

Reviewed Preprint

v1 • September 9, 2026

Not revised

✉ For correspondence:

narun.pat@otago.ac.nz

Competing interests: No

competing interests declared

Funding: See [page 19](#)

Reviewing editor: Alex Fornito,
Monash University, Australia

© 2026, Lal Khakpoor et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Supervised Domain Adaptation Mitigates Cross-Ethnicity Prediction Error in Neuroimaging-Based Cognitive Prediction

Farzane Lal Khakpoor¹, William van der Vliet¹, Jeremiah Deng², Yue Wang¹, Narun Pat¹ ✉

¹Department of Psychology, University of Otago, Dunedin, New Zealand • ²School of Computing, University of Otago, Dunedin, New Zealand

eLife Assessment

This study provides a **useful** investigation of machine learning approaches that can lessen potential gaps in the prediction of behaviour from brain imaging data across majority and minority samples. The authors provide **incomplete** evidence to suggest that domain adaptation methods can mitigate these gaps. The analysis would benefit from further testing of model generalisability and the inclusion of recommended workflows for how the tested approaches can be used in future research. This work will be of interest to scientists using machine learning in brain imaging.

<https://doi.org/10.7554/eLife.112354.1.sa3>

Abstract

Machine-learning models are increasingly used to predict cognitive and clinical outcomes from neuroimaging data, yet challenges in fairness and generalizability remain. Large-scale datasets are often racially and ethnically imbalanced, leading to systematic performance disparities, with models typically achieving higher accuracy for majority populations represented in the training data. In this study, we evaluated whether supervised domain adaptation methods—including balanced weighting, two-stage TrAdaBoost, feature augmentation with SrcOnly prediction, and linear interpolation—can mitigate these biases. Using the ABCD dataset, we assessed whether models trained on 80 different MRI measures from White American participants could generalize more effectively to African American participants. All domain adaptation methods reduced prediction error for African American participants, particularly for MRI modalities with large baseline disparities (e.g., structural MRI), while offering limited improvements where initial gaps were smaller (e.g., functional connectivity). Among the approaches, balanced weighting performed best and remained stable and beneficial even when only 10 African American participants were used to adapt the original model trained exclusively on White American participants. These findings suggest that simple, low-cost strategies can effectively reduce cross-ethnic performance gaps and improve equity in predictive neuroimaging, offering a practical path forward for future neuroimaging predictive biomarkers.

Significant Statement

Large-scale neuroimaging datasets increasingly enable machine-learning models to predict cognitive and clinical outcomes; however, these datasets are often ethnically/racially imbalanced. As a result, predictive models tend to generalize poorly to underrepresented populations. We demonstrate that, across 80 MRI phenotypes, a class of machine-learning approaches collectively known as supervised domain adaptation can substantially reduce cross-ethnicity disparities in neuroimaging-based cognitive prediction, even when only limited data from underrepresented

groups are available. Among the methods evaluated, balanced weighting achieved the best performance while maintaining low computational cost. Together, these findings provide a practical and scalable framework for improving fairness and generalizability in neuroimaging-based machine learning under realistic conditions of ethnic/racial imbalance.

Introduction

Recent progress in machine learning, combined with increasing access to large-scale neuroimaging datasets, has facilitated a growing number of applications using neuroimaging-based models to predict cognitive and clinical outcomes (Alhuwaydi, 2024 [↗](#); Finn et al., 2015 [↗](#); Sripada, Angstadt, et al., 2020; Sripada, Rutherford, et al., 2020; Sui et al., 2020 [↗](#); Teterova et al., 2022 [↗](#), 2025 [↗](#); Vieira et al., 2022 [↗](#)). Yet, most large-scale datasets consist of White participants from Europe and North America, raising concerns about fairness and generalisability to different ethnicities. Studies have documented performance disparities across ethnic/racial groups when machine-learning models are trained on imbalanced samples, demonstrating that training data imbalances and underlying population structure can lead to systematically poorer performance for underrepresented groups (Carey et al., 2024 [↗](#); Cross et al., 2024 [↗](#); Hasanzadeh et al., 2025 [↗](#); Ueda et al., 2024 [↗](#)).

In the context of brain-behaviour modelling, such disparities are particularly concerning, as neuroimaging data are high-dimensional, noisy, and costly to acquire, and bias may reinforce existing inequities in neuroscience research and downstream applications. Prior work in large-scale neuroimaging cohorts has demonstrated that brain MRI-based models showed systematic differences in prediction error across ethnic groups when predicting cognitive-related outcomes (Adkins & Hanson, 2025 [↗](#); Li et al., 2022 [↗](#)). These biases were not uniform across neuroimaging phenotypes. For instance, structural MRI showed larger cross-ethnic performance gaps than functional MRI (Lal Khakpoor et al., 2025 [↗](#)).

Bias in predictive modelling can arise at multiple stages, spanning data collection, pre-processing, feature construction, model development, and post-deployment evaluation (Hasanzadeh et al., 2025 [↗](#)). In principle, early-stage interventions—such as more diverse data collection or unbiased pre-processing pipelines—are likely to provide the most fundamental solutions. In practice, however, these approaches are often slow to implement, resource-intensive, or unavailable for large existing datasets. For instance, datasets from countries with more limited research budgets are likely to be smaller. As a result, there is a need to understand what can be achieved using currently available imbalanced data and standard pre-processing pipelines at later stages of the pipeline.

In our prior work (Lal Khakpoor et al., 2025 [↗](#)), we addressed representation bias through prior-driven balanced and oversampling strategies. We showed that balancing sample size, by subsampling majority-group participants to match the size of minority groups, narrowed the gap in performance but left disparities significant among MRI phenotypes with greater bias, such as structural MRI. This suggests that the issue is not just "how many" samples, but rather differences in the feature-space relationship between ethnicity and cognition. Therefore, ensuring balanced samples in neuroimaging datasets improves performance but does not guarantee fairness across groups. Moreover, subsampling reduces statistical power, increases variance, and may compromise model stability (Rajput et al., 2023 [↗](#); Susan & Kumar, 2021 [↗](#))—an especially critical issue for neuroimaging data, where reliable estimation already requires large samples (Rajput et al., 2023 [↗](#)). Balancing is also often impractical or infeasible when minority-group sample sizes are very small, a frequent scenario in real-world applications. This motivates the exploration of alternative methodological frameworks that can operate effectively under extreme sample imbalance and explicitly account for cross-group differences in data distributions.

One such framework is domain adaptation (DA). Domain adaptation methods are designed to improve model performance when training (source) and test (target) data come from different distributions (Farahani et al., 2021 [↗](#)). In neuroimaging, DA techniques have been widely applied to address site effects, scanner differences, and protocol variability in multi-site studies as well as

cross-task and cross-dataset classifications (Bayer et al., 2022 [↗](#); Cheng et al., 2015 [↗](#); Kushol et al., 2024 [↗](#); Liu et al., 2019 [↗](#); Panda et al., 2022 [↗](#); Wang et al., 2022 [↗](#); Zhang et al., 2020 [↗](#)). These methods aim to reduce performance degradation when models trained on one domain are applied to another, without requiring full retraining on large target datasets. However, despite their extensive use in cross-site harmonization and generalization problems, domain adaptation approaches have not been examined as tools for mitigating ethnicity-related prediction gap in neuroimaging-based predictive models, particularly for regression-based prediction of continuous outcomes such as cognitive or behavioural measures.

Within a domain adaptation framework, ethnic groups can be conceptualized as distinct domains characterized by partially overlapping but non-identical distributions (Mo & Siepel, 2023 [↗](#); Zhou et al., 2021 [↗](#)). In this view, the majority group can be treated as the source domain, while the minority group participants—constitutes the target domain. The primary objective here is not to equalize performance across ethnic groups, but to improve predictive accuracy for the target domain (i.e., the minority participants) using information learned from the source (i.e., the majority participants). This can be achieved by using a small number of labelled target samples—meaning target-domain examples that include the observed output (observed cognitive functioning scores, in the case of using MRI to predict cognitive functioning). These labelled examples from the target domain help “nudge” the source-trained model toward understanding the target data distribution (Guan & Liu, 2022 [↗](#)). An effective adapted model should therefore be able to encode what generally works and how it differs for the target domain.

Beyond performance gain for the target domain, domain adaptation methods can also provide insight into the mechanisms underlying the performance gap, as comparing feature importance between adapted and non-adapted models can help identify which features change most in their contribution to prediction when adapting to the target domain.

In this study, we addressed the ethnicity-related performance gap in neuroimaging-based cognitive prediction as a domain adaptation problem using the Adolescent Brain Cognitive Development Study (ABCD) dataset. Treating White American (WA) participants as the source domain and African American (AA) participants as the target domain, we evaluate whether domain adaptation can improve neuroimaging-based predictions of cognitive functioning for the underrepresented group when only limited labelled target data are available—a setting in which balanced sampling is infeasible or undesirable. Here, we analysed 80 unimodal neuroimaging phenotypes and focused on those producing the largest and smallest performance gaps, defined as the difference in prediction error when a model trained on only white samples is tested on White-American vs. African-American samples. We systematically evaluated supervised domain adaptation methods that incorporated increasing numbers of labelled target samples and compared their performance against non-adapted models trained with the same data. We utilised four methods applicable to the regression task, including balanced weighting, two-staged TrAdaBoost, feature augmentation with SrcOnly prediction and linear interpolation. Figure 1 [↗](#) shows the overall design of the study.

By systematic evaluation of domain adaptation performance, this work aims to clarify whether and for which phenotypes domain adaptation can mitigate ethnicity-related performance gap in neuroimaging-based prediction of cognition. More broadly, it seeks to assess the practical utility of domain adaptation in neuroimaging-based predictive modelling with underrepresented populations.

Methods

Participants

Data were drawn from the Adolescent Brain Cognitive Development study (ABCD-5.1, DOI:10.15154/z563-zd24), a longitudinal neuroimaging project with a diverse U.S. cohort. The baseline cohort included 11,878 participants (48% female; 52% male) aged 9–10 years (Jernigan et al., 2018 [↗](#)). As shown in Figure 1A [↗](#), to investigate ethnicity-related generalisation, we focused on White-American and African-American participants drawn from the two largest matched

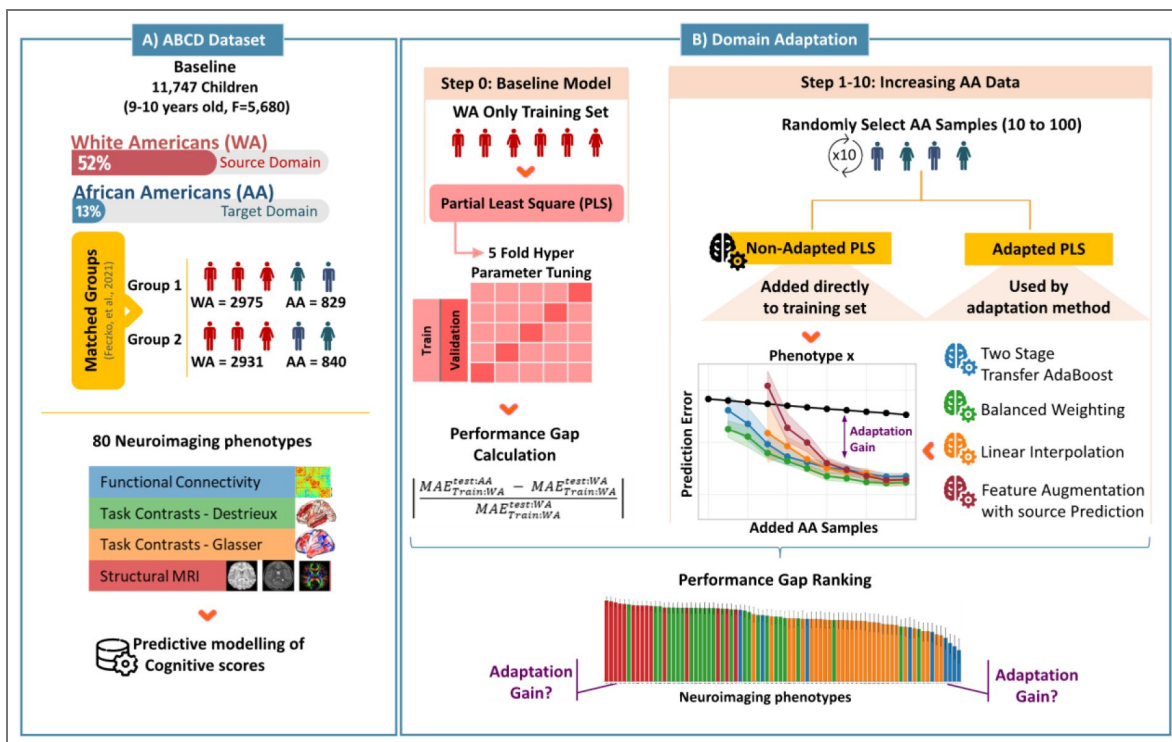


Figure 1. Overall Design.

(A) Participants, matched groups and data extracted from the ABCD dataset. (B) Domain adaptation procedure.

groups within the ABCD-BIDS Community Collection (ABCC-2.0) (ABCD-BIDS Community). The ABCC provides matched groups based on clinical history (e.g., anaesthesia exposure), demographic variables (age, sex, race/ethnicity), and socioeconomic indicators (family income, household size, and family structure) (Feczko et al., 2021 [DOI](#)). For the purposes of the current study, we used the terms “race” and “ethnicity” interchangeably to refer to the characterisation of WA and AA groups.

Cognitive measures

Cognitive performance was indexed using the NIH Toolbox total cognition composite score provided in the ABCD dataset (Akshoomoff et al., 2013 [DOI](#); Jernigan et al., 2024 [DOI](#)) derived from seven tasks: Picture Vocabulary (Gershon et al., 2013 [DOI](#)), Flanker Inhibitory Control & Attention (Zelazo et al., 2013 [DOI](#)), List Sorting Working Memory (Tulsky et al., 2013 [DOI](#)), Dimensional Change Card Sort (Zelazo et al., 2013 [DOI](#)), Pattern Comparison Processing Speed (Carlozzi et al., 2013 [DOI](#)), Picture Sequence Memory (Bauer et al., 2013 [DOI](#)), and Oral Reading Recognition (Gershon et al., 2013 [DOI](#)). Together, these tasks provide a comprehensive assessment of cognitive function across multiple domains.

Neuroimaging Data and Processing

Data Sources

Neuroimaging phenotypes were derived from the Adolescent Brain Cognitive Development (ABCD) Study and integrated across two complementary sources: (i) post-processed tabulated summary measures released by the ABCD consortium (version 5.1) (Jernigan et al., 2024 [DOI](#)) and (ii) minimally pre-processed data from the ABCD-BIDS Community Collection (ABCC-2.0) formatted according to the Brain Imaging Data Structure (BIDS) standard (ABCD-BIDS Community).

The ABCD tabulated datasets provide quality-controlled regional measures and task contrasts parcellated using the Destrieux cortical atlas (Destrieux et al., 2010 [DOI](#)) and FreeSurfer subcortical segmentation (Fischl et al., 2002 [DOI](#)), as described in Hagler et al. (2019 [DOI](#)) and Jernigan et al. (2024 [DOI](#)).

To extend spatial resolution and broaden structural and connectivity coverage beyond these summaries, we conducted additional processing on ABCC data using outputs from fMRIPrep (v20.2.0) and the Human Connectome Project (HCP) minimal pre-processing pipeline. The ABCC includes pre-processed MRI images, confounds tables generated by fMRIPrep, and CIFTI-formatted surface data from the HCP pipeline.

In total, this pipeline yielded 80 distinct neuroimaging phenotypes (Supplementary Figure 1), spanning task contrasts, functional connectivity, structural MRI, and diffusion imaging.

Quality Control and Exclusion Procedures

Quality control followed ABCD-recommended criteria (Jernigan et al., 2024 [DOI](#)). Exclusion flags were applied at the level of individual neuroimaging features based on image quality metrics, MR neurological screening, and task performance indicators (Hagler et al., 2019 [DOI](#)).

We did not perform listwise deletion. Instead, exclusions were applied at the level of the neuroimaging phenotype. For example, if a participant failed quality criteria for the N-back task fMRI but met criteria for other modalities, only the N-back phenotype was excluded.

We excluded Site 22 due to limited sample size and corrected site ID errors (Garavan et al., 2018 [DOI](#)). Additionally, 58 participants with reported vision problems were excluded (Luciana et al., 2018 [DOI](#)). Final sample sizes ranged from approximately 4,000 to 11,177 participants, depending on modality and phenotype.

Neuroimaging Phenotypes

Functional MRI (fMRI) - Task-Evoked Activation (56 Sets)

Task-evoked BOLD responses were derived from three ABCD tasks: Emotional N-back task (Cohen et al., 2016 [↗](#)), Monetary Incentive Delay (MID) task (Knutson et al., 2000 [↗](#)), and Stop Signal Task (SST) (Logan, 1994 [↗](#)). We used 26 precomputed contrasts from the ABCD tabulated releases (9 N-back, 10 MID, 7 SST), parcellated using the Destrieux cortical atlas (Destrieux et al., 2010 [↗](#)) and FreeSurfer subcortical segmentation (Fischl et al., 2002 [↗](#); Hagler et al., 2019 [↗](#)). We computed an additional 30 contrasts (10 per task) using ABCC data and a first-level General Linear Model (GLM) implemented primarily in Nilearn (Abraham et al., 2014 [↗](#); Nilearn contributors et al., 2026) and Nibabel (Brett et al., 2024 [↗](#)).

Event Processing

Task event files in E-Prime format were converted to BIDS-compatible TSV files using the eprimetotsv.py script (Demidenko, 2024 [↗](#)). Event onsets were aligned relative to post-calibration task start times.

Pre-processing and Nuisance Regression

Nuisance regressors were selected based on prior recommendations (Circic et al., 2017 [↗](#); Gotts et al., 2013 [↗](#)). Included confounds were Cosine drift regressors to remove low-frequency noise (Circic et al., 2017 [↗](#)), ten anatomical CompCor components derived from white matter and CSF masks (Muschelli et al., 2014 [↗](#)), and six rigid-body motion parameters (translations and rotations) and their first derivatives (Maknojia et al., 2019 [↗](#)). Specific motion regressors included rot_x, rot_y, rot_z, trans_x, trans_y, trans_z and their first derivative terms. Non-steady-state volumes were removed prior to modeling. The number of removed TRs depended on scanner type: Siemens and Philips (8 TRs), GE DV25 (5 TRs), GE DV26 (16 TRs) (Jernigan et al., 2024 [↗](#)).

General Linear Model Estimation

First-level design matrices were constructed using the SPM canonical hemodynamic response function. Time series were standardized, and General Linear Models (GLMs) were estimated using ordinary least squares (OLS). Contrast matrices were defined to isolate task-specific cognitive processes. Run-level beta coefficients were estimated and combined across runs using variance-weighted fixed effects. Final contrast maps were parcellated into 379 regions (360 Glasser cortical parcels (Glasser et al., 2016 [↗](#)) and 19 FreeSurfer subcortical regions) to yield regional effect sizes.

Functional Connectivity (9 Sets)

Functional connectivity (FC) phenotypes were derived from both ABCD tabulated releases and ABCC processing. From ABCD Tabulated Resting-State FC data, we included: 1) Temporal variance (amplitude of low-frequency fluctuations) across 333 Gordon cortical parcels (Gordon et al., 2016 [↗](#)) and 19 subcortical regions. 2) Subcortical-to-network connectivity (average correlations between 19 subcortical regions and 13 cortical networks). 3) Within- and between-network cortical connectivity across 13 large-scale networks (Jernigan et al., 2024 [↗](#)). Using ABCC data, we computed six 379-region FC matrices for Rest, N-bac, MID, SST, Multitask FC (concatenation across tasks) (Cole et al., 2021 [↗](#)), and General FC (rest + all tasks concatenated) (Elliott et al., 2019 [↗](#)). Concatenation-based FC measures were motivated by evidence that aggregating across tasks improves reliability and behavioural prediction (Cole et al., 2014 [↗](#), 2019 [↗](#); Elliott et al., 2019 [↗](#)).

Motion and Outlier Handling

High-motion TRs were identified using Framewise displacement (FD > 0.5) (Power et al., 2013 [↗](#)) and DVARS > 1.5 (Afyouni & Nichols, 2018 [↗](#); Pham et al., 2023 [↗](#)). Flagged TRs were included as nuisance regressors (Power et al., 2014 [↗](#)). Participants were excluded if more than 50% of timepoints were flagged. A relatively lenient FD threshold was adopted to balance artifact mitigation with signal preservation in this developmental, multiband dataset (Addeh et al., 2023 [↗](#); Pham et al., 2023 [↗](#); Power et al., 2019 [↗](#)).

Connectivity Estimation

Nuisance regressors were selected and included in the design matrix, similar to task contrasts. For task FCs, the events were also included in the design matrix. After nuisance regression, residual time series were extracted, de-spiked (values >3 SD replaced with mean; (Costabel & Müller-Petke, 2014 [↗](#)), concatenated across runs, and parcellated into 379 regions. Functional connectivity was computed using Pearson correlation, followed by Fisher z-transformation and vectorization.

Structural MRI (sMRI) and Diffusion MRI (16 Sets)

Structural phenotypes were derived primarily from FreeSurfer processing of T1- and T2-weighted images (Fischl et al., 2002 [↗](#)). Cortical regions were parcellated using the Destrieux atlas, and subcortical regions using FreeSurfer segmentation. We included 15 structural phenotypes that captured complementary aspects of morphology and tissue properties. FreeSurfer summations comprised five global metrics: estimated intracranial volume, total cortical gray matter volume, total cortical white matter volume, total subcortical gray matter volume, and brain segmentation volume-to-intracranial volume ratio (Tetereva et al., 2022 [↗](#)). T1 and T2 summary phenotypes capture broader regional and global signal characteristics sensitive to tissue composition, including myelination and water content (Jørgensen et al., 2016 [↗](#); Xu et al., 2024 [↗](#)), complementing localized surface-based intensity measures.

Diffusion phenotypes consisted of fractional anisotropy (FA) extracted from 23 major white-matter tracts identified using AtlasTrack (Hagler Jr et al., 2009). FA quantifies directional water diffusion and indexes white-matter microstructural organization (Alexander et al., 2007 [↗](#)) (Alexander et al., 2007 [↗](#)). The processing steps are explained in Hagler and colleagues (2019) [↗](#).

See Supplementary Figure 1 for the list of phenotypes and their corresponding sample sizes.

Ethnicity-related Performance Gap

To quantify cross-ethnicity misprediction for each neuroimaging phenotype, we computed a performance gap based on differences in prediction error when models trained on one ethnicity were evaluated on the same-ethnicity versus different-ethnicity test samples. Specifically, this metric captures the relative increase in prediction error when a model trained on White Americans is applied to African Americans, compared to its performance on White Americans test data (higher values indicate more misprediction for AA). This measure provides an index of cross-ethnicity generalisation error and was used to stratify phenotypes by baseline bias, enabling evaluation of domain adaptation under varying levels of misprediction. The performance gap was defined as:

$$\text{Performance Gap} = \left| \frac{MAE_{\text{Train: WA}}^{\text{Test: AA}} - MAE_{\text{Train: WA}}^{\text{Test: WA}}}{MAE_{\text{Train: WA}}^{\text{Test: WA}}} \right| \quad (1)$$

Superscript shows the test group that MAE is calculated for, and subscript indicates the group that the model is trained on.

Domain Adaptation

We evaluated four machine learning-based supervised domain adaptation methods implemented in the ADAPT Python library (de Mathelin et al., 2021 [↗](#)), all of which are applicable to regression settings and allow explicit incorporation of labelled target-domain samples. Unlike unsupervised methods that only align feature distributions, these supervised approaches utilize labelled target data to address shifts in outcome (cognitive scores) distributions $p(y)$, marginal shifts in feature space (input MRI data) $p(x)$, and feature-outcome conditional shifts $p(y|x)$ either directly or indirectly (Farahani et al., 2021 [↗](#); Katz et al., 2024 [↗](#)). To ensure computational feasibility and consistency across analyses, adaptation hyperparameters were fixed across steps and phenotypes.

These values were selected based on an initial exploratory grid search on validation performance. We selected the most frequently optimal parameter setting and applied it uniformly throughout the study.

Balanced Weighting

Balanced Weighting adjusts the importance of source and target samples during training by assigning higher weights to target-domain instances relative to source-domain instances. By increasing the influence of target samples, Balanced Weighting encourages the learned model to better reflect the target-domain distribution without discarding source-domain data (Wu & Dietterich, 2004 [↗](#)). In this study, Balanced Weighting allows examination of whether reweighting a small number of AA samples can improve prediction relative to naïve inclusion in non-adapted training. The ratio parameter, which corresponds to the importance given to the target labelled data, was set to 0.8. This was fixed across adaptation steps and was decided based on an exploratory grid search.

Two-stage TrAdaBoost

a sophisticated instance-based boosting algorithm that iteratively reweights samples. It identifies and down-weights misleading source samples—those whose brain-cognition associations contradict the patterns observed in the target AA domain—while boosting the influence of informative target samples (Pardoe & Stone, 2010 [↗](#)). The number of boosting iterations for the first and second stages was fixed at 10 and 20, respectively, with a learning rate of 0.1.

Feature Augmentation with SrcOnly Prediction (PRED)

A feature-based adaptation method where a model is first pretrained on the source domain. The resulting source-only predictions of target labelled data are then used as an augmented feature to train a second-stage model on the target data. This approach allows the model to leverage source-group knowledge as an informative prior while specifically learning target-group residuals (Daumé, 2009 [↗](#)).

Linear Interpolation

A parameter-based adaptation strategy that creates a new model by interpolating between the parameters of a source-only model and a target-only model. By weighting contributions from both domains, this method seeks an optimal parameter set that preserves generalizable source knowledge while accounting for target-specific signals. This approach does not explicitly reweight individual samples but instead interpolates between source and target solutions as the proportion of labelled target data increases (Daumé, 2009 [↗](#)). The proportion of target-domain data used to train the target-only model was set to 0.7. The remaining target data were used to estimate the interpolation parameter.

Baseline Estimator

Partial Least Squares (PLS) was used as the estimator for adaptation methods and as the standard non-adapted model. PLS offers several advantages that are particularly relevant for the present study. First, it is well suited to high-dimensional, highly collinear neuroimaging data, such as functional connectivity features, where the number of predictors greatly exceeds the number of observations (Krishnan et al., 2011 [↗](#)). Second, PLS provides a low-dimensional latent representation that is directly optimized for predicting the outcome (Krishnan et al., 2011 [↗](#)), which facilitates interpretability of feature reweighting across domains. Third, PLS is computationally efficient, given the large number of phenotypes, adaptation methods, and resampling steps evaluated in this study. To optimize the number of PLS components, we performed a grid search with five-fold cross-validation, capping the number of components at 30. The optimal model was selected based on the negative mean squared error (NMSE).

Experimental design

The experimental design followed a structured, stepwise adaptation framework to evaluate how model performance on AA participants changes as increasing amounts of labelled target data are incorporated (Figure 1B [↗](#)).

In Step 0, a baseline PLS regression model was trained using only WA participants from the training set (i.e., either matched group 1 or 2). This model represents the source-only scenario and serves as the reference point for all subsequent comparisons.

In Steps 1–10, labelled AA samples were incrementally added to the training process. At each step, a fixed number of AA participants (ranging from 10 to 100 in increments of 10) were randomly selected from the training pool (i.e. either matched group one or two). These samples were either (i) included directly in the training of a non-adapted PLS model or (ii) provided as labelled target-domain data to the domain adaptation methods. This design allows direct comparison between naïve data inclusion and adaptation-based modelling under matched conditions.

For each step, the random selection of AA samples was repeated 10 times to account for sampling variability. Model performance was evaluated on a held-out AA test set that was not used during training or adaptation (e.g., from matched group 2 when the model was trained on matched group 1). Predicted values, error metrics, and model coefficients were saved for each repetition and step for subsequent analysis.

Evaluation metrics

For each adaptation step, MAE values were obtained across repeated random selections of labelled AA samples. To compare adapted models with non-adapted training at each step, paired t-tests were conducted across matched repetitions, comparing MAE values from adapted and non-adapted models trained on the same randomly selected AA samples.

For Linear Interpolation and PRED methods, statistical comparisons were restricted to adaptation steps with $N = 30$ –100, as these approaches involved training target-only models and were unstable at smaller target sample sizes. In contrast, Balanced Weighting and Two-stage TrAdaBoost were stable from $N = 10$.

To summarize adaptation benefits across target sample sizes, we computed the area under the improvement curve (AUC), defined as the integral of $\Delta\text{MAE}_{(N)} = \text{MAE}_{\text{non-adapted}(N)} - \text{MAE}_{\text{adapted}(N)}$ across adaptation steps starting from $N=30$. Starting at $N=30$ allowed us to compare AUC across all four adaptation methods. This scalar measure provides a compact summary of cumulative adaptation benefit across increasing availability of labelled AA data, without assuming independence between steps. The same cumulative adaptation benefit was also calculated using R-squared values. We computed Pearson's correlation between performance gap scores and cumulative adaptation AUC gains (averaged across adaptation methods) to test whether adaptation benefit scales with performance gap severity. This analysis evaluates whether adaptation acts primarily as a gap-mitigation mechanism rather than providing uniform gains across phenotypes.

To compare the effectiveness of the four domain adaptation approaches in reducing the performance gap, cumulative AUC gains were compared across phenotypes using repeated-measures nonparametric analyses. An omnibus Friedman test was first performed to assess overall differences between methods. Pairwise post-hoc comparisons were then conducted using two-sided Wilcoxon signed-rank tests with Holm correction for multiple comparisons. Effect sizes for pairwise comparisons were quantified using paired Cohen's d .

Model interpretation

To examine how adaptation alters outcome-relevant feature weighting, we tracked changes in PLS regression coefficients across adaptation steps for the adaptation method with the largest gain value. For each MRI phenotype, we computed coefficient differences between models trained with no AA data (Step 0) and models trained with 100 labelled AA samples (step 10).

To isolate adaptation-specific effects, we employed a difference-of-differences formulation: (i) the change in coefficients for the non-adapted PLS model after adding 100 AA samples was computed, (ii) the corresponding change for an adapted model was computed, and (iii) the difference between these two changes was taken as an index of adaptation-related reweighting beyond naïve data inclusion. Higher values reflect larger adaptation-specific shifts in the feature's signed contribution to the predicted outcome (PLS coefficient), beyond changes induced by adding AA data alone.

Coefficient changes were visualized by mapping features back onto brain space for each phenotype. These analyses are descriptive and intended to characterize patterns of feature reweighting associated with adaptation; they do not support causal inference regarding the sources of bias.

Results

Performance Gap Ranking

The ranking of phenotypes based on the gap metric revealed systematic differences across modalities, as shown in [Figure 2](#). Functional connectivity measures exhibited the smallest gap values, indicating that models trained exclusively on WA participants showed the least degradation in performance when applied to AA participants relative to WA test performance.

This pattern was followed by task-based fMRI contrast measures derived from the Glasser parcellation, particularly Nback task contrasts, which also demonstrated relatively low gap values. Across task-based contrasts more broadly, most phenotypes occupied an intermediate range of the ranking, with Glasser-parcellated phenotypes exhibiting lower gap values compared to Destrieux-based phenotypes.

In contrast, sMRI and DTI phenotypes consistently showed the largest gap values, indicating greater disparity in model performance between AA and WA participants.

Domain adaptation performance

We first evaluated whether supervised domain adaptation improves prediction performance for AA participants relative to non-adapted PLS models trained with naïve inclusion of AA samples.

To summarize adaptation effects across target sample sizes, we computed the area under the improvement curve (AUC) for each of 80 MRI phenotypes and adaptation methods. [Figure 3a](#) illustrates adaptation gains of each method across phenotypes ranked by their performance gap score. AUC values revealed a clear structure across phenotypes: phenotypes with a larger gap showed larger cumulative gains, whereas those with a smaller gap showed smaller improvements. Connectivity phenotypes consistently showed near-zero or negative AUC values, indicating minimal or absent cumulative benefit from adaptation. A similar pattern was observed when we calculated the adaptation benefit based on changes in the coefficient of determination (R^2), as shown in [Supplementary Figure 2](#). The performance gap score was significantly correlated with average gain from adaptation ($r=0.79$, $p<0.001$) ([Figure 4b](#)).

Across MRI phenotypes with a larger gap, all adaptation methods showed significant reductions in MAE relative to non-adapted training, as shown in [Figure 4a](#), and these reductions persisted as additional target samples were incorporated ($p < 0.05$). Across MRI phenotypes with a lower gap, task-based contrasts also showed improvements with adaptation; however, the magnitude of MAE reduction was smaller, and adapted and non-adapted performance lines were closer throughout the adaptation range. Nevertheless, adapted models often achieved modest but consistent reductions in MAE. In contrast, connectivity phenotypes showed little to no benefit from adaptation, where performance curves for adapted and non-adapted models largely overlapped across adaptation steps, and in some cases, adapted models exhibited slightly higher MAE before converging to similar values at larger sample sizes. The change in prediction errors across adaptation steps for each of 80 phenotypes is depicted in [Supplementary Figure 3](#).

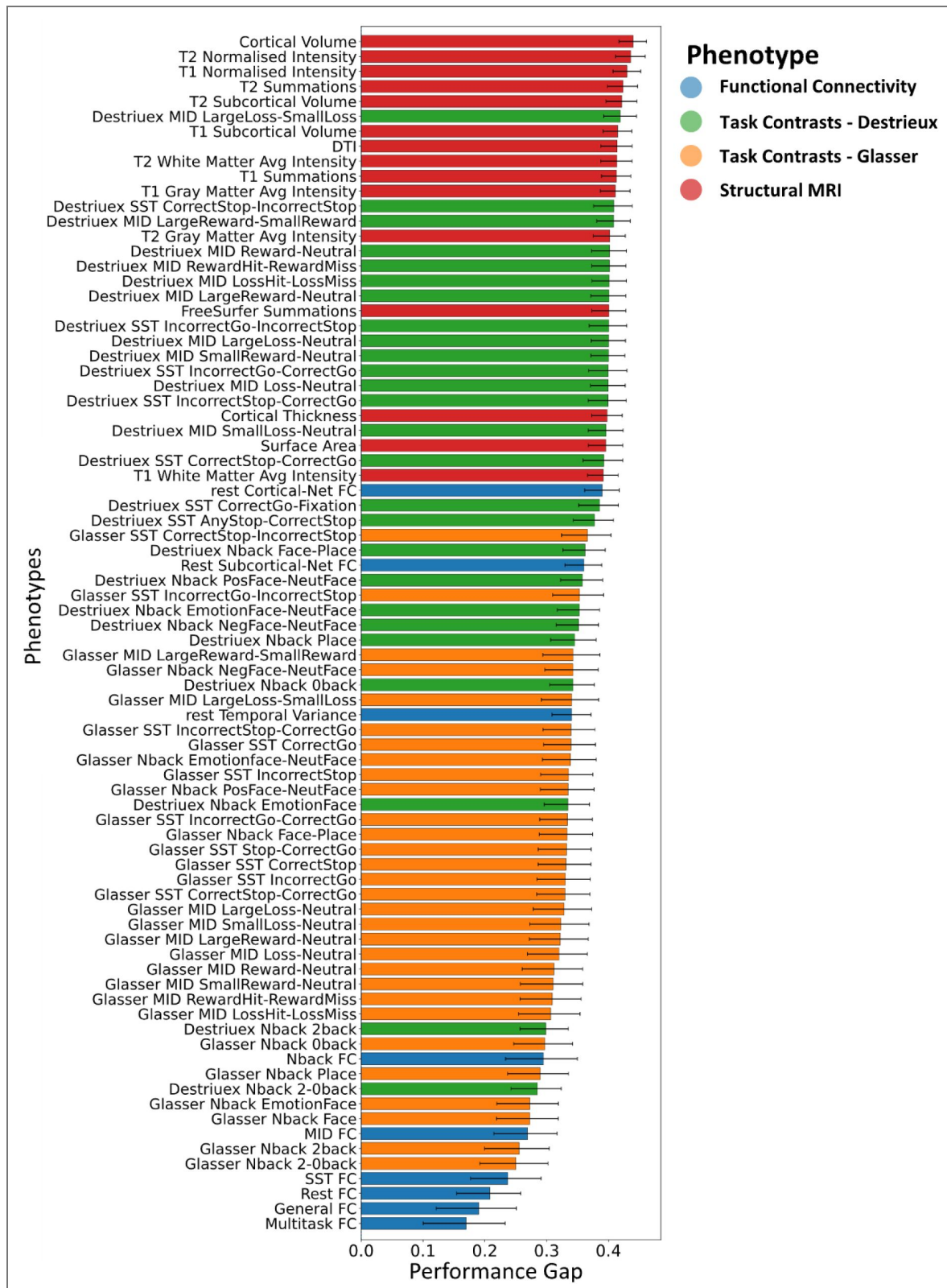


Figure 2. The gap in predictive performance (MAE) of the Baseline Model (trained only on WA) when tested on WA versus AA participants in the test set.

Confidence intervals were estimated using non-parametric bootstrap resampling (5,000 iterations) at the subject level, with resampling performed independently within AA and WA groups. For each bootstrap sample, MAEs and the gap metric were recomputed, and 95% confidence intervals were derived using the percentile method.

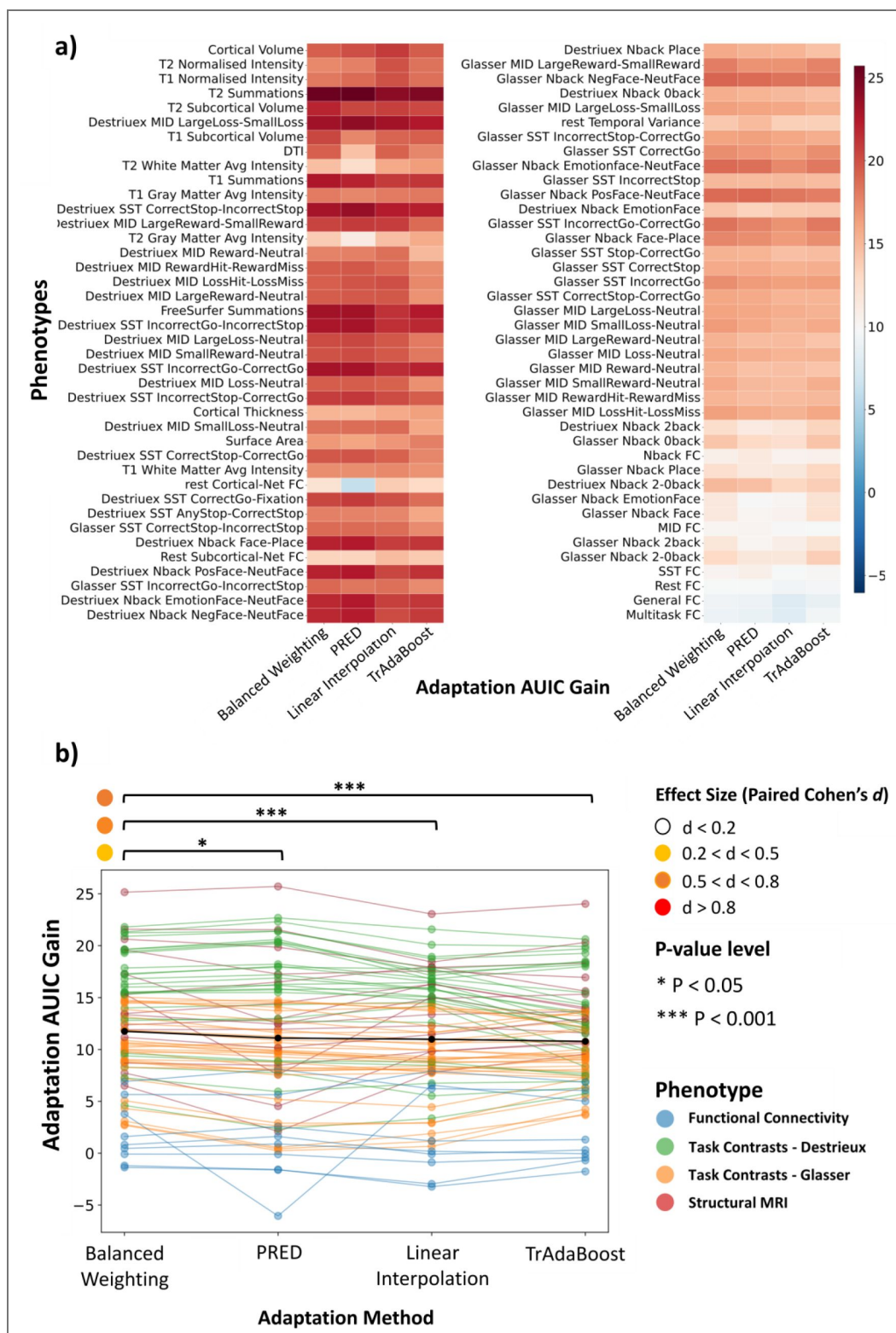


Figure 3. Adaptation Benefits.

(a) The cumulative AUIC gains of each adaptation method across all phenotypes, sorted by their performance gap scores, starting from phenotypes with the largest gap at the top left to phenotypes with the smallest gap at the bottom right. A positive AUIC means the adapted model produced a cumulative reduction in MAE as AA labels increased. (b) Repeated-measures comparison of adaptation-method AUIC gain across neuroimaging phenotypes. Each line represents an individual phenotype tracked across adaptation methods, illustrating within-phenotype changes in performance gain. The black line shows the average gain across phenotypes for each method. Colored points indicate phenotype-specific AUIC values for each method. Significance levels from Pairwise Wilcoxon signed-rank tests with Holm correction and paired Cohen's *d* are shown.

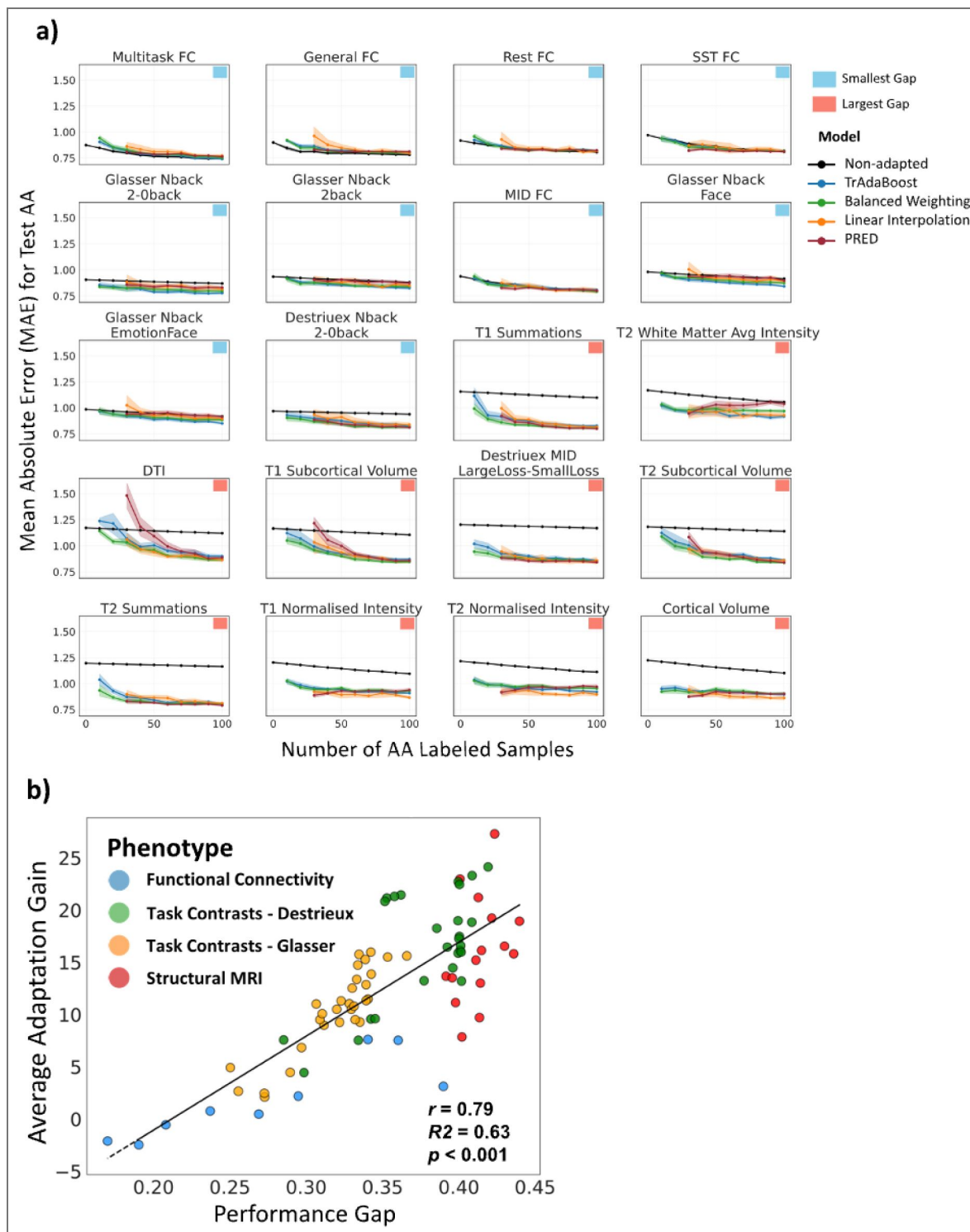


Figure 4.

(a) Changes in AA test prediction error as increasing numbers of labelled AA participants were incorporated into model training, either by direct inclusion (non-adapted) or via adaptation. Points show the mean MAE across repeated random AA subsamples at each target sample size, and shaded bands indicate approximate 95% confidence intervals across repetitions ($\pm 1.96 \text{ SD}/n$). Results are shown for the ten phenotypes with the largest and smallest gap scores. **(b)** The scatter plot of the correlation between average cumulative AUIC gain and performance gap across phenotypes.

Comparison across domain adaptation methods

We used a Friedman test to determine whether the largest cumulative reduction in MAE, as quantified by the Area Under the Improvement Curve (AUC), differed significantly across Adaptation methods. The results showed a significant main effect of Adaptation methods, $\chi^2(3) = 40.85$, $p = 7.05 \times 10^{-9}$ (Figure 3b [↗](#)). Post hoc comparisons using the Wilcoxon signed-rank tests with Holm correction revealed that Balanced Weighting ($M = 11.75$, $SD = 6.04$) lead to significantly higher AUC than PRED ($M = 11.10$, $SD = 6.76$, corrected $p = .02$, paired Cohen's $d = .36$), Linear Interpolation ($M = 10.99$, $SD = 5.99$ corrected $p < .001$, paired Cohen's $d = .54$), and Two-stage TrAdaBoost ($M = 10.78$, $SD = 5.36$ corrected $p < .001$, paired Cohen's $d = .61$). However, there was no statistically significant difference in AUC between PRED and Linear Interpolation (corrected $p = .07$, paired Cohen's $d = .05$) and between PRED and Two-stage TrAdaBoost (corrected $p = .07$, paired Cohen's $d = .12$) and between Linear Interpolation and and Two-stage TrAdaBoost (corrected $p = .74$, paired Cohen's $d = .11$).

It is important to note that when the number of target samples was fewer than 30, PRED and Linear Interpolation were unstable and could not be reliably trained. In contrast, under the constant-strategy setting, Balanced Weighting and Two-stage TrAdaBoost remained stable and could be applied with as few as 10 target samples. Therefore, AUC was computed starting from 30 target samples onward.

Adaptation-related changes in PLS feature weights

Finally, we examined how domain adaptation altered outcome-relevant feature weighting by comparing changes in PLS coefficients between Step 0 and Step 100 for Balanced Weighting as the method with the highest average gain. Across phenotypes, the magnitude of adaptation-related coefficient changes varied substantially, as indicated by the range of color bars in Figure 5 [↗](#). Phenotypes with the largest gap exhibited the largest coefficient differences, while task-based phenotypes with smaller gap scores showed smaller changes, and connectivity phenotypes exhibited minimal coefficient differences, with maximum values below 0.001.

Discussion

This study represents one of the first efforts to apply supervised domain adaptation to regression-based neuroimaging prediction to reduce ethnicity bias in neuroimaging-based cognitive prediction. We examined whether these methods could reduce ethnicity-related performance disparities in cognitive prediction, and identified the conditions under which they are most effective. Our systematic comparison of domain-adaptation strategies demonstrates that adaptation can improve prediction accuracy for AA participants even with limited labelled data, although the magnitude of improvement varies across phenotypes. Among the four strategies evaluated, balanced weighting achieved the best overall performance while maintaining low computational cost, highlighting its promise as a practical approach for reducing ethnicity bias in neuroimaging-based cognitive prediction.

Phenotypes differ in performance gap across ethnicities

The ranking of performance gap scores indicates that functional connectivity phenotypes are comparatively more robust to cross-ethnicity generalisation when models are trained on WA participants. In contrast, structural and diffusion-derived features show greater sensitivity to cross-group differences, exhibiting larger increases in prediction error under cross-ethnicity evaluation.

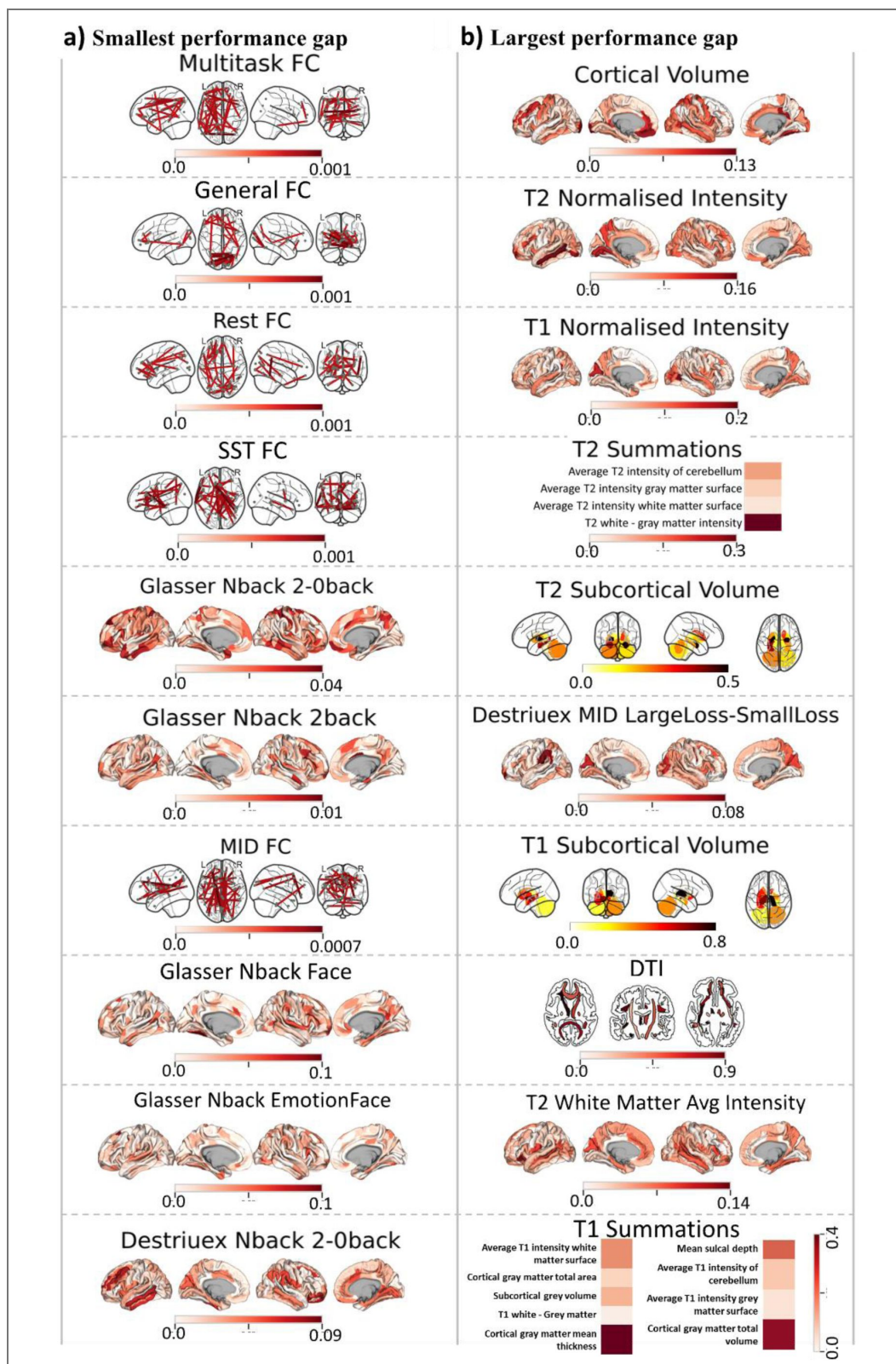


Figure 5. Adaptation-specific feature re-weighting across top ten phenotypes with the (a) smallest and (b) largest performance gap.

This was estimated via a difference-of-differences framework comparing coefficient changes in adapted (Balanced Weighting) and non-adapted PLS models after adding 100 AA samples relative to the baseline (no AA added).

Supervised domain adaptation improves AA prediction with limited labelled data

Across most MRI phenotypes—particularly those with the largest gap—supervised domain adaptation led to significant reductions in prediction error on held-out AA test samples compared with naïve inclusion of the same labelled AA data in non-adapted PLS models. Notably, for Balanced Weighting and Two-stage TrAdaBoost, improvements were evident with as few as 10 labelled AA samples and increased as more target data became available.

At the same time, improvement trajectories varied substantially across MRI phenotypes. Some showed steep early declines in error, while others exhibited gradual or negligible change. This heterogeneity indicates that adaptation effectiveness cannot be inferred from representing sample size alone and instead reflects phenotype-specific properties of the underlying data-generating process.

The strongest overall improvements were observed for the instance-based adaptation approach of Balanced Weighting. This method leverages a small amount of labelled target-domain data to reweight source samples during training, allowing the model to place greater emphasis on source instances that are more representative of the target population. Despite its simplicity, Balanced Weighting achieved the largest overall reduction in performance disparities across phenotypes. Unlike Two-stage TrAdaBoost, which iteratively updates sample weights over multiple boosting rounds, Balanced Weighting applies a direct reweighting strategy, making it computationally efficient and straightforward to implement. Although the absolute performance differences between adaptation methods were modest, Balanced Weighting overall outperformed the other approaches and required substantially less computational overhead than Two-stage TrAdaBoost. In contrast, methods relying primarily on source-informed target predictions or parameter interpolation (PRED and Linear Interpolation) were not applicable in very small-sample settings, requiring more labelled target samples (at least 30) before adaptation could be performed. However, once sufficient target data were available, these approaches achieved performance improvements comparable to those of the instance-based methods.

Adaptation reweights feature contributions

Analysis of PLS coefficient changes further supports this interpretation. Phenotypes that benefited most from adaptation exhibited larger adaptation-specific changes in regression coefficients beyond those induced by naïve inclusion of AA data. In contrast, phenotypes with little or no adaptation gain showed minimal additional changes in the coefficient.

This correspondence suggests that successful adaptation meaningfully alters the effective weighting of features toward the target domain when a source–target mismatch exists, resulting in both larger coefficient shifts and improved prediction. When the non-adapted model is already close to target-optimal, little reweighting occurs, and little improvement is observed. Adaptation-related changes in PLS weights were most consistently observed in orbitofrontal and cingulate cortices, with more distributed effects in visual and sensorimotor regions, based on cortical volume and normalized intensity measures. Subcortically (T1 and T2 subcortical volumes), the largest effects were concentrated in striatal structures, with additional involvement of thalamic and medial temporal regions (hippocampus and amygdala). These analyses are descriptive and do not imply causal involvement of specific brain regions or connections; rather, they characterize how adaptation redistributes predictive emphasis across features as a function of phenotype-specific mismatch.

Practical implications

These findings have direct implications for fairness and generalisability in neuroimaging-based predictive modelling. They indicate that ethnicity-related performance gap is heterogeneous in origin and that mitigation strategies must be matched to the structure of domain differences. Supervised domain adaptation is most effective for phenotypes with ethnicity-specific predictive disparities. Although this study focused on ethnicity-related differences, the same framework

holds promise for other underrepresented groups, including clinical populations, rare conditions, and sociodemographically stratified cohorts, particularly for continuous behavioural or cognitive outcomes for which model generalisability is rarely investigated.

In applied neuroimaging settings, obtaining balanced, exhaustive training data is often infeasible. Imaging protocols, scanners, and acquisition characteristics evolve over time; sensitive or clinical-labelled outcomes are costly and limited; and models trained on highly heterogeneous data may sacrifice accuracy by relying on overly coarse features to ensure robustness (Wang et al., 2022 [↗](#)). Within these constraints, domain adaptation provides a data-efficient means of improving prediction for underrepresented groups by explicitly preserving transferable knowledge from the source domain while adjusting for target-specific shifts. Under scarce target-sample regimes, instance-based approaches, particularly, balanced weighting offer robust gains with minimal overhead. Moreover, our prior work has shown that structural features continue to favour WA populations even under balanced sampling (Lal Khakpoor et al., 2025 [↗](#)). Given the widespread clinical availability of structural MRI, the ability of adaptation techniques to improve performance using small, labelled target samples is therefore particularly promising for real-world deployment.

Rather than framing fairness solely as equalizing performance across groups, these techniques enable the development of models optimized for peak performance within specific target groups, particularly when biological, environmental, or pre-processing factors inherently amplify domain distinctions in feature or outcome distributions that might not otherwise exist. This approach serves as a pragmatic solution in the absence of methods to fully eliminate such confounding factors.

Limitations and future directions

Several limitations warrant consideration. First, we deliberately constructed a highly skewed adaptation scenario by restricting the number of labelled AA samples used for training to between 10 and 100, despite the availability of substantially larger AA samples in the full dataset. This design choice was intentional and motivated by our aim to evaluate the effectiveness of domain adaptation in settings where balanced sampling is infeasible or statistically undesirable. In real-world neuroimaging studies, minority groups are often severely underrepresented, and subsampling majority groups to achieve balance can lead to substantial loss of power and model instability. Our results, therefore, speak specifically to the utility of adaptation under extreme imbalance, rather than to the limits of performance achievable with fully balanced data.

Second, our analyses focused on unimodal neuroimaging phenotypes and linear regression models. This choice reflects both the high dimensionality of MRI data and the regression nature of the cognitive prediction task, and it enabled us to evaluate domain adaptation using well-established, widely validated supervised methods with strong theoretical and empirical grounding. However, this design limits direct generalization to multimodal or nonlinear modelling frameworks. More complex deep learning-based adaptation and representation learning approaches were therefore not explored and remain an important direction for future work (Liang et al., 2023 [↗](#); Stan & Rostami, 2023 [↗](#)), with the present study providing a principled baseline for such extensions.

Finally, we examined a single composite cognitive score, derived from multiple cognitive tasks, within one large cohort. Extension of this framework to other behavioural, clinical, or developmental targets—particularly those with differing noise characteristics or domain sensitivities—will be important for assessing the utility of domain adaptation.

Conclusion

In summary, this study demonstrates that supervised domain adaptation can substantially mitigate ethnicity-related performance disparities in regression-based neuroimaging prediction while requiring only a small amount of labelled target-population data.

Across diverse MRI phenotypes, adaptation-based models consistently produced larger improvements in predictive performance than simply adding the same target samples into conventional training pipelines. These improvements were achieved using computationally efficient methods—particularly balanced weighting—without requiring fully balanced multi-ethnic training datasets or more computationally intensive mitigation strategies, such as ethnicity-specific or ethnicity-neutral template construction, specialised pre-processing pipelines, or large-scale deep learning fine-tuning. Notably, the largest adaptation benefits were observed for MRI phenotypes showing the greatest baseline cross-ethnicity performance disparities, suggesting that supervised adaptation may be especially valuable for high-bias predictive features.

Overall, these findings highlight the value of targeted adaptation strategies for improving model performance in underrepresented groups without requiring fully balanced datasets. More broadly, they support the use of domain adaptation as a principled framework for enhancing fairness and generalisability in neuroimaging-based prediction of continuous behavioral outcomes.

Data availability

ABCD Study data are accessible via <https://nda.nih.gov/abcd> upon NDA approval. All analysis code is available at <https://github.com/HAM-lab-Otago-University/Ethnicity-DA>

Acknowledgements

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (<https://abcdstudy.org>), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is available at <https://abcdstudy.org/federal-partners.html>. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from <https://doi.org/10.15154/z563-zd24>. DOIs can be found at <https://nda.nih.gov/abcd/abcd-annual-releases>. The authors acknowledge the use of New Zealand eScience Infrastructure (NeSI) high-performance computing facilities and support services, funded by NeSI collaborator institutions and the Ministry of Business, Innovation & Employment. This research was supported by the New Zealand Health Research Council (Grant numbers: 21/618 and 24/838; awarded to Narun Pat), the University of Otago (awarded to Farzane Lal Khakpoor), the Neurological Foundation of New Zealand (Grant number: 2350 PRG; awarded to Jeremiah Deng and Narun Pat) and the Ministry of Business, Innovation and Employment (grants UOA2421 and RTVU2403 to Narun Pat).

Additional information

Author contributions

Khakpoor and Pat conceived and designed the study. Data acquisition, analysis and interpretation were conducted by Khakpoor, Pat, Deng and van der Vliet and Yue Wang. Khakpoor and Pat drafted the manuscript and performed statistical analyses. All authors critically revised the manuscript. Funding was acquired by Pat. Pat also supervised the project. All authors approved

the final version and are accountable for the accuracy and integrity of the work. This manuscript reflects the authors' views and does not necessarily represent those of any funding body or the ABCD Consortium.

AI Use Disclosure Statement

During manuscript preparation, the first author used ChatGPT (GPT-5.2, OpenAI) to assist with language editing and refinement. The AI tool was used solely for editorial support. All scientific content, analyses, interpretations, and conclusions were developed by the authors.

Ethics approval statement

The ethical considerations of the Adolescent Brain Cognitive Development (ABCD) Study, including informed consent and assent procedures, participant confidentiality, and communication of assessment results, have been described in detail elsewhere (Clark et al., 2018 [↗](#)). The Institutional Review Board (IRB) at the University of California, San Diego (UCSD; IRB #160091, approved September 13, 2016), serves as the central IRB, with reliance agreements in place across all 21 data collection sites. Prior to enrolment, written informed consent was obtained from a parent or legal guardian, and assent was obtained from all child participants.

Funding

Funder	Grant reference number	Author
Neurological Foundation of New Zealand (Neurological Foundation)	2350 PRG	Narun Pat Jeremiah D Deng
minhealthnz Health Research Council of New Zealand (HRC)	21/618	Narun Pat
minhealthnz Health Research Council of New Zealand (HRC)	24/838	Narun Pat
University of Otago (Te Whare Wānanga o Ōtāgo)		Narun Pat
University of Otago (Te Whare Wānanga o Ōtāgo)		Farzane Lal Khakpoor
Ministry of Business, Innovation and Employment (MBIE)	UOA2421	Narun Pat
Ministry of Business, Innovation and Employment (MBIE)	RTVU2403	Narun Pat

Author ORCID iDs

Jeremiah Deng:  <https://orcid.org/0000-0003-3727-4403>

Narun Pat:  <https://orcid.org/0000-0003-1459-5255>

Additional files

[Supplementary Information.](#) [↗](#)

References

1. **ABCD-BIDS Community** (2019) ABCD-BIDS Community Collection (ABCC) [Dataset]. NIMH Data Repositories. https://nda.nih.gov/edit_collection.html?id=3165
2. **Abraham A**, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G (2014) Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics* **8**:14-14 <https://doi.org/10.3389/fninf.2014.00014> | [PubMed](#)

3. **Addeh A**, Vega F, Medi PR, Williams RJ, Pike GB, MacDonald ME (2023) Direct machine learning reconstruction of respiratory variation waveforms from resting state fMRI data in a pediatric population. *NeuroImage* **269**:119904-119904 <https://doi.org/10.1016/j.neuroimage.2023.119904> | PubMed
4. **Adkins DJ**, Hanson JL (2025) Racial and Ethnic Disparities in Brain Age Algorithm Performance: Investigating Bias Across Six Popular Methods. *medRxiv* <https://doi.org/10.1101/2025.09.18.25336117> | PubMed
5. **Afyouni S**, Nichols TE (2018) Insight and inference for DVARS. *NeuroImage* **172**:291-312 <https://doi.org/10.1016/j.neuroimage.2017.12.098>
6. **Akshoomoff N**, Beaumont JL, Bauer PJ, Dikmen SS, Gershon RC, Mungas D, Slotkin J, Tulskey D, Weintraub S, Zelazo PD, *et al.* (2013) NIH Toolbox Cognition Battery (Cb): Composite scores of crystallized, fluid, and overall cognition. *Monographs of the Society for Research in Child Development* **78**:119-132 <https://doi.org/10.1111/mono.12038>
7. **Alexander AL**, Lee JE, Lazar M, Field AS (2007) Diffusion Tensor Imaging of the Brain. *Neurotherapeutics* **4**:3 <https://doi.org/10.1016/j.nurt.2007.05.011> | PubMed
8. **Alhuwaydi AM** (2024) Exploring the Role of Artificial Intelligence in Mental Healthcare: Current Trends and Future Directions—A Narrative Review for a Comprehensive Insight. *Risk Management and Healthcare Policy* **17**:1339-1348 <https://doi.org/10.2147/RMHP.S461562> | PubMed
9. **Bauer PJ**, Dikmen SS, Heaton RK, Mungas D, Slotkin J, Beaumont JL (2013) NIH Toolbox Cognition Battery (Cb): Measuring episodic memory. *Monographs of the Society for Research in Child Development* **78**:34-48 <https://doi.org/10.1111/mono.12033>
10. **Bayer JMM**, Thompson PM, Ching CRK, Liu M, Chen A, Panzenhagen AC, Jahanshad N, Marquand A, Schmaal L, Sämann PG (2022) Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. *Frontiers in Neurology* <https://doi.org/10.3389/fneur.2022.923988> | PubMed
11. **Brett M**, Markiewicz CJ, Hanke M, Côté M.-A, Cipollini B, McCarthy P, Jarecka D, Cheng CP, Larson E, Halchenko YO, *et al.* (2024) nipy/nibabel: 5.2.1. Zenodo. <https://doi.org/10.5281/zenodo.10714563>
12. **Carey S**, Pang A, Kamps M. de (2024) Fairness in AI for healthcare. *Future Healthcare Journal* **11**:100177 <https://doi.org/10.1016/j.fhj.2024.100177> | PubMed
13. **Carlozzi NE**, Tulskey DS, Kail RV, Beaumont JL (2013) NIH Toolbox Cognition Battery (Cb): Measuring processing speed. *Monographs of the Society for Research in Child Development* **78**:88-102 <https://doi.org/10.1111/mono.12036>
14. **Cheng B**, Liu M, Zhang D, Munsell BC, Shen D (2015) Domain Transfer Learning for MCI Conversion Prediction. *IEEE Transactions on Bio-Medical Engineering* **62**:1805-1817 <https://doi.org/10.1109/TBME.2015.2404809> | PubMed
15. **Ciric R**, Wolf DH, Power JD, Roalf DR, Baum GL, Ruparel K, Shinohara RT, Elliott MA, Eickhoff SB, Davatzikos C, *et al.* (2017) Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage* **154**:174-187 <https://doi.org/10.1016/j.neuroimage.2017.03.020> | PubMed
16. **Clark DB**, Fisher CB, Bookheimer S, Brown SA, Evans JH, Hopfer C, Hudziak J, Montoya I, Murray M, Pfefferbaum A, *et al.* (2018) Biomedical ethics and clinical oversight in multisite observational neuroimaging studies with children and adolescents: The ABCD experience. *Developmental Cognitive Neuroscience* **32**:143-154 <https://doi.org/10.1016/j.dcn.2017.06.005> | PubMed
17. **Cohen AO**, Dellarco DV, Breiner K, Helion C, Heller AS, Rahdar A, Pedersen G, Chein J, Dyke JP, Galvan A, *et al.* (2016) The Impact of Emotional States on Cognitive Control Circuitry and Function. *Journal of Cognitive Neuroscience* **28**:446-459 https://doi.org/10.1162/jocn_a_00906 | PubMed
18. **Cole MW**, Bassett DS, Power JD, Braver TS, Petersen SE (2014) Intrinsic and Task-Evoked Network Architectures of the Human Brain. *Neuron* **83**:238-251 <https://doi.org/10.1016/j.neuron.2014.05.014> | PubMed

19. Cole MW, Ito T, Cocuzza C, Sanchez-Romero R (2021) The functional relevance of task-state functional connectivity. *The Journal of Neuroscience* **41**:2684-2702 <https://doi.org/10.1523/JNEUROSCI.1713-20.2021> | PubMed
20. Cole MW, Ito T, Schultz D, Mill R, Chen R, Cocuzza C (2019) Task activations produce spurious but systematic inflation of task functional connectivity estimates. *NeuroImage* **189**:1-18 <https://doi.org/10.1016/j.neuroimage.2018.12.054> | PubMed
21. Costabel S, Müller-Petke M (2014) Despiking of magnetic resonance signals in time and wavelet domains. *Near Surface Geophysics* **12**:185-198 <https://doi.org/10.3997/1873-0604.2013027>
22. Cross JL, Choma MA, Onofrey JA (2024) Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health* **3**:e0000651 <https://doi.org/10.1371/journal.pdig.0000651> | PubMed
23. Daumé H (2009) Frustratingly Easy Domain Adaptation. *arXiv* <https://doi.org/10.48550/arxiv.0907.1815>
24. de Mathelin A, Atiq M, Richard G, de la Concha A, Yachouti M, Deheeger F, Mougeot M, Vayatis N (2021) Adapt: Awesome domain adaptation python toolbox. *arXiv* <https://doi.org/10.48550/arxiv.2107.03049>
25. Demidenko N (2024) Eprimetotsv.py. GitHub. https://github.com/demidenm/abcc_datapre/blob/main/scripts/eprimetotsv.py
26. Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* **53**:1-15 <https://doi.org/10.1016/j.neuroimage.2010.06.010> | PubMed
27. Elliott ML, Knodt AR, Cooke M, Kim MJ, Melzer TR, Keenan R, Ireland D, Ramrakha S, Poulton R, Caspi A, et al. (2019) General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *NeuroImage* **189**:516-532 <https://doi.org/10.1016/j.neuroimage.2019.01.068> | PubMed
28. Farahani A, Voghoei S, Rasheed K, Arabnia HR (2021) A Brief Review of Domain Adaptation. In: Stahlbock R, Weiss GM, Abou-Nasr M, Yang C.-Y, Arabnia HR, Deligiannidis L (Eds). *Advances in Data Science and Information Engineering* Springer International Publishing. pp. 877-894 https://doi.org/10.1007/978-3-030-71704-9_65
29. Feczko E, Conan G, Marek S, Tervo-Clemmens B, Cordova M, Doyle O, Earl E, Perrone A, Sturgeon D, Klein R, et al. (2021) Adolescent Brain Cognitive Development (ABCD) Community MRI Collection and Utilities. *bioRxiv* <https://doi.org/10.1101/2021.07.09.451638>
30. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT (2015) Functional connectome fingerprinting: Identifying individuals based on patterns of brain connectivity. *Nature Neuroscience* **18**:1664-1671 <https://doi.org/10.1038/nn.4135> | PubMed
31. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, et al. (2002) Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron* **33**:341-355 [https://doi.org/10.1016/S0896-6273\(02\)00569-X](https://doi.org/10.1016/S0896-6273(02)00569-X)
32. Garavan H, Bartsch H, Conway K, Decastro A, Goldstein RZ, Heeringa S, Jernigan T, Potter A, Thompson W, Zahs D (2018) Recruiting the ABCD sample: Design considerations and procedures. *Developmental Cognitive Neuroscience* **32**:16-22 <https://doi.org/10.1016/j.dcn.2018.04.004> | PubMed
33. Gershon RC, Slotkin J, Manly JJ, Blitz DL, Beaumont JL, Schnipke D, Wallner-Allen K, Golinkoff RM, Gleason JB, Hirsh-Pasek K, et al. (2013) NIH Toolbox Cognition Battery (Cb): Measuring language (vocabulary comprehension and reading decoding). *Monographs of the Society for Research in Child Development* **78**:49-69 <https://doi.org/10.1111/mono.12034>
34. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, et al. (2016) A multi-modal parcellation of human cerebral cortex. *Nature* **536**:171-178 <https://doi.org/10.1038/nature18933> | PubMed

35. Gordon EM, Laumann TO, Adeyemo B, Huckins JF, Kelley WM, Petersen SE (2016) Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cerebral Cortex* **26**:288-303 <https://doi.org/10.1093/cercor/bhu239> | PubMed
36. Gotts SJ, Jo HJ, Reynolds RC, Bandettini PA, Martin A, Cox RW, Saad ZS, Im C.-H (2013) Effective Preprocessing Procedures Virtually Eliminate Distance-Dependent Motion Artifacts in Resting State fMRI. *Journal of Applied Mathematics* **2013** <https://doi.org/10.1155/2013/935154> | PubMed
37. Guan H, Liu M (2022) Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Bio-Medical Engineering* **69**:1173-1185 <https://doi.org/10.1109/TBME.2021.3117407> | PubMed
38. Hagler DJ, Cornejo MD, Makowski C, Dick AS, Sutherland MT, Harms MP, Watts R, Bjork JM, Garavan HP, Hilmer L, et al. (2019) Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage (Orlando, Fla.)* **202**:116091-116091 <https://doi.org/10.1016/j.neuroimage.2019.116091> | PubMed
39. Hagler DJ, Ahmadi ME, Kuperman J, Holland D, McDonald CR, Halgren E, Dale AM (2009) Automated white-matter tractography using a probabilistic diffusion tensor atlas: Application to temporal lobe epilepsy. *Human Brain Mapping* **30**:1535-1547 <https://doi.org/10.1002/hbm.20619> | PubMed
40. Hasanzadeh F, Josephson CB, Waters G, Adedinsowo D, Azizi Z, White JA (2025) Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Npj Digital Medicine* **8**:154 <https://doi.org/10.1038/s41746-025-01503-7> | PubMed
41. Jernigan TL, Brown SA, Dale AM, Tapert S, Sowell ES, Herting M, Laird A, Gonzalez R, Squeglia L, Gray K, et al. (2024) Adolescent Brain Cognitive Development Study (ABCD)—Annual Release 5.1 [Dataset]. Adolescent Brain Cognitive Development (ABCD) Study. <https://doi.org/10.15154/z563-zd24>
42. Jernigan TL, Brown SA, Dowling GJ (2018) The Adolescent Brain Cognitive Development Study. *Journal of Research on Adolescence : The Official Journal of the Society for Research on Adolescence* **28**:154-156 <https://doi.org/10.1111/jora.12374> | PubMed
43. Jørgensen KN, Nerland S, Norbom LB, Doan NT, Nesvåg R, Mørch-Johnsen L, Haukvik UK, Melle I, Andreassen OA, Westlye LT, et al. (2016) Increased MRI-based cortical grey/white-matter contrast in sensory and motor regions in schizophrenia and bipolar disorder. *Psychological Medicine* **46**:1971-1985 <https://doi.org/10.1017/S0033291716000593> | PubMed
44. Katz O, Talmon R, Shaham U (2024) Supervised Domain Adaptation Based on Marginal and Conditional Distributions Alignment. In: Trans. Mach. Learn. Res. **2024**
45. Knutson B, Westdorp A, Kaiser E, Hommer D (2000) fMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. *NeuroImage* **12**:20-27 <https://doi.org/10.1006/nimg.2000.0593>
46. Krishnan A, Williams LJ, McIntosh AR, Abdi H (2011) Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage* **56**:455-475 <https://doi.org/10.1016/j.neuroimage.2010.07.034> | PubMed
47. Kushol R, Frayne R, Graham SJ, Wilman AH, Kalra S, Yang Y.-H (2024) Domain Adaptation of MRI Scanners as an Alternative to MRI Harmonization. In: Koch L, Cardoso MJ, Ferrante E, Kamnitsas K, Islam M, Jiang M, Rieke N, Tsaftaris SA, Yang D (Eds). *Domain Adaptation and Representation Transfer* Springer Nature Switzerland.
48. Lal Khakpoor F, Van Der Vliet W, Deng J, Pat N (2025) Brain predictive models of cognition fail to generalize across ethnicities: Modality-dependent bias in MRI-based prediction. *bioRxiv* <https://doi.org/10.1101/2025.11.12.688133> | PubMed
49. Li J, Bzdok D, Chen J, Tam A, Ooi LQR, Holmes AJ, Ge T, Patil KR, Jabbi M, Eickhoff SB, et al. (2022) Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances* **8**:eabj1812 <https://doi.org/10.1126/sciadv.abj1812> | PubMed
50. Liang Y, Chen C, Tian T, Shu K (2023) Fair classification via domain adaptation: A dual adversarial learning approach. *Frontiers in Big Data* **5**:1049565 <https://doi.org/10.3389/fdata.2022.1049565> | PubMed

51. Liu M, Zhang J, Adeli E, Shen D (2019) Joint Classification and Regression via Deep Multi-Task Multi-Channel Learning for Alzheimer's Disease Diagnosis. *IEEE Transactions on Biomedical Engineering* **66**:1195-1206 <https://doi.org/10.1109/tbme.2018.2869989> | PubMed
52. Logan GD (1994) On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In: Dagenbach D, Carr TH (Eds). *Inhibitory processes in attention, memory, and language* Academic Press. pp. 189-239
53. Luciana M, Bjork JM, Nagel BJ, Barch DM, Gonzalez R, Nixon SJ, Banich MT (2018) Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (ABCD) baseline neurocognition battery. *Developmental Cognitive Neuroscience* **32**:67-79 <https://doi.org/10.1016/j.dcn.2018.02.006> | PubMed
54. Maknojia S, Churchill NW, Schweizer TA, Graham SJ (2019) Resting state fMRI: Going through the motions. *Frontiers in Neuroscience* **13**:825-825 <https://doi.org/10.3389/fnins.2019.00825> | PubMed
55. Mo Z, Siepel A (2023) Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. *PLOS Genetics* **19**:1-22 <https://doi.org/10.1371/journal.pgen.1011032> | PubMed
56. Muschelli J, Nebel MB, Caffo BS, Barber AD, Pekar JJ, Mostofsky SH (2014) Reduction of motion-related artifacts in resting state fMRI using aCompCor. *NeuroImage (Orlando, Fla.)* **96**:22-35 <https://doi.org/10.1016/j.neuroimage.2014.03.028> | PubMed
57. Nilearn contributors, Chamma A, Frau-Pascual A, Rothberg A, Abadie A, Abraham A, Gramfort A, Savio A, Cionca A, Sayal A, et al. (2026) Nilearn [Computer software]. Zenodo. <https://doi.org/10.5281/ZENODO.8397156>
58. Panda R, Kalmady SV, Greiner R (2022) Multi-Source Domain Adaptation Techniques for Mitigating Batch Effects: A Comparative Study. *Frontiers in Neuroinformatics* <https://doi.org/10.3389/fninf.2022.805117> | PubMed
59. Pardoe D, Stone P (2010) Boosting for regression transfer. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. pp. 863-870
60. Phạm DD, McDonald DJ, Ding L, Nebel MB, Mejia AF (2023) Less is more: Balancing noise reduction and data retention in fMRI with data-driven scrubbing. *NeuroImage* **270**:119972-119972 <https://doi.org/10.1016/j.neuroimage.2023.119972> | PubMed
61. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2013) Steps toward optimizing motion artifact removal in functional connectivity MRI; a reply to Carp. *NeuroImage (Orlando, Fla.)* **76**:439-441 <https://doi.org/10.1016/j.neuroimage.2012.03.017> | PubMed
62. Power JD, Lynch CJ, Silver BM, Dubin MJ, Martin A, Jones RM (2019) Distinctions among real and apparent respiratory motions in human fMRI data. *NeuroImage (Orlando, Fla.)* **201**:116041-116041 <https://doi.org/10.1016/j.neuroimage.2019.116041> | PubMed
63. Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE (2014) Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage (Orlando, Fla.)* **84**:320-341 <https://doi.org/10.1016/j.neuroimage.2013.08.048> | PubMed
64. Rajput D, Wang W.-J, Chen C.-C (2023) Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics* **24**:48 <https://doi.org/10.1186/s12859-023-05156-9> | PubMed
65. Sripada C, Angstadt M, Rutherford S, Taxali A, Shedden K (2020) Toward a "treadmill test" for cognition: Improved prediction of general cognitive ability from the task activated brain. *Human Brain Mapping* **41**:3186-3197 <https://doi.org/10.1002/hbm.25007> | PubMed
66. Sripada C, Rutherford S, Angstadt M, Thompson WK, Luciana M, Weigard A, Hyde LH, Heitzeg M (2020) Prediction of neurocognition in youth from resting state fMRI. *Molecular Psychiatry* **25**:3413-3421 <https://doi.org/10.1038/s41380-019-0481-6> | PubMed
67. Stan S, Rostami M (2023) Preserving Fairness in AI under Domain Shift. *arXiv* <https://doi.org/10.48550/ARXIV.2301.12369>

68. Sui J, Jiang R, Bustillo J, Calhoun V (2020) Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biological Psychiatry* (1969) 818-828 <https://doi.org/10.1016/j.biopsych.2020.02.016> | PubMed
69. Susan S, Kumar A (2021) The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. *Engineering Reports* 3:e12298 <https://doi.org/10.1002/eng2.12298>
70. Tetereva A, Knodt A, Melzer T, Van Der Vliet W, Gibson B, Hariri A, Whitman ET, Li J, Khakpoor Farzane Lal, Deng J, et al. (2025) Improving Predictability, Test-Retest Reliability and Generalisability of Brain-Wide Associations for Cognitive Abilities via Multimodal Stacking. *bioRxiv* <https://doi.org/10.1101/2024.05.03.589404> | PubMed
71. Tetereva A, Li J, Deng JD, Stringaris A, Pat N (2022) Capturing brain-cognition relationship: Integrating task-based fMRI across tasks markedly boosts prediction and test-retest reliability. *NeuroImage* <https://doi.org/10.1016/j.neuroimage.2022.119588>
72. Tulsky DS, Carlozzi NE, Chevalier N, Espy KA, Beaumont JL, Mungas D (2013) NIH Toolbox Cognition Battery (Cb): Measuring working memory. *Monographs of the Society for Research in Child Development* 78:70-87 <https://doi.org/10.1111/mono.12035>
73. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, Matsui Y, Nozaki T, Nakaura T, Fujima N, et al. (2024) Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology* 42:3-15 <https://doi.org/10.1007/s11604-023-01474-3> | PubMed
74. Vieira BH, Pamplona GSP, Fachinello K, Silva AK, Foss MP, Salmon CEG (2022) On the prediction of human intelligence from neuroimaging: A systematic review of methods and reporting. *Intelligence* 93:101654 <https://doi.org/10.1016/j.intell.2022.101654>
75. Wang R, Chaudhari P, Davatzikos C (2022) Embracing the disharmony in medical imaging: A Simple and effective framework for domain adaptation. *Medical Image Analysis* 76:102309-102309 <https://doi.org/10.1016/j.media.2021.102309> | PubMed
76. Wu P, Dietterich TG (2004) Improving SVM accuracy by training on auxiliary data sources. In: Proceedings of the Twenty-First International Conference on Machine Learning. pp. 110 <https://doi.org/10.1145/1015330.1015436>
77. Xu X, Jang I, Zhang J, Zhang M, Wang L, Ye G, Zhao A, Zhang Y, Li B, Liu J, et al. (2024) Cortical gray to white matter signal intensity ratio as a sign of neurodegeneration and cognition independent of β -amyloid in dementia. *Human Brain Mapping* 45:e26532 <https://doi.org/10.1002/hbm.26532> | PubMed
78. Zelazo PD, Anderson JE, Richler J, Wallner-Allen K, Beaumont JL, Weintraub S (2013) NIH Toolbox Cognition Battery (Cb): Measuring executive function and attention. *Monographs of the Society for Research in Child Development* 78:16-33 <https://doi.org/10.1111/mono.12032>
79. Zhang Y, Chen Y, Yu H, Lv Z, Li Q, Yang X (2020) Bridging Cross-Tasks Gap for Cognitive Assessment via Fine-Grained Domain Adaptation. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence: Special track on AI for CompSust and Human well-being. pp. 4330-4337 <https://doi.org/10.24963/ijcai.2020/597>
80. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC (2021) Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45:4396-4415 <https://doi.org/10.1109/tpami.2022.3195549> | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

The present report describes an investigation into the use of machine learning techniques to improve cross-racial/ethnic performance of brain models of cognitive function. The authors

tested several approaches to boost prediction of NIH cognitive toolbox scores using brain imaging data (function, structure) for minoritized (Black) participants in the ABCD Study sample compared to white (majority) participants. Structural (e.g., volume) measures showed the greatest performance gap, and a balanced weighting method showed the greatest performance gain across features. The authors conclude that supervised domain adaptive methods can improve models for cognitive prediction and mitigate cross-racial/ethnic performance disparities.

Strengths:

This investigation makes some headway into issues by identifying computational methods that may help to improve some models for limited outcome variables (i.e., general cognitive performance). Addressing racial/ethnic disparities in brain imaging research has significant implications for generalizability of findings and for the practical utility of imaging findings in the wider population. A comparative approach to evaluate the improvements in a "prediction gap" across various methods could have benefits for neuroimaging beyond racial/ethnic disparities. The use of the ABCD Study, given its deep phenotyping of individuals, is also a benefit.

Weaknesses:

Despite its strengths, there are several large conceptual and related methodological issues that impact its conclusions and the overall utility of the approach. The sample selection approach limits insight into likely drivers of the performance gap (e.g., socioenvironmental disparities known to exist between groups and associated with neurodevelopment), and in so doing ignores a critical component of understanding racial brain differences, particularly in relation to cognitive functions. Further, while the relative gaps in performance of a single cognitive score across features are well described, the actual performance (and therefore relative benefit to these techniques) is unclear. Specific examples include the following.

(1) The overarching conceptual issue with the manuscript is a lack of engagement with a substantial and growing evidence base on the drivers of racial disparities in brain imaging which impact model performance. Racial/ethnic groups in the US (and other regions of the world) are not equivalent in terms of developmental environments that shape brain function and structure (see Harnett et al., 2023, *Neuropsychopharmacology*; Ricard et al., 2023, *Nature Neuroscience*; Cardenas-Iniguez & Gonzalez, 2024, *Nature Neuroscience* for some overview here). The socioenvironmental disparities inherent to race in the US further shape cognitive development and brain associations with cognitive performance (e.g., Marek et al., 2025, *Science*). The framing of the manuscript focuses almost exclusively on broad sampling issues, and in doing so treats racial/ethnic variability as if it reflects statistical abnormality rather than a critical component of understanding human brain development. This lack of contextualizing racial disparities significantly impacts the overall utility of the proposed approach and the conclusions of the manuscript.

(2) In relation to the above, another conceptual issue in this approach of using a majority to inform minority brain associations with cognitive variables is an assumption that minority brain patterns should match the majority, rather than developmental stressors inducing alternative brain-weighting to predict outcomes. This framework does not assess this possibility and may in fact obscure such an outcome, limiting our inferences into neurodevelopment.

(3) Another conceptual/methodological issue here is the use of "matched groups" for analysis. The specifics of matching are fairly vague, but given the description one would assume the w/B groups are matched on a number of behavioral/socioenvironmental variables, which is a significant issue for interpretability and applicability. As noted, w/B groups in the US (and the ABCD Study) differ substantially across variables; matching has the likely consequence of

creating a highly non-generalizable sample, particularly when the minority group is restricted to $N = 10$.

<https://doi.org/10.7554/eLife.112354.1.sa2>

Reviewer #2 (Public review):

In the manuscript "Supervised domain adaptation mitigates cross-ethnicity prediction errors in neuroimaging-based cognitive prediction", the authors investigated the efficacy of data adaptation techniques to reduce ethnicity-related prediction bias in neuroimaging-based cognitive prediction. They found that data adaptation algorithms, particularly balanced weighting, contributed to mitigating ethnicity-related performance disparities. Furthermore, these bias mitigations could be achieved without requiring a large set of data from the underrepresented ethnic group. This study addressed an important concern in the field of neuroimaging-based behaviour prediction, providing many intriguing results. Nevertheless, the manuscript also suffers from a lack of coherent methods design, the unorganised presentation of information, and the lack of in-depth discussion of results.

The conclusions claimed by the authors are sometimes over-generalised and not fully supported by the study outcomes. Overall, this study demonstrated strong technical designs and convincing statistical analysis for the main outcomes, although clearer presentation would be needed to convey the messages in the manuscript.

The central investigation of this study is whether domain adaptation techniques improve ethnicity-related performance disparities. However, these improvements were only measured against a very weak baseline model, where a small set of African American (AA) subjects were added to the training sample consisting purely of White American (WA) subjects. While the authors recognised that balancing the training sample could already mitigate the ethnicity-related disparities, they considered that such approaches are unfeasible in their experimental scenario, where only a small amount of AA data were available. However, as Li et al. (2022) showed, a balanced sample of around 90-150 AA subjects could already reduce the ethnicity-related bias. Even from a practical standpoint, this balanced sample approach would be a more valid baseline for domain adaptation models to compare against.

The authors made two main conclusions: that domain adaptation methods reduced ethnicity-related bias, and that balanced weighting performed the best and the most stably. Both claims were over-generalised to some extent. First, the adaptation benefit claimed in the first conclusion is not seen in the functional connectivity (FC) modality, which is the most popular modality for neuroimaging-based prediction of behaviour. This difference in adaptation benefit across modalities is an important finding that is meaningful for future studies, the omission of which also removes interesting insights that the audience could take away from this article.

Second, the judgement of prediction performance is based on the area under the improvement curve (AUIC) metric, which summarises a model's performance across different availability of labelled AA data. As a result, the analysis of prediction performance naturally favours algorithms that could perform well with a small amount of added AA data. On the one hand, this provides an easy decision point for users to pick an algorithm to use without being concerned about data availability. On the other hand, important insights could be overlooked with the oversimplified recommendation of balanced weighting. As the authors have also observed, in some cases, domain adaptation strategies do not improve ethnicity-related bias more than the non-adaptation baseline. If the message is to recommend simple, low-cost strategies to reduce ethnicity-related prediction bias, it would be misleading not to note that the simplest and lowest-cost strategy could also be non-adaptation methods sometimes.

Regardless, for the general audience, the underlying assumptions when interpreting the AUIC metric are not immediately clear, which could cause the conclusions to be misleading. Apart from aggregating over different amounts of available AA data, the statistical comparison of AUIC gain across data adaptation algorithms also did not account for the impact of brain phenotype modalities. Even though the upstream analyses have confirmed that adaptation benefits vary greatly across brain modalities, this major observation was not followed in the final analysis where conclusions were made about which algorithm performed the best. Based on visual inspection of Figure 3b, it may be suspected that PRED performed better than or comparably to balanced weighting when task contrasts based on the Destrieux atlas were used.

Finally, the findings from this study align with the common hypothesis that ethnicity-related prediction bias originates from disparities already manifested during data collection and preprocessing. As the authors have noted, the modalities with the most tendency for ethnicity-related bias are the anatomical ones, including all three volume-based modalities (cortical volume, T1 and T2 subcortical volume) in the top ten phenotypes with the largest performance gap. Most prominently, brain features in the occipital pole, frontal pole, and a range of subcortical areas were found to contribute highly to adaptation gain. Subcortical areas are often reported to show noisier measurements compared to cortical areas, whereas the poles of the brain are likely more strongly warped/distorted during alignment to a standard template. From a data quality perspective, these results support the interpretation that ethnicity-related prediction bias may stem from loss of data quality during data collection or preprocessing. In the prediction models based on anatomical brain features, data adaptation methods may have helped to address these disparities in the data, without the more resource-intensive need to improve the bias in preprocessing pipelines.

Li, J., Bzdok, D., Chen, J., ... Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, 8(11), eabj1812.

<https://doi.org/10.7554/eLife.112354.1.sa1>

Reviewer #3 (Public review):

The manuscript frames its work in fairness and disparities but does not show or directly test that its approach decreases differences between White and African Americans. While it is stated that the objective is not to equalize performance across groups, large parts of the paper repeatedly claim that the methods mitigate cross-ethnicity disparities and improve fairness. Improving prediction in African American participants relative to a non-adapted model is not necessarily the same as reducing the disparity between African American and White American participants. The adapted model should be evaluated in both groups, and the post-adaptation performance gap should be reported directly.

Additional prediction performance measures are needed. For example, in Li et al, different results and conclusions are made with MSE and the correlation between observed and predicted variables. In that paper particularly, aggression measures showed better correlation in African Americans but better MSE in White Americans. Such differences are important to note as they likely suggest different mechanisms.

Similarly, characteristics of the cognitive outcome need to be understood. For example, differences in MSE or MAE may reflect a difference in variance between the groups. The group with a larger variance will have a larger MSE. Correlation or other performance measures that are invariant to different mean or variance shifts can be helpful here.

While the authors note that for the paper they treat racial and ethnic backgrounds interchangeably, I do not think that is the best given the differences between them and the impact and history they have in American culture. Overall, the authors likely need to do a better job conceptualizing their results in the history of minoritized populations in the United States. It is immensely important not to treat them as biological domains without considerable qualification and to avoid language implying that observed domain differences are intrinsic properties of racial groups.

Changes in feature weight are not a proper way to identify the mechanisms of improved performance. At most, these analyses characterize how model coefficients change when target-group data are incorporated or upweighted.

The cross-validation strategy is suboptimal. First, the use of the matched splits of the ABCD data introduces data leakage. To match a validation set to the training set in such a manner requires that each split knows about the other split's characteristics. That is data leakage. Though the impact could be small. Second, African American breakdowns are not balanced across sites and scanners. Domain adaption methods may be learning a shortcut or proxy for African American like site, scanner, or something else. A likely better approach would be some sort of leave X sites out approach, where a model is trained on White Americans from a set of sites, adapted with African Americans from those sites, and applied (with and without adaptation) to the White and African Americans from the left-out sites.

Baseline models for comparisons to the domain adaptation are missing. Some simpler ones include a target-only model trained on the same 10-100 African American participants and a pooled model with a group indicator and group-by-feature interaction. Without these comparisons, it is difficult to know whether balanced weighting is learning target-specific neurobiological information or merely recalibrating the prediction distribution.

There are a few statistical issues:

- (1) The repeated MAE estimates are therefore not independent observations. Paired t-tests cannot be applied across repetitions. Subject-level bootstrap or permutation procedures that repeat the complete training and testing process are needed
- (2) The caption describes approximate 95% confidence intervals as $\{\text{plus minus}\}1.96 \times \text{SD}/n$. Conventionally, the standard error would involve $\text{SD}/\sqrt{n}$. However, even if corrected,

there would still be issues about the dependence among the overlapping resamples.

(3) Ten repetitions are likely insufficient, especially in the case of ten target participants. Results at $n = 10$ may be extremely sensitive to which children are selected. The authors should use substantially more repetitions and report the full distribution of results.

(4) The Friedman and Wilcoxon comparisons treat the 80 imaging phenotypes as the observational units. These phenotypes are highly dependent because they are derived from the same participants, many use overlapping images, and numerous task contrasts and structural measures are strongly correlated. This non-independence can make the comparison among adaptation methods look much more precise than it is. A hierarchical analysis by modality or a resampling strategy that preserves dependence among phenotypes would be more appropriate.

(5) The gap metric and AUIC are difficult to interpret. Gap is the absolute relative difference between target-group and source-group MAE, normalized by source-group MAE. It is sensitive to the denominator and may produce large values whenever source-group MAE is relatively small. Reporting signed raw MAE differences and MAE ratios alongside this derived score would help. Similarly, the AUIC combines errors with an arbitrary sequence of target-sample sizes. Statistically significant differences in AUIC do not necessarily indicate practically meaningful differences among methods.

(6) The correlation between baseline gap and adaptation gain is partly tautological. Those with the widest gaps have the most room for improvement and likely thus show the greatest improvement. While still of value, the authors may want to tone down their interpretation of the correlation and describe its limitation.

(7) Given that the sample sizes vary from approximately 4,000 to more than 11,000 depending on modality, the authors may want to consider a reduced sample matched in size across modalities. It is hard to fully know if the conclusion that connectivity is more robust given the wide-scale differences in sample size and feature dimensionality.

(8) PLS are sensitive to many factors like scaling and collinearity. Many recent papers have been written about their limitations when used for subtyping. Some of these hold for prediction too. I think showing the results are consistent with different prediction algorithms is needed. SVR and ridge regression are two common methods for regression prediction with neuroimaging data.

(9) The feature interpretation is partly circular. The method with the largest performance gain is selected, and its coefficient changes are then used to explain that gain. A method designed to give target observations greater influence will unsurprisingly change its coefficients more than naïve inclusion.

(10) The practical and ethical deployment scenario is underspecified. Supervised adaptation requires labelled cognitive outcomes from the target population and, as currently framed, may require choosing a model based on an individual's racial category. What are the implications of deploying race-specific models that need to be considered? It is not self-evident that this approach is preferable to developing a broadly representative model or directly modeling the social and technical sources of distribution shift.

(11) The paper is worded and interpreted much too strongly. The current study supports the conclusion that, within ABCD, giving a small labelled target-group sample greater influence can sometimes improve held-out target-group MAE relative to naïvely adding the same participants. It does not yet establish that the method improves fairness or identifies mechanisms of racial bias. Likewise, the abstract and conclusion overstate the results. The abstract states that all adaptation methods reduced target-group prediction error, while the Results show near-zero or negative benefits for several functional-connectivity phenotypes

and instability of PRED and interpolation below 30 target participants. Similarly, "substantially reduce disparities," "improve equity," "consistently," and "practical path forward" are stronger than the analyses support. Finally, the limitations section is incomplete and omits the more consequential limitations.

(11) That only ten labelled participants are needed to change the results is troubling. This is a shockingly low number. Giving ten target observations disproportionate influence can move the fitted model, particularly when the balanced-weighting ratio is high. A measurable MAE change is therefore possible, but it may reflect a shift in intercept or slope rather than learning a stable target-group brain-cognition relationship. Further, the manuscript does not report the numerical improvement for the $n = 10$ condition in the text or a table. Visual inspection of Figure 4 suggests reductions of roughly 0.10-0.25 standardized MAE units for some high-gap structural phenotypes, approximately 10-20%, while low-gap connectivity phenotypes show little or no gain.

(12) The study lacks genuine external validation, which may be needed to fully convince readers that such a low number of subjects is needed to reduce biases.

<https://doi.org/10.7554/eLife.112354.1.sa0>