prospects of brain-based prediction of behaviour. In *Nature Human Behaviour* (Vol. 7, Issue 8). https://doi.org/10.1038/s41562-023-01670-1

Yu, J., & Fischer, N. L. (2022). Age-specificity and generalization of behavior-associated structural and functional networks and their relevance to behavioral domains. *Human Brain Mapping*, *43*(8). https://doi.org/10.1002/hbm.25759

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2). https://doi.org/10.1111/j.1467-9868.2005.00503.x