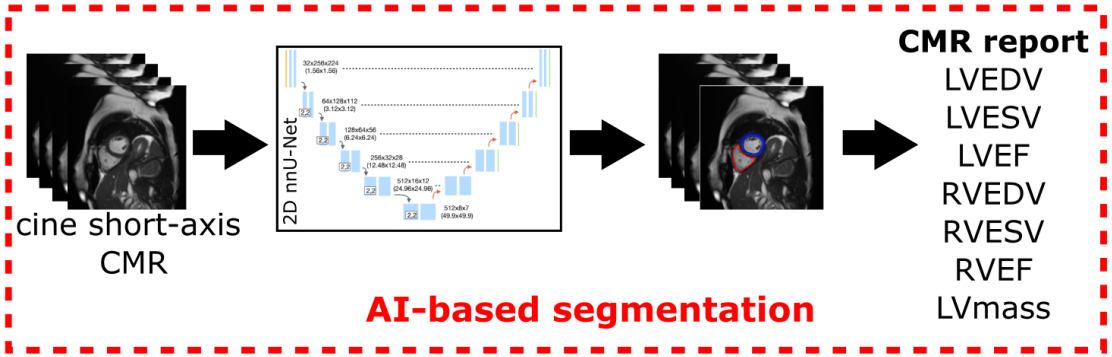
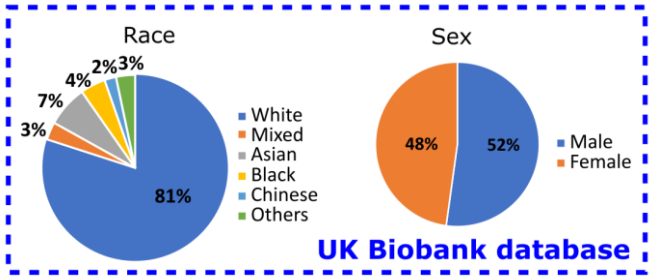
Fairness in AI: Are deep learning-based CMR segmentation algorithms biased?

E. Puyol-Antón[1*], B. Ruijsink[1], SK. Piechnik[2], S. Neubauer[2], SE. Petersen[3], R. Razavi[1] and AP. King[1]
1. King's College London, UK; 2. Division of Cardiovascular Medicine and Oxford NIHR Biomedical Research Centre, UK;
3. William Harvey Research Institute, London, UK. *Contact: esther.puyol_anton@kcl.ac.uk, website: kclmmag.org

Purpose: To perform the first analysis of sex/racial bias in AI-based cine CMR segmentation using a large-scale database.

Methods: 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 *co-variates*, a multivariate linear regression and ANCOVA were performed.





Co-variates	
Age	LVEDV
Sex	LVESV
Weight	RVEDV
Height	RVESV
BMI	LVmass
Heart rate	Hypertension
Systolic BP	Hypercholesteremia
Diastolic BP	Diabetes
Centre	Smoking

Results:

N = 1,250	Dice similarity	Absolute difference			
		LVEDV	LVESV	LVmass	RVEDV
Total	93.0 (3.8)	4.6 (3.0)	3.7 (3.1)	7.4 (5.6)	6.2 (4.7)
Male	93.0 (3.6)	4.7 (3.0)	3.7 (2.9)	7.9 (6.2)*	6.1 (4.6)
Female	93.1 (4.0)	4.6 (3.0)	3.6 (3.2)	6.8 (5.0)*	6.3 (4.7)
White	93.9 (3.1)	4.2 (2.7)*	3.3 (2.8)*	7.1 (5.9)*	5.9 (4.7)*
Mixed	86.7 (2.1)	7.1 (3.5)*	6.2 (2.9)*	7.7 (4.3)	8.5 (3.1)*
Asian	89.8 (4.4)	6.1 (3.5)*	4.9 (4.1)*	8.7 (4.3)*	8.2 (4.3)*
Black	89.9 (2.6)	6.2 (3.3)*	4.3 (3.8)	7.3 (3.7)	7.9 (2.7)*
Chinese	86.3 (5.5)	8.0 (3.9)*	6.4 (4.1)*	10.6 (4.8)*	8.2 (4.0)
Others	88.8 (2.8)	6.3 (3.2)*	5.7 (4.0)	7.6 (3.6)	7.3 (5.7)

Influence of co-variates

* Standardised multivariate regression analysis:
  - No covariate can explain the DSC bias between racial groups.
  - Mixed and Black race groups, gender shows a weak positive association with DSC.
* ANCOVA analysis:
  - Race is the main factor that explain overall differences between racial groups.

Conclusions: Racial bias can exist in DL-based cine CMR segmentation models. *Reason*: Unbalanced nature of the training data (combined with physiological differences)