To which reference class do you belong? Measuring racial fairness of reference classes with normative modeling

Saige Rutherford   SAIGE.RUTHERFORD@DONDERS.RU.NL
Department of Cognitive Neuroscience
Radboud University Medical Center
Nijmegen, the Netherlands

Thomas Wolfers   DR.THOMAS.WOLFERS@GMAIL.COM
Department of Psychiatry
University of Tuebingen
Tuebingen, Germany

Charlotte Fraza   CHARLOTTE.FRAZA@DONDERS.RU.NL
Department of Cognitive Neuroscience
Radboud University Medical Center
Nijmegen, the Netherlands

Nathaniel G. Harnett   NHARNETT@MCLEAN.HARVARD.EDU
Department of Psychiatry
Harvard Medical School
Boston, MA, USA

Christian F. Beckmann   CHRISTIAN.BECKMANN@DONDERS.RU.NL
Department of Cognitive Neuroscience
Radboud University Medical Center
Nijmegen, the Netherlands

Henricus G. Ruhe   ERIC.RUHE@RADBOUDUMC.NL
Department of Psychiatry
Radboud University Medical Center
Nijmegen, the Netherlands

Andre F. Marquand   ANDRE.MARQUAND@DONDERS.RU.NL
Department of Cognitive Neuroscience
Radboud University Medical Center
Nijmegen, the Netherlands

Abstract

Reference classes in healthcare establish healthy norms, such as pediatric growth charts of height and weight, and are used to chart deviations from these norms which represent potential clinical risk. How the demographics of the reference class influence clinical interpretation of deviations is unknown. Using normative modeling, a method for building reference classes, we evaluate the fairness (racial bias) in reference models of structural brain images that are widely used in psychiatry and neurology. We test whether including "race" in the model creates fairer models. We predict self-reported race using the deviation scores from three different reference class normative models, to better understand bias in an integrated, multivariate sense. Across all of these tasks, we uncover racial disparities

that are not easily addressed with existing data or commonly used modeling techniques. Our work suggests that deviations from the norm could be due to demographic mismatch with the reference class, and assigning clinical meaning to these deviations should be done with caution. Our approach also suggests that acquiring more representative samples is an urgent research priority.

1. Introduction

Reference classes can be used to define health and disease in medicine and also to estimate patient risk. Determining the probability of a patient possessing a particular attribute, such as a disease or prognosis, is critical for risk screening and treatment planning. Through direct inference (that is, making inferences about the individual from the population Thorn (2012), probabilities for individual patients can be adjusted based on estimates derived from reference classes to which the patient belongs. Determining which reference class is most useful can be challenging when a patient fits into multiple reference classes, each with differing risk probabilities. This is known as the reference class problem Hájek (2007); Wallmann and Williamson (2017); Venn (1888); Reichenbach (1949). Due to the difficult nature of the reference class problem, most demographic information (excluding perhaps age and sex) is ignored when calculating clinical risk. While ignoring the reference class problem may allow easier group identification, doing so can also lead to worse (i.e., less accurate, more disparate[1]) risk prediction in high-stake settings Suriyakumar et al. (2023); Khor et al. (2023); Pfohl et al. (2021); Zink et al. (2023). For example, consider two hypothetical patients with renal cell carcinoma with brain metastases, John, who is a 34-year-old, white male, unmarried with no family close by, living in rural Michigan, low-income, exercises regularly, non-drinker, who occasionally smokes, and Jim, who is a 68-year-old black male, married, living in New York City, high-income, does not exercise regularly, drinks heavily, and does not smoke. Each of these demographics creates its own reference class that may contradict each other (e.g., regular exercise and smoking). When calculating the chance of survival for each man, the Graded Prognostic Assessment tool is used Sperduto et al. (2020). Even if both men have the same score, their respective reference classes (socioeconomic status, family support, exercise levels, alcohol intake, smoker/non-smoker) will likely influence their survival probabilities.

Normative modeling is a framework for building reference class models and is, therefore, ideal for studying the reference class problem. Normative modeling has been applied across many healthcare contexts, including the most well-known use case in pediatrics - growth charting of height, weight, and head circumference Borghi et al. (2006). Fields of medicine that involve heterogeneous disease categories and complex measurements of biology, such as psychiatry Wolfers et al. (2018); Zabihi et al. (2019); Lv et al. (2020); Elad et al. (2021) and neurology Bhome et al. (2023); Verdi et al. (2021, 2023); Itälinna et al. (2022), have begun to use normative modeling to move away from standard case vs. control methods which mainly consider group mean effects Marquand et al. (2016, 2019); Rutherford et al. (2022b); Bethlehem et al. (2022) towards individualized, or precision, medicine. In addition to clinical applications, there has been considerable methodological improvements in nor-

---

1. Throughout this work, we use the terms fairness, bias, and disparity interchangeably. We consider them all to refer to testing for equal model performance across different self-reported racial categories.

mative modeling, including cross-sectional vs. longitudinal modeling Bučková et al. (2024); Di Biase et al. (2023), handling site effects Bayer et al. (2021); Kia et al. (2022, 2021), predictive modeling using outputs of normative models Rutherford et al. (2023), heterogeneity quantification Nunes et al. (2020a,b,c), extreme value statistics Fraza et al. (2022), and modeling non-Gaussianity Boer et al. (2022).

An important concept in normative modeling is untangling various sources of heterogeneity (variability) within the population. Normative models aim to distinguish between 'healthy' and clinical variation, as well as other factors like demographic or confounding variables (e.g., site/scanner). Typically, 'healthy' variability is assumed to fall within the 95th percentiles, while deviations associated with illness are assumed to fall within the outer percentiles (Figure 1A). However, it is also possible that deviations in the outer percentiles could relate to demographics rather than being clinically relevant, which is usually not empirically tested. Despite the increasing use of normative modeling in both machine learning and healthcare domains, the fairness implications of these models have not been thoroughly investigated. Considerations for determining who should be included in the reference group often revolve around data availability and increasing sample sizes. However, an essential unresolved question is how the demographics of this reference class affect the interpretation of deviation scores. Before addressing concerns regarding fairness in reference classes as operationalized in normative models, it is important to first quantify them.

Generalizable Insights about Machine Learning in the Context of Healthcare

- We first quantify the racial bias in existing (pre-trained) normative models of structural brain images that are widely used in psychiatry and neurology Rutherford et al. (2022a). Racial demographics of the reference class used to train these existing normative models are unknown because many of the samples used for the training set did not collect or share race or ethnicity data. We probe these models, using a sample where race is known, by summarizing deviations and residual errors across Asian, Black, and White individuals and testing for group differences (Figure 1A-C).

- Next, we train new normative models using reference class data where race information is available during training. In this setting, we test whether including self-reported race variables into the normative model in an identical cohort (race included and race not included) creates fairer models, by summarizing deviations and residual errors across Asian, Black, and White individuals and testing for group differences (Figure 1A-C).

- Finally, we use the deviation scores from three different reference class normative models (pre-trained, race included, race not included) to predict self-reported race. Normative models are fit in a univariate manner, meaning one brain region is predicted per model, and there are models fit across many brain regions. The race prediction uses deviation scores from all brain regions combined into the same model to better understand bias in an integrated, multivariate sense (Figure 1D).
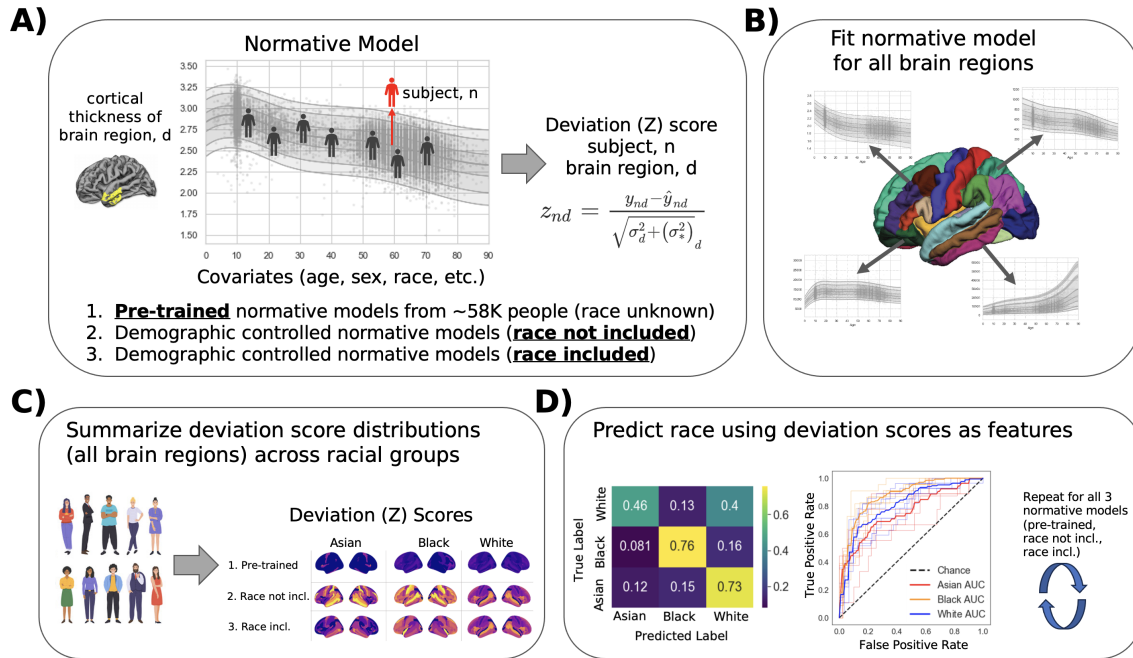
Figure 1: Overview of analysis workflow. A) Normative models of brain structure were used to generate deviation scores. Three normative models were fit (pre-trained, race not included, and race included) representing two different reference classes and two sets of covariates. B) Normative models were estimated for all regions in the Destrieux atlas Destrieux et al. (2010), a commonly used anatomical brain parcellation. C) The effect of self-reported race on the distribution of normative modeling deviation scores was quantified across all three normative models. D) Self-reported race was predicted using normative modeling deviation scores as features.

2. Background and Problem Formulation

2.1. Fairness in Machine Learning for Healthcare Setting

Prior work on fairness in machine learning considers three criteria: independence, separation, and sufficiency Barocas et al. (2023). For independence (often called demographic parity), the predictions should be statistically independent of the sensitive attribute (e.g. race) Corbett-Davies et al. (2017); Kamiran and Calders (2009). In separation (often called equalized odds), sensitive attributes are statistically independent of the prediction given the ground truth Hardt et al. (2016). For sufficiency (often called predictive parity), sensitive attributes are statistically independent of the ground truth given the prediction (similar to group calibration) Berk et al. (2017); Chouldechova (2016). There are also three stages of the machine learning lifecycle where fairness concerns are often addressed: pre-processing, training time, and post-processing. Data acquisition is also an obvious source of bias. However, this stage is often disregarded as it is considered outside control of the machine learning community who are typically performing secondary data analysis and not directly involved in data collection Paullada et al. (2021); Sambasivan et al. (2021); Roh et al. (2021). In

pre-processing, the training data are modified such that the information correlated to the sensitive attribute is removed. For training time fairness, there is typically some penalty added to the loss function (e.g., joint optimization of a fairness metric and the loss function). Post-processing fairness is often performed by applying a transformation to the model output to reduce unfairness. There are numerous fairness metrics used in machine learning for healthcare and they vary depending on the context. There are too many fairness metrics to review in this work but we refer to recent work on this topic in the classification Franklin et al. (2022) and regression setting Franklin et al. (2023).

Across all of these definitions of fairness and implementation stages, it becomes clear that much of the research on fairness in machine learning for healthcare is trying to solve fairness by eliminating bias with some new method or metric. An underlying assumption seems to be that models can be made perfectly fair with the right optimization or algorithm. Similarly, research on the reference class problem is also often focused on developing more "correct" reference classes through various approaches such as the digital twin Turab and Jamil (2023); Bruynseels et al. (2018) or improved modeling of overlapping probabilities Wallmann and Williamson (2017). Less attention is given to simply measuring rather than eliminating bias in widely used datasets, reference classes, and predictive models. Quantifying bias in predictive healthcare models rather than focusing on eliminating it might be as straightforward as creating guidelines for their use based on performance across different groups.

2.1.1. Racial Fairness

When considering the fairness of machine learning models with respect to racial identity, extensive research has been conducted on challenging the inclusion of race as a predictor (race corrections[2] Vyas et al. (2020); Cerdeña et al. (2020); Borrell et al. (2021); Ioannidis et al. (2021); Khor et al. (2023); Yang et al. (2023); Suriyakumar et al. (2023). Racial corrections in medicine can be traced back to the practice of using white male bodies as the reference, or norm, against which other bodies and physiological functions are measured. Due to biased data collection, the white male norm principally reflects data availability, as has been the case in many biomedical domains (e.g. genetics or brain imaging), but it does not mean that this reference class should be the gold standard of health against which all others are measured. The consensus on race correction suggests avoiding a binary approach (always perform or always avoid), but rather in every context carefully considering the rationale behind its use to filter out scientifically debunked beliefs about biological differences between racial groups Roberts (2021), and calls for further research into the multifaceted interactions among genetic ancestry, race, racism, socioeconomic status, and environmental factors Vyas et al. (2020). **It is important to explicitly state that race is not simply a phenotype related to skin color but is necessarily concomitant with a complex and interactive set of influences from society and the environment.**

There has been important work documenting existing racial bias in clinical predictive models, for example in explainable AI methods Balagopalan et al. (2022), clinical record

---

2. We use "race correction" as it is standard terminology Ioannidis et al. (2021); Vyas et al. (2020), however, we disagree with this name because it is impossible to correct properly if you have insufficient data to define an alternative reference class.

de-identification Xiao et al. (2024), and underdiagnosis in chest X-ray pathology Seyyed-Kalantari et al. (2021). Studies have shown that race can be predicted from clinical notes Adam et al. (2022) and medical images (chest x-rays, chest CTs, mammograms, and spine x-rays) Gichoya et al. (2022). In work that directly predicted race from medical images, detection of race was not due to proxies for race (body-mass index, disease distribution, breast density) and race could be predicted from all anatomical regions and frequency spectrums of the images Gichoya et al. (2022). Another topic of work has looked at why racial disparities exist in clinical predictive models. This includes studying label misspecification Pierson et al. (2021); Obermeyer et al. (2019), undertesting rates of certain minority groups Chang et al. (2022); Wu et al. (2023); Akpinar et al. (2024), and testing whether disparities are due to differences in outcome frequency, feature distributions or feature-outcome relationships Movva et al. (2023). The work on underdiagnosis in chest X-ray pathology Seyyed-Kalantari et al. (2021) also opened further conversations on understanding the sources of racial bias, including population, prevalence, and annotation shifts Bernhardt et al. (2022); Mukherjee et al. (2022); Seyyed-Kalantari et al. (2022).

2.2. Normative Model Setting

The normative model setting is a natural approach for studying fairness of the reference class problem because it can be seen as a statistical model of a particular reference class. When selecting a specific normative model, we essentially define a concrete reference group, which can then be used to quantitatively evaluate fairness. It is nevertheless somewhat distinctive from many machine learning in healthcare settings. Therefore, we briefly describe normative modeling at a conceptual level and in section 4.1, we formulate the normative model mathematically. A normative model is most often used to model population variation in biological measurements across a particular reference class. For example, pediatric growth charts are a well-known example that plot centiles of variation in height or weight as a function of age. Normative modeling can also be thought of as a feature generating step (i.e., creating features for downstream clinical decisions) whereas most machine learning in healthcare models operate directly at the clinical decision-making step (e.g., predicting clinical outcomes such as hospitalization, critical outcomes (ICU transfer), or re-admission Movva et al. (2023); Xie et al. (2022). However, it is important to recognize that if the features are biased, then it is likely that downstream algorithms will be too Chang et al. (2022). More concretely, we can employ a pre-trained predictive regression model (Figure 1A) estimated on the basis of a set of binary and continuous features, X_{nd} from n subjects, to predict a continuous response variable, y_d, representing a biological measurement (the thickness of the cerebral cortex in brain region d). In the setting of neuroimaging, many normative models are fit for each biological feature (e.g. for each anatomical brain region derived from brain imaging). The model is typically a population reference model that pools data from many different study sites. Covariates X_{nd} are variables that are considered to be important in predicting the biological response variable, yet in practice are constrained to include only the overlapping available variables across sites (typically age and sex for most neuroimaging studies). A central goal of these 'population' reference class normative models is to be openly shared and to be able to transfer the models to new sites or datasets (e.g. which might contain individuals with a given medical condition). At some of these

transfer sites, there are additional variables relating to demographics (race) that we would like to use to test the model's fairness across different racial groups (Figure 1C). These demographic variables are only available during model transfer and are typically inaccessible before or during training, because, surprisingly, many of the samples that were pooled for training did not collect or share race or ethnicity data. The outputs of fitting a normative model are statistics that measure the overall fit of the model and a set of deviation scores (Z_{nd}-scores), which is as an expression of how far a given person (e.g. a patient) deviates from the reference class population mean (Figure 1A) and is often interpreted as risk of an adverse clinical event Alexander-Bloch et al. (2014); Marquand et al. (2019). Clinical labels are often used to make statements such as 'patients in group A have a greater number of extreme deviations than controls.' Lv et al. (2020); Pinaya et al. (2021).

3. Cohort

3.1. Cohort Selection and Inclusion Criteria

We used data from two publicly available datasets of neuroimaging and phenotypic data for our experiments, namely the Human Connectome Project (HCP) Glasser et al. (2013) and the UK Biobank (UKB) Alfaro-Almagro et al. (2018). Inclusion criteria for all samples used in our analysis participants having basic demographic information (age, sex, and race) and a high quality (defined in section 3.3) T1-weighted MRI volume. Importantly, both HCP and UKB collected self-reported race variables. Due to the low number of people in some racial groups, we combined UKB categories into Asian (Indian, Pakistani, Asian or Asian British, Bangladeshi, Chinese), Black (Caribbean, African, Black or Black British), and White (White, British, Irish). In HCP, categories included: White, Black or African American, Native American/Native Alaskan, and Asian/Native Hawaiian/Other Pacific Islander. The HCP Native American/Native Alaskan category was not large enough to stand alone as a category, did not fit into other categories, and was not included in our analyses. We did not include UKB or HCP subjects with unknown, unreported, or mixed race. We acknowledge that the Asian group is heterogeneous and quite different across HCP and UKB samples. UKB is more representative of South Asian population (i.e., India, Bangladeshi), whereas HCP is likely more representative of east Asian populations (i.e., Chinese). We also acknowledge that these coarse representations of race are not ideal and we expand on this further in the discussion. Sample demographics of each data set are described in Tables 1 and 2.

3.2. Train/test split - Normative models

Slightly different train/test splits were used for the pre-trained models (Table 1) and the normative models trained in this work (race-included and race-not-included, Table 2). For the pre-trained models, we followed the train/test split of the original paper (split-half, stratified by site) Rutherford et al. (2022a). For the race-included and race-not-included models, after learning of the racial bias in the pre-trained models, we had a hypothesis that the pre-trained models' training set was primarily composed of white individuals. However, because we did not have access to race information for individuals in the training set of the pre-trained models, we could not assess this. This is when we decided to train our

Table 1: Test set demographics of pre-trained normative models. Demographics of the training set are unknown because many of the samples that were pooled for the training set did not collect or share race or ethnicity data. A = Asian, B = Black, W = White.

	HCP test	UKB test
N	533	13416
Sex (F%, M%)	53.7%, 46.3%	52.2%, 47.8%
Age (M, S.D)	28.9, 3.6	63.5, 7.5
Race (A%, B%, W%)	5.5%, 15%, 79.5%	1%, 1%, 98%

Table 2: Train and test set demographics of race not included, and race included normative models. Models were fit separately for HCP and UKB data sets. The train/test split was the same across the race not included and the race-included models. A = Asian, B = Black, W = White.

	HCP train	HCP test	UKB train	UKB test
N	710	353	29553	1472
Sex (F%, M%)	54%, 46%	56%, 44%	53%, 47%	50%, 50%
Age (M, S.D)	29.0, 3.5	28.6, 4.0	63.7, 7.5	62.2, 7.8
Race (A%, B%, W%)	2%, 5%, 93%	15%, 38%, 47%	0.2%, 0.2%, 99.6%	19%, 19%, 62%

own models in samples where we had access to self-reported race for all individuals in the train/test set. We wanted to test if having a primarily white training set led to the same racial bias we observed in the pre-trained models. So, we stratified the train/test split (Table 2) on race and only included a very small number of Black and Asian individuals in the training set. Due to the very low sample size of Black and Asian individuals in HCP and UKB, we wanted to have most of these people in the test set to increase our power. However, to learn the race term during training, we had to include a small subset of these groups in the training sets (2% Asian, 5% Black in HCP and 0.2% Asian and 0.2% Black in UKB).

3.3. Feature Extraction

Freesurfer image analysis software (version 6.0) was used to reconstruct the volumetric MRI neuroimages into a surface (Figure 1B), which better represents the brain's gyri and sulci folding patterns. The boundary between white matter and cortical gray matter is mapped during Freesurfer's surface reconstruction which allows cortical thickness (measured in millimeters) to be calculated. Cortical thickness values were extracted for all brain regions in the Destrieux parcellation Destrieux et al. (2010), a commonly used brain atlas spanning 148 brain regions (74 left hemisphere, 74 right hemisphere). Further technical details of these procedures are described in prior publications Fischl and Dale (2000); Fischl et al. (2002). We adopted an automated quality control procedure that quantifies image quality based on the Freesurfer Euler Characteristic, which has been shown to be an excellent proxy for manual labeling of scan quality Monereo-Sánchez et al. (2021); Rosen et al. (2018) and is the most important feature in automated scan quality classifiers Klapwijk et al. (2019).

4. Methods

4.1. Normative Model Estimation

Normative models were estimated using Bayesian Linear Regression with likelihood warping to predict cortical thickness from a vector of covariates Fraza et al. (2021). For each d brain region of interest, y_d is predicted as:

$$y_d = w^T \phi(x) + \epsilon \tag{1}$$

Where w^T is the estimated weight vector, $\phi(x)$ is a basis expansion of the of covariate vector x, consisting of a B-spline basis expansion (cubic spline with 5 evenly spaced knots) to model non-linear effects of age, and $\epsilon = \eta(0, \beta)$ a Gaussian noise distribution with mean zero and noise precision term β (the inverse variance). A likelihood warping approach Rios and Tobar (2019); Snelson et al. (2003) was used to model non-Gaussian effects. This involves applying a bijective nonlinear warping function to the non-Gaussian response variables to map them to a Gaussian latent space where inference can be performed in closed form. We used a 'sinarcsinsh' warping function, equivalent to the SHASH distribution used in the generalized additive modeling literature Jones and Pewsey (2009). A fast numerical optimization algorithm was used to optimize hyperparameters (L-BFGS) and computational complexity of hyperparameter optimization was controlled by minimizing the negative log likelihood. Deviation scores (Z-scores) are calculated in the latent Gaussian space for the n_{th} subject, and d_{th} brain area, in the test set as:

$$Z_{nd} = \frac{y_{nd} - \hat{y}_{nd}}{\sqrt{\sigma_d^2 + (\sigma_*^2)_d}} \tag{2}$$

Where y_{nd} is the true response, $\hat{y}_nd$ is the predicted mean, σ_d^2 is the estimated noise variance (reflecting uncertainty in the data), and $(\sigma_*^2)_d$ is the variance attributed to modeling uncertainty.

Error is calculated as the model residual:

$$E_{nd} = y_{nd} - \hat{y}_{nd} \tag{3}$$

Model fit was evaluated by explained variance, mean squared log-loss, skew, and kurtosis (Appendix Figure 5, Appendix Figure 6).

Three normative models were used in this work, all of them following the same modeling and evaluation procedure described above but differing in the training and testing data. First, we wanted to quantify racial bias in existing normative models that are publicly available and are currently being used in practice in the fields of psychiatry and neurology. These pre-trained models were trained on large (58K people) data sets and limited demographic information was used during training (age and sex). The racial demographics of the reference class used in the pre-trained models are unknown. Next, to study the effects of race in normative models more carefully, we trained two normative models using a sample where the racial demographics are known. We trained one model with race included as 'one hot' dummy variable encoded predictors and one model without race included. The training and testing samples were held constant across both models, the only difference is

that one model included race as a predictor, and one did not. To summarize, we wanted to study the effects of race in three different settings, namely: (i) where we first transfer from a large pre-trained model derived from a sample where we cannot determine the demographic distribution (ii) train from scratch in a more carefully controlled subsample where racial information is present and (iii) train from scratch but additionally accounting for the effect of race during training using the same data as and training/test split used in (ii). The differences between these three normative models are summarized in Table 3. All analysis code is available on GitHub.[3]

Table 3: Summary of Normative Models.

Model Name	Model Equation	Training Data	Testing Data
Pre-trained	Y (brain region) = Age + Sex + Site	~58,000 people, race is unknown.	HCP & UKB (table 1), race is known and used post-hoc to quantify bias.
Race not included	Y (brain region) = Age + Sex	HCP & UKB, race is known but not included.	HCP & UKB (table 2), race is known and used post-hoc to quantify bias.
Race included	Y (brain region) = Age + Sex + Race	HCP & UKB, race is known and included.	HCP & UKB (table 2), race is known and used post-hoc to quantify bias.

4.2. Evaluating Fairness of Normative Models

4.2.1. QUALITATIVE EVALUATION: SUMMARIZING MODELS FOR EACH RACIAL GROUP

To evaluate the normative models in different racial groups, we calculated the average deviation score at every brain region for each group (Figure 2A). We summarize patterns of extreme deviation ($|Z| > 2$) for each group by counting how many subjects had an extreme deviation ($|Z| > 2$) at a given brain region and dividing by the group size to show the percentage of individuals with extreme deviations at that brain area (Figure 2B).

4.2.2. QUANTITATIVE EVALUATION: TESTING FOR GROUP DIFFERENCES

To test for statistically significant group difference in the normative models, we performed group difference testing on the deviation scores and residual errors, thresholding the results at a Benjamini-Hochberg Hochberg and Benjamini (1990) false discovery rate (FDR) of $p < 0.05$ to correct for multiple comparisons across all brain regions (Figure 3, Table 4).

4.3. Predicting Race

To further test if race is encoded in the normative models, we used deviation scores as features (separately for each normative model described in Table 3) to predict self-reported race using logistic regression one vs. rest framework (i.e., Black vs. White + Asian, Asian vs. White + Black, and White vs. Asian + Black). For this, we used penalized logistic

---

3. https://github.com/saigerutherford/nm_demographics/tree/main

regression implemented in scikit-learn Pedregosa et al. (2011) with default settings (L2 penalty, L-BFGS solver).

Data (Tables 1 and 2) were divided into training and testing, stratified by race (so the test sets had approximately the same number of Asian, Black, White individuals), using an 80/20 split and generalization was assessed with 5-fold cross validation. Area under the curve, precision, recall, and F-score were determined for each class (Table 5), as well as visualizing the receiver operating characteristic curves and confusion matrices (Figure 4). All metrics were averaged across folds and the mean and standard deviation are reported (Table 5).

4.4. Fairness Metrics

We use the concept of demographic parity as our main fairness metric. Parity is often used in decision making binary classification settings (e.g., did you get the loan?) In these yes or no settings, there is a "positive" outcome (got the loan) and parity states that all groups should receive the positive outcome at equal rates. In this work, we have (i) a regression problem (the normative model) and (ii) a multi-class classification problem that directly predicts the sensitive attribute, thus there is no decision or positive outcome. Therefore, we use slightly different notions of parity in our assessments. We define parity as the normative model providing equal performance for all groups. Note that in the settings where this is not met, the resulting bias (e.g. a poorer fit in some groups than in others) will simply be propagated to the downstream analysis step. In the classification step, we define parity as equal classification performance across racial groups.

5. Results

5.1. Racial bias is present in normative models

5.1.1. QUALITATIVE EVALUATION OF RACIAL BIAS

The average deviation summaries (Figure 2A) measure the average deviation across each racial group. These show that white individuals (i.e. the majority class) are centered around zero across all reference class normative model (and across the whole brain) in both HCP and UKB datasets. The pre-trained average deviations show that for Asian individuals and Black individuals' cortical thickness is overestimated, meaning there are negative deviation scores across most of the brain, with the exception of the prefrontal cortex in HCP Black individuals, UKB Black individuals, and UKB Asian individuals, and in the UKB Asian individuals' temporal and visual areas. The 'race not included' average deviations follow a similar pattern as shown in the pre-trained average deviations, although there is somewhat stronger overestimation (i.e. more negative deviations) of Asian individuals and Black individuals in the race not included models. When race is included, the overestimation of HCP Asian individuals, HCP Black individuals, and UKB Black individuals disappears. This is replaced by an underestimation of cortical thickness in HCP Asian individuals, HCP Black individuals, and UKB Black individuals in temporal and visual areas, but mostly the models are centered around zero and these group differences, for the most part, disappear. In the UKB Asian individuals, an interesting phenomenon occurs where the patterns of average deviations in the pre-train and race not included seem to flip signs in the race

11

included models. We interpret this as being because either (i) there are insufficient numbers of Asian individuals in the training set to get a good estimate, or (ii) encoding race using one-hot dummy variables does not provide sufficient flexibility to model race differences across demographic groups. In the pre-trained and race not included average deviations, UKB Asian individuals are overestimated in motor cortex, but not the rest of the brain. However, in the race included models, the motor cortex is underestimated, and the rest of the brain is underestimated.

The extreme deviation summaries (Figure 2B) measure the number of individuals in the tail of the normative distribution at each region and are therefore more sensitive to differences in the shape of the distribution rather than the mean. These show several interesting features. In HCP, the pre-trained normative models seem to be relatively fair in terms of identifying extreme deviations. There is a constant pattern of extreme deviations across all groups in the pre-trained models. One could interpret this as a particular type of parity, as noted above. Also in the HCP data, the number of extreme deviations in the White group does not change much between the race included and race not included models. In HCP Asian and Black groups, there are more extreme deviations in the race not included models than in the race included models. In UKB, across all three normative models, Asian individuals seems to have more extreme deviations than the Black and White individuals. In UKB Asian individuals, there is not much difference in the extreme deviations across normative models, but there are slightly more extreme deviations in the race included normative models. In UKB Black individuals, the most extreme deviations are seen in the pre-trained normative models and there is not much difference between race included and race not included extreme deviations. In UKB White individuals, there is also not much difference across normative models in the extreme deviations.

Taken together, these results indicate that the distribution of different racial groups have a different mean, and that the shape of the distribution is different.

5.1.2. Quantitative evaluation of racial bias

There were significant group differences in residual errors (equation 3) (Figure 3A) and deviation scores (equation 2) (Figure 3B) in all normative models after multiple comparison correction was performed (Table 4). The lowest percentage of group differences was consistently observed in the race included normative models, with the exception of HCP residual errors and UKB deviations. In HCP residual errors, White vs. Black showed the least differences in the race not included normative models and in UKB deviations, White vs. Asian showed the least differences in the pre-trained normative models. An interesting result is that the residual errors are higher in the White group in almost every model (Figure 3A), except for HCP race included where there is a strong lateralization effect where the left hemisphere residual errors are greater in the Asian and Black groups than in the White group. UKB race included also shows a few regions where residual errors in the Asian group and in the Black group are greater than the White group residual errors. The residual errors being higher in the White group may be due to the much larger sample size (power) in this group. The direction of group differences, $White > Asian$ and $White > Black$, in the deviation scores (Figure 3B) makes sense, as we show above in the average deviations (Figure 2A) that the models are often underestimated for Asian and Black groups.

A) Average Deviations
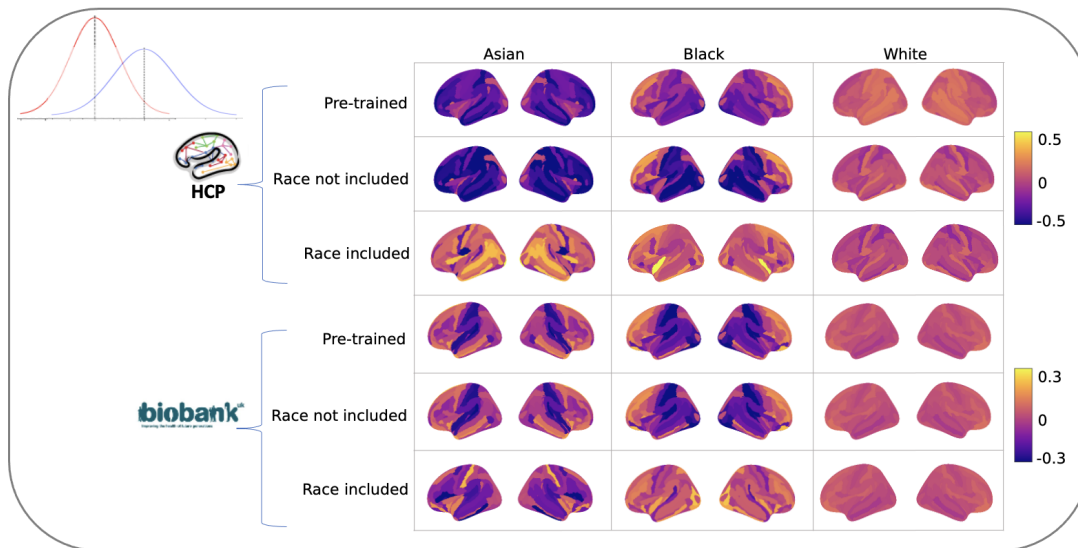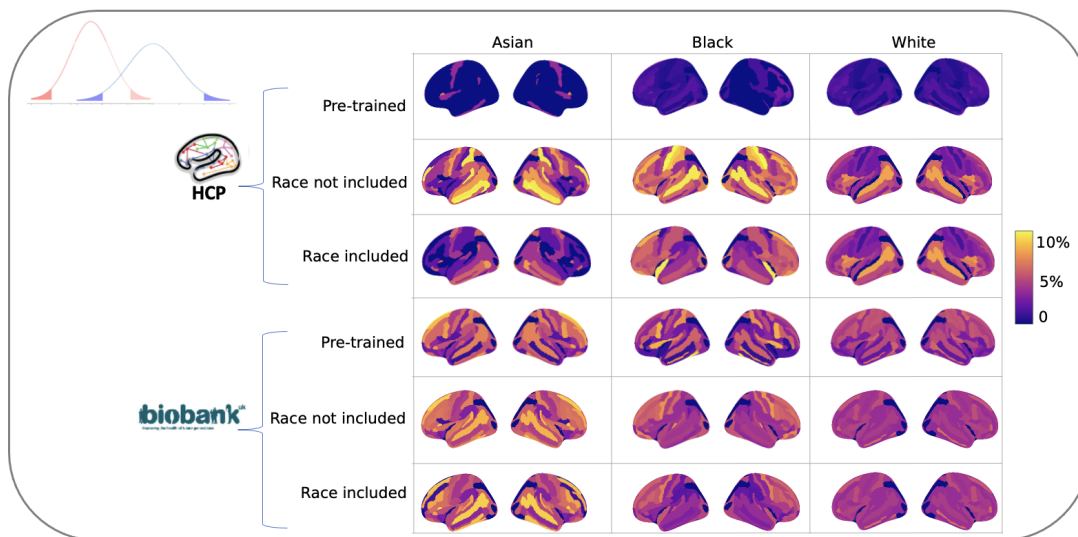

B) Percentage of Extreme (|Z|>2) Deviations

Figure 2: Summary of normative model deviation scores across all three reference classes (pre-trained, race not included, and race included) in HCP and UKB datasets. A) Average (mean) deviations for all brain regions within all racial groups (columns). B) Percentage of extreme deviations (positive and negative) for all brain regions within all racial groups (columns).

5.2. Race can be identified with high accuracy from normative models

In HCP, the pre-trained and race included models show similar average AUC and the race not included model had the lowest average AUC. White and Black individuals had the highest AUC in the pre-trained models and lowest AUC in the race included models. Asian

A) Group Difference Testing: Residual Errors
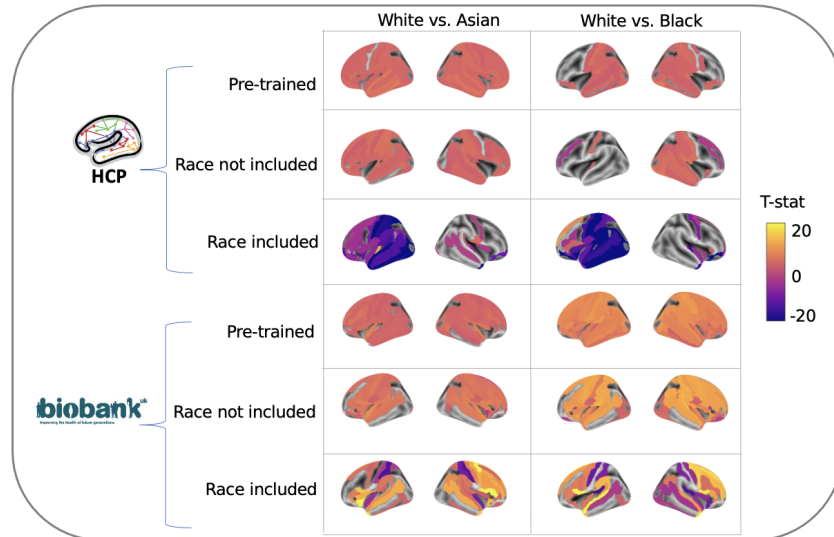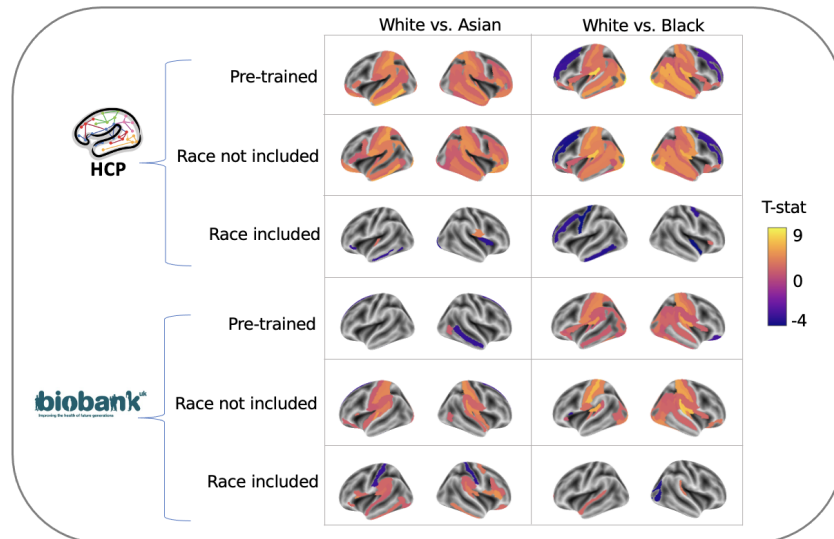

B) Group Difference Testing: Deviation Scores

Figure 3: Group differences in A) residual errors and B) deviation scores across all three reference classes (pre-trained, race not modeled, and race modeled) in HCP and UKB. The t-statistic is plotted where White individuals are group one, and Asian or Black individuals are group two. Light colors (positive t-stat) represent larger residual errors or deviations in White individuals and dark colors (negative t-stat) represent larger residual errors or deviations in Asian or Black individuals. Brain regions with statistically significant group differences after multiple comparison correction (FDRcorr $p < 0.05$) are shown. The number of brain regions showing group differences for each model is shown in Table 4.

Table 4: Group differences in deviations and residual errors across all three reference classes (pre-trained, race not included, and race included) in HCP and UKB. We show the percentage of models with statistically significant differences after multiple comparison correction ($FDR_{corr}p < 0.05$). In the Group column, W vs. A = White vs. Asian and W vs. B = White vs. Black.

Dataset	Group	Metric	Pre-train	Race not included	Race included
HCP	W vs. A	deviations	40%	49%	9%
HCP	W vs. B	deviations	55%	51%	5%
UKB	W vs. A	deviations	17%	25%	19%
UKB	W vs. B	deviations	45%	37%	7%
HCP	W vs. A	error	74%	64%	55%
HCP	W vs. B	error	56%	28%	54%
UKB	W vs. A	error	71%	56%	53%
UKB	W vs. B	error	87%	73%	51%

individuals had the highest AUC in the race included and lowest in the race not included models. Pre-trained models have the lowest average precision, recall, and F-score but there is larger variance across racial groups. White individuals had the highest precision, recall, and F-score in the pre-trained models and lowest in the race included models. Black individuals had the highest precision, recall, and F-score in the race not included models, the lowest precision and F-score in the race included models, and the lowest recall in the pre-trained models. Asian individuals had the highest precision, recall, and F-score in the race included models and lowest in the pre-trained models.

In UKB, the pre-trained models showed the highest average and per group AUC. The race not included models had the lowest average, Asian group, and Black group AUC. The lowest AUC for the White group was in the race included models. As observed in HCP, the pre-trained models had the lowest average precision, recall, and F-score but increased variance across racial groups. White individuals had the highest precision, recall, and F-score in the pre-trained models and the lowest in the race not included models. Black and Asian individuals had the highest precision, recall, and F-score in the race included models and the lowest in the pre-trained models.

Taken together, these results show that even after accounting for differences across racial groups in the normative model estimation, sufficient information remains in the deviation scores to accurately predict self-reported race.

6. Discussion

In this study, we used two independent datasets to document the racial bias present in reference class normative models of neuroimaging-derived brain structure. Across three different normative models, we uncovered differences in the way deviations were identified across racial groups. Racial bias was still present even when including race as a predictor in the normative model. This is an important finding because (i) it suggests that the modeling approach used here is not sufficiently flexible to account for racial differences,

Table 5: Race prediction using normative model (pre-train, race not included, and race included) deviation scores as features. One vs Rest classification was used within 5-fold cross validation. Values shown are the average across all folds (±s.d.). The mean across all race categories for each data set and model is shown in the mean (M) columns (mean of the column to the left). In the Group column A = Asian, B = Black and W = White.

|	Data	Group	Pre-train	M	Race not included	M	Race included	M
AUC	HCP	A	0.80±0.15		0.76±0.06		0.95±0.03	
	HCP	B	0.92±0.05	0.86±0.05	0.87±0.03	0.81±0.04	0.81±0.06	0.85±0.07
	HCP	W	0.87±0.04		0.81±0.05		0.80±0.05	
	UKB	A	0.84±0.03	0.83±0.01	0.73±0.02		0.80±0.05	
	UKB	B	0.82±0.04		0.73±0.04	0.73±0.0	0.80±0.03	0.77±0.04
	UKB	W	0.82±0.03		0.73±0.02		0.72±0.05	
Precision	HCP	A	0.31±0.20		0.43±0.10		0.79±0.16	
	HCP	B	0.70±0.04	0.64±0.24	0.76±0.07	0.64±0.15	0.65±0.09	0.71±0.06
	HCP	W	0.90±0.02		0.74±0.09		0.70±0.08	
	UKB	A	0.21±0.08		0.40±0.03		0.52±0.11	
	UKB	B	0.35±0.34	0.32±0.33	0.45±0.05	0.52±0.14	0.57±0.07	0.61±0.09
	UKB	W	0.98±0.0		0.72±0.02		0.73±0.03	
Recall	HCP	A	0.26±0.23		0.46±0.15		0.69±0.14	
	HCP	B	0.63±0.08	0.61±0.27	0.76±0.07	0.65±0.13	0.65±0.12	0.69±0.03
	HCP	W	0.93±0.02		0.73±0.08		0.72±0.09	
	UKB	A	0.09±0.04		0.32±0.04		0.44±0.10	
	UKB	B	0.07±0.06	0.38±0.43	0.39±0.05	0.50±0.21	0.47±0.05	0.57±0.17
	UKB	W	0.99±0.0		0.80±0.03		0.81±0.05	
F-score	HCP	A	0.27±0.20		0.44±0.12	0.64±0.14	0.74±0.14	
	HCP	B	0.66±0.05	0.61±0.26	0.76±0.06		0.64±0.09	0.70±0.04
	HCP	W	0.91±0.01		0.73±0.08		0.71±0.07	
	UKB	A	0.12±0.05		0.35±0.03		0.47±0.10	
	UKB	B	0.12±0.10	0.41±0.41	0.42±0.05	0.51±0.18	0.52±0.05	0.58±0.13
	UKB	W	0.99±0.0		0.76±0.02		0.76±0.03	
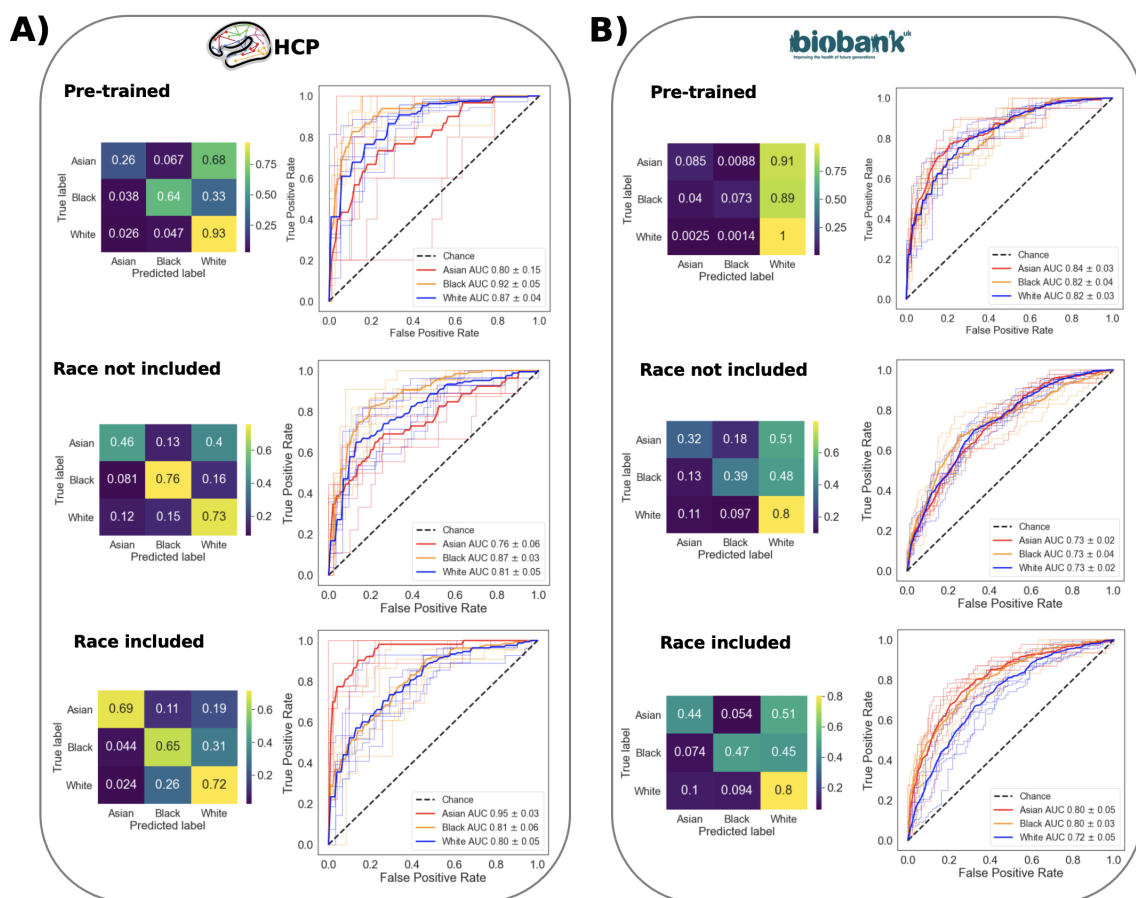
Figure 4: Prediction of self-reported race in A) HCP and B) UKB datasets using deviation scores from three different reference class normative models (pre-trained, race not included, and race included) as features. Performance is evaluated with confusion matrices, receiver operator characteristic (ROC) curves, and Area under the ROC curve (AUC). For the confusion matrix interpretation, the diagonal elements show where predicted label == true label, and the off-diagonal elements show mislabeled (predicted label != true label). The confusion matrices were normalized by the true labels to show ratios rather than counts. For interpreting the receiver operator characteristic (ROC) curves, we plot the performance across 5-fold cross validation (lighter colors, thin lines) and the also the mean across all folds (darker colors, thicker lines).

and (ii) deviations from the reference class "norm" (average/mean) are often interpreted with clinical meaning (i.e., representing biological dysfunction) when in reality we show that deviations could be due to an individual's demographic mismatch with the reference class. However, including race as a covariate in normative models did increase the parity of classifier predictions. We also show that self-reported race can be accurately predicted from normative model features well above chance. This discovery is surprising because this

17

is considered to be a challenging task for human experts (i.e., a neuroradiologist could not easily identify a patient's race from looking at an MRI of their brain).

Identifying self-reported race from neuroimaging-derived features is not necessarily a meaningful goal on its own. Our results in the normative model setting align with work using deep learning models Gichoya et al. (2022); Seyyed-Kalantari et al. (2021) to predict race directly from medical images (i.e., not reference class normalized images) to show that racial information cannot easily be isolated or removed from imaging data. The purpose of this work is not to point out new cases of racial disparities in model performance. Rather, it is important to view this within the broader context of employing predictive models in real-world clinical scenarios (where there are known racial disparities Adam et al. (2022); Seyyed-Kalantari et al. (2021); Xiao et al. (2024); Chang et al. (2022); Pierson et al. (2021); Li et al. (2022) and recognize that these models are not colorblind Watson-Daniels (2024). When predicting race based on medical images or other healthcare data, it's essential to ask: What exactly are we predicting? Are we predicting race as in someone's skin color? Or (more likely) are we predicting an identity intertwined with unmeasured risk factors or exposures, including the sociohistorical pressures (individual and structural racism) that shape group differences. We cannot ignore the fact that racism exists and is coded in technologies. This contemplation naturally segues into a discussion about race and genetic ancestry. Race is a social construct, categorizing individuals based on perceived physical traits, while genetic ancestry is contextualized within reference populations, indicating the genetic heritage of a population Borrell et al. (2021); Yudell et al. (2016); Maglo et al. (2016). Differences in clinical measures and outcomes among racial groups are frequently depicted as inherent biological distinctions Vyas et al. (2020). However, these differences are typically studied using self-reported race (social construct) not genetic ancestry Vyas et al. (2019); Kowalsky et al. (2020). **Given this knowledge, it is critical to acknowledge that the racial group differences we observe in this work cannot be interpreted biologically.**

Our results show that the linear approach we use to modeling race is not sufficient to accommodate racial differences. While it is likely that these can be more adequately modeled using non-linear techniques such as neural networks, it is crucial to recognize that – at least in brain imaging – we lack sufficiently representative cohorts to be able to estimate a reference class that faithfully reflects variation across different demographic groups. While pooling of neuroimaging datasets is becoming more prevalent in the field Rutherford et al. (2022a); Bethlehem et al. (2022), most legacy datasets do not contain racial information making it impossible to even assess racial bias in these models. While more recent cohort studies such as UKB and HCP do acquire this information, these samples are still heavily biased toward WEIRD (Western, Educated, Industrialized, Rich, Democratic individuals Henrich et al. (2010) (as shown in Table 1 and 2). Therefore, acquiring more representative cohorts should be seen as an urgent research priority.

Limitations There are several limitations of this work concerning the data that was used. Compared to other machine learning in healthcare settings that use different types of data such as electronic medical records (e.g., the MIMIC dataset) that contain 425,000 patient visits Johnson et al. (2023), the sample size used in this study is rather modest (Table 1 and 2) but it is in line with sample sizes used in neuroimaging studies Greene et al.

(2022); Rosenberg and Finn (2022). Regarding the representation of the neuroimaging data, we used features that were extracted from a group average template space that is defined based on a non-diverse sample (129 White individuals, 15 Asian individuals, and 1 person of mixed race) Fonov et al. (2009). Mapping individual's neuroimaging data into this space may create or conceal structural brain differences across individuals. Regarding the demographic variables, we only considered coarse race labels (Asian, Black, White) due to data availability. There is a need to question how race is defined in the algorithmic fairness community and how this impacts the utility of algorithmic fairness in real-world equity goals Abdu et al. (2023). Recent work has also shown in the performance metrics of several clinical prediction tasks, there is greater variability within (coarse) racial group labels than between (coarse) racial groups Movva et al. (2023). Thus, considering more granular racial categories would likely help the interpretation of our group difference findings. In addition to including more granular racial groups, intersectionality of sensitive groups is also an important (e.g., Black woman) consideration Tolbert and Diana (2023) as well as including socioeconomic indicators such as education and income Yang et al. (2023).

Conclusion It is not sufficient to model all racial groups in the same reference class, given the current frequencies available in the training data where there is over-representation of White individuals. We observed racial disparities in the normative models, in the univariate (group difference testing) and multivariate (race prediction) sense, that likely arise due to lack of data diversity. We reiterate that these findings cannot be interpreted biologically. Rather, race is a proxy for the multifaceted interactions among genetic ancestry, race, racism, socioeconomic status, and other environmental factors. These factors have not been well documented in existing datasets. In other words, we need to collect not just bigger data, but more diverse and representative data Kopal et al. (2023) which has been recognized with research initiatives such as the NIHBridge2AI. This should include granular measures of race and ethnicity, as well as additional information on experiences of racism, socioeconomic status, and environmental factors in order to begin disentangling the mechanisms of heterogeneity Carter et al. (2022); Cardenas-Iniguez and Gonzalez (2023); Yang et al. (2023).

Transparency should be a priority, given it is unlikely to achieve perfectly fair, or unbiased predictive models. This means communicating known biases (i.e., well documented sample demographics) of the training, validation, and testing sets. There are existing tools to help with this communication, including datasheets for datasets Gebru et al. (2021) and model cards for model reporting Mitchell et al. (2019). Subgroup performance audits are a key first step in revealing underlying issues that should be addressed before model integration. **Due to unmeasured dimensions of bias, acknowledging that predicted phenotypes reflect complex combinations of variables should also be part of routine model interpretation.**

A recent editorial on biased machine learning in healthcare shifts the fairness framework from viewing biased health data solely as detrimental to considering it as informative artifacts. The authors propose that AI's powerful pattern recognition abilities can be leveraged to detect exclusion from preventive healthcare and can serve as a hypothesis-generating tool that motivates new research on health inequities in healthcare Ferryman et al. (2023). We agree with this change of perspective on biased data and models. While these results reveal

uncomfortable truths about our data (and societies in which data was collected), they also align with public health goals regarding health equity Lin et al. (2024) which recognizes that acknowledging and understanding population inequities is an essential first step to making progress.

Acknowledgments

References

Amina A. Abdu, Irene V. Pasquetto, and Abigail Z. Jacobs. An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature, September 2023. URL http://arxiv.org/abs/2309.06607. arXiv:2309.06607 [cs].

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 7–21, July 2022. doi: 10.1145/3514094.3534203. URL http://arxiv.org/abs/2205.03931. arXiv:2205.03931 [cs].

Nil-Jana Akpinar, Zachary C. Lipton, and Alexandra Chouldechova. The Impact of Differential Feature Under-reporting on Algorithmic Fairness, January 2024. URL http://arxiv.org/abs/2401.08788. arXiv:2401.08788 [cs, stat].

Aaron F. Alexander-Bloch, Philip T. Reiss, Judith Rapoport, Harry McAdams, Jay N. Giedd, Ed T. Bullmore, and Nitin Gogtay. Abnormal Cortical Growth in Schizophrenia Targets Normative Modules of Synchronized Development. *Biological Psychiatry*, 76(6): 438–446, September 2014. ISSN 0006-3223. doi: 10.1016/j.biopsych.2014.02.010. URL https://www.sciencedirect.com/science/article/pii/S000632231400105X.

Fidel Alfaro-Almagro, Mark Jenkinson, Neal K. Bangerter, Jesper L. R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N. Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, Diego Vidaurre, Matthew Webster, Paul McCarthy, Christopher Rorden, Alessandro Daducci, Daniel C. Alexander, Hui Zhang, Iulius Dragonu, Paul M. Matthews, Karla L. Miller, and Stephen M. Smith. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166:400–424, 2018. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2017.10.034.

Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1194–1206, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533179. URL https://dl.acm.org/doi/10.1145/3531146.3533179.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Johanna M. M. Bayer, Richard Dinga, Seyed Mostafa Kia, Akhil R. Kottaram, Thomas Wolfers, Jinglei Lv, Andrew Zalesky, Lianne Schmaal, and Andre Marquand. Accommodating site variation in neuroimaging data using normative and hierarchical Bayesian models. *bioRxiv*, page 2021.02.09.430363, May 2021. doi: 10.1101/2021.02.09.430363. URL https://www.biorxiv.org/content/10.1101/2021.02.09.430363v2. Publisher: Cold Spring Harbor Laboratory Section: Confirmatory Results.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art, May 2017. URL http://arxiv.org/abs/1703.09207. arXiv:1703.09207 [stat].

Mélanie Bernhardt, Charles Jones, and Ben Glocker. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, 28(6):1157–1158, June 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01846-8. URL https://www.nature.com/articles/s41591-022-01846-8. Publisher: Nature Publishing Group.

R. a. I. Bethlehem, J. Seidlitz, S. R. White, J. W. Vogel, K. M. Anderson, C. Adamson, S. Adler, G. S. Alexopoulos, E. Anagnostou, A. Areces-Gonzalez, D. E. Astle, B. Auyeung, M. Ayub, J. Bae, G. Ball, S. Baron-Cohen, R. Beare, S. A. Bedford, V. Benegal, F. Beyer, J. Blangero, M. Blesa Cábez, J. P. Boardman, M. Borzage, J. F. Bosch-Bayard, N. Bourke, V. D. Calhoun, M. M. Chakravarty, C. Chen, C. Chertavian, G. Chetelat, Y. S. Chong, J. H. Cole, A. Corvin, M. Costantino, E. Courchesne, F. Crivello, V. L. Cropley, J. Crosbie, N. Crossley, M. Delarue, R. Delorme, S. Desrivieres, G. A. Devenyi, M. A. Di Biase, R. Dolan, K. A. Donald, G. Donohoe, K. Dunlop, A. D. Edwards, J. T. Elison, C. T. Ellis, J. A. Elman, L. Eyler, D. A. Fair, E. Feczko, P. C. Fletcher, P. Fonagy, C. E. Franz, L. Galan-Garcia, A. Gholipour, J. Giedd, J. H. Gilmore, D. C. Glahn, I. M. Goodyer, P. E. Grant, N. A. Groenewold, F. M. Gunning, R. E. Gur, R. C. Gur, C. F. Hammill, O. Hansson, T. Hedden, A. Heinz, R. N. Henson, K. Heuer, J. Hoare, B. Holla, A. J. Holmes, R. Holt, H. Huang, K. Im, J. Ipser, C. R. Jack, A. P. Jackowski, T. Jia, K. A. Johnson, P. B. Jones, D. T. Jones, R. S. Kahn, H. Karlsson, L. Karlsson, R. Kawashima, E. A. Kelley, S. Kern, K. W. Kim, M. G. Kitzbichler, W. S. Kremen, F. Lalonde, B. Landeau, S. Lee, J. Lerch, J. D. Lewis, J. Li, W. Liao, C. Liston, M. V. Lombardo, J. Lv, C. Lynch, T. T. Mallard, M. Marcelis, R. D. Markello, S. R. Mathias, B. Mazoyer, P. McGuire, M. J. Meaney, A. Mechelli, N. Medic, B. Misic, S. E. Morgan, D. Mothersill, J. Nigg, M. Q. W. Ong, C. Ortinau, R. Ossenkoppele, M. Ouyang, L. Palaniyappan, L. Paly, P. M. Pan, C. Pantelis, M. M. Park, T. Paus,

Z. Pausova, D. Paz-Linares, A. Pichet Binette, K. Pierce, X. Qian, J. Qiu, A. Qiu, A. Raznahan, T. Rittman, A. Rodrigue, C. K. Rollins, R. Romero-Garcia, L. Ronan, M. D. Rosenberg, D. H. Rowitch, G. A. Salum, T. D. Satterthwaite, H. L. Schaare, R. J. Schachar, A. P. Schultz, G. Schumann, M. Schöll, D. Sharp, R. T. Shinohara, I. Skoog, C. D. Smyser, R. A. Sperling, D. J. Stein, A. Stolicyn, J. Suckling, G. Sullivan, Y. Taki, B. Thyreau, R. Toro, N. Traut, K. A. Tsvetanov, N. B. Turk-Browne, J. J. Tuulari, C. Tzourio, É Vachon-Presseau, M. J. Valdes-Sosa, P. A. Valdes-Sosa, S. L. Valk, T. van Amelsvoort, S. N. Vandekar, L. Vasung, L. W. Victoria, S. Villeneuve, A. Villringer, P. E. Vértes, K. Wagstyl, Y. S. Wang, S. K. Warfield, V. Warrier, E. Westman, M. L. Westwater, H. C. Whalley, A. V. Witte, N. Yang, B. Yeo, H. Yun, A. Zalesky, H. J. Zar, A. Zettergren, J. H. Zhou, H. Ziauddeen, A. Zugman, X. N. Zuo, E. T. Bullmore, and A. F. Alexander-Bloch. Brain charts for the human lifespan. *Nature*, 604 (7906):525–533, April 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04554-y. URL https://www.nature.com/articles/s41586-022-04554-y. Number: 7906 Publisher: Nature Publishing Group.

R. Bhome, S. Verdi, S. A. Martin, N. Hannaway, I. Dobreva, N. P. Oxtoby, Castro Leal G, S. Rutherford, A. F. Marquand, R. S. Weil, and J. H. Cole. A neuroimaging measure to capture heterogeneous patterns of atrophy in Parkinson's disease and dementia with Lewy bodies, August 2023. URL https://www.medrxiv.org/content/10.1101/2023.08.01.23293480v1. Pages: 2023.08.01.23293480.

Augustijn A. A. de Boer, Seyed Mostafa Kia, Saige Rutherford, Mariam Zabihi, Charlotte Fraza, Pieter Barkema, Lars T. Westlye, Ole A. Andreassen, Max Hinne, Christian F. Beckmann, and Andre Marquand. Non-Gaussian Normative Modelling With Hierarchical Bayesian Regression, October 2022. URL https://www.biorxiv.org/content/10.1101/2022.10.05.510988v1. Pages: 2022.10.05.510988 Section: New Results.

E. Borghi, M. de Onis, C. Garza, J. Van den Broeck, E. A. Frongillo, L. Grummer-Strawn, S. Van Buuren, H. Pan, L. Molinari, R. Martorell, A. W. Onyango, and J. C. Martines. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. *Statistics in Medicine*, 25(2):247–265, 2006. ISSN 1097-0258. doi: 10.1002/sim.2227. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2227. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2227.

Luisa N. Borrell, Jennifer R. Elhawary, Elena Fuentes-Afflick, Jonathan Witonsky, Nirav Bhakta, Alan H.B. Wu, Kirsten Bibbins-Domingo, José R. Rodríguez-Santana, Michael A. Lenoir, James R. Gavin, Rick A. Kittles, Noah A. Zaitlen, David S. Wilkes, Neil R. Powe, Elad Ziv, and Esteban G. Burchard. Race and Genetic Ancestry in Medicine — A Time for Reckoning with Racism. *New England Journal of Medicine*, 384(5):474–480, February 2021. ISSN 0028-4793. doi: 10.1056/NEJMms2029562. URL https://doi.org/10.1056/NEJMms2029562. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMms2029562.

Koen Bruynseels, Filippo Santoni de Sio, and Jeroen van den Hoven. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Frontiers in*

Genetics, 9:31, February 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00031. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5816748/.

Barbora Rehák Bučková, Charlotte Fraza, Rastislav Rehák, Marián Kolenič, Christian Beckmann, Filip Španiel, Andre Marquand, and Jaroslav Hlinka. Using normative models pre-trained on cross-sectional data to evaluate longitudinal changes in neuroimaging data, January 2024. URL https://www.biorxiv.org/content/10.1101/2023.06.09.544217v2. Pages: 2023.06.09.544217 Section: New Results.

Carlos Cardenas-Iniguez and Marybel Robledo Gonzalez. "We controlled for race and ethnicity..." Considerations for the use and communication of race and ethnicity in neuroimaging research, June 2023. URL https://psyarxiv.com/xs8t3/.

Sierra E. Carter, Yara Mekawi, and Nathaniel G. Harnett. It's about racism, not race: a call to purge oppressive practices from neuropsychiatry and scientific discovery. *Neuropsychopharmacology*, 47(13):2179–2180, December 2022. ISSN 1740-634X. doi: 10.1038/s41386-022-01367-5. URL https://www.nature.com/articles/s41386-022-01367-5. Number: 13 Publisher: Nature Publishing Group.

Jessica P. Cerdeña, Marie V. Plaisime, and Jennifer Tsai. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *The Lancet*, 396(10257):1125–1128, October 2020. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(20)32076-6. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32076-6/fulltext. Publisher: Elsevier.

Trenton Chang, Michael W. Sjoding, and Jenna Wiens. Disparate Censorship & Undertesting: A Source of Label Bias in Clinical Machine Learning, August 2022. URL http://arxiv.org/abs/2208.01127. arXiv:2208.01127 [cs].

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, October 2016. URL http://arxiv.org/abs/1610.07524. arXiv:1610.07524 [cs, stat].

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness, June 2017. URL http://arxiv.org/abs/1701.08230. arXiv:1701.08230 [cs, stat].

Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, October 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.06.010.

Maria A. Di Biase, Ye Ella Tian, Richard A. I. Bethlehem, Jakob Seidlitz, Aaron. F. Alexander-Bloch, B.T. Thomas Yeo, and Andrew Zalesky. Mapping human brain charts cross-sectionally and longitudinally. *Proceedings of the National Academy of Sciences*, 120 (20):e2216798120, May 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2216798120. URL https://pnas.org/doi/10.1073/pnas.2216798120.

Doron Elad, Suheyla Cetin-Karayumak, Fan Zhang, Kang Ik K. Cho, Amanda E. Lyall, Johanna Seitz-Holland, Rami Ben-Ari, Godfrey D. Pearlson, Carol A. Tamminga,

John A. Sweeney, Brett A. Clementz, David J. Schretlen, Petra Verena Viher, Katharina Stegmayer, Sebastian Walther, Jungsun Lee, Tim J. Crow, Anthony James, Aristotle N. Voineskos, Robert W. Buchanan, Philip R. Szeszko, Anil K. Malhotra, Matcheri S. Keshavan, Martha E. Shenton, Yogesh Rathi, Sylvain Bouix, Nir Sochen, Marek R. Kubicki, and Ofer Pasternak. Improving the predictive potential of diffusion MRI in schizophrenia using normative models-Towards subject-level classification. *Human Brain Mapping*, 42 (14):4658–4670, October 2021. ISSN 1097-0193. doi: 10.1002/hbm.25574.

Kadija Ferryman, Maxine Mackintosh, and Marzyeh Ghassemi. Considering Biased Data as Informative Artifacts in AI-Assisted Health Care. *New England Journal of Medicine*, 389 (9):833–838, August 2023. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMra2214964. URL http://www.nejm.org/doi/10.1056/NEJMra2214964.

B Fischl and A M Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl. Acad. Sci. U. S. A.*, 97(20):11050–11055, September 2000. ISSN 0027-8424. doi: 10.1073/pnas.200033797. URL http://dx.doi.org/10.1073/pnas.200033797.

Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre van der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, Albert Montillo, Nikos Makris, Bruce Rosen, and Anders M Dale. Whole Brain Segmentation. *Neuron*, 33(3):341–355, 2002. doi: 10.1016/s0896-6273(02)00569-x. URL http://dx.doi.org/10.1016/s0896-6273(02)00569-x.

VS Fonov, AC Evans, RC McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, July 2009. ISSN 1053-8119. doi: 10.1016/S1053-8119(09)70884-5. URL https://www.sciencedirect.com/science/article/pii/S1053811909708845.

Jade S. Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P. Bennett, Jamie McCusker, and Deborah L. McGuinness. An Ontology for Fairness Metrics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pages 265–275, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534137. URL https://doi.org/10.1145/3514094.3534137.

Jade S. Franklin, Hannah Powers, John S. Erickson, Jamie McCusker, Deborah L. McGuinness, and Kristin P. Bennett. An Ontology for Reasoning About Fairness in Regression and Machine Learning. In Fernando Ortiz-Rodriguez, Boris Villazón-Terrazas, Sanju Tiwari, and Carlos Bobed, editors, *Knowledge Graphs and Semantic Web*, Lecture Notes in Computer Science, pages 243–261, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47745-4. doi: 10.1007/978-3-031-47745-4_18.

Charlotte Fraza, Mariam Zabihi, Christian F. Beckmann, and Andre F. Marquand. The Extremes of Normative Modelling, August 2022. URL https://www.biorxiv.org/content/10.1101/2022.08.23.505049v1. Pages: 2022.08.23.505049 Section: New Results.

Charlotte J. Fraza, Richard Dinga, Christian F. Beckmann, and Andre F. Marquand. Warped Bayesian linear regression for normative modelling of big data. *NeuroImage*, 245:118715, December 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2021.118715. URL https://www.sciencedirect.com/science/article/pii/S1053811921009873.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, December 2021. URL http://arxiv.org/abs/1803.09010. arXiv:1803.09010 [cs].

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis T. Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, June 2022. ISSN 2589-7500. doi: 10.1016/S2589-7500(22)00063-2. URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00063-2/fulltext. Publisher: Elsevier.

Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, David C. Van Essen, and Mark Jenkinson. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124, October 2013. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2013.04.127. URL http://www.sciencedirect.com/science/article/pii/S1053811913005053.

Abigail S. Greene, Xilin Shen, Stephanie Noble, Corey Horien, C. Alice Hahn, Jagriti Arora, Fuyuze Tokoglu, Marisa N. Spann, Carmen I. Carrión, Daniel S. Barron, Gerard Sanacora, Vinod H. Srihari, Scott W. Woods, Dustin Scheinost, and R. Todd Constable. Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature*, 609 (7925):109–118, September 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05118-w. URL https://www.nature.com/articles/s41586-022-05118-w. Number: 7925 Publisher: Nature Publishing Group.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning, October 2016. URL http://arxiv.org/abs/1610.02413. arXiv:1610.02413 [cs].

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, June 2010. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X0999152X. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/weirdest-people-in-the-world/BF84F7517D56AFF7B7EB58411A554C17. Publisher: Cambridge University Press.

Y. Hochberg and Y. Benjamini. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7):811–818, July 1990. ISSN 0277-6715. doi: 10.1002/sim.4780090710.

Alan Hájek. The reference class problem is your problem too. *Synthese*, 156(3):563–585, June 2007. ISSN 1573-0964. doi: 10.1007/s11229-006-9138-5. URL https://doi.org/10.1007/s11229-006-9138-5.

John P. A. Ioannidis, Neil R. Powe, and Clyde Yancy. Recalibrating the Use of Race in Medical Research. *JAMA*, 325(7):623–624, February 2021. ISSN 0098-7484. doi: 10.1001/jama.2021.0003. URL https://doi.org/10.1001/jama.2021.0003.

Veera Itälinna, Hanna Kaltiainen, Nina Forss, Mia Liljeström, and Lauri Parkkonen. Detecting mild traumatic brain injury with MEG, normative modelling and machine learning, September 2022. URL https://www.medrxiv.org/content/10.1101/2022.09.29.22280521v1. Pages: 2022.09.29.22280521.

A Johnson, L Bulgarelli, L.A. Celi, R Mark, and S Horng. MIMIC-IV-ED, January 2023. URL https://physionet.org/content/mimic-iv-ed/2.2/.

M. C. Jones and Arthur Pewsey. Sinh-arcsinh distributions. *Biometrika*, 96(4):761–780, 2009. ISSN 0006-3444. URL https://www.jstor.org/stable/27798865.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6, Karachi, Pakistan, February 2009. IEEE. ISBN 978-1-4244-3313-1. doi: 10.1109/IC4.2009.4909197. URL http://ieeexplore.ieee.org/document/4909197/.

Sara Khor, Eric C. Haupt, Erin E. Hahn, Lindsay Joe L. Lyons, Veena Shankaran, and Aasthaa Bansal. Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors. *JAMA Network Open*, 6(6):e2318495, June 2023. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2023.18495. URL https://doi.org/10.1001/jamanetworkopen.2023.18495.

Seyed Mostafa Kia, Hester Huijsdens, Saige Rutherford, Richard Dinga, Thomas Wolfers, Maarten Mennes, Ole A. Andreassen, Lars T. Westlye, Christian F. Beckmann, and Andre F. Marquand. Federated Multi-Site Normative Modeling using Hierarchical Bayesian Regression. *bioRxiv*, page 2021.05.28.446120, May 2021. doi: 10.1101/2021.05.28.446120. URL https://www.biorxiv.org/content/10.1101/2021.05.28.446120v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Seyed Mostafa Kia, Hester Huijsdens, Saige Rutherford, Augustijn de Boer, Richard Dinga, Thomas Wolfers, Pierre Berthet, Maarten Mennes, Ole A. Andreassen, Lars T. Westlye, Christian F. Beckmann, and Andre F. Marquand. Closing the life-cycle of normative modeling using federated hierarchical Bayesian regression. *PLOS ONE*, 17(12):e0278776, December 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0278776. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0278776. Publisher: Public Library of Science.

Eduard T. Klapwijk, Ferdi van de Kamp, Mara van der Meulen, Sabine Peters, and Lara M. Wierenga. Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data. *NeuroImage*, 189:116–129, April 2019. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2019.01.014.

Jakub Kopal, Lucina Q. Uddin, and Danilo Bzdok. The end game: respecting major sources of population diversity. *Nature Methods*, pages 1–7, March 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01812-3. URL https://www.nature.com/articles/s41592-023-01812-3. Publisher: Nature Publishing Group.

Rachel H. Kowalsky, Ashley C. Rondini, and Shari L. Platt. The Case for Removing Race From the American Academy of Pediatrics Clinical Practice Guideline for Urinary Tract Infection in Infants and Young Children With Fever. *JAMA Pediatrics*, 174(3): 229–230, March 2020. ISSN 2168-6203. doi: 10.1001/jamapediatrics.2019.5242. URL https://doi.org/10.1001/jamapediatrics.2019.5242.

Jingwei Li, Danilo Bzdok, Jianzhong Chen, Angela Tam, Leon Qi Rong Ooi, Avram J. Holmes, Tian Ge, Kaustubh R. Patil, Mbemba Jabbi, Simon B. Eickhoff, B. T. Thomas Yeo, and Sarah Genon. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, 8(11):eabj1812, March 2022. doi: 10.1126/sciadv.abj1812. URL https://www.science.org/doi/full/10.1126/sciadv.abj1812. Publisher: American Association for the Advancement of Science.

Jennifer S. Lin, Elizabeth M. Webber, Sarah I. Bean, and Corinne V. Evans. Development of a Health Equity Framework for the US Preventive Services Task Force. *JAMA Network Open*, 7(3):e241875, March 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.1875. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2815866.

Jinglei Lv, Maria Di Biase, Robin F. H. Cash, Luca Cocchi, Vanessa L. Cropley, Paul Klauser, Ye Tian, Johanna Bayer, Lianne Schmaal, Suheyla Cetin-Karayumak, Yogesh Rathi, Ofer Pasternak, Chad Bousman, Christos Pantelis, Fernando Calamante, and Andrew Zalesky. Individual deviations from normative models of brain structure in a large cross-sectional schizophrenia cohort. *Molecular Psychiatry*, pages 1–12, September 2020. ISSN 1476-5578. doi: 10.1038/s41380-020-00882-5. URL https://www.nature.com/articles/s41380-020-00882-5. Publisher: Nature Publishing Group.

Koffi N. Maglo, Tesfaye B. Mersha, and Lisa J. Martin. Population Genomics and the Statistical Values of Race: An Interdisciplinary Perspective on the Biological Classification of Human Populations and Implications for Clinical Genetic Epidemiological Research. *Frontiers in Genetics*, 7:22, February 2016. ISSN 1664-8021. doi: 10.3389/fgene.2016.00022. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4756148/.

Andre F. Marquand, Thomas Wolfers, Maarten Mennes, Jan Buitelaar, and Christian F. Beckmann. Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(5):433–447, September 2016. ISSN 2451-9022. doi: 10.1016/j.bpsc.2016.04.002. URL http://www.sciencedirect.com/science/article/pii/S2451902216300301.

Andre F. Marquand, Seyed Mostafa Kia, Mariam Zabihi, Thomas Wolfers, Jan K. Buitelaar, and Christian F. Beckmann. Conceptualizing mental disorders as deviations from

normative functioning. *Molecular Psychiatry*, 24(10):1415–1424, October 2019. ISSN 1476-5578. doi: 10.1038/s41380-019-0441-1. URL https://www.nature.com/articles/s41380-019-0441-1. Number: 10 Publisher: Nature Publishing Group.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, January 2019. doi: 10.1145/3287560.3287596. URL http://arxiv.org/abs/1810.03993. arXiv: 1810.03993.

Jennifer Monereo-Sánchez, Joost J. A. de Jong, Gerhard S. Drenthen, Magdalena Beran, Walter H. Backes, Coen D. A. Stehouwer, Miranda T. Schram, David E. J. Linden, and Jacobus F. A. Jansen. Quality control strategies for brain MRI segmentation and parcellation: Practical approaches and recommendations - insights from the Maastricht study. *NeuroImage*, 237:118174, August 2021. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2021.118174.

Rajiv Movva, Divya Shanmugam, Kaihua Hou, Priya Pathak, John Guttag, Nikhil Garg, and Emma Pierson. Coarse race data conceals disparities in clinical risk score performance, August 2023. URL http://arxiv.org/abs/2304.09270. arXiv:2304.09270 [cs, stat].

Pritam Mukherjee, Thomas C. Shen, Jianfei Liu, Tejas Mathai, Omid Shafaat, and Ronald M. Summers. Confounding factors need to be accounted for in assessing bias by machine learning algorithms. *Nature Medicine*, 28(6):1159–1160, June 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01847-7. URL https://www.nature.com/articles/s41591-022-01847-7. Publisher: Nature Publishing Group.

Abraham Nunes, Thomas Trappenberg, and Martin Alda. The definition and measurement of heterogeneity. *Translational Psychiatry*, 10(1):1–13, August 2020a. ISSN 2158-3188. doi: 10.1038/s41398-020-00986-0. URL https://www.nature.com/articles/s41398-020-00986-0. Number: 1 Publisher: Nature Publishing Group.

Abraham Nunes, Thomas Trappenberg, and Martin Alda. Measuring heterogeneity in normative models as the effective number of deviation patterns. *PLOS ONE*, 15(11):e0242320, November 2020b. ISSN 1932-6203. doi: 10.1371/journal.pone.0242320. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0242320. Publisher: Public Library of Science.

Abraham Nunes, Thomas Trappenberg, and Martin Alda. We need an operational framework for heterogeneity in psychiatric research. *Journal of Psychiatry & Neuroscience : JPN*, 45(1):3–6, January 2020c. ISSN 1180-4882. doi: 10.1503.jpn/190198. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6919921/.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, October 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/abs/10.1126/science.aax2342. Publisher: American Association for the Advancement of Science.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, November 2021. ISSN 26663899. doi: 10.1016/j.patter.2021.100336. URL http://arxiv.org/abs/2012.05345. arXiv:2012.05345 [cs].

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Stephen R. Pfohl, Agata Foryciarz, and Nigam H. Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, 113:103621, January 2021. ISSN 1532-0464. doi: 10.1016/j.jbi.2020.103621. URL https://www.sciencedirect.com/science/article/pii/S1532046420302495.

Emma Pierson, David M. Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, January 2021. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-020-01192-7. URL http://www.nature.com/articles/s41591-020-01192-7.

Walter H. L. Pinaya, Cristina Scarpazza, Rafael Garcia-Dias, Sandra Vieira, Lea Baecker, Pedro F da Costa, Alberto Redolfi, Giovanni B. Frisoni, Michela Pievani, Vince D. Calhoun, João R. Sato, and Andrea Mechelli. Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multi-cohort study. *Scientific Reports*, 11(1):15746, August 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-95098-0. URL https://www.nature.com/articles/s41598-021-95098-0. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Alzheimer's disease;Biomedical engineering;Computer science;Diagnostic markers Subject_term_id: alzheimers-disease;biomedical-engineering;computer-science;diagnostic-markers.

Hans Reichenbach. *The theory of probability*. University of California Press, 1949.

Gonzalo Rios and Felipe Tobar. Compositionally-warped Gaussian processes. *Neural Networks*, 118:235–246, October 2019. ISSN 08936080. doi: 10.1016/j.neunet.2019.06.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608019301856.

Dorothy E. Roberts. Abolish race correction. *The Lancet*, 397(10268):17–18, January 2021. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(20)32716-1. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32716-1/fulltext. Publisher: Elsevier.

Yuji Roh, Geon Heo, and Steven Euijong Whang. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions*

on Knowledge and Data Engineering, 33(4):1328–1347, April 2021. ISSN 1558-2191. doi: 10.1109/TKDE.2019.2946162. URL https://ieeexplore.ieee.org/document/8862913/citations#citations. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Adon F. G. Rosen, David R. Roalf, Kosha Ruparel, Jason Blake, Kevin Seelaus, Lakshmi P. Villa, Rastko Ciric, Philip A. Cook, Christos Davatzikos, Mark A. Elliott, Angel Garcia de La Garza, Efstathios D. Gennatas, Megan Quarmley, J. Eric Schmitt, Russell T. Shinohara, M. Dylan Tisdall, R. Cameron Craddock, Raquel E. Gur, Ruben C. Gur, and Theodore D. Satterthwaite. Quantitative assessment of structural image quality. *NeuroImage*, 169:407–418, April 2018. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2017.12.059.

Monica D. Rosenberg and Emily S. Finn. How to establish robust brain–behavior relationships without thousands of individuals. *Nature Neuroscience*, 25(7):835–837, July 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01110-9. URL https://www.nature.com/articles/s41593-022-01110-9. Number: 7 Publisher: Nature Publishing Group.

Saige Rutherford, Charlotte Fraza, Richard Dinga, Seyed Mostafa Kia, Thomas Wolfers, Mariam Zabihi, Pierre Berthet, Amanda Worker, Serena Verdi, Derek Andrews, Laura KM Han, Johanna MM Bayer, Paola Dazzan, Phillip McGuire, Roel T. Mocking, Aart Schene, Chandra Sripada, Ivy F. Tso, Elizabeth R. Duval, Soo-Eun Chang, Brenda WJH Penninx, Mary M. Heitzeg, S. Alexandra Burt, Luke W. Hyde, David Amaral, Christine Wu Nordahl, Ole A. Andreasssen, Lars T. Westlye, Roland Zahn, Henricus G. Ruhe, Christian Beckmann, and Andre F. Marquand. Charting brain growth and aging at high spatial precision, February 2022a. URL https://elifesciences.org/articles/72904/figures. Publisher: eLife Sciences Publications Limited.

Saige Rutherford, Seyed Mostafa Kia, Thomas Wolfers, Charlotte Fraza, Mariam Zabihi, Richard Dinga, Pierre Berthet, Amanda Worker, Serena Verdi, Henricus G. Ruhe, Christian F. Beckmann, and Andre F. Marquand. The normative modeling framework for computational psychiatry. *Nature Protocols*, pages 1–24, June 2022b. ISSN 1750-2799. doi: 10.1038/s41596-022-00696-5. URL https://www.nature.com/articles/s41596-022-00696-5. Publisher: Nature Publishing Group.

Saige Rutherford, Pieter Barkema, Ivy F Tso, Chandra Sripada, Christian F Beckmann, Henricus G Ruhe, and Andre F Marquand. Evidence for embracing normative modeling. *eLife*, 12:e85082, March 2023. ISSN 2050-084X. doi: 10.7554/eLife.85082. URL https://doi.org/10.7554/eLife.85082. Publisher: eLife Sciences Publications, Ltd.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445518. URL https://dl.acm.org/doi/10.1145/3411764.3445518.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, pages 1–7, December 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01595-0. URL https://www.nature.com/articles/s41591-021-01595-0. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Machine learning;Medical imaging Subject_term_id: machine-learning;medical-imaging.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Reply to: 'Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms' and 'Confounding factors need to be accounted for in assessing bias by machine learning algorithms'. *Nature Medicine*, 28 (6):1161–1162, June 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01854-8. URL https://www.nature.com/articles/s41591-022-01854-8. Publisher: Nature Publishing Group.

Edward Snelson, Carl Edward Rasmussen, and Zoubin Ghahramani. Warped Gaussian processes. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 337–344, Whistler, British Columbia, Canada, December 2003. MIT Press.

Paul W. Sperduto, Shane Mesko, Jing Li, Daniel Cagney, Ayal Aizer, Nancy U. Lin, Eric Nesbit, Tim J. Kruser, Jason Chan, Steve Braunstein, Jessica Lee, John P. Kirkpatrick, Will Breen, Paul D. Brown, Diana Shi, Helen A. Shih, Hany Soliman, Arjun Sahgal, Ryan Shanley, William A. Sperduto, Emil Lou, Ashlyn Everett, Drexell H. Boggs, Laura Masucci, David Roberge, Jill Remick, Kristin Plichta, John M. Buatti, Supriya Jain, Laurie E. Gaspar, Cheng-Chia Wu, Tony J.C. Wang, John Bryant, Michael Chuong, Yi An, Veronica Chiang, Toshimichi Nakano, Hidefumi Aoyama, and Minesh P. Mehta. Survival in Patients With Brain Metastases: Summary Report on the Updated Diagnosis-Specific Graded Prognostic Assessment and Definition of the Eligibility Quotient. *Journal of Clinical Oncology*, 38(32):3773–3784, November 2020. ISSN 0732-183X. doi: 10.1200/ JCO.20.01255. URL https://ascopubs.org/doi/10.1200/JCO.20.01255. Publisher: Wolters Kluwer.

Vinith M. Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction, July 2023. URL http://arxiv.org/abs/2206.02058. arXiv:2206.02058 [cs, stat].

Paul D. Thorn. Two Problems of Direct Inference. *Erkenntnis*, 76(3):299–318, May 2012. ISSN 1572-8420. doi: 10.1007/s10670-011-9319-6. URL https://doi.org/10.1007/s10670-011-9319-6.

Alexander Williams Tolbert and Emily Diana. Correcting Underrepresentation and Intersectional Bias for Fair Classification, September 2023. URL http://arxiv.org/abs/2306.11112. arXiv:2306.11112 [cs, stat].

Muhammad Turab and Sonain Jamil. A Comprehensive Survey of Digital Twins in Healthcare in the Era of Metaverse. *BioMedInformatics*, 3(3):563–584, September 2023. ISSN 2673-7426. doi: 10.3390/biomedinformatics3030039. URL https://www.mdpi.com/2673-7426/3/3/39. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

John Venn. *The Logic of Chance*. Macmillian and Co., 3rd edition, 1888. URL https://www.gutenberg.org/ebooks/57359/pg57359-images.html.

Serena Verdi, Andre F Marquand, Jonathan M Schott, and James H Cole. Beyond the average patient: how neuroimaging models can address heterogeneity in dementia. *Brain*, April 2021. ISSN 0006-8950. doi: 10.1093/brain/awab165. URL https://doi.org/10.1093/brain/awab165.

Serena Verdi, Saige Rutherford, Charlotte Fraza, Duygu Tosun, Andre Altmann, Lars Lau Raket, Jonathan M. Schott, Andre F. Marquand, and James H. Cole. Personalising Alzheimer's Disease progression using brain atrophy markers, June 2023. URL https://www.medrxiv.org/content/10.1101/2023.06.15.23291418v1. Pages: 2023.06.15.23291418.

Darshali A. Vyas, David S. Jones, Audra R. Meadows, Khady Diouf, Nawal M. Nour, and Julianna Schantz-Dunn. Challenging the Use of Race in the Vaginal Birth after Cesarean Section Calculator. *Women's Health Issues*, 29(3):201–204, May 2019. ISSN 1049-3867, 1878-4321. doi: 10.1016/j.whi.2019.04.007. URL https://www.whijournal.com/article/S1049-3867(19)30098-2/fulltext. Publisher: Elsevier.

Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine*, 383(9):874–882, August 2020. ISSN 0028-4793. doi: 10.1056/NEJMms2004740. URL https://doi.org/10.1056/NEJMms2004740. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMms2004740.

Christian Wallmann and Jon Williamson. Four Approaches to the Reference Class Problem. In Gábor Hofer-Szabó and Leszek Wroński, editors, *Making it Formally Explicit*, European Studies in Philosophy of Science, pages 61–81, Cham, 2017. Springer International Publishing. ISBN 978-3-319-55486-0. doi: 10.1007/978-3-319-55486-0_4.

Jamelle Watson-Daniels. Algorithmic Fairness and Color-blind Racism: Navigating the Intersection, February 2024. URL http://arxiv.org/abs/2402.07778. arXiv:2402.07778 [cs].

Thomas Wolfers, Nhat Trung Doan, Tobias Kaufmann, Dag Alnæs, Torgeir Moberget, Ingrid Agartz, Jan K. Buitelaar, Torill Ueland, Ingrid Melle, Barbara Franke, Ole A. Andreassen, Christian F. Beckmann, Lars T. Westlye, and Andre F. Marquand. Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA Psychiatry*, 75(11):1146–1155, November 2018. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2018.2467. URL https://doi.org/10.1001/jamapsychiatry.2018.2467.

Kevin Wu, Dominik Dahlem, Christopher Hane, Eran Halperin, and James Zou. Collecting data when missingness is unknown: a method for improving model performance given under-reporting in patient populations. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 229–242. PMLR, June 2023. URL https://proceedings.mlr.press/v209/wu23b.html. ISSN: 2640-3498.

Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification, January 2024. URL http://arxiv.org/abs/2305.11348. arXiv:2305.11348 [cs].

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Eng Hock Ong, Fei Gao, and Nan Liu. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9 (1):658, October 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01782-9. URL https://www.nature.com/articles/s41597-022-01782-9. Number: 1 Publisher: Nature Publishing Group.

Ming Ying Yang, Gloria Hyunjung Kwak, Tom Pollard, Leo Anthony Celi, and Marzyeh Ghassemi. Evaluating the Impact of Social Determinants on Health Prediction in the Intensive Care Unit. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 333–350, August 2023. doi: 10.1145/3600211.3604719. URL http://arxiv.org/abs/2305.12622. arXiv:2305.12622 [cs].

Michael Yudell, Dorothy Roberts, Rob DeSalle, and Sarah Tishkoff. Taking race out of human genetics. *Science*, 351(6273):564–565, February 2016. doi: 10.1126/science.aac4951. URL https://www.science.org/doi/10.1126/science.aac4951. Publisher: American Association for the Advancement of Science.

Mariam Zabihi, Marianne Oldehinkel, Thomas Wolfers, Vincent Frouin, David Goyard, Eva Loth, Tony Charman, Julian Tillmann, Tobias Banaschewski, Guillaume Dumas, Rosemary Holt, Simon Baron-Cohen, Sarah Durston, Sven Bölte, Declan Murphy, Christine Ecker, Jan K. Buitelaar, Christian F. Beckmann, and Andre F. Marquand. Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4 (6):567–578, June 2019. ISSN 2451-9022. doi: 10.1016/j.bpsc.2018.11.013. URL https://www.sciencedirect.com/science/article/pii/S245190221830329X.

Anna Zink, Ziad Obermeyer, and Emma Pierson. Race Corrections in Clinical Algorithms Can Help Correct for Racial Disparities in Data Quality, October 2023. URL https://www.medrxiv.org/content/10.1101/2023.03.31.23287926v3. ISSN: 2328-7926 Pages: 2023.03.31.23287926.
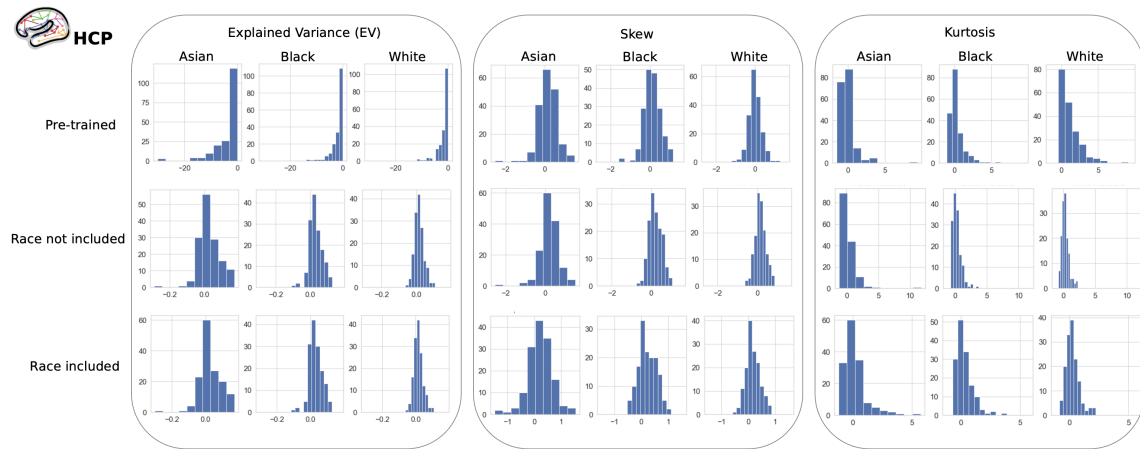
Appendix A.

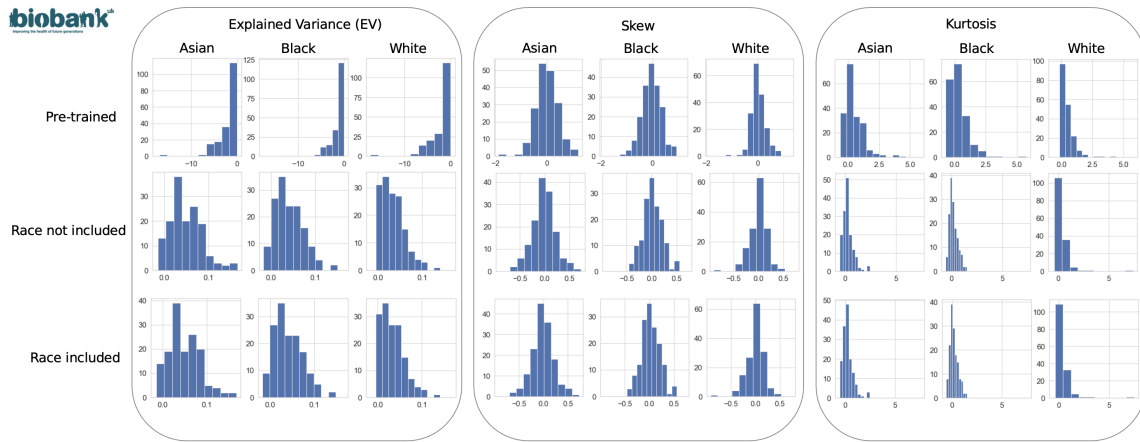Figure 5: Evaluation metrics for normative models in HCP dataset.

Figure 6: Evaluation metrics for normative models in UKB dataset.