When Brain Models Aren't Universal: Benchmarking of Ethnic Bias in MRI-Based Cognitive Prediction Across Modalities

Farzane Lal Khakpoor,[1] William van der Vliet,[1] Jeremiah Deng,[2] Narun Pat[1]

[1]Department of Psychology, University of Otago, Dunedin, New Zealand

[2]School of Computing, University of Otago, Dunedin, New Zealand

Corresponding Author:

Farzane Lal Khakpoor

Department of Psychology, University of Otago

William James Building, 275 Leith Walk, Dunedin 9016, New Zealand

Email: farzanelk@gmail.com

Abstract

Predictive neuroimaging models promise precision medicine but risk exacerbating health inequities if they perform unevenly across ethnic/racial groups. Using the Adolescent Brain Cognitive Development data, we benchmarked ethnic/racial bias in models predicting cognitive functioning from 81 MRI phenotypes across four training strategies. Models trained on one ethnicity performed best within that group. Models trained on participants sampled without regard to ethnicity, a common practice, performed better on White participants, likely because the ABCD sample was predominantly White. Training on equal-sized White and

African American subsamples reduced disparities without accuracy loss. Structural MRI exhibited the greatest bias, whereas task-based fMRI phenotypes were more equitable. Stronger brain–cognition associations generalized more equitably, but multimodal stacking—despite enhancing prediction—did not improve fairness. Increasing representation of African American participants improved performance up to balanced sampling, with diminishing returns beyond. This first modality-wide benchmark reveals pervasive, modality-dependent ethnic bias in cognitive prediction and identifies key factors shaping equity in neuroimaging models.

Introduction

Using machine learning in predictive modelling enables neuroscientists to move beyond associating individual brain features to a single variable of interest, such as cognitive functioning, and instead predict that variable for a new, unseen participant/patient based on a large set of brain features. This approach has accelerated biomarker discovery and paved the way for precision medicine (Sui et al., 2020) and could eventually support earlier detection of risk and targeted intervention (Alhuwaydi, 2024). For instance, predictive modelling has transformed how we study individual differences in cognitive functioning (Finn et al., 2015; Sripada, Angstadt, et al., 2020; Sripada, Rutherford, et al., 2020; Vieira et al., 2022), as a key psychiatric phenotype that predicts daily functioning and treatment outcomes (Abramovitch et al., 2021). However, a critical translational challenge remains: ensuring these models perform similarly across diverse populations.

Ethnic/racial bias in predictive models is well-documented across multiple biomedical domains (Carey et al., 2024; Cross et al., 2024; Hasanzadeh et al., 2025; Ueda et al., 2024). In genomics, for instance, polygenic risk scores trained predominantly on European-ancestry cohorts often exhibit reduced accuracy in individuals from other ancestries. This illustrates how the composition of a training cohort, referred to as the discovery group in genomics, can systematically disadvantage underrepresented populations (Duncan et al., 2019; Kachuri et al., 2024; Y. Wang et al., 2022). Such disparities in predictive performance raise concerns that these tools may inadvertently exacerbate existing health inequities (Lewis & Green, 2021; Martin et al., 2019).

Neuroimaging could face similar challenges. Large-scale imaging datasets are disproportionately composed of White participants from North America and Europe (Horien

et al., 2021), and models trained on such data may learn associations that are less stable or less relevant in underrepresented populations. Recent work has begun to quantify this risk. For example, Li and Colleagues (2022) reported that resting-state functional connectivity models trained on majority White American samples from the Adolescent Brain Cognitive Development (ABCD) study generalized poorly to African American participants, typically favouring White Americans in predicting a range of cognitive and behavioral measures. Another recent study found that brain age algorithms based on structural data exhibit reduced accuracy for African American participants (Adkins & Hanson, 2025). These findings underscored that ethnic/racial biases are not confined to genomics, but are also relevant to neuroimaging-based prediction.

However, several open questions remain. First, the scale and scope of ethnic/racial bias in predictive neuroimaging are not well characterized. It is unclear whether poor cross-ethnicity generalizability is confined to specific MRI-derived phenotypes or extends across the full spectrum of neuroimaging types. Different MRI-based representations—ranging from task-related contrasts to connectivity and structural phenotypes—may vary in their bias and susceptibility to training sample composition, such as models trained on a single ethnicity, balanced datasets including equal samples of each ethnicity, or imbalanced majority–minority combinations (Li et al., 2022). A systematic benchmarking across neuroimaging phenotypes is therefore needed to determine which representations are more or less sensitive to training composition and which yield more equitable predictions.

Second, the factors influencing such disparities remain poorly understood. Given that neuroimaging phenotypes differ in their predictive utility (Pat et al., 2023; Tetereva et al., 2022, 2025), It is important to determine whether neuroimaging phenotypes that capture stronger brain–behavior relationships also generalize more equitably across populations, or

whether consistent performance gaps between ethnic/racial groups persist regardless of overall predictive accuracy.

Third, potential strategies for mitigating these disparities remain largely untested. Multimodal "stacking" approaches (Engemann et al., 2020; Tetereva et al., 2022) have been shown to reliably enhance predictive accuracy for traits such as cognitive functioning by integrating information across multiple neuroimaging phenotypes (Marek et al., 2022; Rasero et al., 2021; Rosenberg et al., 2018; Tetereva et al., 2022, 2025). However, it remains unclear whether such integration mitigates ethnic/racial bias. Furthermore, the extent of representation required from underrepresented groups to minimize these disparities is unknown, as is whether synthetically changing the number of training data from underrepresented groups through oversampling yields additional improvements.

Here, we address these gaps by systematically benchmarking cross-ethnicity generalizability across a broad array of MRI-derived neuroimaging phenotypes in predicting cognitive functioning. We trained predictive models of cognitive functioning on 81 unimodal and 11 multimodal neuroimaging phenotypes. For each neuroimaging phenotype, we compared four training strategies: models trained on (1) All available data (majority White American), (2) Randomly Selected White Americans only (RandWA-only), matched in size to the available African American sample, (3) African Americans only (AA-only), and (4) Balanced African American + Randomly Selected White American subsamples (Balanced AA+RandWA). To assess fairness, we quantified ethnicity-specific prediction errors by calculating the mean absolute error (MAE) separately for White American (WA) and African American (AA) participants in the test sets. A higher MAE in a test group indicated poorer prediction performance for that group. This design allowed us to directly compare performance disparities between groups across models and feature sets. We derived an Ethnicity Bias

Index based on the difference in MAE between WA and AA participants for models trained on RandWA-only and AA-only data, and tested whether neuroimaging phenotypes with higher predictive power exhibited reduced bias. Finally, we extended the analysis to multimodal stacked models to evaluate whether integration of multiple feature types improves both accuracy and fairness. In addition to benchmarking existing training strategies, we also explored how incremental increases in AA representation influence prediction accuracy and whether synthetically increasing the number of AA participants through oversampling helps.

Our overarching aim was to provide the first systematic, modality-wide evaluation of ethnicity-related bias in predictive neuroimaging models of cognitive functioning and to test whether boosting predictive accuracy—through feature selection, multimodal integration, or synthetic oversampling—reduces or perpetuates disparities.

Results

To evaluate ethnicity-related bias, we analyzed data from the Adolescent Brain Cognitive Development ABCD-BIDS Community Collection (ABCC) (ABCD-BIDS Community), focusing on White and African American participants in groups matched across demographic and socioeconomic factors (Feczko et al., 2021) (Figure 1A). Cognitive functioning was measured using the NIH Toolbox total score (Bauer & Zelazo, 2013; Jernigan et al., 2024). We derived 81 unimodal neuroimaging phenotypes from structural MRI (sMRI), diffusion tensor imaging (DTI), functional MRI (fMRI) task contrasts, and functional connectivity (FC), plus 11 multimodal stacked combinations. Models were trained under four strategies: All (majority WA), RandWA-only, AA-only, and Balanced AA+RandWA (Figure 1B). Performance was assessed via mean absolute error (MAE; lower values indicate better

predictions), with disparities quantified using an Ethnicity Bias Index (difference in MAE gaps between RandWA-only and AA-only models; lower absolute values indicate less bias). We also tested the effect of incremental and oversampling of AA on model predictions (Figure 1C). Full details are in the Methods and Supplementary Information.

Prediction Error Across Modalities and Training Strategies

sMRI and DTI

For sMRI features, the RandWA-only models performed better on WA test participants, as indicated by lower MAE compared to AA test participants. The All models, trained on the full dataset with a majority of WA participants, achieved overall higher accuracy than the RandWA-only model but continued to perform better for WA than AA participants. The AA-only models showed the opposite trend, performing better for AA than WA participants. The Balanced models had better accuracy relative to the ethnicity-specific models and performed on par to the All model, but still yielded significantly lower MAE for WA than for AA participants (Figure 2A).

Functional Connectivity (FC)

For FC features, RandWA-only models predicted WA participants more accurately, while AA-only models predicted AA participants more accurately. For task-related functional connectivity (Nback, SST, Multitask and General FC), the All and Balanced models achieved similar MAE across WA and AA test groups with no significant difference revealed. Their prediction errors for AA were on par or lower than their AA-only counterparts. Among resting-state phenotypes, this pattern of similar performance across AA and WA participants only held for the balanced model of rest FC Figure 2B).

Task-based fMRI contrasts

For task-based contrasts, the RandWA-only models performed better on WA test participants, as indicated by lower mean absolute errors (MAE) compared to AA participants. The All model reduced overall errors but continued to perform better for WA than AA participants. In contrast, the AA-only model performed better for AA participants. For contrast features parcellated with the Glasser atlas, the Balanced model achieved comparable MAEs across WA and AA groups, showing no significant performance differences for most task contrasts (26 out of 30) (Figure 2C). However, when using the Destrieux atlas, this pattern of comparable performance was observed in only seven out of 26 contrasts (Figure 2D).

Multimodal models

Multimodal stacked models that combined multiple neuroimaging phenotypes showed a similar pattern of ethnicity-specific models having lower error (better predictions) in their own group. Using the All model, MAE remained lower for WA than AA participants (Figure 3). The Balanced model showed the similar pattern: only five multimodal models which combined predictions from task-specific neuroimaging phenotypes showed non-significant differences between AA and WA.

Ethnicity Bias Index

The Ethnicity Bias Index was derived for each of the 81 unimodal and 11 multimodal neuroimaging phenotypes by subtracting the group-specific MAE differences observed in RandWA-only and AA-only models. Smaller absolute Ethnicity Bias Index values indicated less ethnicity-related bias. As shown in Figure 4A, ranking features by Ethnicity Bias Index showed that functional measures, particularly Nback 2-0back contrast and task and rest functional connectivity were among the least biased. Structural MRI features consistently

showed the largest Ethnicity Bias Index values, indicating greater disparities. A similar pattern was seen when ranking multimodal neuroimaging phenotypes: models combining predictions from Nback-related neuroimaging phenotypes and different functional connectivity showed lowest bias and combinations of different sMRI showed the highest. The least biased multimodal feature set had a higher bias score compared to its unimodal counterpart. The most biased multimodal model had a smaller bias score than its unimodal counterpart (Figure 4B).

Relation Between Performance and Bias

Across unimodal neuroimaging phenotypes, lower predictive error—calcultaed by applying the All model on the whole test set, was associated with lower Ethnicity Bias Index. When comparing All model performance (MAE) on the whole test set with corresponding Ethnicity Bias Index values, the correlation was 0.62, indicating that features yielding stronger predictions tended to also yield smaller disparities between AA and WA (Figure 4C).

Multimodal stacked models, integrating predictions across multiple neuroimaging phenotypes, achieved lower predictive error than unimodal models. However, bias values in multimodal models were not improved beyond unimodal neuroimaging phenotypes and rather were in the mid-range (Figure 4C). The association between All model performance (MAE) and Ethnicity Bias Index was weaker, with a correlation of 0.3.

Effects of Incremental AA Representation and Synthetical Oversampling

The incremental sampling analysis revealed that increasing the proportion of AA participants in the training set reduced MAE for AA test participants up to approximately the 50% balance point. Beyond this threshold, additional AA representation through synthetical

oversampling did not further improve AA performance; in some cases, performance plateaued or worsened. For sMRI features, a slight decrease in AA test prediction error was observed with additional AA sampling, but this was accompanied by increased error for WA test participants. For task-based and functional connectivity features, increasing AA representation beyond 50% did not yield further benefits and in some cases increased error (Figure 5).

Discussion

This study provides a systematic benchmark of ethnicity-related generalization in MRI-based predictive modelling of cognition. Several key findings emerged:

MRI Neuroimaging phenotypes differ in sensitivity to training composition and generalizability

For all MRI modalities, models trained on a single ethnicity performed best within that group. The All models—trained on participants sampled without regard to ethnicity, yielding predominantly White samples—performed better on White participants, consistent with prior findings for resting-state fMRI (Li et al., 2022). This pattern indicates that the brain–cognition associations learned by these models are influenced by the racial/ethnic composition of the training sample.

Balanced AA+RandWA training showed promise in reducing performance disparities between AA and WA test participants, notably without compromising performance for WA participants—a common limitation of fairness optimization methods (Chen et al., 2023; Pfohl et al., 2021). However, with structural MRI phenotypes, prediction accuracy for AA participants still lagged behind that of WA participants. In contrast, task-based fMRI

phenotypes, such as Nback 2–0back and 2back contrasts, were more equitable, with balanced training often eliminating group differences. Task functional connectivity was least sensitive to the ethnic composition of the training set, as both the Balanced and All (majority-White) models performed similarly for AA test participants. Overall, these results highlight that fairness in predictive performance is strongly shaped by both imaging modality and the composition of the training data.

The Ethnicity Bias Index provided a systematic way to rank MRI neuroimaging phenotypes. Task-based contrasts and functional connectivity consistently fell among the least biased neuroimaging phenotypes, while sMRI measures were the most biased. Importantly, within a modality, neuroimaging phenotypes varied: for example, Nback contrasts (0back, 2back and 2-0back) which manipulated working memory load were less biased than other task contrasts.

Predictive strength is linked to lower bias

Neuroimaging phenotypes with higher predictive accuracy tended to show lower bias, with a moderate correlation between MAE and the Ethnicity Bias Index. This suggests that, although ethnic/racial bias was not fully eliminated, robust brain–behavior associations generalized more fairly across populations and at least some disparities reflect model fragility rather than intrinsic group differences.

Predictive gains do not mitigate ethnic/racial bias

The stacked-all model, which combined predictions from all neuroimaging phenotypes, achieved higher predictive performance than any unimodal model when trained on all participants. More broadly, most multimodal models outperformed unimodal models in terms of predictive accuracy across the test set. However, these accuracy gains did not translate into greater fairness. Ethnicity-specific multimodal models favoured their own group, while All

and Balanced multimodal models favoured WA participants. Across feature combinations, multimodal models showed intermediate Ethnicity Bias Index values, indicating that their fairness largely depended on the properties of the constituent features. Thus, while multimodal stacking boosted predictive performance, it did not resolve disparities, paralleling lessons from genomics that gains in predictive power do not guarantee equitable outcomes (Xiang et al., 2024).

Incremental AA Representation Yields Diminishing Returns Beyond Balance

The incremental sampling analysis provides additional insight into the relationship between dataset composition and bias. As expected, increasing AA representation improved performance for AA test participants up to a balanced training distribution. However, adding AA participants beyond this balance point—via repeated sampling—did not produce further gains. For sMRI, oversampling improved AA predictions slightly but reduced WA performance, suggesting models may be learning group-specific structural patterns in opposing directions. These findings suggest diminishing returns to oversampling beyond balanced representation as a potential solution.

Factors Contributing to the Differences in Prediction Performance

These results extend prior literature (Li et al., 2022) by showing that ethnicity-related bias is modality-dependent. A consistent trend was that neuroimaging phenotypes with higher predictive strength also tended to exhibit lower bias, suggesting that robust brain–behavior associations may generalize more equitably across populations, whereas weaker predictors may be more fragile to sample composition. This raises the possibility that methodological improvements at the levels of data collection, preprocessing, and algorithm design could simultaneously enhance both accuracy and fairness. However, our findings also demonstrate

that fairness does not automatically follow from improved accuracy—multimodal models increased predictive performance but left disparities largely intact.

The observed modality differences—task-based features being less sensitive to training composition, while structural features remained more vulnerable—point toward additional methodological sources of bias. Widely used volumetric and surface templates (e.g., MNI152, fsaverage, fs_LR) were constructed primarily from White/European-descendant samples. Applying these templates to non-European populations may increase deformation or segmentation errors, amplifying disparities in sMRI-based prediction. Prior work has shown that population-specific templates can improve segmentation accuracy and reduce deformations in Chinese cohorts (Jao et al., 2009; Liang et al., 2015; G. Yang et al., 2020), yet comparable African-specific templates have not been developed. In contrast, task contrasts and task-regressed functional connectivity reduce reliance on structural alignment by emphasizing condition-specific within-subject differences or residual fluctuations in the time series, thereby attenuating biases introduced by structural templates or atlas boundaries. Further evidence comes from our comparison of parcellation schemes: Glasser parcellation (Glasser et al., 2016) yielded both stronger predictions and markedly lower bias compared to Destrieux (Destrieux et al., 2010). Although both atlases were derived from a predominantly White HCP sample, their design differs in ways that may influence cross-ethnic generalizability. The Glasser atlas integrates multiple modalities, including myelin maps, task-related response, and connectivity, to create functionally homogeneous parcels (Glasser et al., 2016; C. Wang et al., 2018). This multimodal basis may make it less sensitive to ancestry-related morphological variation. In contrast, the Destrieux atlas is based primarily on cortical folding patterns such as sulci and gyri (Destrieux et al., 2010), which are known to vary across populations (Chee et al., 2011; Tang et al., 2018; G. Yang et al., 2023). This

reliance on anatomical landmarks may explain why Destrieux features exhibited larger bias despite only modest differences in predictive performance. Further supporting this view: for Glasser-based contrast and functional connectivity neuroimaging phenotypes, a balanced training sample largely eliminated performance differences between AA and WA test groups. This suggests that, although recent work has documented population differences in functional network topography (Yang et al., 2023), such variability may play a smaller role in cognitive prediction than structural factors embedded in preprocessing and atlas choice.

Together, these observations suggest that modality-dependent disparities reflect both the predictive robustness of features and the ways in which preprocessing pipelines embed structural assumptions.

Limitations

Our analyses were limited to WA and AA participants in the ABCD dataset and may not generalize to other underrepresented groups. Also, we relied on widely-used preprocessing pipelines and standard templates—such as the MNI152 brain template—which may embed structural biases that disproportionately affect sMRI-based prediction. The MNI152 template was constructed from 152 healthy young adult MRI scans, primarily from Western populations (Montreal Neurological Institute and Hospital, 2009). The Human Connectome Project (HCP) database—which many surface/atlas pipelines draw from—reports that ~74.7% of its full sample identified as White/Caucasian (Glasser et al., 2016; Owens et al., 2019).

Recommendations for Future Use of Predictive Models for Neuroimaging

From a translational perspective, these findings underscore the risks of deploying predictive neuroimaging models without explicit attention to data composition. Unequal accuracy across

groups could reinforce existing health disparities—a concern long recognized in genomics, where polygenic scores trained in European ancestry cohorts show limited transferability to other populations (Duncan et al., 2019; Martin et al., 2019; Peterson et al., 2019). Similar principles apply to neuroimaging: expanding dataset diversity and adopting balanced training designs are essential for equitable generalization, particularly since balanced training sample as a bias-mitigation technique did not exhibit the accuracy–fairness trade-off cautioned in fairness studies (Chen et al., 2023; Pfohl et al., 2021).

Modality choice also matters. Selecting neuroimaging phenotypes less sensitive to preprocessing biases—such as task-based contrasts or task-regressed functional connectivity—may improve fairness. Therefore, when predictive models are applied as research or clinical tools, it may be more appropriate to prioritise such phenotypes that demonstrate lower sensitivity to training-sample composition.

Developing ethnicity-specific models, analogous to subgroup-specific polygenic scores, might improve within-group predictions but risks institutionalizing segregated tools and diverting attention from inclusivity. If models trained on large White-majority datasets cannot generalize, this reflects deeper representational limits in how current neuroimaging frameworks capture human diversity. The priority should be developing processing pipelines and models that are both accurate and demographically inclusive.

Algorithmic fairness interventions offer complementary solutions. These techniques aim to reduce model dependence on sensitive attributes such as ethnicity and improve generalization across population by removing information about group membership from model representations (adversarial debiasing), aligning data distributions across group (domain adaptation) or adjusting sample importance during training (reweighting) (Chen et al., 2023;

Hasanzadeh et al., 2025; Singhal et al., 2025; Zhang et al., 2018). Beyond algorithmic solutions, advances in more inclusive brain templates and functionally representative atlases could further mitigate structural biases embedded in current preprocessing standards. Together, these efforts represent necessary steps toward equitable and generalizable neuroscience.

In sum, this study provides the first modality-wide benchmark of ethnicity-related bias in predictive neuroimaging. Disparities varied across modalities and model types, with structural features consistently more biased and task-based and connectivity features more equitable. Predictive strength was moderately associated with fairness, but stacking did not resolve disparities. Balanced training improved fairness, but oversampling beyond balance yielded diminishing returns. These results underscore that ensuring equitable neuroimaging-based prediction will require diverse datasets, careful feature selection, and explicit bias-mitigation strategies. As neuroimaging-based prediction advances toward clinical translation, addressing bias is both a scientific necessity and an ethical imperative to ensure that progress in precision neuroscience benefits all populations.

Methods

Participants

Data were drawn from the Adolescent Brain Cognitive Development (ABCD-5.1) study (Jernigan et al., 2024), a longitudinal neuroimaging project with a diverse U.S. cohort. To examine ethnicity-related generalization, we selected WA and AA participants from the two bigger matched groups of the ABCD-BIDS Community Collection (ABCC) (ABCD-BIDS Community), where participants were matched across 21 sites based on a combination of

clinical (history of anesthesia), demographic (age, sex, race/ethnicity) and socioeconomic variables (family income, family type, household size) to ensure comparability between groups (Feczko et al., 2021). The terminology for racial and ethnic categories in this study followed the National Institutes of Health (NIH) definitions of racial and ethnic categories https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-089.html.

Training Strategies

We created four ethnicity-based training strategies to evaluate the effect of ethnicity on model performance. The "All model" included all (majority White) participants in the training matched group. The "RandWA-only model" trained on White American participants, randomly selecting a number of White Americans equal to the number of African American participants to ensure comparability. The "AA-only model" trained exclusively on African American participants. The "Balanced model" included all African American participants along with an equal number of randomly selected White Americans, creating an ethnically balanced training set.

The overall study design, including the cross-validated training–testing framework, the ethnicity-based training strategies, and the corresponding sample sizes for each group, is illustrated in Figure 1.

Cognitive Measures

The cognitive phenotype of interest was the total cognitive score from the NIH Toolbox (Bauer & Zelazo, 2013; Jernigan et al., 2024). See Supplementary Information for details of each cognitive test.

Neuroimaging

For the ABCD, we used pre-processed tabulated data provided by the ABCD study (R. Yang & Jernigan) and a subset of fMRIPrep-processed data from the ABCD-BIDS Community Collection (ABCC) (ABCD-BIDS Community). In total, 81 distinct neuroimaging phenotypes were constructed across modalities. See (Hagler et al., 2019) and Supplementary Information for more details about how we processed the neuroimaging features.

Functional MRI (fMRI) Task Contrasts: 56 sets

We computed GLM contrasts, reflecting BOLD activation related to different events in three tasks: Emotional Nback (Barch et al., 2013), reward-related Monetary Incentive Delay (MID) (Knutson et al., 2000) and inhibitory-controlled Stop Signal Task (SST) (Whelan et al., 2012) (see Supplementary Information). There were 26 contrasts provided by the ABCD study and 30 contrasts computed by us from the ABCC (Hagler et al., 2019) (see Supplementary Fig. 1).

Functional MRI Connectivity (fMRI FC): 9 sets

Functional Connectivity (FC) was based on fMRI time series during rest and tasks. The ABCD provided 3 sets of rest FC: 1) cortical FC (the mean values of the correlations between pairs of regions within the same large-scale cortical network), 2) subcortical-to-network FC (the mean values of the correlations between subcortical regions and large-scale cortical networks) and 3) temporal variance of each brain region. Using the ABCC, we added region-to-region FC during rest and the three tasks, resulting in four more sets. Note that for the task FC, we regressed out the HRF-convolved events from the BOLD time series. Finally, following previous work (Cole et al., 2014, 2019; Elliott et al., 2019), we created two sets: 1)

Multitask FC (region-to-region FC of all tasks combined) and 2) General FC (region-to-region FC of rest and all tasks combined) (see Supplementary Information).

Structural MRI (sMRI) and Diffusion Tensor Imaging (DTI): 16 sets

sMRI reflects anatomical morphology. Using the ABCD and ABCC, we extracted 15 neuroimaging phenotypes: cortical thickness, surface area, FreeSurfer summations, cortical volume, sulcal depth, white-matter averaged intensity, grey-matter averaged intensity, T1 normalized intensity, T1 subcortical averaged intensity, T1 subcortical volume, T1 summations, T2 normalized intensity, T2 subcortical averaged intensity, T2 subcortical volume and T2 summations. DTI reflects white matter integrity. We focused on fractional anisotropy (FA) of 23 white matter tracts (Hagler Jr et al., 2009) (see Supplementary Information).

Predictive Modelling

Models were trained on one matched group and tested on the other, ensuring that test participants were entirely independent of the training set.

All neuroimaging features were standardized (z-scored) based on the training data, and the same standardization parameters were applied to the test set. Hyperparameters were optimized using a five-fold cross-validation, with coefficient of determination (R2) as the evaluation metric.

Prediction models were developed in two frameworks: unimodal models and multimodal models. Unimodal models used Partial Least Squares regression (PLS) to predict cognitive functioning scores from each of the 81 neuroimaging phenotypes derived from task-based fMRI, resting-state fMRI, sMRI, and DTI. Multimodal models combined predictions from the unimodal models across neuroimaging phenotypes into multiple configurations. For

example, the “stacked all” model included all 81 neuroimaging phenotypes, whereas the “stacked task contrasts” model combined task-based fMRI contrasts such as Nback, MID, and SST. Multimodal models employed Random Forest regression with 1,000 decision trees and tuned parameters (e.g., maximum tree depth, number of features per split) to capture non-linear relationships and enhance predictive accuracy (Breiman, 2001). We trained 368 models in total, spanning 92 neuroimaging phenotypes (81 unimodal, 11 multimodal) across 4 training strategies (All, RandWA-only, AA-only, Balanced). See Supplementary Information and Supplementary Figure 1 for a full list of unimodal and multimodal models.

Model Performance

Predicted values from test sets were concatenated across matched groups, and model performance was evaluated using MAE, R2 and Pearson’s r for each training strategy and feature set. For each model, to assess differences in predictive performance between AA and WA participants in the test set, we used a permutation test. Predicted scores were shuffled 1,000 times to generate a null distribution of performance differences between the two groups. The observed difference was then compared against this null distribution to calculate a p-value. Here, we only focus on MAE results. See the supplementary Figures 2-5 for the other metrics.

Ethnicity Bias Index

To assess ethnicity-specific prediction errors, we computed the difference in MAE between AA and WA in the test set when using the model trained on White Americans (RandWA-only). The same was done for the AA-only model. The Ethnicity Bias Index was then obtained by subtracting the difference observed in the RandWA-only model from the difference observed in the AA-only model. Lower absolute Ethnicity Bias Index values

indicate that model predictions are less influenced by participant ethnicity, reflecting less bias. To investigate the relationship between overall predictive performance and bias, we then calculated Pearson's correlation between Ethnicity Bias Index scores and the All model performance on the whole test set for each unimodal and multimodal feature set.

Incremental sampling and synthetic oversampling

To assess the effect of training sample composition on prediction accuracy, we implemented an incremental sampling procedure. First, we randomly selected a subset of WA participants equal in size to the available AA participants in the training set. Models were trained on this WA set, and MAEs were calculated for WA and AA participants in the independent test set. We then progressively increased the proportion of AA participants by randomly adding one AA participant at a time to the training set, repeating the process until AA participants comprised 75% of the training set. Beyond the 50:50 balance point, we conducted synthetic oversampling. AA participants were randomly re-sampled with replacement to increase their proportion. For each step, the model was retrained, and prediction errors for both the WA and AA test groups were recorded. For the stacked stage, unimodal predictions derived from the Balanced AA+RandWA strategy were used to create stacked neuroimaging phenotypes.

Acknowledgments

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 aged 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, and the Neurological Foundation of New Zealand (Grant number: 2350 PRG; awarded to Narun Pat).

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. 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.

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-Benchmark

Disclosures: Farzane Lal Khakpoor, William van der Vliet, Jeremiah Deng, and Narun Pat report no biomedical financial interests or potential conflicts of interest.

Figure Legends

Fig. 1 Overall design and information about participants. (a) Matched groups and data extracted from the ABCD. **(b)** The pipeline for building the cognitive prediction models and training strategies. **(c)** Evaluation of prediction error across different neuroimaging phenotypes and training strategies in the test set.

Fig. 2 Average Prediction Error for Test AA and WA Participants Across Different Neuroimaging phenotypes: (a) sMRI and DTI, **(b)** FC, **(c)** Task Contrast Parcellated with Destrieux, and **(d)** Task Contrast Parcellated with Glasser atlas. For a given model, higher MAE reflects worse predictive performance. Blue stars indicate significantly better performance in tests AA (i.e. lower MAE) and red stars indicate significantly better performance (i.e. lower MAE) in test WA based on the permutation tests (*p-value* = 0.05).

Fig. 3 Average Prediction Error for Test AA and WA Participants Across Multimodal Stacked Neuroimaging phenotypes. For a given model, higher MAE reflects worse predictive performance. Blue stars indicate significantly better performance in tests AA (i.e. lower MAE) and red stars indicate significantly better performance (i.e. lower MAE) in test WA based on the permutation tests (*p-value* = 0.05).

Fig. 4 Ranking of (a) Unimodal and (b) Multimodal neuroimaging phenotypes based on Ethnicity Bias Index. (c) The Correlation Between Test MAE of the Model Trained on All Participants and The Ethnicity Bias Index Score Across Both Unimodal and Multimodal Neuroimaging phenotypes. A lower ethnicity bias score for a phenotype indicates reduced favoritism toward any one ethnic group over the other.

Fig. 5 The Effect of Incremental Addiction of AA to Training Set on Average Prediction Errors in Test AA and WA for the Least and Most Biased Among Unimodal and

Multimodal Neuroimaging phenotypes. A lower MAE reflects improved predictive performance.

References

ABCD-BIDS Community. *ABCD-BIDS Community Collection (ABCC)* [Dataset]. NIMH Data Repositories. https://nda.nih.gov/edit_collection.html?id=3165

Abramovitch, A., Short, T., & Schweiger, A. (2021). The C Factor: Cognitive dysfunction as a transdiagnostic dimension in psychopathology. *Clinical Psychology Review.*, *86*, 102007. https://doi.org/10.1016/j.cpr.2021.102007

Adkins, D. J., & Hanson, J. L. (2025). Racial and Ethnic Disparities in Brain Age Algorithm Performance: Investigating Bias Across Six Popular Methods. *medRxiv*, 2025.09.18.25336117. https://doi.org/10.1101/2025.09.18.25336117

Alhuwaydi, A. M. (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

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., & Van Essen, D. C. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage (Orlando, Fla.)*, *80*, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Bauer, P. J., & Zelazo, P. D. (2013). NIH TOOLBOX COGNITION BATTERY (CB): SUMMARY, CONCLUSIONS, AND IMPLICATIONS FOR COGNITIVE DEVELOPMENT. *Monographs of the Society for Research in Child Development*, *78*(4), 133–146. https://doi.org/10.1111/mono.12039

Carey, S., Pang, A., & Kamps, M. de. (2024). Fairness in AI for healthcare. *Future Healthcare Journal*, *11*(3), 100177. https://doi.org/10.1016/j.fhj.2024.100177

Chee, M. W. L., Zheng, H., Goh, J. O. S., Park, D., & Sutton, B. P. (2011). Brain Structure in Young and Old East Asians and Westerners: Comparisons of Structural Volume and Cortical Thickness. *Journal of Cognitive Neuroscience*, *23*(5), 1065–1079. https://doi.org/10.1162/jocn.2010.21513

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, *7*(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., & Petersen, S. E. (2014). Intrinsic and Task-Evoked Network Architectures of the Human Brain. *Neuron (Cambridge, Mass.)*, *83*(1), Article 1. https://doi.org/10.1016/j.neuron.2014.05.014

Cole, M. W., Ito, T., Schultz, D., Mill, R., Chen, R., & Cocuzza, C. (2019). Task activations produce spurious but systematic inflation of task functional connectivity estimates. *NeuroImage (Orlando, Fla.)*, *189*, 1–18. https://doi.org/10.1016/j.neuroimage.2018.12.054

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, *3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage (Orlando, Fla.)*, *53*(1), Article 1. https://doi.org/10.1016/j.neuroimage.2010.06.010

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, *10*(1), 3328. https://doi.org/10.1038/s41467-019-11112-0

Elliott, M. L., Knodt, A. R., Cooke, M., Kim, M. J., Melzer, T. R., Keenan, R., Ireland, D., Ramrakha, S., Poulton, R., Caspi, A., Moffitt, T. E., & Hariri, A. R. (2019). General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *NeuroImage (Orlando, Fla.)*, *189*, 516–532. https://doi.org/10.1016/j.neuroimage.2019.01.068

Engemann, D. A., Kozynets, O., Sabbagh, D., Lemaitre, G., Varoquaux, G., Liem, F., & Gramfort, A. (2020). Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *eLife*, *9*, 1–33. https://doi.org/10.7554/eLife.54055

Feczko, E., Conan, G., Marek, S., Tervo-Clemmens, B., Cordova, M., Doyle, O., Earl, E., Perrone, A., Sturgeon, D., Klein, R., Harman, G., Kilamovich, D., Hermosillo, R., Miranda-Dominguez, O., Adebimpe, A., Bertolero, M., Cieslak, M., Covitz, S., Hendrickson, T., … Fair, D. A. (2021). Adolescent Brain Cognitive Development (ABCD) Community MRI Collection and Utilities. *bioRxiv*. https://doi.org/10.1101/2021.07.09.451638

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals based on patterns of brain connectivity. *Nature Neuroscience*, *18*(11), Article 11. https://doi.org/10.1038/nn.4135

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature (London)*, *536*(7615), Article 7615. https://doi.org/10.1038/nature18933

Hagler, D. J., Cornejo, M. D., Makowski, C., Dick, A. S., Sutherland, M. T., Harms, M. P., Watts, R., Bjork, J. M., Garavan, H. P., Hilmer, L., Pung, C. J., Sicat, C. S., Kuperman, J., Bartsch, H., Xue, F., Heitzeg, M. M., Laird, A. R., Tapert, S. F., Hyde, L. W., … Dale, A. M. (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

Hagler Jr, D. J., Ahmadi, M. E., Kuperman, J., Holland, D., McDonald, C. R., Halgren, E., & Dale, A. M. (2009). Automated white-matter tractography using a probabilistic diffusion tensor atlas: Application to temporal lobe epilepsy. *Human Brain Mapping*, *30*(5), Article 5. https://doi.org/10.1002/hbm.20619

Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Npj Digital Medicine*, *8*(1), 154. https://doi.org/10.1038/s41746-025-01503-7

Horien, C., Noble, S., Greene, A. S., Lee, K., Barron, D. S., Gao, S., O'Connor, D., Salehi, M., Dadashkarimi, J., Shen, X., Lake, E. M. R., Constable, R. T., & Scheinost, D. (2021). A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nature Human Behaviour*, *5*(2), 185–193. https://doi.org/10.1038/s41562-020-01005-4

Jao, T., Chang, C.-Y., Li, C.-W., Chen, D.-Y., Wu, E., Wu, C.-W., Tsou, C.-H., Ho, C.-C., & Chen, J.-H. (2009). Development of NTU standard Chinese Brain Template: Morphologic and functional comparison with MNI template using magnetic resonance imaging. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2009*, 4779–4782. https://doi.org/10.1109/IEMBS.2009.5334211

Jernigan, T. L., Brown, S. A., Dale, A. M., Tapert, S., Sowell, E. S., Herting, M., Laird, A., Gonzalez, R., Squeglia, L., Gray, K., Paulus, M. P., Aupperle, R., Feldstein Ewing, S. W., Nagel, B. J., Fair, D. A., Baker, F., Colrain, I. M., Bookheimer, S. Y., Dapretto, M., … Gee, D. (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

Kachuri, L., Chatterjee, N., Hirbo, J., Schaid, D. J., Martin, I., Kullo, I. J., Kenny, E. E., Pasaniuc, B., Auer, P. L., Conomos, M. P., Conti, D. V., Ding, Y., Wang, Y., Zhang, H., Zhang, Y., Witte, J. S., Ge, T., & Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium Methods Working Group. (2024). Principles and methods for transferring polygenic risk scores across global populations. *Nature Reviews Genetics*, *25*(1), 8–25. https://doi.org/10.1038/s41576-023-00637-2

Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). FMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. *NeuroImage (Orlando, Fla.)*, *12*(1), 20–27. https://doi.org/10.1006/nimg.2000.0593

Lewis, A. C. F., & Green, R. C. (2021). Polygenic risk scores in the clinic: New perspectives needed on familiar ethical issues. *Genome Medicine*, *13*(1), 14. https://doi.org/10.1186/s13073-021-00829-7

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., Ge, T., Patil, K. R., Jabbi, M., Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, *8*(11), eabj1812. https://doi.org/10.1126/sciadv.abj1812

Liang, P., Shi, L., Chen, N., Luo, Y., Wang, X., Liu, K., Mok, V. C. T., Chu, W. C. W., Wang, D., & Li, K. (2015). Construction of brain atlases based on a multi-center MRI dataset of 2020 Chinese adults. *Scientific Reports*, *5*, 18216. https://doi.org/10.1038/srep18216

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature (London)*, *603*(7902), Article 7902. https://doi.org/10.1038/s41586-022-04492-9

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, *51*(4), 584–591. https://doi.org/10.1038/s41588-019-0379-x

Montreal Neurological Institute and Hospital. (2009). *ICBM152 Nonlinear Atlases (MNI152 Template)*. McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University. https://nist.mni.mcgill.ca/icbm-152lin/

Owens, M. M., Hyatt, C. S., Gray, J. C., Carter, N. T., MacKillop, J., Miller, J. D., & Sweet, L. H. (2019). Cortical morphometry of the five-factor model of personality: Findings from the Human Connectome Project full sample. *Social Cognitive and Affective Neuroscience*, *14*(4), 381–395. https://doi.org/10.1093/scan/nsz017

Pat, N., Wang, Y., Bartonicek, A., Candia, J., & Stringaris, A. (2023). *Explainable machine learning approach to predict and explain the relationship between task-based fMRI and individual differences in cognition*.

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C.-Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., … Duncan, L. E. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*, *179*(3), 589–603. https://doi.org/10.1016/j.cell.2019.08.051

Pfohl, S. R., Foryciarz, A., & Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, *113*, 103621. https://doi.org/10.1016/j.jbi.2020.103621

Rasero, J., Sentis, A. I., Yeh, F. C., & Verstynen, T. (2021). Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *PLoS Computational Biology*, *17*(3), Article 3. https://doi.org/10.1371/journal.pcbi.1008347

Rosenberg, M. D., Casey, B. J., & Holmes, A. J. (2018). Prediction complements explanation in understanding the developing brain. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-02887-9

Singhal, P., Gokhale, S., Shah, A., Jain, D. K., Walambe, R., Ekart, A., & Kotecha, K. (2025). Domain adaptation for bias mitigation in affective computing: Use cases for facial emotion recognition and sentiment analysis systems. *Discover Applied Sciences*, *7*(4), 229. https://doi.org/10.1007/s42452-025-06659-1

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*(12), Article 12. https://doi.org/10.1002/hbm.25007

Sripada, C., Rutherford, S., Angstadt, M., Thompson, W. K., Luciana, M., Weigard, A., Hyde, L. H., & Heitzeg, M. (2020). Prediction of neurocognition in youth from resting state fMRI. *Molecular Psychiatry.*, *25*(12), Article 12. https://doi.org/10.1038/s41380-019-0481-6

Tang, Y., Zhao, L., Lou, Y., Shi, Y., Fang, R., Lin, X., Liu, S., & Toga, A. (2018). Brain structure differences between Chinese and Caucasian cohorts: A comprehensive morphometry study. *Human Brain Mapping*, *39*(5), 2147–2155. https://doi.org/10.1002/hbm.23994

Tetereva, A., Knodt, A., Melzer, T., Van Der Vliet, W., Gibson, B., Hariri, A., Whitman, E. T., Li, J., Farzane Lal Khakpoor, Deng, J., Ireland, D., Ramrakha, S., & Narun Pat. (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

Tetereva, A., Li, J., Deng, J. D., Stringaris, A., & Pat, N. (2022). *Capturing brain‐cognition relationship: Integrating task‐based fMRI across tasks markedly boosts prediction and test‐retest reliability*.

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology*, *42*(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3

Vieira, B. H., Pamplona, G. S. P., Fachinello, K., Silva, A. K., Foss, M. P., & Salmon, C. E. G. (2022). On the prediction of human intelligence from neuroimaging: A systematic review of methods and reporting. *Intelligence (Norwood)*, *93*, 101654. https://doi.org/10.1016/j.intell.2022.101654

Wang, C., Ng, B., & Garbi, R. (2018). Multimodal Brain Parcellation Based on Functional and Anatomical Connectivity. *Brain Connectivity*. https://doi.org/10.1089/brain.2017.0576

Wang, Y., Tsuo, K., Kanai, M., Neale, B. M., & Martin, A. R. (2022). Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. In *Annual Review of Biomedical Data Science* (Vol. 5, Issue Volume 5, 2022, pp. 293–320). Annual Reviews. https://doi.org/10.1146/annurev-biodatasci-111721-074830

Whelan, R., Conrod, P. J., Poline, J.-B., Lourdusamy, A., Banaschewski, T., Barker, G. J., Bellgrove, M. A., Büchel, C., Byrne, M., Cummins, T. D. R., Fauth-Bühler, M., Flor, H., Gallinat, J., Heinz, A., Ittermann, B., Mann, K., Martinot, J.-L., Lalor, E. C., Lathrop, M., … Garavan, H. (2012). Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neuroscience*, *15*(6), 920–925. https://doi.org/10.1038/nn.3092

Xiang, R., Kelemen, M., Xu, Y., Harris, L. W., Parkinson, H., Inouye, M., & Lambert, S. A. (2024). Recent advances in polygenic scores: Translation, equitability, methods and

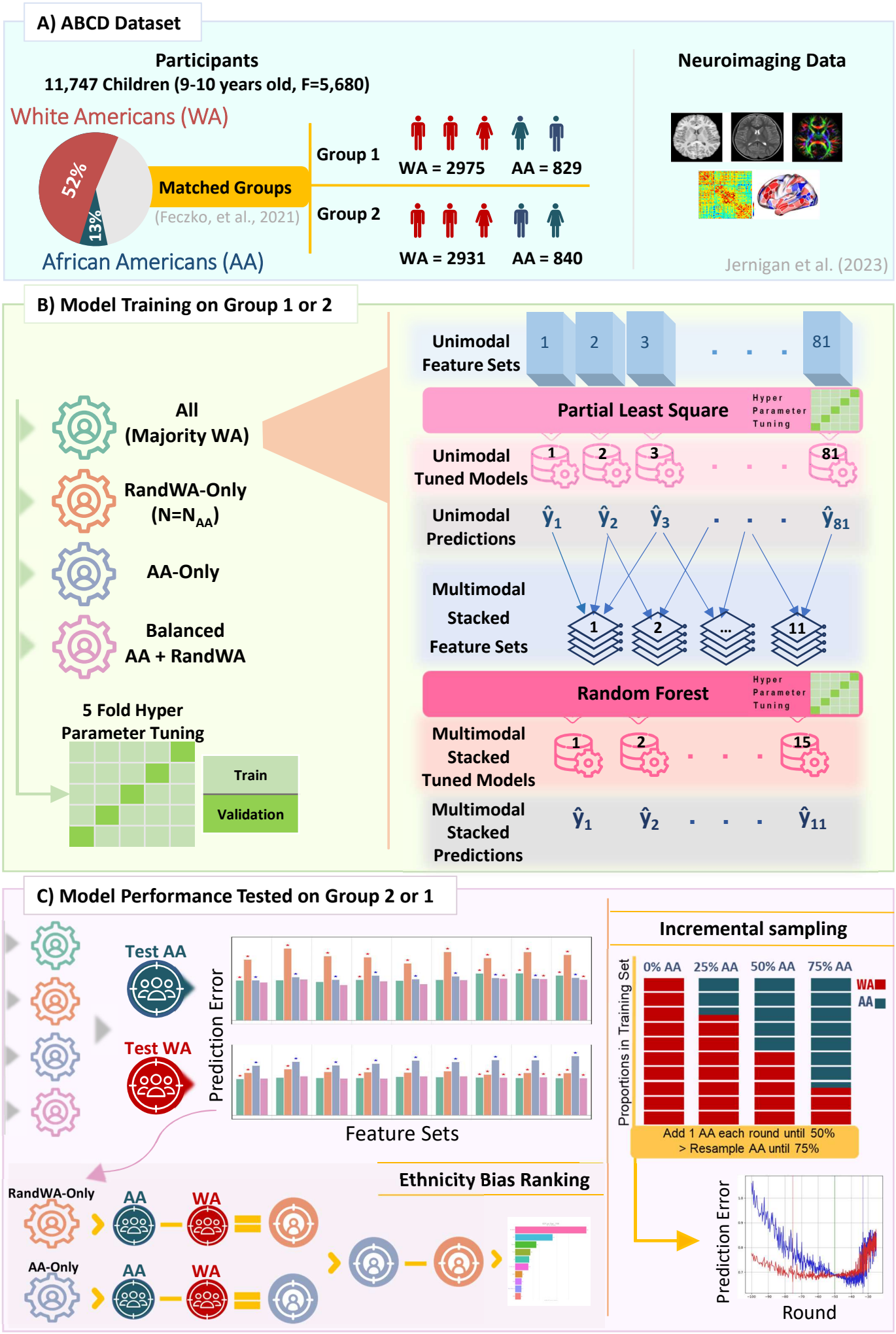
FAIR tools. *Genome Medicine*, *16*(1), 33. https://doi.org/10.1186/s13073-024-01304-9

Yang, G., Bozek, J., Noble, S., Han, M., Wu, X., Xue, M., Kang, J., Jia, T., Fu, J., Ge, J., Cui, Z., Li, X., Feng, J., & Gao, J.-H. (2023). Global diversity in individualized cortical network topography. *Cerebral Cortex*, *33*(11), 6803–6817. https://doi.org/10.1093/cercor/bhad002

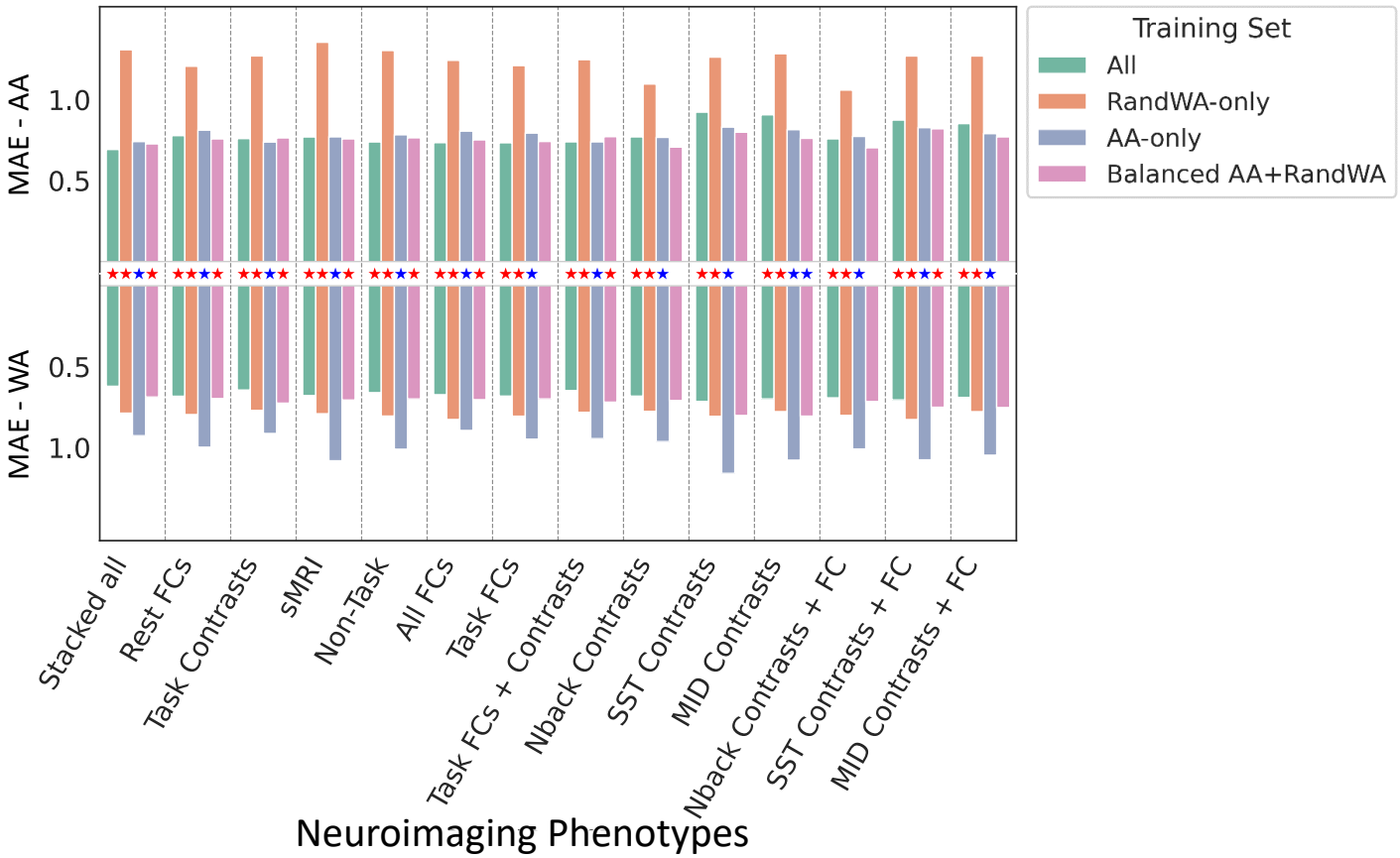
Yang, G., Zhou, S., Bozek, J., Dong, H.-M., Han, M., Zuo, X.-N., Liu, H., & Gao, J.-H. (2020). Sample sizes and population differences in brain template construction. *NeuroImage*, *206*, 116318. https://doi.org/10.1016/j.neuroimage.2019.116318

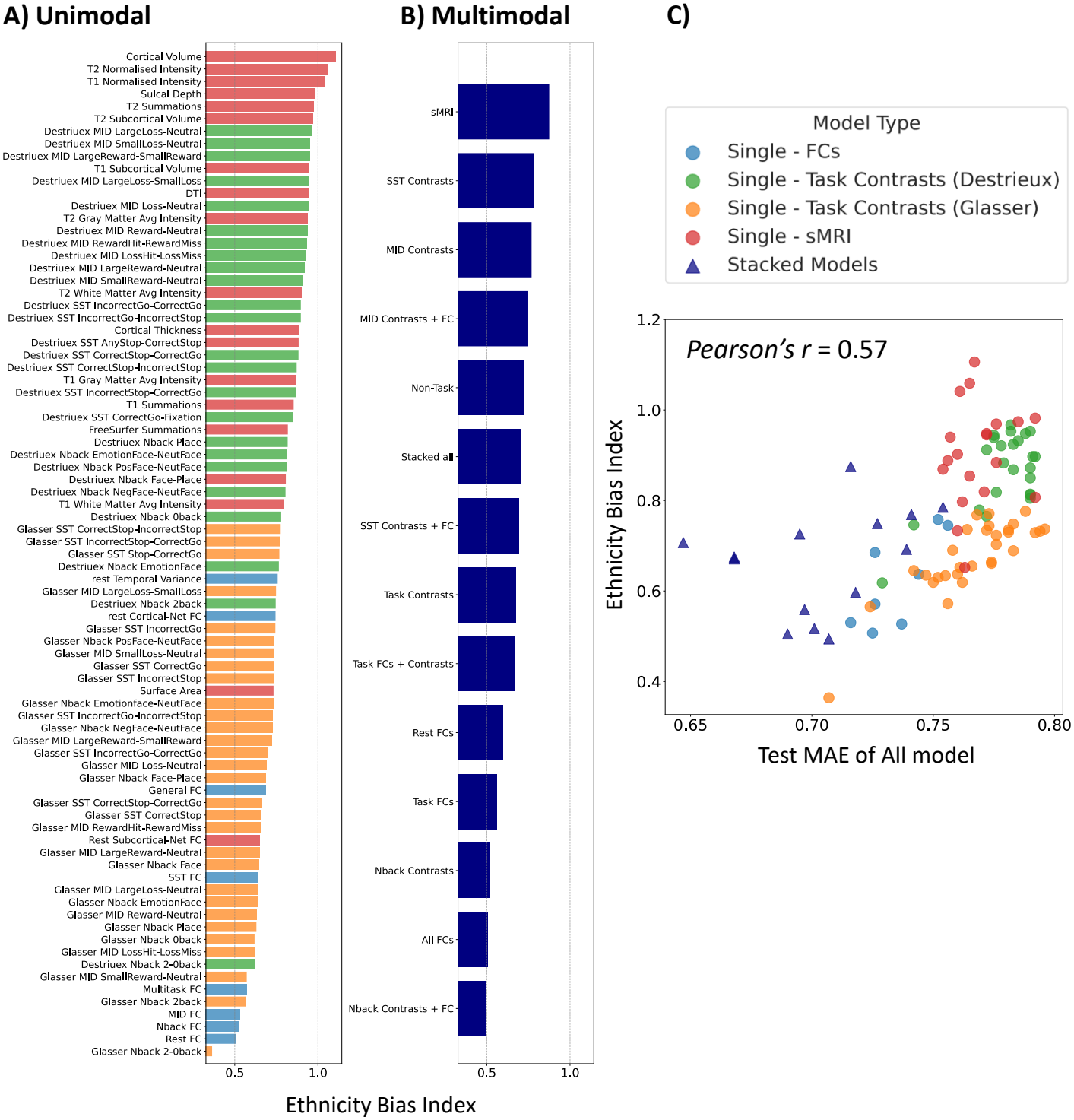
Yang, R., & Jernigan, T. *Adolescent Brain Cognitive Development Study (ABCD)—Annual Release 4.0* [Dataset]. NIMH Data Repositories. https://nda.nih.gov/edit_collection.html?id=2573

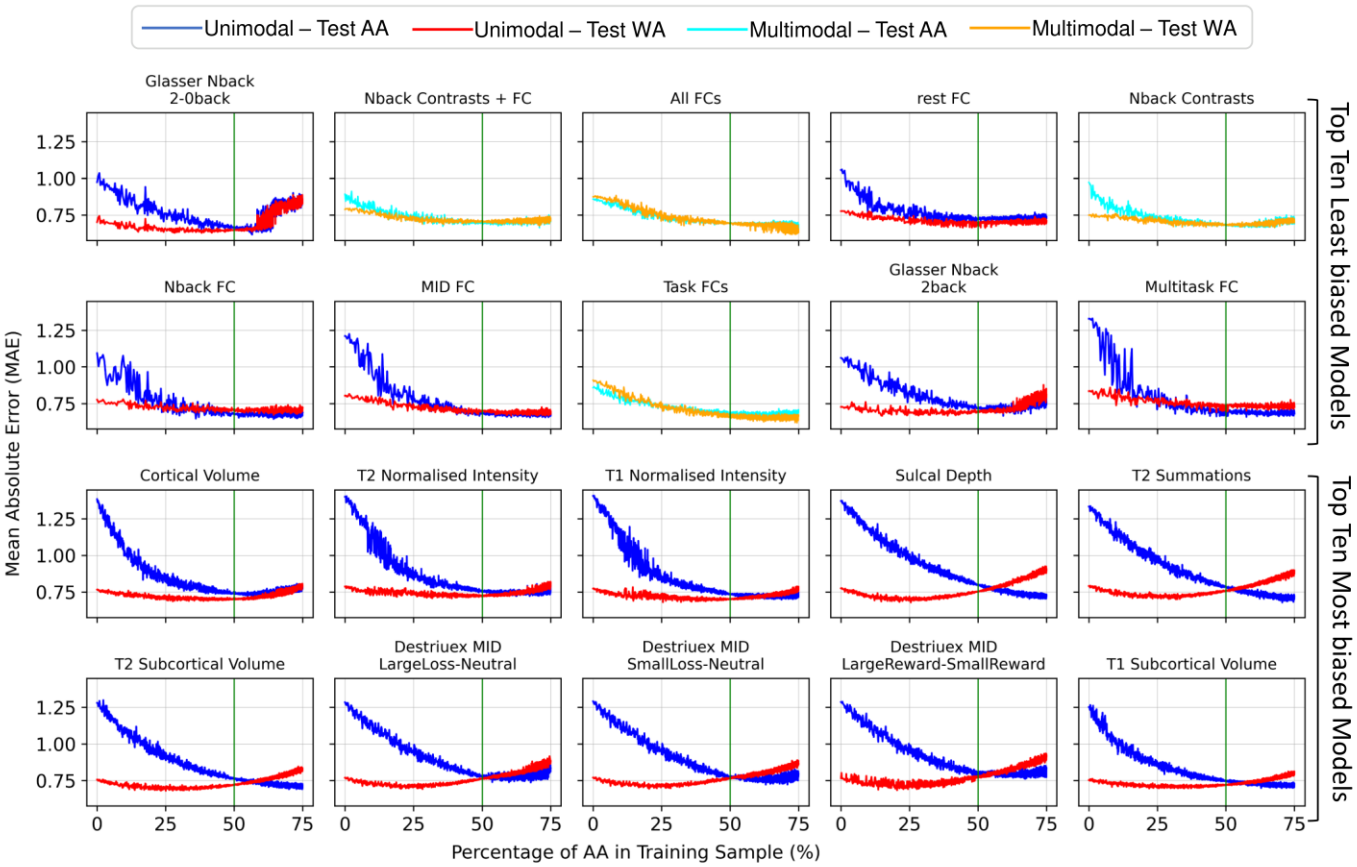
Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779


A) ABCD Dataset
Participants
11,747 Children (9-10 years old, F=5,680)
White Americans (WA)
52%
13%
African Americans (AA)
Matched Groups
(Feczko, et al., 2021)
Group 1
WA = 2975
AA = 829
Group 2
WA = 2931
AA = 840
Neuroimaging Data
Jernigan et al. (2023)
B) Model Training on Group 1 or 2
All
(Majority WA)
RandWA-Only
(N=N_AA)
AA-Only
Balanced
AA + RandWA
5 Fold Hyper
Parameter Tuning
Train
Validation
Unimodal
Feature Sets
1
2
3
81
Partial Least Square
Hyper
Parameter
Tuning
Unimodal
Tuned Models
Unimodal
Predictions
Multimodal
Stacked
Feature Sets
11
Random Forest
Multimodal
Stacked
Tuned Models
15
Multimodal
Stacked
Predictions
C) Model Performance Tested on Group 2 or 1
Test AA
Test WA
Prediction Error
Feature Sets
Incremental sampling
0% AA
25% AA
50% AA
75% AA
WA
AA
Proportions in Training Set
Add 1 AA each round until 50%
> Resample AA until 75%
Prediction Error
Round
Ethnicity Bias Ranking
RandWA-Only
AA-Only
AA
WA

A) sMRI
B) FC
MAE - AA
MAE - WA
1.5
1.0
0.5
T2 Gray Matter Avg Intensity
T1 Gray Matter Avg Intensity
T2 White Matter Avg Intensity
T1 Normalised Intensity
T1 White Matter Avg Intensity
T2 Normalised Intensity
Surface Area
T1 Summations
FreeSurfer Summations
DTI
T2 Summations
Sulcal Depth
T1 Subcortical Volume
Cortical Thickness
T2 Subcortical Volume
Cortical Volume
MID FC
General FC
Multitask FC
Rest FC
Nback FC
SST FC
Rest Subcortical-Net FC
Rest cortical-Net FC
Rest Temporal Variance
Training Set
All
RandWA-only
AA-only
Balanced AA+RandWA
C) Task Contrasts - Destrieux
MID LargeReward-SmallReward
MID LossHit-LossMiss
SST IncorrectGo-IncorrectStop
SST AnyStop-CorrectStop
SST IncorrectStop-CorrectGo
Nback EmotionFace-NeutFace
SST IncorrectGo-CorrectGo
SST CorrectGo-Fixation
MID Reward-Neutral
Nback 2back
Nback Face-Place
MID SmallLoss-Neutral
Nback NegFace-NeutFace
MID LargeLoss-Neutral
Nback EmotionFace
SST CorrectStop-IncorrectStop
MID SmallReward-Neutral
MID LargeReward-Neutral
MID Loss-Neutral
SST CorrectStop-CorrectGo
Nback PosFace-NeutFace
Nback 0back
Nback Place
MID LargeLoss-SmallLoss
Nback 2-0back
MID RewardHit-RewardMiss
D) Task Contrasts - Glasser
SST Stop-CorrectGo
SST IncorrectStop-CorrectGo
SST IncorrectGo
SST IncorrectStop
SST IncorrectGo-CorrectGo
SST CorrectStop
SST CorrectStop-CorrectGo
SST IncorrectGo-IncorrectStop
SST CorrectGo
SST CorrectStop-IncorrectStop
Nback 2back
MID LossHit-LossMiss
MID SmallLoss-Neutral
MID LargeReward-SmallReward
MID Loss-Neutral
Nback PosFace-NeutFace
MID LargeLoss-Neutral
Nback NegFace-NeutFace
Nback Face-Place
MID LargeReward-Neutral
MID RewardHit-RewardMiss
Nback Place
Nback 2-0back
MID LargeLoss-SmallLoss
Nback Face
Nback EmotionFace
Nback EmotionFace-NeutFace
Nback 0back
MID SmallReward-Neutral
MID Reward-Neutral
Neuroimaging Phenotypes


Training Set
All
RandWA-only
AA-only
Balanced AA+RandWA
MAE - AA
1.0
0.5
MAE - WA
0.5
1.0
Stacked all
Rest FCs
Task Contrasts
sMRI
Non-Task
All FCs
Task FCs
Task FCs + Contrasts
Nback Contrasts
SST Contrasts
MID Contrasts
Nback Contrasts + FC
SST Contrasts + FC
MID Contrasts + FC
Neuroimaging Phenotypes


A) Unimodal
Cortical Volume
T2 Normalised Intensity
T1 Normalised Intensity
Sulcal Depth
T2 Summations
T2 Subcortical Volume
Destriuex MID LargeLoss-Neutral
Destriuex MID SmallLoss-Neutral
Destriuex MID LargeReward-SmallReward
T1 Subcortical Volume
Destriuex MID LargeLoss-SmallLoss
DTI
Destriuex MID Loss-Neutral
T2 Gray Matter Avg Intensity
Destriuex MID Reward-Neutral
Destriuex MID RewardHit-RewardMiss
Destriuex MID LossHit-LossMiss
Destriuex MID LargeReward-Neutral
Destriuex MID SmallReward-Neutral
T2 White Matter Avg Intensity
Destriuex SST IncorrectGo-CorrectGo
Destriuex SST IncorrectGo-IncorrectStop
Cortical Thickness
Destriuex SST AnyStop-CorrectStop
Destriuex SST CorrectStop-CorrectGo
Destriuex SST CorrectStop-IncorrectStop
T1 Gray Matter Avg Intensity
Destriuex SST IncorrectStop-CorrectGo
T1 Summations
Destriuex SST CorrectGo-Fixation
FreeSurfer Summations
Destriuex Nback Place
Destriuex Nback EmotionFace-NeutFace
Destriuex Nback PosFace-NeutFace
Destriuex Nback Face-Place
Destriuex Nback NegFace-NeutFace
T1 White Matter Avg Intensity
Destriuex Nback 0back
Glasser SST CorrectStop-IncorrectStop
Glasser SST IncorrectStop-CorrectGo
Glasser SST Stop-CorrectGo
Destriuex Nback EmotionFace
rest Temporal Variance
Glasser MID LargeLoss-SmallLoss
Destriuex Nback 2back
rest Cortical-Net FC
Glasser SST IncorrectGo
Glasser Nback PosFace-NeutFace
Glasser MID SmallLoss-Neutral
Glasser SST CorrectGo
Glasser SST IncorrectStop
Surface Area
Glasser Nback Emotionface-NeutFace
Glasser SST IncorrectGo-IncorrectStop
Glasser Nback NegFace-NeutFace
Glasser MID LargeReward-SmallReward
Glasser SST IncorrectGo-CorrectGo
Glasser MID Loss-Neutral
Glasser Nback Face-Place
General FC
Glasser SST CorrectStop-CorrectGo
Glasser SST CorrectStop
Glasser MID RewardHit-RewardMiss
Rest Subcortical-Net FC
Glasser MID LargeReward-Neutral
Glasser Nback Face
SST FC
Glasser MID LargeLoss-Neutral
Glasser Nback EmotionFace
Glasser MID Reward-Neutral
Glasser Nback Place
Glasser Nback 0back
Glasser MID LossHit-LossMiss
Destriuex Nback 2-0back
Glasser MID SmallReward-Neutral
Multitask FC
Glasser Nback 2back
MID FC
Nback FC
Rest FC
Glasser Nback 2-0back
0.5
1.0
B) Multimodal
sMRI
SST Contrasts
MID Contrasts
MID Contrasts + FC
Non-Task
Stacked all
SST Contrasts + FC
Task Contrasts
Task FCs + Contrasts
Rest FCs
Task FCs
Nback Contrasts
All FCs
Nback Contrasts + FC
0.5
1.0
Ethnicity Bias Index
C)
Model Type
Single - FCs
Single - Task Contrasts (Destrieux)
Single - Task Contrasts (Glasser)
Single - sMRI
Stacked Models
Pearson's r = 0.57
1.2
1.0
0.8
0.6
0.4
Ethnicity Bias Index
0.65
0.70
0.75
0.80
Test MAE of All model


Unimodal – Test AA
Unimodal – Test WA
Multimodal – Test AA
Multimodal – Test WA
Glasser Nback 2-0back
Nback Contrasts + FC
All FCs
rest FC
Nback Contrasts
Nback FC
MID FC
Task FCs
Glasser Nback 2back
Multitask FC
Cortical Volume
T2 Normalised Intensity
T1 Normalised Intensity
Sulcal Depth
T2 Summations
T2 Subcortical Volume
Destriuex MID LargeLoss-Neutral
Destriuex MID SmallLoss-Neutral
Destriuex MID LargeReward-SmallReward
T1 Subcortical Volume
Top Ten Least biased Models
Top Ten Most biased Models
Mean Absolute Error (MAE)
1.25
1.00
0.75
0
25
50
75
Percentage of AA in Training Sample (%)