Brain predictive models of cognition fail to generalize across ethnicities: Modality-dependent bias in MRI-based prediction

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

Keywords: neuroimaging, predictive modelling, bias, cognitive prediction, population generalizability, model fairness

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 91 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, emerging as the upper bound for both accuracy and fairness. 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

The use of machine learning in predictive modelling enables neuroscientists to move beyond associating individual brain features to variables of interest, such as cognitive functioning, toward predicting these variables in new unseen participants or patients based on high-dimensional brain features. This approach has accelerated biomarker discovery and paved the way for the development of precision medicine [1] with the potential to support earlier risk detection and more targeted intervention [2]. For instance, predictive modelling has transformed how we study individual differences in cognitive functioning [3–6], a core psychiatric phenotype that is strongly associated with daily functioning and treatment outcomes [7]. Despite these advances, a critical translational challenge remains: ensuring that predictive models generalise and perform consistently across diverse populations.

Ethnic/racial bias in predictive models is well-documented across multiple biomedical domains [8–11]. 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 [12–14]. Such disparities in predictive performance raise concerns that these tools may inadvertently exacerbate existing health inequities [15,16].

Neuroimaging could face similar challenges. Large-scale imaging datasets are disproportionately composed of White participants from North America and Europe [17], 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 [18] 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 behavioural measures. Another recent study found that brain-age algorithms based on structural data exhibit lower accuracy among African American participants [19]. These findings underscore 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 [18]. 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 [20–22], it is important to determine whether neuroimaging phenotypes that capture stronger brain–behaviour 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 [21,23] have been shown to reliably enhance predictive accuracy for traits such as cognitive functioning by integrating information across multiple neuroimaging phenotypes [21,24–26]. 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 80 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 neuroimaging phenotypes. 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 neuroimaging phenotypes 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.

Figure 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.

Results

To evaluate ethnicity-related bias, we analysed data from the ABCD dataset, focusing on White and African American participants in groups matched across demographic and socioeconomic factors [27] (Figure 1A). Cognitive functioning was measured using the NIH Toolbox total score [28,29]. We derived 80 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 phenotypes, the RandWA-only models performed significantly better on WA test participants, as indicated by lower MAE compared to AA test participants, based on permutation tests. The All models, trained on the full dataset with a majority WA participants, achieved overall higher accuracy than the RandWA-only model but continued to perform significantly better for WA than AA participants. The AA-only models showed the opposite trend, performing significantly 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). See Supplementary Figures 2A and 4A for coefficient of determination (*R2*) and Pearson correlation (*r*) across training strategies.

Functional Connectivity (FC)

For FC phenotypes, 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 numerically similar to 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). See Supplementary Figures 2B and 4B for coefficient of determination (*R2*) and Pearson correlation (*r*) across training strategies.

Task-based fMRI contrasts

For task-based contrasts, the RandWA-only models performed significantly better on WA test participants, as indicated by lower MAE compared to AA participants. The All model reduced overall errors but continued to perform significantly better for WA than AA participants. In contrast, the AA-only model performed significantly better for AA participants. For contrast phenotypes parcellated with the Glasser atlas [30], 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 [31], this pattern of comparable performance was observed in only seven out of 26 contrasts (Figure 2D). See Supplementary Figures 2C&D and 4 C&D for coefficient of determination (*R2*) and Pearson correlation (*r*) across training strategies.

Figure 2 Average prediction error (MAE) for test AA and WA participants across different neuroimaging phenotypes: **(A)** sMRI and DTI, **(B)** FC, **(C)** Task contrasts parcellated with Destrieux [31], and **(D)** Task contrasts parcellated with Glasser atlas [30]. 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).

Multimodal models

Multimodal models were implemented as stacked models that integrated multiple neuroimaging phenotypes. These models exhibited a pattern similar to ethnicity-specific models, with significantly lower prediction error (i.e., better performance) within the ethnic group on which they were trained. When applying the All model, MAE remained significantly lower for WA than AA participants (Figure 3). The Balanced model showed a comparable pattern: only five multimodal models, those combining predictions from task-specific neuroimaging phenotypes, demonstrated non-significant differences between AA and WA. See Supplementary Figures 3 and 5 for coefficient of determination (*R2*) and Pearson correlation (*r*) across training strategies.

Figure 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).

Ethnicity Bias Index

The Ethnicity Bias Index was derived for each of the 80 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 neuroimaging phenotypes 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 phenotypes 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 phenotype had a higher bias score compared to its unimodal counterpart. The most biased multimodal phenotype had a smaller bias score than its unimodal counterpart (Figure 4B).

Relation between performance and bias

Across all unimodal and multimodal neuroimaging phenotypes, lower predictive error—calculated by applying the All model on the whole test set, was associated with lower Ethnicity Bias Index score ($r(89) = .57$, 95% bootstrap CI [.45, .68], $p = 0.0000$, Figure 4C). For unimodal phenotypes, the correlation between the All model performance on the whole test set and the corresponding Ethnicity Bias Index values was $r(78) = .62$, 95% bootstrap CI [.44, .75], $p = 0.0000$, indicating that models yielding stronger predictions tended to also yield smaller disparities between AA and WA.

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 $r(9) =.3$, 95% bootstrap CI [-.1, .64], $p = .3$.

Figure 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 favouritism toward any one ethnic group over the other.

To examine how feature contributions to predictions vary with the ethnicity of the training sample, we computed, for each unimodal phenotype, the difference in Partial Least Square (PLS) regression coefficients between models trained on WA-only and AA-only samples. These coefficient differences quantify how the contribution of each brain region or network shifts when models are trained on AA versus WA samples. In Figure 5, we mapped these feature weights back onto brain space for the ten least and most biased unimodal phenotypes.

Figure 5 The difference in PLS coefficients when training sample changes from WA-only to AA-only for the ten **(A)** least biased and **(B)** most biased unimodal phenotypes.

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 phenotypes, 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 phenotypes, increasing AA representation beyond 50% did not yield further benefits and in some cases increased error (Figure 6).

Figure 6 The effect of incremental addition of AA to training set on average prediction errors in test AA and WA for the ten least and most biased among unimodal and multimodal neuroimaging phenotypes. The green vertical line indicates balanced training sample including same number of AA and WA participants. A lower MAE reflects improved predictive performance.

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

Across all MRI modalities, models trained on a single ethnic group achieved their best performance when applied to that same group. The All models—trained on participants sampled without regard to ethnicity and therefore dominated by White participants—performed better on White participants, consistent with prior findings for resting-state fMRI [18]. Together, these results suggest that the brain–cognition associations captured by predictive models are systematically shaped 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 [32,33]. 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–behaviour 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 [34].

Balanced sampling as the fairness ceiling

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. Across all 80 unimodal and 11 multimodal phenotypes, no strategy achieved lower ethnicity-specific prediction error than balanced sampling. This includes models trained on the full imbalanced dataset, ethnicity-specific models, multimodal stacked ensembles, and synthetic oversampling. The superiority of balanced sampling stems from its preservation of the full target distribution for the underrepresented group, eliminating extrapolation error without sacrificing exposure to the majority distribution. In the absence of new data collection, balanced subsampling is the optimal mitigation strategy for ethnic bias in neuroimaging-based cognitive prediction.

Factors contributing to the differences in prediction performance

These results extend prior literature [18] 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–behaviour 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, pre-processing, 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 phenotypes being less sensitive to training composition, while structural phenotypes 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 [35–37], 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 [30] yielded both stronger

predictions and markedly lower bias compared to Destrieux [31]. 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 [30,38]. 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 [31], which are known to vary across populations [39–41]. This reliance on anatomical landmarks may explain why Destrieux-based phenotypes 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 pre-processing and atlas choice.

Together, these observations suggest that modality-dependent disparities reflect both the predictive robustness of phenotypes and the ways in which pre-processing 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 pre-processing 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 [42]. 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 [30,43].

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 [12,16,44]. 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 [32,33]. Balanced sampling thus represents the practical upper bound for equitable prediction — a simple, cost-free strategy that outperforms complex multimodal integration and data augmentation.

Modality choice also matters. Selecting neuroimaging phenotypes less sensitive to pre-processing 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) [10,32,45,46]. Beyond algorithmic solutions, advances in more inclusive brain templates and functionally representative atlases could further mitigate structural biases embedded in current pre-processing standards. Together, these efforts represent necessary steps toward equitable and generalizable neuroscience.

Conclusion

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 phenotypes consistently more biased and task-based and connectivity phenotypes more equitable. Predictive strength was moderately associated with fairness, but stacking did not resolve disparities. Balanced training emerged as the upper bound for equitable

predictions, 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) Study Curated Annual Release 5.1 (DOI:10.15154/z563-zd24), a longitudinal neuroimaging project with a diverse U.S. cohort [47]. The baseline cohort included 11,878 participants (48% female; 52% male) aged 9–10 years. To examine ethnicity-related generalization, we selected WA and AA participants from the two bigger matched groups of the ABCD-BIDS Community Collection (ABCC-2.0) [48], where participants were matched across 21 sites based on a combination of clinical (history of anaesthesia), demographic (age, sex, race/ethnicity) and socioeconomic variables (family income, family type, household size) to ensure comparability between groups [27]. For the purpose of current study, we interchangeably used the word race and ethnic to refer to the characterisation of WA and AA groups.

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 composite score from the NIH Toolbox [28,29]. We selected the total cognition composite as the primary outcome because it exhibits higher test–retest reliability and reduced measurement noise relative to individual task scores [49–51], and has been widely used in prior neuroimaging-based prediction studies within ABCD and related developmental cohorts [52–54]. Aggregating across domains yields a more stable and robust phenotype for predictive modelling, mitigates task-specific idiosyncrasies, and improves sensitivity to distributed brain–behaviour relationships that are unlikely to be captured by single cognitive measures [51,55,56]. See Supplementary Information for details of each cognitive test.

Neuroimaging phenotypes

Neuroimaging data were drawn from the Adolescent Brain Cognitive Development (ABCD) Study, integrating two complementary sources: (i) post-processed tabulated summary measures released by the ABCD consortium (version 5.1, DOI:10.15154/z563-zd24) and (ii) additional phenotypes derived from custom processing of minimally pre-processed data from the ABCD-BIDS Community Collection (ABCC) [48]. The ABCD tabulated datasets provide quality-controlled regional measures and task contrasts parcellated using the Destrieux cortical atlas [31] and FreeSurfer [57] subcortical segmentation. To extend spatial resolution and connectivity and structural coverage beyond these summaries, we further processed ABCC data using outputs from fMRIPrep (v20.2.0) and the Human Connectome Project (HCP) minimal pre-processing pipeline. See Hagler et al. [58] and Supplementary Information for more details about how we processed the neuroimaging data.

After applying ABCD-recommended quality control and exclusion procedures—including removal of Site 22 due to limited sample size, correction of site ID errors, exclusion of participants with reported vision problems, scan and task quality flags—the final neuroimaging sample sizes ranged from approximately 4,000 to 11,177 participants, depending on modality and phenotype. Importantly, exclusions were applied at the phenotype level rather than via listwise deletion, allowing participants to contribute data to phenotypes

that met quality criteria. In total, this pipeline yielded 80 distinct neuroimaging sets (summarized in Supplementary Information, Supplementary Figure 1).

Functional MRI (fMRI) Task Contrasts: 56 sets

Task-evoked fMRI BOLD response were derived from three ABCD tasks: the Emotional N-back task (working memory and affective processing) [59], the Monetary Incentive Delay (MID) task (reward and punishment processing) [60], and the Stop Signal Task (SST; inhibitory control) [61]. From the ABCD tabulated data, we used 26 precomputed contrasts (9 N-back, 10 MID, 7 SST), parcellated with the Destrieux cortical [31] and FreeSurfer subcortical [57] atlases [58].

To complement these measures, we computed an additional 30 task contrasts (10 per task) from ABCC data using a first-level General Linear Model (GLM) framework implemented primarily in Nilearn [62]. Task event files were converted from E-Prime format to BIDS-compatible TSV files, with event onsets aligned relative to post-calibration task start times. Pre-processing incorporated nuisance regression informed by prior recommendations, including cosine drift regressors, ten anatomical CompCor components, and six rigid-body motion parameters with their first temporal derivatives. Non-steady-state volumes were removed in a scanner-specific manner prior to modelling.

GLMs were estimated using ordinary least squares with the SPM canonical hemodynamic response function. For each contrast of interest, beta coefficients were estimated at the run level and combined across runs using variance-weighted fixed effects. Contrasts were designed to isolate specific cognitive processes, such as working memory load (e.g., 2back, 0back, 2-0back), reward magnitude (e.g., large reward vs. neutral), response type (e.g., hit vs. miss), and inhibitory control (e.g., successful stop vs. go). Final contrast maps were parcellated into the 379-region Glasser–FreeSurfer [30,57] atlas to produce regional effect sizes.

Functional MRI Connectivity (fMRI FC): 9 sets

Functional connectivity phenotypes were derived from both ABCD tabulated data and ABCC-based analyses. From the ABCD releases, three resting-state FC sets were included: (i) temporal variance measures reflecting amplitude of low-frequency fluctuations across 333 Gordon cortical parcels [63] and 19 subcortical regions [57] ; (ii) subcortical-to-network connectivity, defined as average correlations between each subcortical region and 13 large-scale cortical networks; and (iii) within- and between-network cortical connectivity across

these networks. Using ABCC data, we additionally computed region-to-region FC matrices at the 379-region resolution Glasser–FreeSurfer [30,57] for resting state and each of the three tasks, yielding four further sets. Nuisance regression and dummy-scan removal followed the same procedures as for task activation analyses. Motion outliers were identified using framewise displacement (>0.5) and DVARS (>1.5), with flagged volumes incorporated as regressors. Participants with more than have flagged volumes were excluded. Given the multiband acquisition and developmental sample, a relatively lenient motion threshold was adopted to balance artifact mitigation with signal preservation. Residual time series were extracted after regressing out nuisance variables and, for task data, task-evoked activity. Extreme values were attenuated via de-spiking prior to connectivity estimation. Residuals were concatenated across runs and parcellated into the 379 regions. Functional connectivity was computed using Pearson correlation, followed by Fisher z-transformation and vectorization. In addition to task-specific and resting-state FC, we derived multitask FC by concatenating residuals across all three tasks and a general FC measure by concatenating residuals across rest and all tasks, consistent with evidence that aggregated data improve reliability and behavioural prediction [64–66] (see Supplementary Information).

Structural MRI (sMRI) and Diffusion Tensor Imaging (DTI): 15 sets

Structural MRI phenotypes were derived primarily from FreeSurfer [57] processing of T1- and T2-weighted images. We included 14 structural phenotypes capturing complementary aspects of brain morphology and tissue properties, comprising regional measures of cortical thickness, surface area, volume, and subcortical volumes, parcellated using the Destrieux cortical atlas [31] and FreeSurfer subcortical segmentation [57]. In addition to morphometric measures, we incorporated gray- and white-matter intensity which reflect local signal properties sampled at cortical surfaces or within subcortical masks. Corresponding normalized intensity measures were also included to reduce scanner- and site-related variability. Beyond these region-specific metrics, we leveraged additional T1- and T2-weighted intensity–based phenotypes provided in the ABCD tabulated datasets. These phenotypes summarize average signal intensities across cortical and subcortical regions, as well as global tissue classes, and include whole-brain “summation” measures that aggregate intensity information across gray matter, white matter, and cerebellar regions, along with contrasts between tissue types (e.g., gray–white matter differences). Compared with gray- and white-matter intensity measures derived directly from FreeSurfer, which index localized tissue signal at specific anatomical surfaces, the T1- and T2-based summaries are designed to

capture broader regional and global signal characteristics sensitive to tissue composition, including myelination, water content, and other microstructural properties [67,68]. We additionally included a global FreeSurfer summations comprising five whole-brain metrics: estimated intracranial volume, total cortical gray matter volume, total cortical white matter volume, total subcortical gray matter volume, and the ratio of brain segmentation volume to intracranial volume. Together, these diverse structural representations were selected to capture both localized and global aspects of brain structure and tissue composition, providing complementary information relevant to individual differences in cognitive development.

The Diffusion phenotype consisted of fractional anisotropy (FA) values extracted from 23 major white-matter tracts identified using AtlasTrack. FA provides an index of directional water diffusion and was included to capture complementary aspects of white-matter microstructural organization relevant to cognitive development [69].

Predictive modelling

Models were trained on one matched group [27] and tested on the other, ensuring that test participants were entirely independent of the training set.

All neuroimaging phenotypes and the cognitive scores 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 within each training set, with negative mean squared error (NMSE) as the evaluation metric.

Prediction models were developed in two frameworks:

Unimodal models (80 models total)

For each of the 80 neuroimaging phenotypes (derived from task-based fMRI, resting-state fMRI, sMRI, and DTI), we trained a separate predictive model using Partial Least Squares (PLS) regression, a supervised dimensionality reduction approach that maximizes covariance between neuroimaging features and the cognitive target [70]. The number of PLS components was tuned via grid search (30 components maximum; for phenotypes with fewer than 30 variables, all possible components were considered). PLS was selected not only for its ability to handle high-dimensional and highly collinear data (e.g., functional connectivity matrices), but also because prior analyses in our primary work demonstrated comparable predictive performance between PLS and Elastic Net regression in this dataset, and independent benchmarking study has shown Elastic Net to perform well in neuroimaging-based cognitive prediction [20]. Given this equivalence in performance, PLS was preferred

here due to its explicit dimensionality reduction, which yields compact latent representations while retaining variance most relevant to the cognitive outcome.

Multimodal (stacked) models (11 configurations)

To integrate information across phenotypes and capture potential non-linear interactions, we trained 11 multimodal models using late fusion [71]: predictions from the relevant unimodal PLS models were combined as inputs (rather than raw feature concatenation). Each configuration integrated predictions from related groups of unimodal models (e.g., “stacked all” used predictions from all 80 phenotypes; “stacked task contrasts” used predictions from task-based fMRI contrast phenotypes such as N-back, MID, and SST; see Supplementary Figure 1 for full list). The multimodal regressor was a Random Forest [72] with 1,000 trees. Hyperparameters were tuned via inner-fold grid search: maximum tree depth (None or 3–10), and number of features considered at each split (sqrt or log2 of total input features). Random Forest was selected for the multimodal layer due to its robustness to mixed-scale inputs, ability to model non-linear relationships and interactions between modalities, and native handling of missing values tested in previous studies [20,21,73]

We trained 368 models in total, spanning 91 neuroimaging phenotypes (80 unimodal, 11 multimodal) across 4 training strategies (All, RandWA-only, AA-only, Balanced). See 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 phenotype. 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 emphasize MAE as the primary metric for group-wise model comparison because it provides a direct and scale-preserving measure of prediction error that is readily interpretable at the individual level. MAE reflects absolute deviations between predicted and observed values and aligns with common practices in model comparison when assessing performance equity [74]. See Supplementary Information and 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 phenotype. Confidence intervals were estimated using nonparametric bootstrap resampling (1,000 iterations).

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 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, 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.

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.

Availability of Data and Materials: 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.

References

1. Sui J, Jiang R, Bustillo J, Calhoun V. Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biol Psychiatry 1969*. 88:818–282020;

2. Alhuwaydi AM. Exploring the Role of Artificial Intelligence in Mental Healthcare: Current Trends and Future Directions - A Narrative Review for a Comprehensive Insight. *Risk Manag Healthc Policy*. 2024; doi: 10.2147/RMHP.S461562.

3. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, et al.. Functional connectome fingerprinting: Identifying individuals based on patterns of brain connectivity. *Nat Neurosci*. 2015; doi: 10.1038/nn.4135.

4. Sripada C, Angstadt M, Rutherford S, Taxali A, Shedden K. Toward a “treadmill test” for cognition: Improved prediction of general cognitive ability from the task activated brain. *Hum Brain Mapp*. 2020; doi: 10.1002/hbm.25007.

5. Sripada C, Rutherford S, Angstadt M, Thompson WK, Luciana M, Weigard A, et al.. Prediction of neurocognition in youth from resting state fMRI. *Mol Psychiatry*. 2020; doi: 10.1038/s41380-019-0481-6.

6. Vieira BH, Pamplona GSP, Fachinello K, Silva AK, Foss MP, Salmon CEG. On the prediction of human intelligence from neuroimaging: A systematic review of methods and reporting. *Intell Norwood*. 2022; doi: 10.1016/j.intell.2022.101654.

7. Abramovitch A, Short T, Schweiger A. The C Factor: Cognitive dysfunction as a transdiagnostic dimension in psychopathology. *Clin Psychol Rev*. 2021; doi: 10.1016/j.cpr.2021.102007.

8. Carey S, Pang A, Kamps M de. Fairness in AI for healthcare. *Future Healthc J*. 2024; doi: 10.1016/j.fhj.2024.100177.

9. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. *PLOS Digit Health*. 2024; doi: 10.1371/journal.pdig.0000651.

10. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Npj Digit Med*. 2025; doi: 10.1038/s41746-025-01503-7.

11. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al.. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol*. 2024; doi: 10.1007/s11604-023-01474-3.

12. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al.. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun*. 2019; doi: 10.1038/s41467-019-11112-0.

13. Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, et al.. Principles and methods for transferring polygenic risk scores across global populations. *Nat Rev Genet*. 2024; doi: 10.1038/s41576-023-00637-2.

14. Wang Y, Tsuo K, Kanai M, Neale BM, Martin AR. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. Annu. Rev. Biomed. Data Sci. Annual Reviews;

15. Lewis ACF, Green RC. Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Med*. 2021; doi: 10.1186/s13073-021-00829-7.

16. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019; doi: 10.1038/s41588-019-0379-x.

17. Horien C, Noble S, Greene AS, Lee K, Barron DS, Gao S, et al.. A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nat Hum Behav*. 2021; doi: 10.1038/s41562-020-01005-4.

18. Li J, Bzdok D, Chen J, Tam A, Ooi LQR, Holmes AJ, et al.. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci Adv*. 2022; doi: 10.1126/sciadv.abj1812.

19. Adkins DJ, Hanson JL. Racial and Ethnic Disparities in Brain Age Algorithm Performance: Investigating Bias Across Six Popular Methods. *medRxiv*. 2025; doi: 10.1101/2025.09.18.25336117.

20. Pat N, Wang Y, Bartonicek A, Candia J, Stringaris A. Explainable machine learning approach to predict and explain the relationship between task-based fMRI and individual differences in cognition. *Cereb Cortex*. 2022; doi: 10.1093/cercor/bhac235.

21. Tetereva A, Li J, Deng JD, Stringaris A, Pat N. Capturing brain□cognition relationship: Integrating task□based fMRI across tasks markedly boosts prediction and test□retest reliability. *NeuroImage Orlando Fla*. 263:119588-2022;

22. Tetereva A, Knodt AR, Melzer TR, van der Vliet W, Gibson B, Hariri AR, et al.. Improving predictability, reliability, and generalizability of brain-wide associations for cognitive abilities via multimodal stacking. *PNAS Nexus*. 2025; doi: 10.1093/pnasnexus/pgaf175.

23. Engemann DA, Kozynets O, Sabbagh D, Lemaitre G, Varoquaux G, Liem F, et al.. Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *eLife*. 2020; doi: 10.7554/eLife.54055.

24. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, et al.. Reproducible brain-wide association studies require thousands of individuals. *Nat Lond*. 2022; doi: 10.1038/s41586-022-04492-9.

25. Rasero J, Sentis AI, Yeh FC, Verstynen T. Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *PLoS Comput Biol*. 2021; doi: 10.1371/journal.pcbi.1008347.

26. Rosenberg MD, Casey BJ, Holmes AJ. Prediction complements explanation in understanding the developing brain. *Nat Commun*. 2018; doi: 10.1038/s41467-018-02887-9.

27. Feczko E, Conan G, Marek S, Tervo-Clemmens B, Cordova M, Doyle O, et al.. Adolescent Brain Cognitive Development (ABCD) Community MRI Collection and Utilities. *bioRxiv*. Cold Spring Harbor Laboratory; 2021; doi: 10.1101/2021.07.09.451638.

28. Jernigan TL, Brown SA, Dale AM, Tapert S, Sowell ES, Herting M, et al.. Adolescent Brain Cognitive Development Study (ABCD) - Annual Release 5.1. Adolescent Brain Cognitive Development (ABCD) Study;

29. Bauer PJ, Zelazo PD. NIH TOOLBOX COGNITION BATTERY (CB): SUMMARY, CONCLUSIONS, AND IMPLICATIONS FOR COGNITIVE DEVELOPMENT. *Monogr Soc Res Child Dev*. 2013; doi: 10.1111/mono.12039.

30. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, et al.. A multi-modal parcellation of human cerebral cortex. *Nat Lond*. 2016; doi: 10.1038/nature18933.

31. Destrieux C, Fischl B, Dale A, Halgren E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage Orlando Fla*. 2010; doi: 10.1016/j.neuroimage.2010.06.010.

32. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al.. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023; doi: 10.1038/s41551-023-01056-8.

33. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform*. 2021; doi: 10.1016/j.jbi.2020.103621.

34. Xiang R, Kelemen M, Xu Y, Harris LW, Parkinson H, Inouye M, et al.. Recent advances in polygenic scores: translation, equitability, methods and FAIR tools. *Genome Med*. 2024; doi: 10.1186/s13073-024-01304-9.

35. Jao T, Chang C-Y, Li C-W, Chen D-Y, Wu E, Wu C-W, et al.. Development of NTU standard Chinese Brain Template: morphologic and functional comparison with MNI template using magnetic resonance imaging. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf*. 2009; doi: 10.1109/IEMBS.2009.5334211.

36. Liang P, Shi L, Chen N, Luo Y, Wang X, Liu K, et al.. Construction of brain atlases based on a multi-center MRI dataset of 2020 Chinese adults. *Sci Rep*. 2015; doi: 10.1038/srep18216.

37. Yang G, Zhou S, Bozek J, Dong H-M, Han M, Zuo X-N, et al.. Sample sizes and population differences in brain template construction. *NeuroImage*. 2020; doi: 10.1016/j.neuroimage.2019.116318.

38. Wang C, Ng B, Garbi R. Multimodal Brain Parcellation Based on Functional and Anatomical Connectivity. *Brain Connect*. 2018; doi: 10.1089/brain.2017.0576.

39. Chee MWL, Zheng H, Goh JOS, Park D, Sutton BP. Brain Structure in Young and Old East Asians and Westerners: Comparisons of Structural Volume and Cortical Thickness. *J Cogn Neurosci*. 2011; doi: 10.1162/jocn.2010.21513.

40. Tang Y, Zhao L, Lou Y, Shi Y, Fang R, Lin X, et al.. Brain structure differences between Chinese and Caucasian cohorts: A comprehensive morphometry study. *Hum Brain Mapp*. 2018; doi: 10.1002/hbm.23994.

41. Yang G, Bozek J, Noble S, Han M, Wu X, Xue M, et al.. Global diversity in individualized cortical network topography. *Cereb Cortex*. 2023; doi: 10.1093/cercor/bhad002.

42. Montreal Neurological Institute and Hospital. ICBM152 Nonlinear Atlases (MNI152 Template). McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University;

43. Owens MM, Hyatt CS, Gray JC, Carter NT, MacKillop J, Miller JD, et al.. Cortical morphometry of the five-factor model of personality: findings from the Human Connectome Project full sample. *Soc Cogn Affect Neurosci*. 2019; doi: 10.1093/scan/nsz017.

44. Peterson RE, Kuchenbaecker K, Walters RK, Chen C-Y, Popejoy AB, Periyasamy S, et al.. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*. 2019; doi: 10.1016/j.cell.2019.08.051.

45. Singhal P, Gokhale S, Shah A, Jain DK, Walambe R, Ekart A, et al.. Domain adaptation for bias mitigation in affective computing: use cases for facial emotion recognition and sentiment analysis systems. *Discov Appl Sci*. 2025; doi: 10.1007/s42452-025-06659-1.

46. Zhang BH, Lemoine B, Mitchell M. Mitigating Unwanted Biases with Adversarial Learning. *Proc 2018 AAAIACM Conf AI Ethics Soc*. New York, NY, USA: Association for Computing Machinery;

47. Jernigan TL, Brown SA, Dowling GJ. The Adolescent Brain Cognitive Development Study. *J Res Adolesc Off J Soc Res Adolesc*. 2018; doi: 10.1111/jora.12374.

48. ABCD-BIDS Community. ABCD-BIDS Community Collection (ABCC). NIMH Data Repositories;

49. Anokhin A, Luciana M, Banich M, Barch D, Bjork J, Gonzalez M, et al.. Age-related changes and longitudinal stability of individual differences in ABCD Neurocognition measures. *Dev Cogn Neurosci*. 2022; doi: 10.1016/j.dcn.2022.101078.

50. Heaton R, Akshoomoff N, Tulsky D, Mungas D, Weintraub S, Dikmen S, et al.. Reliability and Validity of Composite Scores from the NIH Toolbox Cognition Battery in Adults. *J Int Neuropsychol Soc*. 2014; doi: 10.1017/s1355617714000241.

51. Jonaitis E, Koscik R, Clark L, Yue, Betthauser T, Berman S, et al.. Measuring longitudinal cognition: Individual tests versus composites. *Alzheimers Dement Diagn Assess Dis Monit*. 2018; doi: 10.1016/j.dadm.2018.11.006.

52. Byington N, Grimsrud G, Mooney M, Cordova M, Doyle O, Hermosillo R, et al.. Polyneuro risk scores capture widely distributed connectivity patterns of cognition. *Dev Cogn Neurosci*. 2023; doi: 10.1016/j.dcn.2023.101231.

53. Pat N, Wang Y, Anney R, Riglin L, Thapar A, Stringaris A. Longitudinally stable, brain based predictive models mediate the relationships between childhood cognition and socio demographic, psychological and genetic factors. *Hum Brain Mapp*. 2022; doi: 10.1002/hbm.26027.

54. Thompson W, Barch D, Bjork J, Gonzalez R, Nagel B, Nixon S, et al.. The structure of cognition in 9 and 10 year-old children and associations with problem behaviors: Findings from the ABCD study's baseline neurocognitive battery. *Dev Cogn Neurosci*. 2018; doi: 10.1016/j.dcn.2018.12.004.

55. Ooi LQR, Chen J, Zhang S, Kong R, Tam A, Li J, et al.. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *NeuroImage*. 2022; doi: 10.1016/j.neuroimage.2022.119636.

56. Stenberg J, Karr J, Karlsen R, Skandsen T, Silverberg N, Iverson G. Examining Test-Retest Reliability and Reliable Change for Cognition Endpoints for the CENTER-TBI Neuropsychological Test Battery. *Front Neurol*. 2020; doi: 10.3389/fneur.2020.541533.

57. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al.. Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron Camb Mass*. 2002; doi: 10.1016/S0896-6273(02)00569-X.

58. Hagler DJ, Cornejo MD, Makowski C, Dick AS, Sutherland MT, Harms MP, et al.. Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage Orlando Fla*. 2019; doi: 10.1016/j.neuroimage.2019.116091.

59. Cohen AO, Dellarco DV, Breiner K, Helion C, Heller AS, Rahdar A, et al.. The Impact of Emotional States on Cognitive Control Circuitry and Function. *J Cogn Neurosci*. 2016; doi: 10.1162/jocn_a_00906.

60. Knutson B, Westdorp A, Kaiser E, Hommer D. FMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. *NeuroImage*. 2000; doi: https://doi.org/10.1006/nimg.2000.0593.

61. Logan GD. On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. 1994;

62. Nilearn contributors, Chamma A, Frau-Pascual A, Rothberg A, Abadie A, Abraham A, et al.. nilearn. Zenodo;

63. Gordon EM, Laumann TO, Adeyemo B, Huckins JF, Kelley WM, Petersen SE. Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cereb Cortex N Y N 1991*. 2016; doi: 10.1093/cercor/bhu239.

64. Elliott ML, Knodt AR, Cooke M, Kim MJ, Melzer TR, Keenan R, et al.. General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *NeuroImage Orlando Fla*. 2019; doi: 10.1016/j.neuroimage.2019.01.068.

65. Cole MW, Bassett DS, Power JD, Braver TS, Petersen SE. Intrinsic and Task-Evoked Network Architectures of the Human Brain. *Neuron Camb Mass*. 2014; doi: 10.1016/j.neuron.2014.05.014.

66. Cole MW, Ito T, Schultz D, Mill R, Chen R, Cocuzza C. Task activations produce spurious but systematic inflation of task functional connectivity estimates. *NeuroImage Orlando Fla*. 2019; doi: 10.1016/j.neuroimage.2018.12.054.

67. Jørgensen KN, Nerland S, Norbom LB, Doan NT, Nesvåg R, Mørch-Johnsen L, et al.. Increased MRI-based cortical grey/white-matter contrast in sensory and motor regions in schizophrenia and bipolar disorder. *Psychol Med*. 2016; doi: 10.1017/S0033291716000593.

68. Xu X, Jang I, Zhang J, Zhang M, Wang L, Ye G, et al.. Cortical gray to white matter signal intensity ratio as a sign of neurodegeneration and cognition independent of β-amyloid in dementia. *Hum Brain Mapp*. 2024; doi: https://doi.org/10.1002/hbm.26532.

69. Hagler Jr DJ, Ahmadi ME, Kuperman J, Holland D, McDonald CR, Halgren E, et al.. Automated white-matter tractography using a probabilistic diffusion tensor atlas: Application to temporal lobe epilepsy. *Hum Brain Mapp*. 2009; doi: 10.1002/hbm.20619.

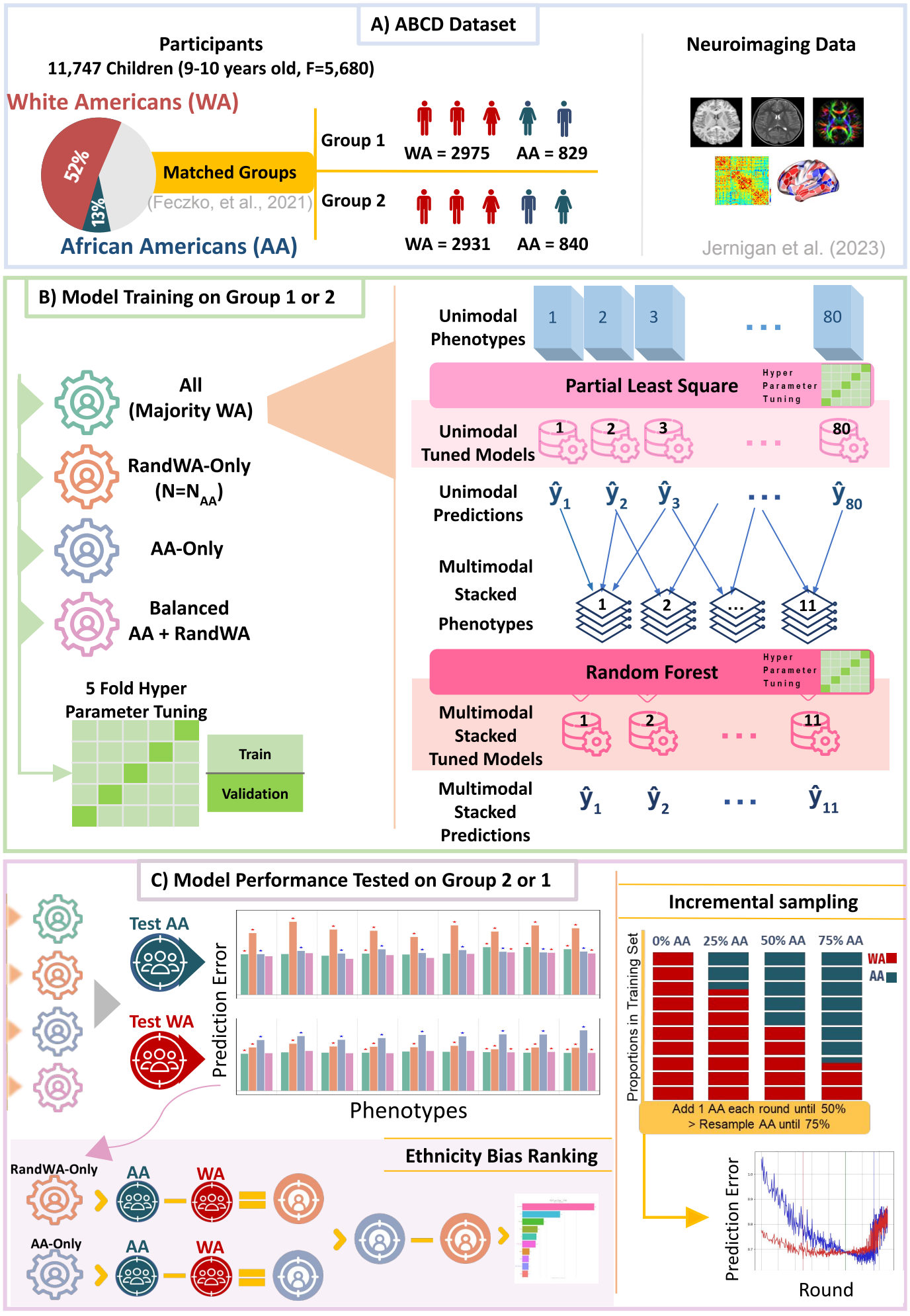
70. Krishnan A, Williams LJ, McIntosh AR, Abdi H. Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*. 2011; doi: https://doi.org/10.1016/j.neuroimage.2010.07.034.

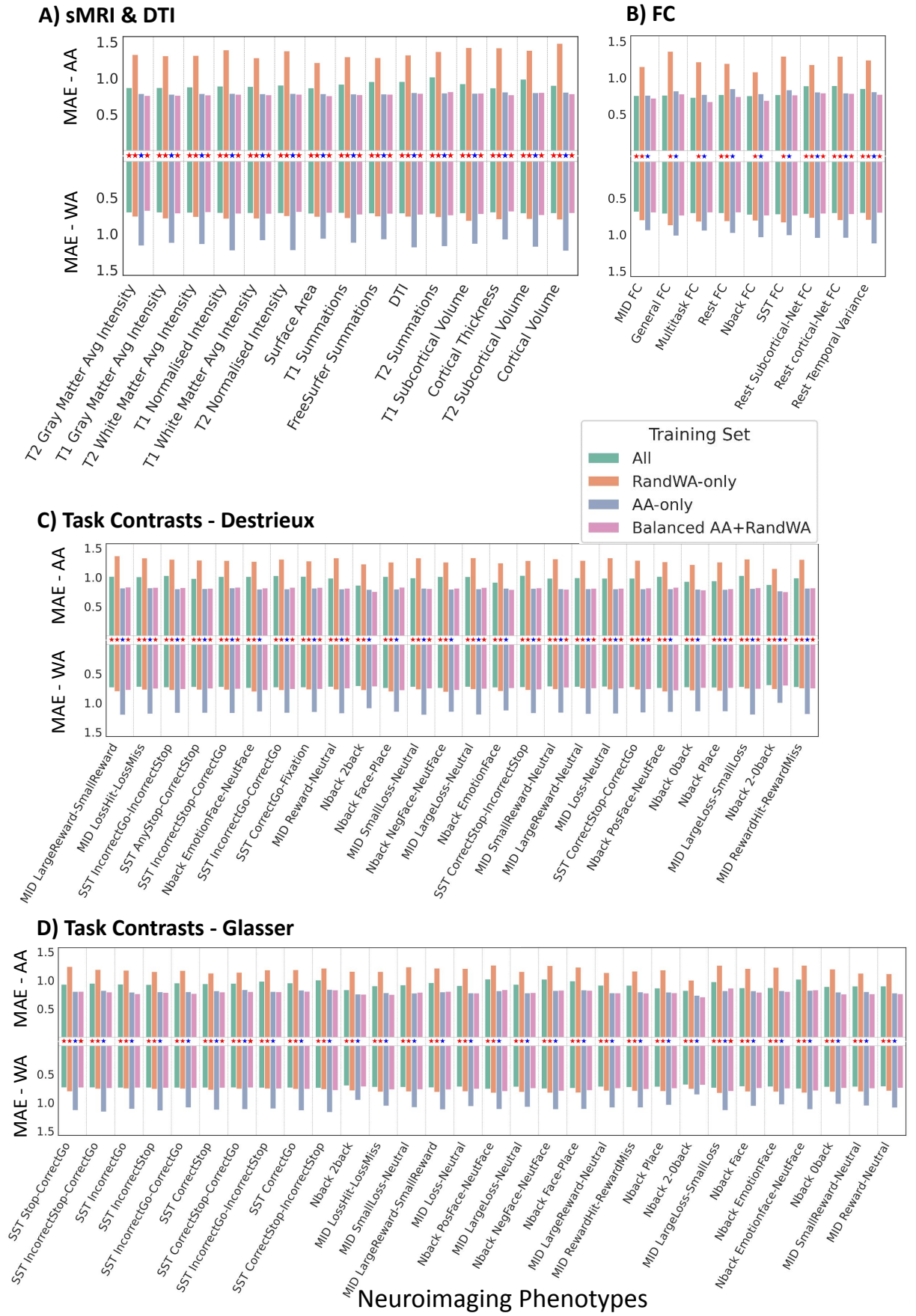
71. Li S, Tang H. Multimodal alignment and fusion: A survey. *ArXiv Prepr ArXiv241117040*. 2024;

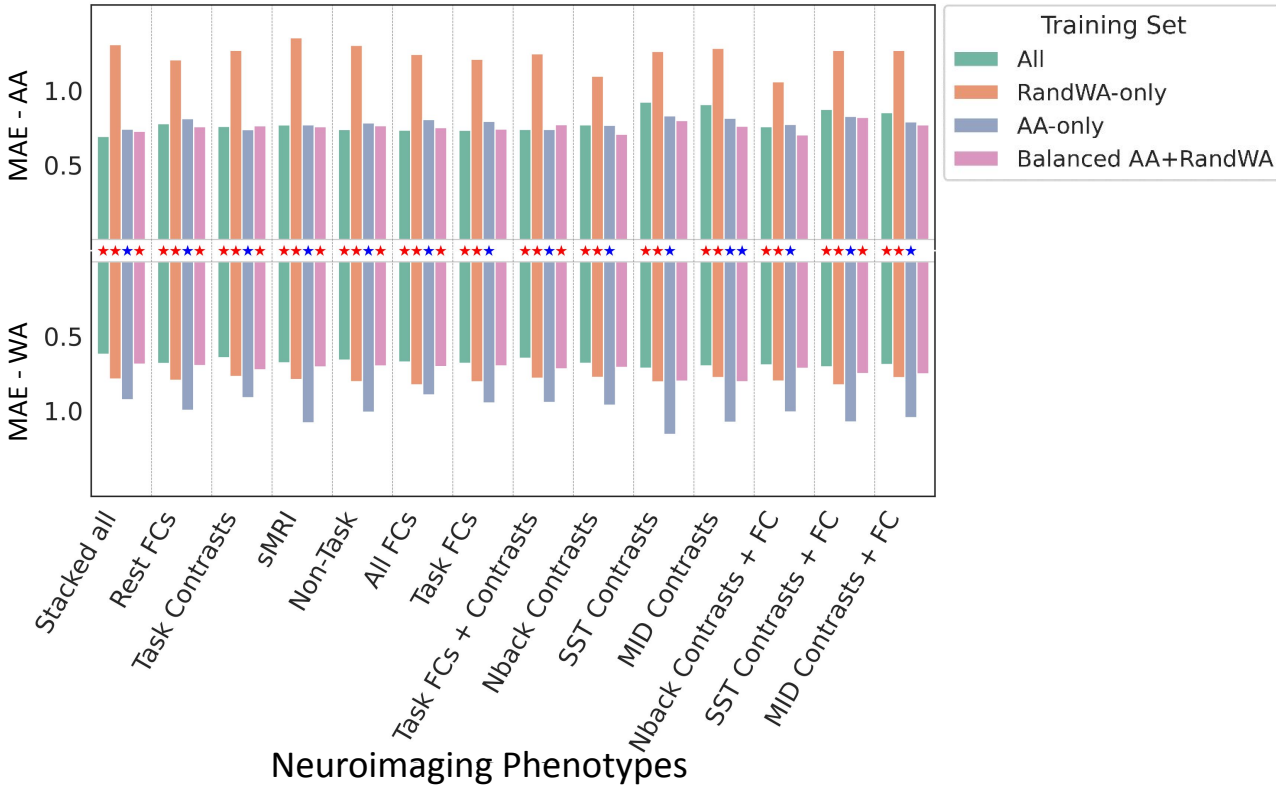
72. Breiman L. Random Forests. *Mach Learn*. 2001; doi: 10.1023/A:1010933404324.

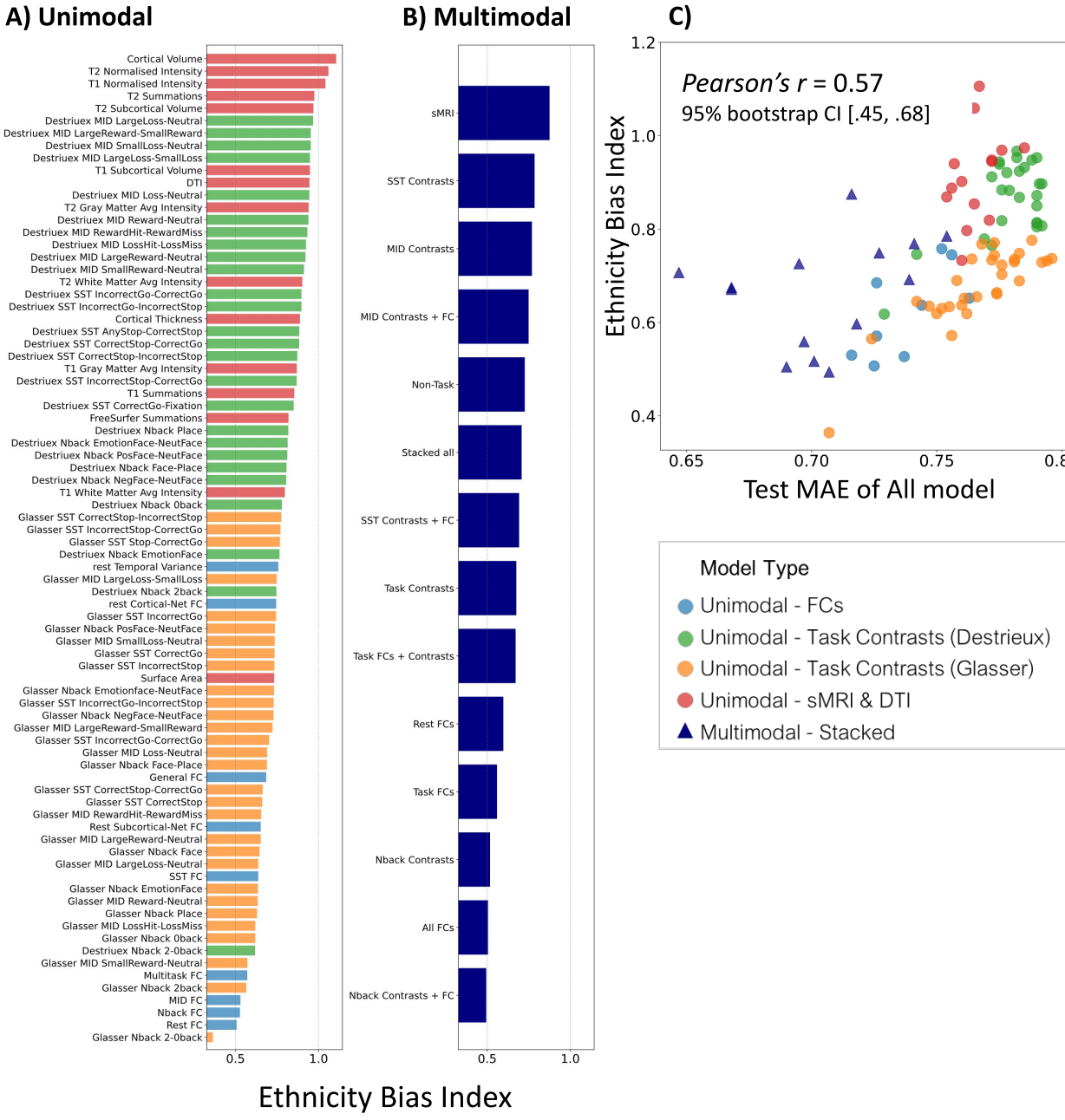
73. Zhu T. Analysis on the Applicability of the Random Forest. *J Phys Conf Ser*. 2020; doi: 10.1088/1742-6596/1607/1/012123.

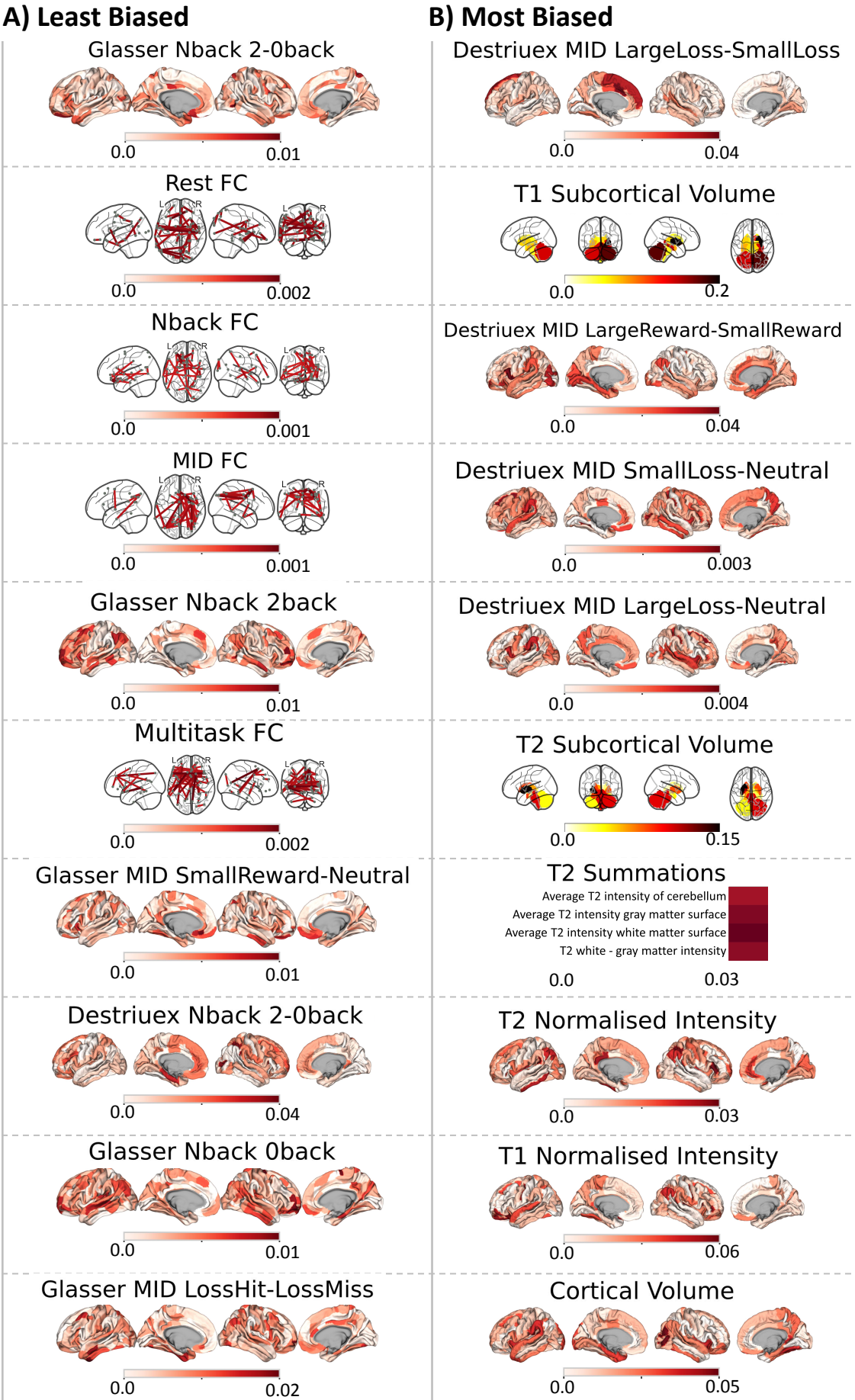
74. Gursoy F, Kakadiaris IA. Error parity fairness: Testing for group fairness in regression tasks. *ArXiv Prepr ArXiv220808279*. 2022;

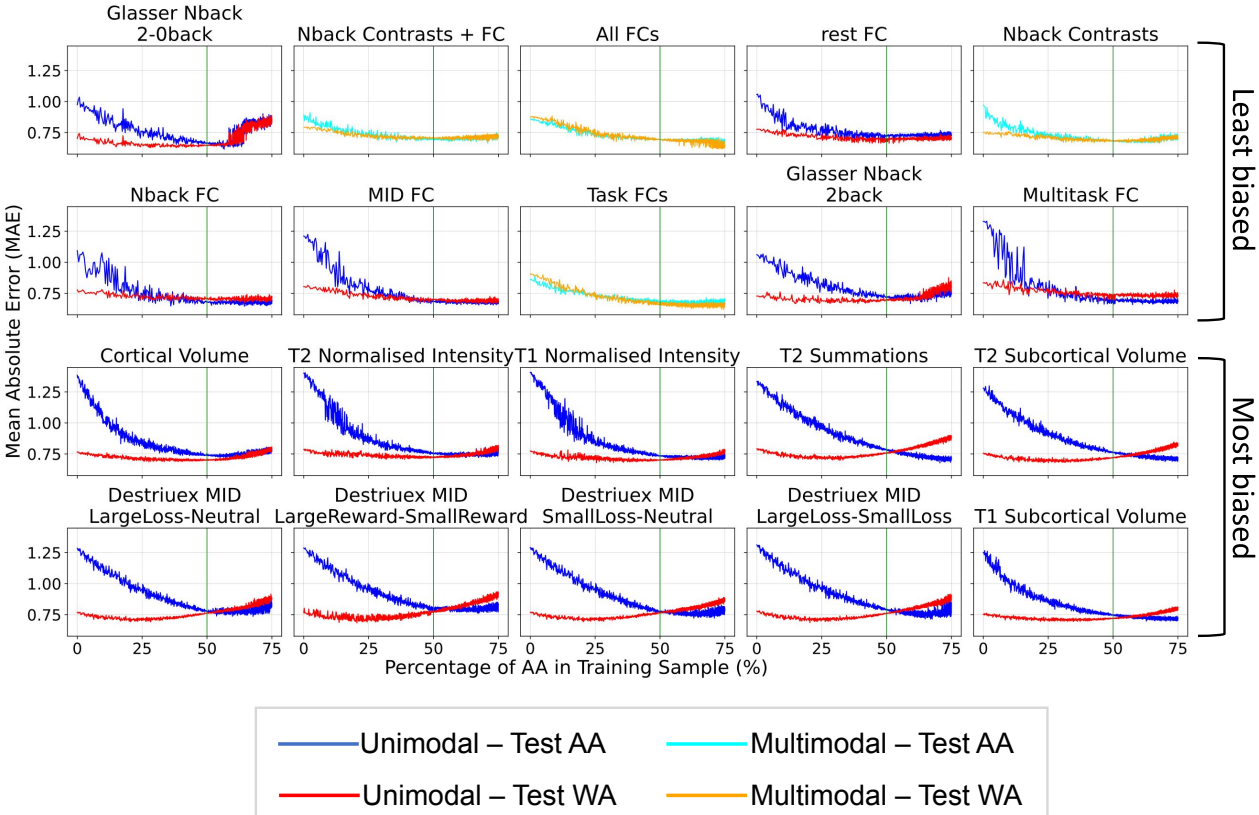

A) ABCD Dataset
Participants
11,747 Children (9-10 years old, F=5,680)
White Americans (WA)
52%
13%
African Americans (AA)
Matched Groups
(Feczko, et al., 2021)
Group 1
WA = 2975
AA = 829
Group 2
WA = 2931
AA = 840
Neuroimaging Data
Jernigan et al. (2023)
B) Model Training on Group 1 or 2
All
(Majority WA)
RandWA-Only
(N=N_AA)
AA-Only
Balanced
AA + RandWA
5 Fold Hyper
Parameter Tuning
Train
Validation
Unimodal
Phenotypes
1
2
3
80
Partial Least Square
Hyper
Parameter
Tuning
Unimodal
Tuned Models
Unimodal
Predictions
Multimodal
Stacked
Phenotypes
11
Random Forest
Multimodal
Stacked
Tuned Models
Multimodal
Stacked
Predictions
C) Model Performance Tested on Group 2 or 1
Test AA
Test WA
Prediction Error
Phenotypes
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 & DTI
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
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
T2 Summations
T2 Subcortical Volume
Destriuex MID LargeLoss-Neutral
Destriuex MID LargeReward-SmallReward
Destriuex MID SmallLoss-Neutral
Destriuex MID LargeLoss-SmallLoss
T1 Subcortical Volume
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
Glasser MID LargeLoss-Neutral
SST FC
Glasser Nback EmotionFace
Glasser MID Reward-Neutral
Glasser Nback Place
Glasser MID LossHit-LossMiss
Glasser Nback 0back
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)
Pearson's r = 0.57
95% bootstrap CI [.45, .68]
Ethnicity Bias Index
1.2
1.0
0.8
0.6
0.4
0.65
0.70
0.75
0.80
Test MAE of All model
Model Type
Unimodal - FCs
Unimodal - Task Contrasts (Destrieux)
Unimodal - Task Contrasts (Glasser)
Unimodal - sMRI & DTI
Multimodal - Stacked


A) Least Biased
Glasser Nback 2-0back
0.0
0.01
Rest FC
0.0
0.002
Nback FC
0.0
0.001
MID FC
0.0
0.001
Glasser Nback 2back
0.0
0.01
Multitask FC
0.0
0.002
Glasser MID SmallReward-Neutral
0.0
0.01
Destriuex Nback 2-0back
0.0
0.04
Glasser Nback 0back
0.0
0.01
Glasser MID LossHit-LossMiss
0.0
0.02
B) Most Biased
Destriuex MID LargeLoss-SmallLoss
0.0
0.04
T1 Subcortical Volume
0.0
0.2
Destriuex MID LargeReward-SmallReward
0.0
0.04
Destriuex MID SmallLoss-Neutral
0.0
0.003
Destriuex MID LargeLoss-Neutral
0.0
0.004
T2 Subcortical Volume
0.0
0.15
T2 Summations
Average T2 intensity of cerebellum
Average T2 intensity gray matter surface
Average T2 intensity white matter surface
T2 white - gray matter intensity
0.0
0.03
T2 Normalised Intensity
0.0
0.03
T1 Normalised Intensity
0.0
0.06
Cortical Volume
0.0
0.05


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
T2 Summations
T2 Subcortical Volume
Destriuex MID LargeLoss-Neutral
Destriuex MID LargeReward-SmallReward
Destriuex MID SmallLoss-Neutral
Destriuex MID LargeLoss-SmallLoss
T1 Subcortical Volume
Least biased
Most biased
Mean Absolute Error (MAE)
1.25
1.00
0.75
0
25
50
75
Percentage of AA in Training Sample (%)
Unimodal – Test AA
Multimodal – Test AA
Unimodal – Test WA
Multimodal – Test WA