frontiers | Frontiers in Cardiovascular Medicine

ORIGINAL RESEARCH
published: 07 April 2022
doi: 10.3389/fcvm.2022.859310

Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation

Esther Puyol-Antón[1], Bram Ruijsink[1,2,3], Jorge Mariscal Harana[1], Stefan K. Piechnik[4], Stefan Neubauer[4], Steffen E. Petersen[5,6,7,8], Reza Razavi[1,2], Phil Chowienczyk[1,9] and Andrew P. King[1]*

[1] School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom, [2] Department of Adult and Paediatric Cardiology, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom, [3] Division of Heart and Lungs, Department of Cardiology, University Medical Centre Utrecht, Utrecht, Netherlands, [4] Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom, [5] National Institute for Health Research (NIHR) Barts Biomedical Research Centre, William Harvey Research Institute, Queen Mary University London, London, United Kingdom, [6] Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, United Kingdom, [7] Health Data Research UK, London, United Kingdom, [8] Alan Turing Institute, London, United Kingdom, [9] British Heart Foundation Centre, King's College London, London, United Kingdom

OPEN ACCESS

Edited by:
Chayakrit Krittanawong, NYU Grossman School of Medicine, United States

Reviewed by:
Vivek Pinakin Jani, The Johns Hopkins Hospital, Johns Hopkins Medicine, United States
Feng Yang, National Institutes of Health (NIH), United States
Anna Baritussio, University of Padua, Italy
Yuki Mori, University of Copenhagen, Denmark

***Correspondence:**
Esther Puyol-Antón
esther.puyol_anton@kcl.ac.uk

Specialty section:
This article was submitted to Cardiovascular Imaging, a section of the journal Frontiers in Cardiovascular Medicine

Received: *21 January 2022*
Accepted: *02 March 2022*
Published: *07 April 2022*

Citation:
Puyol-Antón E, Ruijsink B, Mariscal Harana J, Piechnik SK, Neubauer S, Petersen SE, Razavi R, Chowienczyk P and King AP (2022) Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. Front. Cardiovasc. Med. 9:859310. doi: 10.3389/fcvm.2022.859310

Background: Artificial intelligence (AI) techniques have been proposed for automation of cine CMR segmentation for functional quantification. However, in other applications AI models have been shown to have potential for sex and/or racial bias. The objective of this paper is to perform the first analysis of sex/racial bias in AI-based cine CMR segmentation using a large-scale database.

Methods: A state-of-the-art deep learning (DL) model was used for automatic segmentation of both ventricles and the myocardium from cine short-axis CMR. The dataset consisted of end-diastole and end-systole short-axis cine CMR images of 5,903 subjects from the UK Biobank database (61.5 ± 7.1 years, 52% male, 81% white). To assess sex and racial bias, we compared Dice scores and errors in measurements of biventricular volumes and function between patients grouped by race and sex. To investigate whether segmentation bias could be explained by potential confounders, a multivariate linear regression and ANCOVA were performed.

Results: Results on the overall population showed an excellent agreement between the manual and automatic segmentations. We found statistically significant differences in Dice scores between races (white ∼94% vs. minority ethnic groups 86–89%) as well as in absolute/relative errors in volumetric and functional measures, showing that the AI model was biased against minority racial groups, even after correction for possible confounders. The results of a multivariate linear regression analysis showed that no covariate could explain the Dice score bias between racial groups. However, for the Mixed and Black race groups, sex showed a weak positive association with the Dice

score. The results of an ANCOVA analysis showed that race was the main factor that can explain the overall difference in Dice scores between racial groups.

Conclusion: We have shown that racial bias can exist in DL-based cine CMR segmentation models when training with a database that is sex-balanced but not race-balanced such as the UK Biobank.

Keywords: cardiac magnetic resonance, deep learning, fair AI, segmentation, inequality fairness in deep learning-based CMR segmentation

INTRODUCTION

Artificial intelligence (AI) is a rapidly evolving field in medicine, especially cardiology. AI has the potential to aid cardiologists in making better decisions, improving workflows, productivity, cost-effectiveness, and ultimately patient outcomes (1). Deep learning (DL) is a recent advance in AI which allows computers to learn a task using data instead of being explicitly programmed. Several studies in cardiology and other applications have shown that DL methods can match or even exceed human experts in tasks such as identifying and classifying disease (2–4).

In cardiology, cardiovascular imaging has a pivotal role in diagnostic decision making. Cardiac magnetic resonance (CMR) is the established non-invasive gold-standard modality for quantification of cardiac volumes and ejection fraction (EF). For decades, clinicians have been relying on manual or semi-automatic segmentation approaches to trace the cardiac chamber contours. However, manual expert segmentation of CMR images is tedious, time-consuming and prone to subjective errors. Recently, DL models have shown remarkable success in automating many medical image segmentation tasks. In cardiology, human-level performance in segmenting the main structures of the heart has been reported (5, 6), and researchers have proposed to use these models for tasks such as automating cardiac functional quantification (7). These methods are now starting to move toward broader clinical translation.

In the vast majority of cardiovascular diseases (CVDs), there are known associations between sex/race and epidemiology, pathophysiology, clinical manifestations, effects of therapy, and outcomes (8–10). Furthermore, in clinically asymptomatic individuals the Multi-Ethnic Study of Atherosclerosis (MESA) study showed that men had greater right ventricular (RV) mass and larger RV volumes than women, but had lower RV ejection fraction; African-Americans had lower RV mass than whites, whereas Hispanics had higher RV mass (11); and the LV was more trabeculated in African-American and Hispanic participants than white participants, and smoothest in Chinese-American participants (12), but the greater extent of LV trabeculation was not associated with an absolute decline in LVEF during the approximately 10 years of the MESA study. Similarly, the Coronary Artery Risk Development in Young Adults (CARDIA) study (13) showed differences between races (African American and white) and sexes in LV systolic and diastolic function, which persist after adjustment for established cardiovascular risk factors.

Although these physiological differences are associations and not proven causative links with race/gender, their presence raises a potential concern about the performance of AI models in cardiovascular imaging. Although AI has great potential in this area, no previous work has investigated the fairness of such models. In AI, the concept of "fairness" refers to assessing AI algorithms for potential bias based on demographic characteristics such as race and sex. In general, AI models are trained agnostic to demographic characteristics, and they assume that if the model is unaware of these characteristics while making decisions, the decisions will be fair. However, we have recently shown, for the first time, that using this assumption there exists racial bias in DL-based cine CMR segmentation models when trained using racially imbalanced data (14). The previous study aimed to identify the presence of bias and the technical development of different bias mitigation strategies, in order to reduce the bias effect between different racial groups. The object of this study is to investigate in more detail the origin and the effect of this bias on cardiac structure and function and to assess whether the bias could be explained by any confounder and therefore be linked with changes in subject characteristics, anatomy or cardiovascular risk factors.

MATERIALS AND METHODS

Participants

The UK Biobank is a prospective cohort study with more than 500,000 participants aged 40–69 years of age conducted in the United Kingdom (15). This study complies with the Declaration of Helsinki; the work was covered by the ethical approval for UK Biobank studies from the NHS National Research Ethics Service on 17th June 2011 (Ref 11/NW/0382) and extended on 18th June 2021 (Ref 21/NW/0157) with written informed consent obtained from all participants. The present study was performed using a sub-cohort of the UK Biobank imaging database, for whom CMR imaging and ground truth manual segmentations were available. In this study, in order to minimize the effects of physiological differences due to cardiovascular and other related diseases, we only focus on the healthy population of the UK Biobank database and analyze possible confounders that can explain racial and sex bias.

Therefore, we excluded any subjects with known cardiovascular disease, respiratory disease, hematological disease, renal disease, rheumatic disease, malignancies, symptoms of chest pain, respiratory symptoms or other diseases

impacting the cardiovascular system, except for diabetes mellitus, hypercholesterolemia and hypertension (see all exclusion criteria in **Supplementary List 1**). We included these cardiovascular risk factors to evaluate if or to what degree different cardiovascular risk in otherwise healthy patients could explain a potential bias in segmentation performance. We used the ICD-9 and ICD-10 codes and self-reported detailed health questionnaires and medication history for the selection process.

In this paper, race was assumed to align with self-reported ethnicity, which was the data collected in the UK Biobank. From the total UK Biobank database (N = 501,642), the race distribution is as follows: White 94.3%, Mixed 0.6%, Asian, 1.9%, Black 1.6%, Chinese: 0.9%, Other: 0.4%. The UK Biobank cohort has a similar ethnic distribution to the national population of the same age range in the 2011 UK Census (16). The imaging cohort used in this study (N = 5,660) has a slightly different racial distribution (White 81%, Mixed 3%, Asian, 7%, Black 4%, Chinese: 2%, Other: 3%), but it is still predominantly White race, in line with the full cohort of the UK Biobank database. Imaging centers of the UK Biobank are in Newcastle upon Tyne, Stockport, Reading and Bristol. The same imaging protocol was used in all imaging centers and no racial distribution difference was found between them. More details of the image acquisition protocol can be found in Petersen et al. (17).

Subject characteristics obtained were age, binary sex category, race, body measures (height; weight; body mass index, BMI; and body surface area, BSA), and smoker status (smoker was defined as a subject smoking or smoked daily for over 25 years in the previous 35 years). We also obtained the average heart rate (HR) and brachial systolic and diastolic blood pressure (SBP and DBP) measured during the CMR exam. These subject characteristics were considered as possible confounders in the statistical analysis, as they are directly or indirectly related to the measurements made and therefore plausibly associated with the accuracy of the measurements.

Automated Image Analysis

A state-of-the-art DL based segmentation model, the "nnU-Net" framework (18), was used for automatic segmentation of the left ventricle blood pool (LVBP), left ventricular myocardium (LVMyo) and right ventricle blood pool (RVBP) from cine short-axis CMR slices at end-diastole (ED) and end-systole (ES). This model was chosen as it has performed well across a range of segmentation challenges and was the top-performing model in the "ACDC" CMR segmentation challenge (6). For training and testing the segmentation model, we used a random split of 4,410 and 1,250 subjects, respectively, each with similar sex and racial distributions. We refer the reader to our previous paper (14) for further details of the model architecture and training.

Evaluation of the Method

For quantitative assessment of the image segmentation model, we used the Dice similarity coefficient (DSC), which quantifies the overlap between an automated segmentation and a ground truth segmentation. DSC has values between 0 and 100%, where 0 denotes no overlap, and 100% denotes perfect agreement. From the manual and automated image segmentations, we calculated the LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV), and RV end-diastolic volume (RVEDV) and end-systolic volume (RVESV) by summing the number of voxels belonging to the corresponding label classes in the segmentation and multiplying this by the volume per voxel. The LV myocardial mass (LVmass) was calculated by multiplying the LV myocardial volume by a density of 1.05 g/mL. Derived from the LV and RV volumes, we also computed LV ejection fraction (LVEF) and RV ejection fraction (RVEF). We evaluated the accuracy of these volumetric and functional measures by computing the absolute and relative differences between automated and manual measurements. We define the absolute and relative error as $\varepsilon_{absolute} = |v_{manual} - v_{auto}|$) and $\varepsilon_{relative}(\%) = 100 * |v_{manual} - v_{auto}|/v_{manual}$, where v corresponds to each clinical measure.

Analysis of the Influence of Confounders

To investigate whether a true bias between racial and/or sex groups exists for automated DL-based cine CMR segmentation, we conducted a statistical analysis to investigate if the observed bias could be explained by the most common confounders. In this study, we use as possible confounders age, sex, body measures (i.e., height, weight and BMI), HR, SBP, DBP, CMR-derived parameters (LVEDV, LVESV, RVEDV, RVESV, LVmass), cardiovascular risk factors (i.e., hypertension, hypercholesteremia, diabetes and smoking) and center (i.e., core lab where most of the segmentations were performed vs. additional lab).

Statistical Analysis

Data analysis was performed using SPSS Statistics (version 27, IBM, United States). Continuous variables are reported as mean ± standard deviation (SD) and tested for normal distributions with the Shapiro–Wilk test. Log transformations were applied to the (1-DSC) values to obtain an approximately normal distribution. After transformation, all continuous variables were normally distributed. Categorical data are presented as absolute counts and percentages. Comparison of variables between groups (i.e., races and sexes) was carried out using an independent Student's t-test.

Independent association between log-transformed DSC values and race was performed using univariate linear regression followed by multivariate adjustment for confounders. All variables in the regression models were standardized by computing the z-score for individual data points.

Finally, the differences in DSC values among different racial groups were initially assessed by a 1-way ANOVA (Model 4) followed by an analysis of covariance—ANCOVA (Model 5) to statistically control the effect of covariates. In addition, we check the assumption concerning regression residuals (19) as follows: (1) Homoscedasticity tested by a Levene's Test of quality of error variance; (2) Normality of residuals tested by the Kolmogorov–Smirnov and Shapiro–Wilk test; (3) Multicollinearity tested by the Durbin Watson Test. For all statistical analysis, the threshold for statistical significance was $p < 0.01$ and confidence intervals (%) were calculated by non-parametric bootstrapping with 1,000 resamples.

Pairwise *post hoc* testing was carried out using Bonferroni correction and Scheffé correction for multiple comparisons on the *t*-test and ANOVA analysis, respectively.

MATERIALS

Subject Characteristics

The dataset used consisted of ED and ES short-axis cine CMR images of 5,660 healthy subjects (with or without cardiovascular risk factors). Subject characteristics for all participants were obtained from the UK Biobank database and are provided in **Table 1**.

For all subjects, the LV endocardial and epicardial borders and the RV endocardial border were manually traced at ED and ES frames using the cvi42 software (version 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Alberta, Canada). 4,975 subjects were previously analyzed by two core laboratories based in London and Oxford (20), the remaining 685 subjects were analyzed by two experienced CMR cardiologists at Guy's and St Thomas' Hospital following the same standard operating procedures described in Petersen et al. (20). For all CMR examinations that underwent manual image analysis, any case with insufficient quality (i.e., presence of artifacts or slice location problems, operator error or evidence of pathology, such as significant shunt or valve regurgitation) were rejected (21). All experts performing the segmentations were blinded to subject characteristics such as race and sex. From our database, 4,410 subjects were used to train and validate the DL-based CMR segmentation model, and 1,250 subjects were used as a test set for the validation of the model and the statistical analysis (split 70/10/20 for training/validation/test set). The train and test sets were stratified to contain approximately the same percentage of samples for each racial group and sex. **Supplementary Figure 1** shows the flow chart for selection of cases for this study.

RESULTS

Deep Learning-Based Image Segmentation Pipeline

Table 2 reports the DSC values between manual and automated segmentations evaluated on the test set of 1,250 subjects which the segmentation model had never seen before. The table shows the mean DSC for LVBP, LVMyo and RVBP for both the full test set and stratified by sex and race. Overall, the average (AVG) DSC was 93.03 $\pm$ 3.83% (94.40 $\pm$ 2.61% for the LVBP, 88.78 $\pm$ 3.08% for the LVMyo and 90.77 $\pm$ 3.96% for the RVBP). **Table 2** shows that the CMR segmentation model had a racial bias for all comparisons but no sex bias (independent Student's *t*-test between each racial group and rest of the population; $p < 0.001$ for LVBP, LVMyo, RVBP and AVG for all races).[1] **Supplementary Figure 2** shows in the first-row visual examples of frames from a cine CMR sequence and their associated ground truth segmentations, and in the two last rows some sample segmentation results (on different frames) for different racial groups with both high and low DSC.

Next, we evaluate the accuracy of the volumetric and functional measures (LVEDV, LVESV, LVEF, LVmass, RVEDV, RESV, RVEF). **Table 3A** reports the mean values based on the manual segmentations, and **Tables 3B,C** report the mean absolute differences and relative differences between automated and manual measurements, respectively. The Bland-Altman plots for agreement between the pipeline and manual analysis are shown in **Supplementary Figure 3**. For the overall population, results are in line with previous reported values (5, 22) and within the inter-observability range (20).

These results show that for sex there is a statistically significant difference in the absolute error for LVEF, LVmass and RVEF

[1] **Table 2** differs from **Table 1** of our previous work (14), as in the present study we have excluded any case with cardiovascular disease.

TABLE 1 | Population characteristics for the train/validation and test sets.

			Train/validation	Test
Continuous variables	Patients, *n*		4,410	1,250
	Age (years; mean, *SD*)		62 (8)	61 (8)
	Height (cm; mean, *SD*)		169 (9)	169 (9)
	Weight (kg; mean, *SD*)		76 (15)	75 (14)
	BMI (kg/m^2; mean, *SD*)		27 (4)	26 (4)
	BSA (m^2; mean, *SD*)		1.86 (0.21)	1.85 (0.20)
	Systolic blood pressure (mmHg; mean, *SD*)		136 (20)	136 (18)
	Diastolic blood pressure (mmHg; mean, SD)		79 (11)	80 (10)
	Heart rate (bpm; mean, SD)		63 (20)	63 (10)
Categorical variables	Sex (males; *n*, %)		2,299 (52)	655 (52)
	Racial group	White (*n*, %)	3,570 (81)	1,025 (81)
		Mixed (*n*, %)	136 (3)	34 (3)
		Asian (*n*, %)	313 (7)	83 (7)
		Black (*n*, %)	190 (4)	47 (4)
		Chinese (*n*, %)	87 (2)	27 (2)
		Other (*n*, %)	144 (3)	34 (3)

All continuous values are reported as mean(SD), while categorical variables are reported as number (percentage). SD, standard deviation.

TABLE 2 | Dice similarity coefficient (DSC) values for the overall test set and by sex and race.

N = 1,250	LVBP	LVMyo	RVBP	AVG
Total	94.39 (2.61)	88.68 (3.06)	90.77 (3.86)	91.28 (3.18)
Male	94.35 (2.55)	89.10 (2.84)	90.61 (3.96)	91.35 (3.12)
Female	94.44 (2.67)	88.59 (3.26)	90.94 (3.94)	91.32 (3.29)
White	95.13 (1.98)***	89.81 (1.48)***	92.24 (2.11)***	92.39 (1.86)***
Mixed	89.79 (1.34)**	80.72 (2.38)**	82.95 (2.53)**	84.49 (2.08)**
Asian	92.15 (2.48)**	86.46 (2.18)*	86.27 (2.63)**	88.29 (2.43)**
Black	91.41 (1.53)***	85.78 (1.73)***	80.88 (2.10)***	86.02 (1.79)***
Chinese	88.98 (2.43)*	79.75 (2.21)*	82.58 (2.32)*	83.77 (2.32)*
Others	90.46 (2.53)*	82.64 (5.44)*	84.77 (3.46)	85.96 (3.81)*

*DSC reported for the LV blood pool (LVBP), LV myocardium (LVMyo) and RV blood pool (RVBP), and average DSC values across LVBP, LVM and RVBP (AVG column). DSC is reported as mean and standard deviation (in parentheses). The first row reports the DSC for the full database, the second and third rows report DSC by sex and the remaining rows report DSC by racial group. Values are reported as mean(SD). Comparison of variables between groups (i.e., male vs. female, white vs. non-white, mixed vs. non-mixed, etc.) was carried out using an independent Student's t-test. Pairwise post hoc testing was carried out using Bonferroni correction for multiple comparisons. Asterisks indicate statistically significant differences between each group and the rest of the test set after correction (28 tests), where *$p < 0.01/28$, **$p < 0.001/28$, ***$p < 0.0001/28$. Exact p-values are reported in* ***Supplementary Table 3****. SD, standard deviation.*

(independent Student's *t*-test $p < 0.001$). For different racial groups, they show that the White and Mixed groups have for all clinical parameters a statistically significant difference in absolute and relative error (except Mixed LVmass $p = 0.66$ and $p = 0.15$ for absolute and relative error, respectively). They also show that there is a statistically significant difference in the absolute and relative errors for LVEDV, LVESV, LVEF (except for absolute error for Black and Other LVESV $p = 0.25$ and $p = 0.01$, respectively, and Black LVEF $p = 0.17$; and relative error for Black LVEDV $p = 0.03$, LVESV $p = 0.53$ and LVEF $p = 0.20$). Interestingly, there is no statistically significant difference in absolute or relative error for RV clinical parameters for the Chinese and Other racial groups.

Multivariable Analysis

To analyze if there is any other factor (i.e., risk factors, patient characteristics) that could explain the bias in DSC between races, we performed a multivariate linear regression between the DSC and race adjusted for patient size, cardiac parameters and cardiovascular risk factors and taking the white group as control. **Table 4** shows the unadjusted [model 1—4(a)] and adjusted [model 2—4(b)] standardized regression beta coefficients [with 95% confidence interval (CI)] for the association between DSC and racial groups. **Supplementary Table 1** shows the full list of standardized regression beta-coefficients from the multivariate analysis for each racial group (model 3), representing the z-score change in variables with the associated factors. Our results show that all associations remained significant after multivariate adjustment and that there is no covariate that can explain the DSC bias between racial groups (see **Table 4B**). For the Mixed and Black race groups, sex shows a weak positive association with DSC (see **Supplementary Table 1**), however, race remains the main factor.

Analysis of Variance

We also compared change of marginal means of DSC between different racial groups using a 1-way ANOVA ($F = 219.43$, $p < 0.0001$, $\eta^2 = 0.47$) and an ANCOVA adjusted for patient size, cardiac parameters and cardiovascular risk factors ($F = 196.237$, < 0.0001, $\eta^2 = 0.44$, see **Supplementary Table 2**). Estimated marginal means are given in **Table 5**, before and after adjustment for the mean of covariates. The results show that there is an overall difference between racial groups, and after adjustment for covariates race still remains the main factor.

Effect of Bias on Heart Failure Diagnosis

The previous experiments have demonstrated that racial bias exists in the DL-based CMR segmentation model. This final experiment aims to provide an example of how this racial bias could potentially have an effect on the diagnosis and characterization of heart failure (HF). To this end, we trained another nnU-Net segmentation model using both healthy and cardiomyopathy subjects from the UK Biobank (training and validation: 4,410 healthy subjects/200 cardiomyopathy subjects and test: 1,250 healthy subjects/150 cardiomyopathy subjects). For the cardiomyopathy test cases, we computed the misclassification rate (MCR, %) between the manual LVEF and the automated LVEF based on the standard classification of HF according to LVEF (23, 24), i.e., HF with reduced EF (HFrEF): HF with an LVEF of $\leq$ 40%; HF with mildly reduced EF (HFmrEF): HF with an LVEF of 41–49%; HF with preserved EF (HFpEF): HF with an LVEF of $\geq$ 50%. The results are presented in **Table 6**. Overall, although the number of subjects in the minority racial groups was relatively small, the misclassification rate using the AI-derived segmentations for White subjects was low, with generally much higher rates for minority races.

DISCUSSION

We have demonstrated for the first time the existence of racial bias in DL-based cine CMR segmentation. The results show that after adjustment for possible confounders such as cardiovascular risk factors the bias persists, suggesting that it is related to the balance of the database used to train the DL model. This conclusion is supported by our earlier work (14), where a model trained with a (much smaller) racially balanced database had

TABLE 3 | Manual clinical measurements (top table) and absolute (middle table) and relative (bottom table) differences in volumetric and functional measures between automated and manual segmentations, overall and by sex and race.

(A) Manual

	iLVEDV (mL/mm^2)	iLVESV (mL/mm^2)	LVEF (%)	iLVmass (g/mm^2)	iRVEDV (mL/mm^2)	iRVESV (mL/mm^2)	RVEF (%)
Total	79 (20)	33 (12)	60 (7)	51 (14)	86 (22)	38 (13)	57 (7)
Male	82 (20)*	36 (12)	59 (7)*	50 (12)	95 (21)*	45 (13)	54 (7)*
Female	72 (14)*	29 (8)	61 (7)*	42 (9)	77 (14)*	32 (8)	58 (6)*
White	83 (20)	35 (12)	59 (6)	51 (14)*	87 (22)*	39 (13)*	56 (6)
Mixed	76 (20)*	27 (9)*	64 (8)*	47 (14)	83 (20)*	35 (10)*	58 (8)*
Asian	70 (18)*	25 (10)*	65 (8)*	48 (12)*	76 (19)*	32 (11)	58 (6)
Black	87 (21)	33 (11)	63 (6)	59 (13)	94 (27)*	41 (14)	56 (6)
Chinese	66 (12)*	22 (7)*	66 (7)*	46 (11)*	75 (16)	32 (8)	58 (6)
Others	77 (19)*	28 (9)	64 (6)*	53 (15)	86 (23)	36 (13)	59 (7)

(B) Absolute difference

	iLVEDV (mL/mm^2)	iLVESV (mL/mm^2)	LVEF (%)	iLVmass (g/mm^2)	iRVEDV (mL/mm^2)	iRVESV (mL/mm^2)	RVEF (%)
Total	2.6 (1.7)	2.1 (1.8)	2.5 (2.4)	3.8 (3.9)	3.5 (2.6)	3.0 (2.2)	3.6 (3.0)
Male	2.7 (1.7)	2.1 (1.7)	2.1 (1.9)*	4.1 (4.2)	3.4 (2.6)	3.0 (2.1)	3.1 (2.7)*
Female	2.6 (1.7)	2.1 (1.8)	2.9 (2.8)*	3.5 (3.4)	3.5 (2.6)	4.6 (2.2)	4.1 (3.3)*
White	2.3 (1.5)	1.9 (1.5)*	2.1 (2.1)*	4.0 (3.3)*	3.2 (2.6)*	2.8 (2.2)	3.4 (2.9)*
Mixed	3.9 (2.1)*	3.4 (1.7)*	4.1 (2.7)	1.9 (1.7)*	4.6 (1.8)*	3.9 (1.8)*	4.9 (2.5)*
Asian	3.4 (1.9)*	2.8 (2.3)*	4.0 (2.9)	2.0 (2.3)*	4.4 (2.4)*	3.4 (1.9)	4.4 (3.3)
Black	3.6 (1.8)*	2.9 (2.8)*	3.3 (3.0)*	2.0 (2.2)*	4.4 (1.6)*	3.5 (1.9)	3.9 (2.6)
Chinese	4.4 (2.2)*	3.4 (2.1)*	4.7 (2.8)*	4.1 (3.6)*	4.8 (2.4)	4.0 (2.9)*	6.4 (5.4)*
Others	3.7 (1.9)	3.1 (2.0)*	4.3 (3.2)	2.3 (2.5)	4.6 (3.4)	3.6 (1.8)*	4.3 (2.8)

(C) Relative difference

	iLVEDV (mL/mm^2)	iLVESV (mL/mm^2)	LVEF (%)	iLVmass (g/mm^2)	iRVEDV (mL/mm^2)	iRVESV (mL/mm^2)	RVEF (%)
Total	3.4 (2.5)	7.1 (7.4)	4.1 (3.9)	8.7 (8.3)	4.3 (3.4)	8.8 (7.5)	6.4 (5.2)
Male	3.0 (2.3)*	6.2 (6.3)*	3.6 (3.1)*	7.8 (6.5)*	3.7 (3.0)*	7.3 (5.9)*	5.8 (5.0)*
Female	3.7 (2.7)*	7.9 (8.2)*	4.6 (4.4)*	9.6 (9.6)*	4.9 (3.7)*	10.2 (8.4)*	7.0 (5.4)*
White	3.0 (2.1)*	6.0 (6.1)*	3.7 (3.6)	8.4 (8.7)	4.0 (3.4)*	8.2 (7.3)	6.0 (5.1)*
Mixed	5.7 (3.1)*	14.1 (8.2)*	6.5 (4.2)*	10.3 (6.1)*	6.2 (2.4)*	13.3 (6.8)*	9.2 (5.1)*
Asian	5.1 (3.2)*	11.8 (11.6)*	5.8 (4.2)*	10.5 (5.4)*	6.1 (3.4)*	11.5 (6.8)	7.2 (4.9)
Black	4.1 (2.3)	7.7 (6.8)	5.1 (4.8)*	7.3 (4.1)	5.1 (2.2)	9.3 (5.9)	7.3 (4.7)
Chinese	7.0 (4.3)*	16.5 (10.6)*	6.9 (3.7)*	13.6 (7.1)*	6.2 (3.2)	13.8 (11.4)*	10.4 (9.4)*
Others	5.0 (2.9)*	12.6 (10.2)	7.7 (5.5)*	8.9 (4.2)	5.2 (3.9)	11.9 (7.0)*	8.1 (4.9)

Clinical measurements for the LV and RV end diastolic volume (EDV), end systolic volume (ESV), ejection fraction (EF), and left ventricular mass (LVmass). All cardiac volumes were indexed to body surface area using the Dubois and Dubois formula (32). We define the absolute and relative errors as $\varepsilon_{absolute} = |v_{manual} - v_{auto}|$ *and* $\varepsilon_{relative}(\%) = 100 * |v_{manual} - v_{auto}|/v_{manual}$*, where* v *corresponds to each clinical measure. Clinical measures are reported as mean and standard deviation (in parentheses). The first row reports the clinical measurements for the full database, the second and third rows report the clinical measurements by sex and the remaining rows report the clinical measurements by racial group. Values are reported as mean(SD). Comparison of variables between groups (i.e., male vs. female, white vs. non-white, mixed vs. non-mixed, etc.) was carried out using an independent Student's t-test. Pairwise post hoc testing was carried out using Bonferroni correction for multiple comparisons. Asterisks indicate statistically significant differences between each group and the rest of the test set after correction (49 tests), i.e.,* $p < 0.01/49$*. Exact p-values are reported on* ***Supplementary Table 4****. SD, standard deviation.*

much reduced bias (although poorer performance overall due to the smaller training database).

Assessment of the Bias in the Deep Learning-Based Cardiac Magnetic Resonance Segmentation Model

For the overall population, the DSC values are in line with previous reported values (5, 22) and with the inter-observer variability range (20). DSC as well as absolute differences and relative differences show a higher bias on the RV, however, this is expected as previous studies have highlighted the difficulty in manual contouring of the RV and the higher variability between observers (20).

The bias we found in segmentation model performance was near-exclusively based on race. Statistically significant differences in some derived volumetric/functional measures (see **Table 3**) were found by sex but these differences were small

TABLE 4 | Associations between average DSC and racial group.

(A) Univariate linear regression

	Standardized beta-coefficients (95% CI)		
	N	**Model 1**	***p*-value**
Mixed	1,250	0.34 (0.30, 0.38)***	6.30E-16
Asian	1,250	0.33 (0.29, 0.37)***	1.57E-12
Black	1,250	0.36 (0.32, 0.40)***	1.30E-19
Chinese	1,250	0.32 (0.28, 0.36)***	1.08E-8
Other	1,250	0.30 (0.26, 0.34)***	4.43E-14

(B) Multivariate linear regression

	Standardized beta-coefficients (95% CI)		
	N	**Model 2**	***p*-values**
Age	1,250	0.03 (–0.02, 0.08)	0.210
Sex	1,250	0.02 (–0.03, 0.08)	0.364
Weight	1,250	0.10 (–0.36, 0.51)	0.699
Height	1,250	0.00 (–0.28, 0.29)	0.972
BMI	1,250	-0.02 (–0.36, 0.36)	0.944
HR	1,250	0.03 (–0.01, 0.07)	0.114
SBP	1,250	-0.01 (–0.07, 0.04)	0.579
DBP	1,250	-0.04 (–0.08, 0.01)	0.114
LVEDV	1,250	-0.02 (–0.21, 0.17)	0.855
LVESV	1,250	-0.07 (–0.20, 0.06)	0.284
RVEDV	1,250	0.12 (–0.09, 0.31)	0.235
RVESV	1,250	-0.11 (–0.24, 0.04)	0.127
Lvmass	1,250	-0.04 (–0.11, 0.02)	0.174
Diabetes	1,250	0.10 (–0.07, 0.27)	0.273
Hypertension	1,250	0.05 (0.00, 0.10)	0.034
Hyper cholesterolemia	1,250	0.00 (–0.04, 0.05)	0.860
Smoking	1,250	0.00 (–0.05, 0.03)	0.812
Center	1,250	0.15 (0.09, 0.21)	9.99E-02
Mixed	1,250	0.38 (0.36, 0.41)**	9.99E-04
Asian	1,250	0.37 (0.34, 0.41)**	9.99E-04
Black	1,250	0.40 (0.38, 0.43)**	9.99E-04
Chinese	1,250	0.36 (0.34, 0.39)**	9.99E-04
Other	1,250	0.34 (0.30, 0.38)**	9.99E-04

*Standardized regression beta-coefficients and CI are shown, representing the z-score change in variables with increasing DSC. The White racial group was selected as control. LV, left ventricle, EDV, end-diastolic volume, ESV, end-systolic volume, SBP, systolic blood pressure, DBP, diastolic blood pressure, CI, confidence interval. Model 1 is unadjusted; Model 2 is adjusted for sex, height, weight, blood pressure at scan-time, heart rate at scan-time, LVEDV, LVESV, RVEDV, RVESV, LVmass, diabetes, hypertension, hypercholesterolemia, smoking and center. *p < 0.01, **p < 0.001, ***p < 0.00001.*

compared to the differences observed in both DSC (**Table 2**) and volumetric/functional measures (**Table 3**) by race. Therefore, none of the confounders used in this study could explain the differences by race. Results from the ANCOVA analysis show that one factor that contributed more to the model was the center where the segmentations were performed. This could be explained by differences in CMR reporting between the core lab and the additional lab. Similarly to the complete UK Biobank database, the subcohort that we used is approximately sex-balanced but not race-balanced, and the highest errors were found for relatively underrepresented racial groups. This phenomenon has been observed before in applications in computer vision (25) and medical imaging (26, 27), but never before reported in CMR image analysis.

We believe that this bias is due to the imbalanced nature of the training data. Combined with previous studies that have shown race-based associations with differences in cardiac physiology using diverse databases (10, 11), the imbalance causes the performance of the DL model to be biased toward the physiology of the majority group (i.e., white subjects), to the detriment of performance on minority racial groups.

Our last experiment showed that using the AI-based predicted EF values will result in higher misclassification rates for the minority races compared to the White subjects, which is in line with the other experiments showing a higher bias for the minority groups.

Consistent Reporting of Sex and Racial Subgroups in Artificial Intelligence Models

It is envisioned that AI will dramatically change the way doctors practice medicine. In the short term, it will assist physicians with easy tasks, such as automating measurements, making predictions based on big data, and putting clinical findings into an evidence-based context. In the long term, it has the potential to significantly optimize patient care, reduce costs, and improve outcomes. With AI models now starting to be deployed in the real world it is essential that the benefits of AI are shared equitably according to race, sex and other demographic characteristics. It has long been known that current medical guidelines have the potential for sex/racial bias due to the imbalanced nature of the cohorts upon which they were based (28, 29). One might think that AI can solve such problems, as they are "neutral" or "blind" to characteristics such as sex and race. However, as we have shown in this paper, when AI models are used naively, they can inherit the bias present in clinical databases. It is important to highlight

TABLE 5 | The comparison of adjusted mean between racial groups based on one-way ANOVA and ANCOVA.

		Mean (95% CI)	
	N	**Model 4**	**Model 5**
White	1,025	0.93 (0.93, 0.93)	0.93 (0.93, 0.93)
Mixed	34	0.84 (0.86, 0.82)	0.83 (0.85, 0.80)
Asian	83	0.89 (0.90, 0.88)	0.88 (0.89, 0.88)
Black	47	0.86 (0.87, 0.85)	0.85 (0.86, 0.83)
Chinese	27	0.84 (0.86, 0.81)	0.82 (0.84, 0.78)
Other	34	0.86 (0.88, 0.85)	0.85 (0.87, 0.83)

Model 4 is unadjusted; Model 5 is adjusted for sex, height, weight, blood pressure at scan-time, heart rate at scan-time, LVEDV, LVESV, RVEDV, RVESV, LVmass, diabetes, hypertension, hypercholesterolemia, smoking, and center. CI, confidence interval. For model 4 and model 5, pairwise post hoc testing was carried out using Scheffé's method.

TABLE 6 | Misclassification rate for HF diagnosis.

		HFrEF LVEF < 40%		HFmrEF LEF 40–49%		HFpEF LVEF ≥ 50%	
	n	n GT	MCR (%)	n GT	MCR	n GT	MCR (%)
White	107	5	3.74	14	5.61	88	7.48
Mixed	11	3	45.45	0	–	8	36.36
Black	8	0	–	4	12.05	4	25.00
Asian	14	4	21.43	2	7.14	8	14.29
Chinese	4	0	–	2	25.00	2	50.00
Other	6	1	33.33	5	16.67	0	–
Minority groups	43	8	23.26	13	9.30	22	23.26

The table summarizes numbers of subjects in each racial group and HF diagnosis (i.e., HFrEF, HFmrEF and HFpEF), as well as the misclassification rate (MCR,%) for each racial group and diagnosis. The row Minority groups combines data from the Mixed, Black, Asian, Chinese and Other groups. The left column (n overall) shows the number of subjects for each racial group used to compute the MCRs. For each HF diagnosis, the first column shows the number of ground truth positive subjects in that group, and the second column shows the MCR. When computing the MCRs, the ground truth negative subjects were all subjects from the other HF diagnoses for that racial group. HFrEF, HF with reduced EF; HFmrEF, HF with mildly reduced EF; HFpEF, HF with preserved EF. Blank cells show regions with missing data.

the potential shortcomings of AI at this stage before AI models become more widely deployed in clinical practice.

For these reasons, we believe that it is necessary that new standards are established to ensure equality between demographic groups in AI model performance, and that there is consistent and rigorous reporting of performance for new AI models that are intended to be integrated into clinical practice. Similar to Noseworthy et al. (30), we would recommend that any new AI-based publication include a report of performance across a range of demographic subgroups, particularly race/sex.

Strategies to Reduce Racial Bias

The obvious way to mitigate bias due to imbalanced datasets (whether in current clinical guidelines or AI models) is to use more balanced datasets. However, this is a multifactorial problem and is associated with many challenges, such as historical discrimination, research design and accessibility (22). We note that AI has the potential to address/mitigate bias without requiring such balanced datasets. A range of bias mitigation strategies have been proposed that either pre-process the dataset to make it less imbalanced, alter the training procedure or post-process the model outputs to reduce bias (31). We have recently proposed three algorithms to mitigate racial bias in CMR image segmentation: (1) train a CMR segmentation algorithm that ensures racial balance during training; (2) add an AI race classifier that helps the segmentation model to capture racial variations; and (3) train a different CMR segmentation model for each racial group. For more detail of these models, we refer to the reader to our previous work (14). All three proposed algorithms result in a fairer segmentation model that aims to ensure that no racial group will be disadvantaged when segmentations of their CMR data are used to inform clinical management. Note that, compared to our previous work (14), in this paper we have excluded all subjects with cardiovascular disease to ensure that racial bias was not influenced by this factor.

Limitations

This study utilizes the imaging cohort from the UK Biobank. UK Biobank is a long-term prospective epidemiology study of over 500,000 persons aged 40–69 years across England, Scotland, and Wales. Therefore, the data are geographically limited to the UK population, which might not reflect geographic, socioeconomic or healthcare differences among other populations. This work uses the UK Biobank participants' self-reported ethnicity, which corresponds to them self-identifying as belonging to ethnic groups based on shared culture and heritage. A possible limitation is that ethnic groups are socially constructed and thus may not serve as reliable proxies for analysis. Future work should aim to perform a similar study using genetic ancestry data, which will make the analysis more generalizable. In addition, Mixed Race was considered to be a single category, whereas in reality this encompasses many different subcategories.

Manual analysis of CMR scans was performed by three independent centers using the same operating procedures for analysis. For the three centers, inter- and intra-observer variability between analysts was assessed by analysis of fifty, randomly selected CMR examinations (20). However, one limitation of this study is that inter- and intra-observer variability was not assessed individually by race and sex. Also, this study is limited by the lack of diversity and relatively small sample sizes for certain racial groups and by the exclusion criteria for comorbid and pre-morbid conditions. The study only includes the following cardiovascular risk factors as confounders: hypertension, hypercholesteremia, diabetes and smoking. However, there are other clinically relevant risk factors such as sedentarism, alcohol consumption or stress that could potentially explain the bias found in our study. For instance, a previous study showed an association between RV size and living in a high traffic area (7). Another limitation is that current analysis does not adjust for any measures of ventricular function, which could explain the structural differences. Future work will aim to extract echocardiographic measures of relaxation to assess whether the current bias could be explained by changes in subclinical diastolic dysfunction.

CONCLUSION

We have demonstrated that a DL-based cine CMR segmentation model derived from an imbalanced database has poor

generalizability across racial groups and has the potential to lead to inequalities in early diagnosis, treatments and outcomes. Therefore, for best practice, we recommend reporting of performance among diverse groups such as those based on sex and race for all new AI tools to ensure responsible use of AI technology in cardiology.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The UK Biobank datasets are publicly available for approved research projects. Requests to access these datasets should be directed to https://www.ukbiobank.ac.uk/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the NHS National Research Ethics Service on 17th June 2011 (Ref 11/NW/0382). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EP-A designed, developed the method, and analyzed the data. AK, RR, BR, JM, PC, and EP-A conceived the study. BR, RR, SKP, SN, and SEP provided the manual segmentation used for the implementation of the method. PC, RR, and AK were part of the supervision of EP-A. AK and EP-A wrote the manuscript with input from all authors.

FUNDING

EP-A and AK were supported by the EPSRC (EP/R005516/1) and by core funding from the Wellcome/EPSRC Centre for Medical Engineering (WT203148/Z/16/Z). This research was funded in whole, or in part, by the Wellcome Trust WT203148/Z/16/Z. For the purpose of open access, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. SEP, AK, and RR acknowledge funding from the EPSRC through the Smart Heart Programme grant (EP/P001009/1). EP-A, BR, JM, AK, and RR acknowledged support from the Wellcome/EPSRC Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z), the NIHR Cardiovascular MedTech Co-operative award to the Guy's and St Thomas' NHS Foundation Trust and the Department of Health National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London. SEP, SN, and SKP acknowledged the British Heart Foundation for funding the manual analysis to create a cardiovascular magnetic resonance imaging reference standard for the UK Biobank imaging resource in 5,000 CMR scans (www.bhf.org.uk; PG/14/89/31194). SEP acknowledged support from the National Institute for Health Research (NIHR) Biomedical Research Centre at Barts. SEP has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 825903 (euCanSHare project). SEP also acknowledged support from the CAP-AI Programme, London's First AI Enabling Programme focused on stimulating growth in the capital's AI Sector. CAP-AI was led by Capital Enterprise in partnership with Barts Health NHS Trust and Digital Catapult and was funded by the European Regional Development Fund and Barts Charity. SEP acknowledged support from the Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. SN and SKP were supported by the Oxford NIHR Biomedical Research Centre and the Oxford British Heart Foundation Centre of Research Excellence.

ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank Resource (application numbers 17,806 and 2,964) on a GPU generously donated by NVIDIA Corporation. The UK Biobank data are available for approved projects from https://www.ukbiobank.ac.uk/.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.859310/full#supplementary-material

REFERENCES

1. Constantinides P, Fitzmaurice DA. Artificial intelligence in cardiology: applications, benefits and challenges. *Br J Cardiol.* (2018) 7:25–86.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* (2017) 542:115–8. doi: 10.1038/nature21056
3. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med.* (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5
4. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. *J Am Coll Cardiol.* (2018) 71:2668–79.
5. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson.* (2018) 20:65. doi: 10.1186/s12968-018-0471-x
6. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng P-A, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging.* (2018) 37:2514–25. doi: 10.1109/TMI.2018.2837502

7. Yoneyama K, Venkatesh BA, Bluemke DA, McClelland RL, Lima JAC. Cardiovascular magnetic resonance in an adult human population: serial observations from the multi-ethnic study of atherosclerosis. *J Cardiovasc Magn Reson.* (2017) 19:52. doi: 10.1186/s12968-017-0367-1
8. Holmes MD. Racial inequalities in the use of procedures for ischemic heart disease. *JAMA.* (1989) 261:3242–3. doi: 10.1001/jama.1989.03420220056014
9. Regitz-Zagrosek V, Oertelt-Prigione S, Prescott E, Franconi F, Gerdts E, Foryst-Ludwig A, et al. Gender in cardiovascular diseases: impact on clinical manifestations, management, and outcomes. *Eur Heart J.* (2016) 37:24–34. doi: 10.1093/eurheartj/ehv598
10. Oertelt-Prigione S, Regitz-Zagrosek V. *Sex and Gender Aspects in Clinical Medicine*. London: Springer (2012).
11. Kawut SM, Lima JAC, Barr RG, Chahal H, Jain A, Tandri H, et al. Sex and race differences in right ventricular structure and function. *Circulation.* (2011) 123:2542–51. doi: 10.1161/CIRCULATIONAHA.110.985515
12. Captur G, Zemrak F, Muthurangu V, Petersen SE, Li C, Bassett P, et al. Fractal analysis of myocardial trabeculations in 2547 study participants: multi-ethnic study of atherosclerosis. *Radiology.* (2015) 277:707–15. doi: 10.1148/radiol.2015142948
13. Kishi S, Reis JP, Venkatesh BA, Gidding SS, Armstrong AC, Jacobs DR, et al. Race–ethnic and sex differences in left ventricular structure and function: the coronary artery risk development in young adults (CARDIA) study. *J Am Heart Assoc.* (2015) 4:e001264. doi: 10.1161/JAHA.114.001264
14. Puyol-Antón E, Ruijsink B, Piechnik SK, Neubauer S, Petersen SE, Razavi R, et al. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*. Cham: Springer (2021). p. 413–23.
15. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* (2015) 12:e1001779. doi: 10.1371/journal.pmed.1001779
16. Office for National Statistics, National Records of Scotland, Northern Ireland Statistics and Research Agency. *2011 Census Aggregate Data* (Edition: February 2017). UK Data Service (2017). doi: 10.5257/census/aggregate-2011-2
17. Petersen SE, Matthews PM, Francis JM, Robson MD, Zemrak F, Boubertakh R, et al. UK Biobank's cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson.* (2015) 18:8. doi: 10.1186/s12968-016-0227-4
18. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z
19. Barker LE, Shaw KM. Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. *Am J Clin Nutr.* (2015) 102:533–9. doi: 10.3945/ajcn.115.113498
20. Petersen SE, Aung N, Sanghvi MM, Zemrak F, Fung K, Paiva JM, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK biobank population cohort. *J Cardiovasc Magn Reson.* (2017) 19:1–19. doi: 10.1186/s12968-017-0327-9
21. Carapella V, Jiménez-Ruiz E, Lukaschuk E, Aung N, Fung K, Paiva J, et al. Towards the semantic enrichment of free-text annotation of image quality assessment for UK biobank cardiac cine MRI scans. In: Carneiro G, Mateus D, Peter L, Bradley A, Tavares JMR, Belagiannis V, et al. editors. *Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science*. (Vol. 10008), Cham: Springer (2016). doi: 10.1007/978-3-319-46976-8_25
22. Ruijsink B, Puyol-Antón E, Oksuz I, Sinclair M, Bai W, Schnabel JA, et al. Fully automated, quality-controlled cardiac analysis from CMR. *JACC Cardiovasc Imaging.* (2020) 13:684–95. doi: 10.1016/j.jcmg.2019.05.030
23. Bozkurt B, Coats AJ, Tsutsui H, Abdelhamid M, Adamopoulos S, Albert N, et al. Universal definition and classification of heart failure. *J Card Fail.* (2021) 27:387–413.
24. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J.* (2016) 37:2129–200.
25. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C editors. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. New York, NY (2018). p. 77—-91.
26. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. *CheXclusion: Fairness Gaps in Deep Chest X-Ray Classifiers*. Singapore: World Scientific (2020).
27. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci USA.* (2020) 117:12592–4. doi: 10.1073/pnas.1919012117
28. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Smedley BD, Stith AY, Nelson AR editors. Washington, DC: National Academies Press (2003).
29. Smith Taylor J. Women's health research: progress, pitfalls, and promise. *Health Care Women Int.* (2011) 32:555–6. doi: 10.17226/12908
30. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, et al. Assessing and mitigating bias in medical artificial intelligence. *Circ Arrhythm Electrophysiol.* (2020) 13:e007988. doi: 10.1161/CIRCEP.119.007988
31. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. In: *Proceedings of the ACM Computing Surveys (CSUR)*. (Vol. 54), New York, NY: Association for Computing Machinery (2019), 1–35. doi: 10.1145/3457607
32. Du Bois D, Du Bois EF. A formula to estimate the approximate surface area if height and weight be known. *Arch Intern Med.* (1916) 17:863–71. doi: 10.1001/archinte.1916.00080130010002

Author Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, EPSRC, or the Department of Health

Conflict of Interest: SEP provided consultancy to and is shareholder of Circle Cardiovascular Imaging, Inc., Calgary, Alberta, Canada.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.