The Role of AI in Image De-Identification

AI's ability to detect demographics

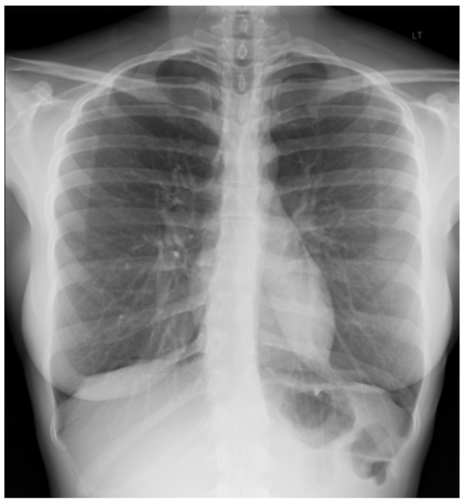
Judy Gichoya, MD MS
Director, HITI Lab
Assistant Professor, Department of Radiology
Emory University

Disclosures

- ACR
  - AI advisory council
- RSNA
  - Associate Editor - Radiology AI Trainee Editorial Board
  - CIRE Committee member
- SIIM
  - Co-chair – Research Committee
  - Board member
- NSF grant
  - FW-HTF-RM: Measuring learning gains in man-machine assemblage when augmenting radiology work with artificial intelligence

- HL7 and AHLI Board member
  - Association for Health Learning and Inference
- Softbrew LTD
  - Consulting on Global Health /Clinical informatics
- Funding –
  - NBIB MIDRC / COVID -19 Data repository
  - Clairity Consortium
  - NIH AIM AHEAD pilot grant
  - RSNA Health disparities grant
  - DeepLook

RECAP

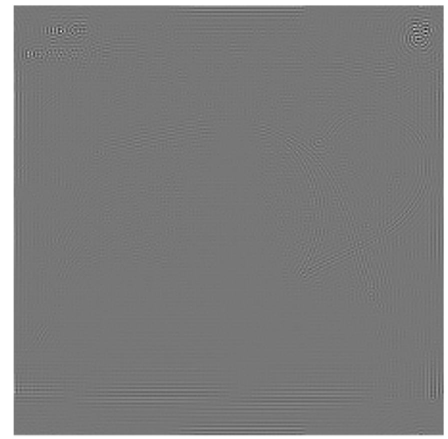


The Lancet Digital Health
Available online 11 May 2022
In Press, Corrected Proof

Articles

AI recognition of patient race in medical imaging: a modelling study

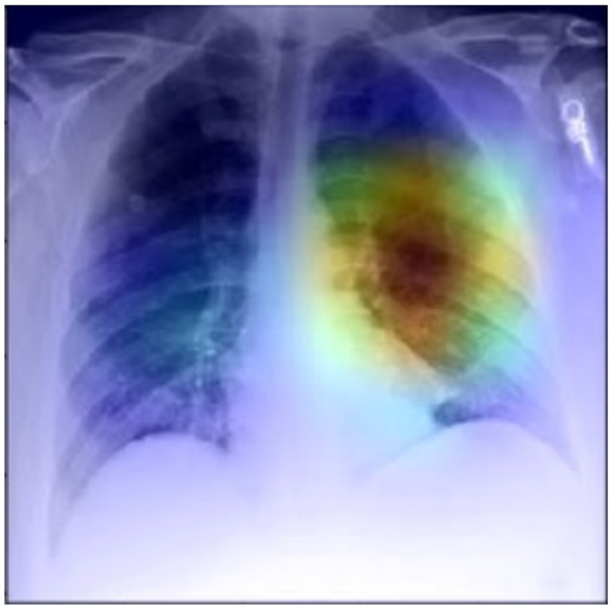
Judy Wawira Gichoya MD [a], Imon Banerjee PhD [c], Ananth Reddy Bhimireddy MS [a], John L Burns MS [d], Leo Anthony Celi MD [e, g], Li-Ching Chen BS [h], Ramon Correa BS [c], Natalie Dullerud MS [i], Marzyeh Ghassemi PhD [e, f], Shih-Cheng Huang [j], Po-Chih Kuo PhD [h], Matthew P Lungren MD [j], Lyle J Palmer PhD [k, l], Brandon J Price MD [m], Saptarshi Purkayastha PhD [d], Ayis T Pyrros MD [n], Lauren Oakden-Rayner MD [k], Chima Okechukwu MS [o] … Haoran Zhang MS [i]



1) Performance of deep learning models to detect race from medical images across modalities and external datasets
2) Assessment of possible anatomic and phenotype confounders such as body habitus and disease distribution
3) Investigation into underlying mechanisms by which AI models can recognize race.

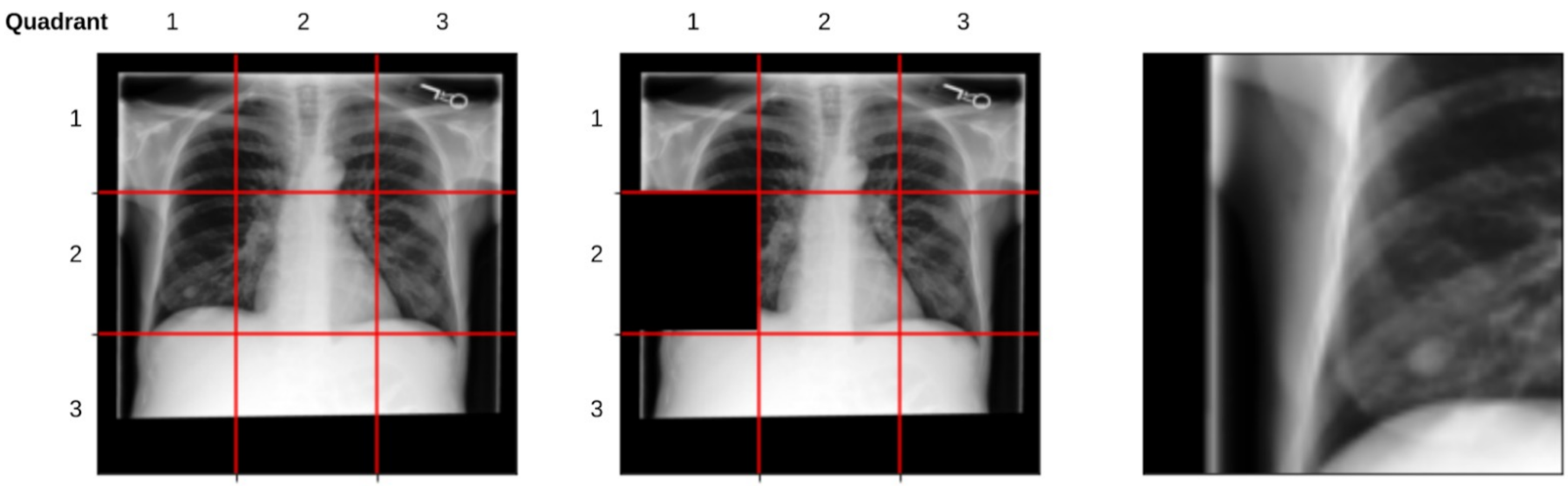
Race detection in radiology imaging	
Chest x-ray (internal validation)*	
MXR (Resnet34, Densenet121)	0·97, 0·94
CXP (Resnet 34)	0·98
EMX (Resnet34, Densenet121, EfficientNet-B0)	0·98, 0·97, 0·99
Chest x-ray (external validation)*	
MXR to CXP, MXR to EMX	0·97, 0·97
CXP to EMX, CXP to MXR	0·97, 0·96
EMX to MXR, EMX to CXP	0·98, 0·98
Chest x-ray (comparison of models)†	
MXR, CXP, EMX	Multiple results (appendix p 26)
CT chest (internal validation)*	
NLST (slice, study)	0·92, 0·96
CT chest (external validation)*	
NLST to EM-CT (slice, study)	0·80, 0·87
NLST to RSPECT (slice, study)	0·83, 0·90
Limb x-ray (internal validation)*	
DHA	0·91
Mammography*	
EM-Mammo (image, study)	0·78, 0·81
Cervical spine x-ray*	
EM-CS	0·92

Image Obscuration



	Asian	Black	White
MXR Densenet121-Original	0.93	0.94	0.94
MXR Densenet121-Masked	0.88	0.79	0.79

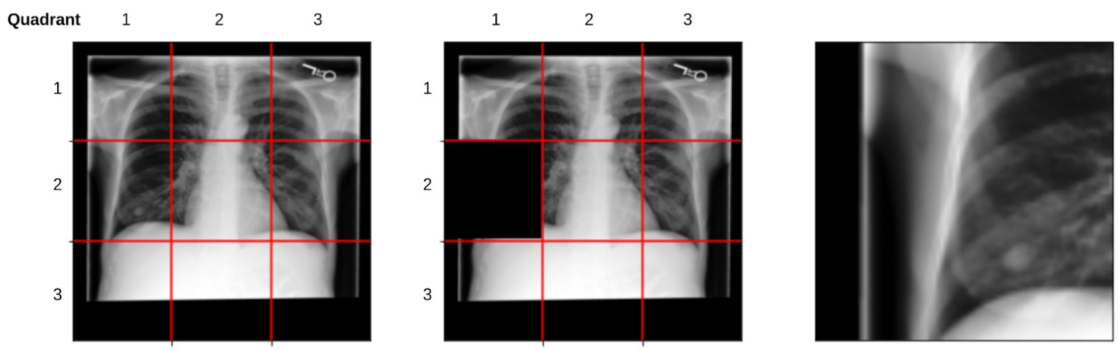
Patch predictions and exclusions



a) Original image b) Removing a patch c) Using only one patch

@judywawira

Patch predictions and exclusions



a) Original image b) Removing a patch c) Using only one patch

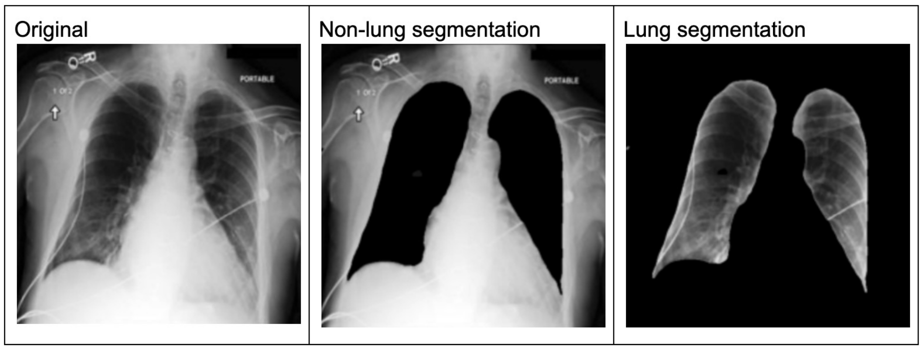
Patch exclusion

Quadrant	1	2	3
1	0.87	0.88	0.87
2	0.81	0.82	0.81
3	0.75	0.60	0.75

Single Patch Training

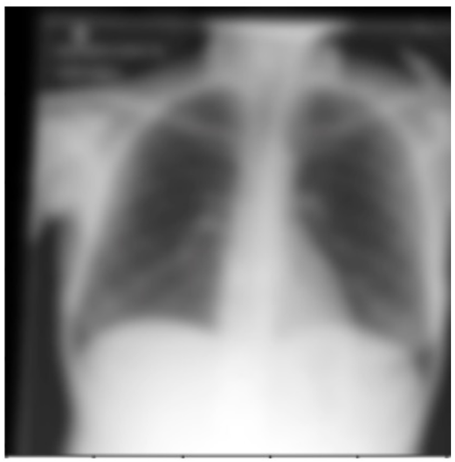
Quadrant	1	2	3
1	0.91	0.90	0.91
2	0.91	0.91	0.91
3	0.91	0.91	0.91

Anatomic Segmentation



	Asian	Black	White
MXR Densenet121-Original	0.93	0.94	0.94
MXR Densenet121-Non lung	0.87	0.85	0.87
MXR Densenet121-Lung	0.68	0.74	0.73

Image Degradation Experiments



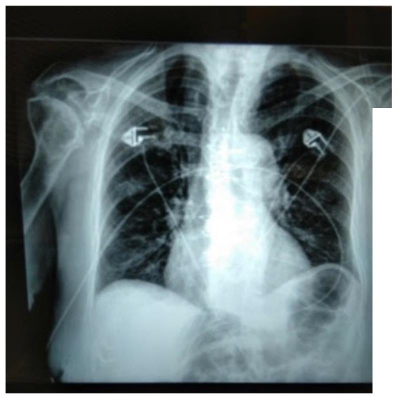
	Asian	Black	White
MXR Densenet121-Original	0.93	0.94	0.94
MXR Densenet121-Noisy	0.64	0.72	0.70
MXR Densenet121-Blurred	0.59	0.64	0.62

Image Degradation Experiments

Model trained on CheXphoto performance compared to CheXpert

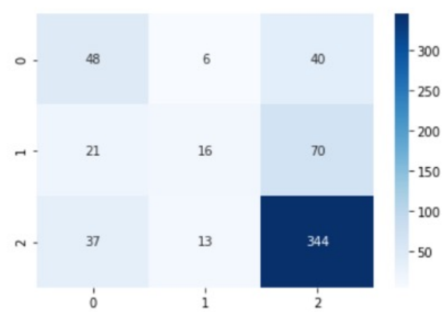
Original (CheXpert)

CheXphoto



CheXphoto

	AUC
Asian	0.81
Black	0.77
White	0.80



CheXpert

	AUC
Asian	0.90
Black	0.94
White	0.89

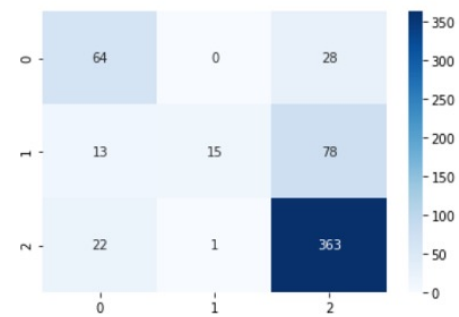
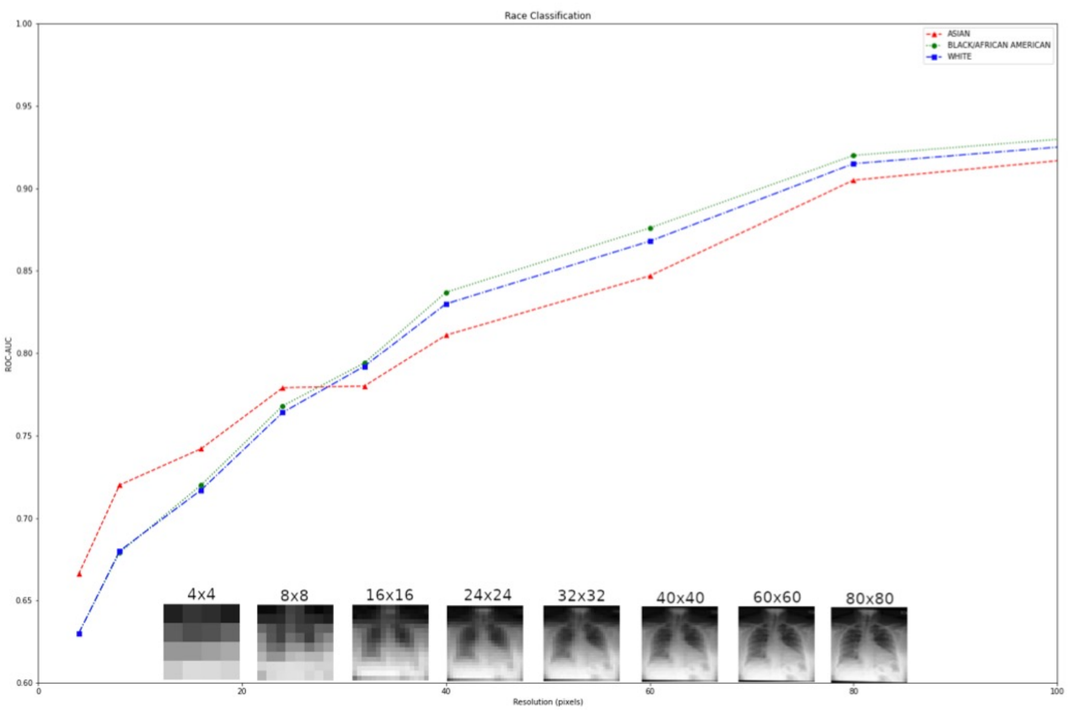
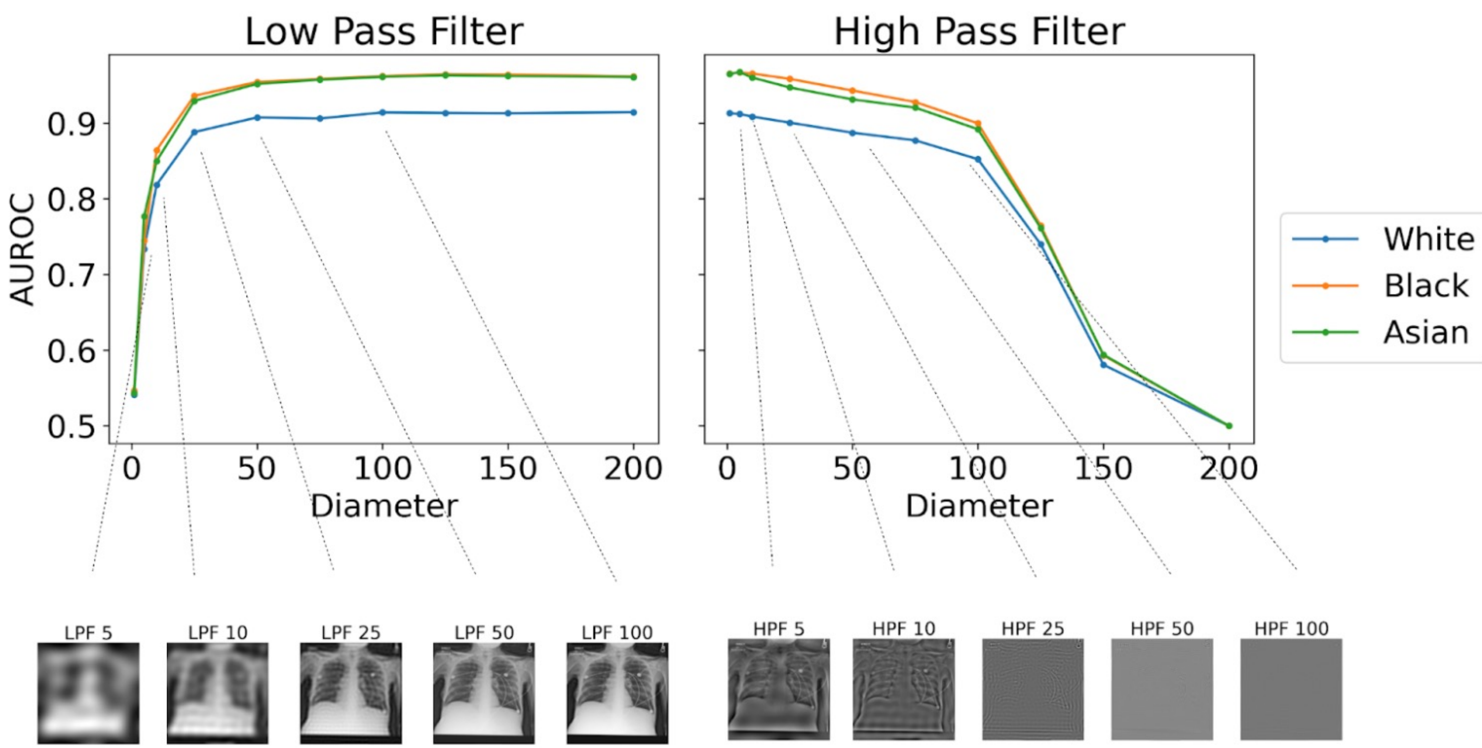


Image Degradation Experiments

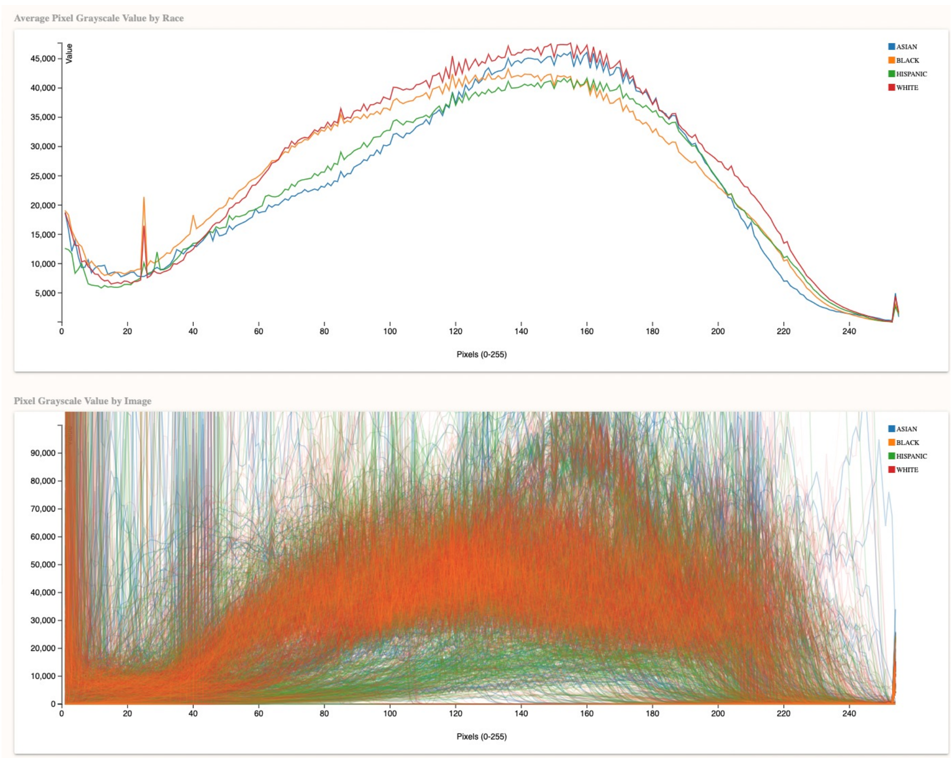


Race	Resolution											
	4	8	16	24	32	40	60	80	160	240	320	512
Asian	0.66	0.72	0.74	0.78	0.78	0.81	0.85	0.90	0.95	0.96	0.97	0.98
Black	0.63	0.68	0.72	0.77	0.79	0.84	0.88	0.92	0.96	0.97	0.97	0.98
White	0.63	0.68	0.72	0.76	0.79	0.83	0.87	0.92	0.96	0.96	0.97	0.97

Image Degradation Experiments



Data Visualization: Pixel Color Averages by Race in Chest X-Ray



https://ai-vengers.web.app

Pixel Intensity Averages by Race in Chest X-Ray

INDIANA UNIVERSITY
SCHOOL OF MEDICINE

John Burns[1], Zachary Zaiman[2], Gaoxiang Luo[3], Le Peng[3], Brandon Price[4], Garric Mathias[1], Vijay Mittal[1], Akshay Sanage[1], Jack Vanschaik[1], Christopher Tignanelli[3], Sunandan Chakraborty[1], Judy Wawira Gichoya[2], Saptarshi Purkayastha[1]

[1] Indiana University Purdue University at Indianapolis, Indianapolis IN 46202, USA; [2] Emory University, Atlanta GA 30322, USA;
[3] University of Minnesota Twin Cities, Minneapolis MN 55455, USA; [4] University of Florida College of Medicine, Gainesville FL 32610, USA

ABSTRACT

Recent work using neural networks has shown that self-reported race is embedded in medical imaging. We seek an understanding of the mechanism that reveals race within medical imaging, investigating the possibility that race is embedded within the individual pixel intensities of grayscale medical images.

Using datasets from 3 institutions and MIMIC-CXR with a combined total of 298,827 images, **we remove all image structure, count how many times each intensity value appears, and standardize to percent per image (PPI)**. Visual analysis, statistical tests, and machine learning processes show that intensity PPI contains self-reported race information. **The best performing model using gradient boosted trees can predict with an AUROC of 77.24.** We investigate confounders of body habitus using Body Mass Index (BMI) and scanner specific settings by limiting to one scanner. Neither factor significantly influences the model.

BACKGROUND

Medical imaging artificial intelligence (AI) models produce racial disparities [2, 3]. There is potential for discriminatory harm if we assume that AI models are agnostic to race. Understanding the relationship between race and medical imaging AI models is important [4]. *AI recognition of patient race in medical imaging: a modelling study* authors found that self-reported race is trivially predictable by AI models. They show that AI models can predict self-reported race across multiple imaging modalities, various datasets, and diverse clinical tasks.

These results are surprising, particularly as this task has not been shown possible for human experts. This capability is trivially learned and therefore likely to be present in many medical image analysis models, providing a direct vector for the reproduction or exacerbation of the racial disparities that already exist in medical practice [1].

MATERIALS and METHODS

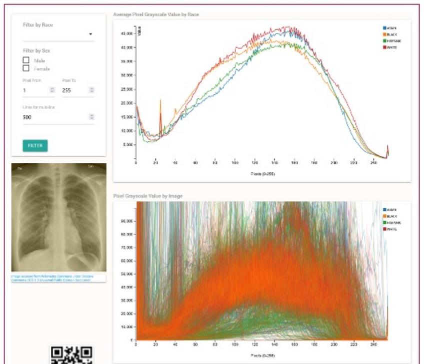
The dataset consists of 3 large academic health centers and one publicly available dataset MIMIC-CXR [5] and is 298,827 images total.

- Institution 1 has 2 datasets: 1 year at hospital W (1.1); and 1 year at hospitals X, Y, Z (1.2) – each limited to the top 10% diverse X-Ray devices.
- Institution 2 has 5 datasets: 1 uncontrolled (2.1); and 4 limited to a single X-Ray device categorized by body habitus using BMI – underweight (2.2), normal (2.3), overweight (2.4), and obese (2.5).
- Institution 3 has 1 uncontrolled dataset (3).

Data is converted from DICOM to 8-bit PNG format, then each intensity value is counted, and then the counts are converted to PPI to normalize for size of image.

Visualizations, MANOVA, and machine learning are applied for individual datasets and all combined

VISUALIZATION RESULTS



Web based visualization - ai-vengers.web.app

STATISTICAL RESULTS

	ANOVA		MANOVA			MANOVA Balanced	
Dataset	# P < 0.05	# F > 2	Df	F-Value	P-Value	F-Value	P-Value
1.1	217	234	762	1.49	0.0000	1.14	0.0119
1.2	184	210	762	1.22	0.0000	1.16	0.0031
1 - All	215	223	762	1.64	0.0000	1.30	0.0000
2.1	245	252	508	3.23	0.0000	2.99	0.0000
2.2	138	158	762	1.38	0.0000	0.93	0.7726
2.3	241	244	762	2.65	0.0000	1.13	0.0140
2.4	241	246	762	2.88	0.0000	1.36	0.0000
2.5	226	239	762	2.73	0.0000	1.18	0.0015
2 - All	245	250	762	7.38	0.0000	2.02	0.0000
3	220	237	508	2.58	0.0000	1.67	0.0000
MIMIC	243	246	762	7.04	0.0000	2.90	0.0000
Combined – No MIMIC	248	249	762	8.63	0.0000	3.37	0.0000
Combined – All	255	255	762	35.64	0.0000	11.07	0.0000

ANOVA –

- With 95% confidence we can reject the null hypothesis for most grayscale values.
- As a combined dataset 255 of 256 values are $p < 0.05$ and 255 of 256 values $F > 2$.
- ANOVA assumes variables are uncorrelated, many pixel counts appear to be highly correlated with other pixel counts.

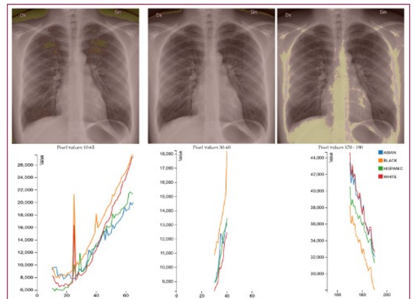
MANOVA with Bonferroni multiple-comparison adjustment at an $\alpha=0.05$ –

- **All MANOVA tests have significant p values, indicating that for all source datasets, the pixel percentage distribution is significantly different across different races.**

MANOVA balanced –

- Datasets 1.1, 2.2, 2.3 cannot reject the null hypothesis when classes are balanced.

REGIONS OF INTEREST



Interesting intensity value relationships and ROI-

- 10-65 have more pixels for Black patients and appear to correlate with background, skin/muscle, and some lung areas
- 30-40 is minimal soft tissue in the example CXR of the ROI above
- 170-190 correlating to bone and some organ systems and less pixels for Black patients

MACHINE LEARNING RESULTS

Dataset	Binary Black or White					Binary Black or all				
	Feed-Forward Network		Decision Tree [RF|GBT|Cart]			Feed-Forward Network		Decision Tree [RF|GBT|Cart]		
	Accuracy	AUROC	Model	Accuracy	AUROC	Accuracy	AUROC	Model	Accuracy	AUROC
1.1	60.9	63.2	RF	57.3	60.6	62.9	57.9	RF	65.3	58.3
1.2	58.8	59.8	GBT	57.1	61.5	63.6	54.1	RF	65.9	62.4
1 – All	57.1	58.2	RF	60.5	66.8	64.5	58.0	RF	66.3	63.2
2.1	60.6	63.5	RF	63.4	67.7	67.5	64.6	RF	66.3	66.6
2.2	60.6	64.6	RF	64.0	67.3	54.5	52.9	RF	61.4	65.1
2.3	63.2	62.0	RF	63.5	67.3	62.5	64.4	RF	65.1	66.8
2.4	59.7	62.2	RF	65.2	65.5	65.3	64.6	RF	66.0	67.6
2.5	62.2	65.3	RF	61.6	64.7	60.4	62.9	RF	62.9	66.4
2 - All	61.7	64.5	RF	64.7	69.6	62.5	63.4	RF	65.9	68.8
3	67.4	67.9	**RF**	**70.5**	**74.1**	68.5	66.1	RF	71.9	72.6
MIMIC	80.5	61.2	GBT	80.4	61.7	82.4	60.2	GBT	82.3	60.0
Combined – No MIMIC	58.4	62.5	GBT	63.0	66.8	61.2	62.7	GBT	64.3	65.8
Combined - All	**75.0**	**69.2**	GBT	75.6	70.4	**77.0**	**68.4**	**GBT**	**68.5**	**77.2**

Feed-Forward Networks (FFN)had some success in predicting race, with the highest accuracy of 77.0% and AUROC of 68.4%. In general, model performance follows dataset size. However, Institution 2 single modality/body habitus has better results than the combined.

Gradient Boosted Trees outperformed FFN and other decision trees. The best accuracy of 68.5% and AUROC of 77.2% shows that self-reported race can be identified without image structure.

SUMMARY

- **ANOVA/MANOVA results show that there is a significant relationship between intensity PPI and self-reported race**
- Visualization of this data proved critical for analysis and idea generation
- FFN were unable to accurately predict self-reported race from PPI
- **Gradient Boosted Trees achieved a best AUROC of 77.24%, showing that PPI can be used to predict race**
- There is little evidence that modality configurations or BMI are correlated to model performance
- Identifying race using pixel gray-scale values without image structure is novel
- **Race-bias information exists in x-ray images, even when image structure is removed**

CONCLUSIONS

Prior studies utilized CNN and the full image to achieve high AUROC in race prediction. We have shown predictive value in gray-scale PPI for self-reported race classification. Future work includes-

- Histogram normalization to remove race from images and evaluating for clinically-relevant information loss
- Review of additional modalities and body parts for similar information
- Using 3D CT, segment body parts and evaluate PPI for regions of images

Diversity of data in medical imaging AI is important. It is unknown how racial bias affects medical imaging AI – though it is shown to affect multiple other medical disciplines. Researchers and vendors should publish training/testing dataset information inclusive of diversity-performance information.

REFERENCE

1. Gichoya, Judy Wawira et al. "AI recognition of patient race in medical imaging: a modelling study." The Lancet. Digital health vol. 4,6 (2022): e406-e414. doi:10.1016/S2589-7500(22)00063-2.
2. E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer, "An algorithmic approach to reducing unexplained pain disparities in underserved populations," (in eng), Nat Med, vol. 27, no. 1, pp. 136-140, Jan 2021, doi: 10.1038/s41591-020-01192-7.
3. L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest X-ray classifiers," (in eng), Pac Symp Biocomput, vol. 26, pp. 232-243, 2021.
4. A. Tariq et al., "Current Clinical Applications of Artificial Intelligence in Radiology and Their Best Supporting Evidence," (in eng), J Am Coll Radiol, vol. 17, no. 11, pp. 1371-1381, Nov 2020, doi: 10.1016/j.jacr.2020.08.018.
5. T. J. P. Alistair E. W. Johnson, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, Steven Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1901.07042.

Confounders mediate AI prediction of demographics in medical imaging

Grant Duffy, Shoa L. Clarke, Matthew Christensen, Bryan He, Neal Yuan, Susan Cheng & David Ouyang

npj Digital Medicine **5**, Article number: 188 (2022) | Cite this article

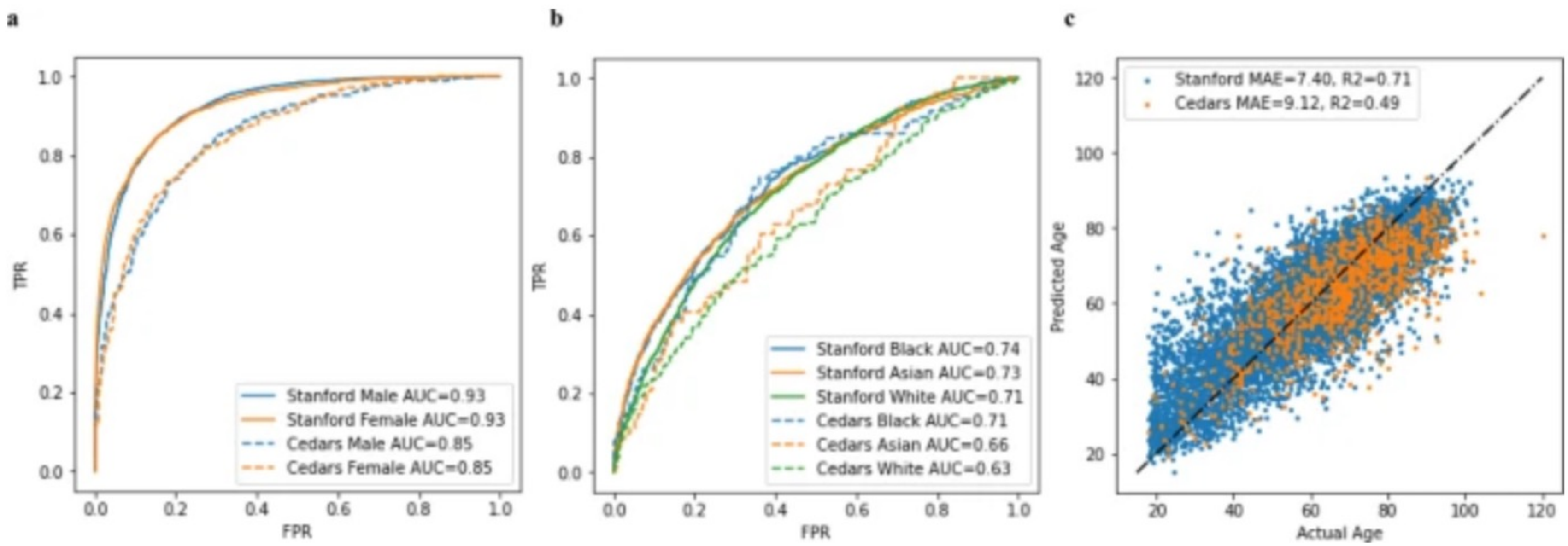
Table 1. Demographic characteristics of study participants.

	CSMC				SHC
	Apical 4 chamber	Apical 2 chamber	Parasternal long axis	Subcostal	Apical 4 chamber
n, patients	28,450	25,502	28,685	23,596	99,909
n, videos	186,426	71,086	110,399	65,558	99,909
Age (mean (SD))	66.5 (±16.5)	66.7 (±16.5)	66.1 (±16.5)	66.2 (±16.4)	59.9 (±17.7)
Male (%)	15,713 (55.2%)	14,093 (55.3%)	15,739 (54.9%)	12,884 (54.6%)	55,610 (55.7%)
Race/ethnicity, *n* (%)					
American Indian	65 (0.2%)	57 (0.2%)	66 (0.2%)	56 (0.2%)	267 (0.3%)
Asian	2162 (7.6%)	1945 (7.6%)	2157 (7.5%)	1808 (7.7%)	14,197 (14.2%)
Black	4058 (14.3%)	3681 (14.4%)	4156 (14.5%)	3322 (14.1%)	4826 (4.8%)
Pacific Islander	87 (0.3%)	82 (0.3%)	86 (0.3%)	75 (0.3%)	1428 (1.4%)
White	19,519 (68.6%)	17,444 (68.4%)	19,595 (68.3%)	16,211 (68.7%)	56,498 (56.5%)
Other	1980 (7.0%)	1790 (7.0%)	2021 (7.0%)	1659 (7.0%)	17,452 (17.5%)
Unknown	579 (2.0%)	503 (2.0%)	604 (2.1%)	465 (2.0%)	5241 (5.2%)

CSMC Cedars-Sinai Medical Center, *SHC* Stanford Healthcare.

% White-male 58.5%
% Black-male 54.1%

Fig. 1: AI model performance in predicting demographics in with unadjusted training and test datasets.

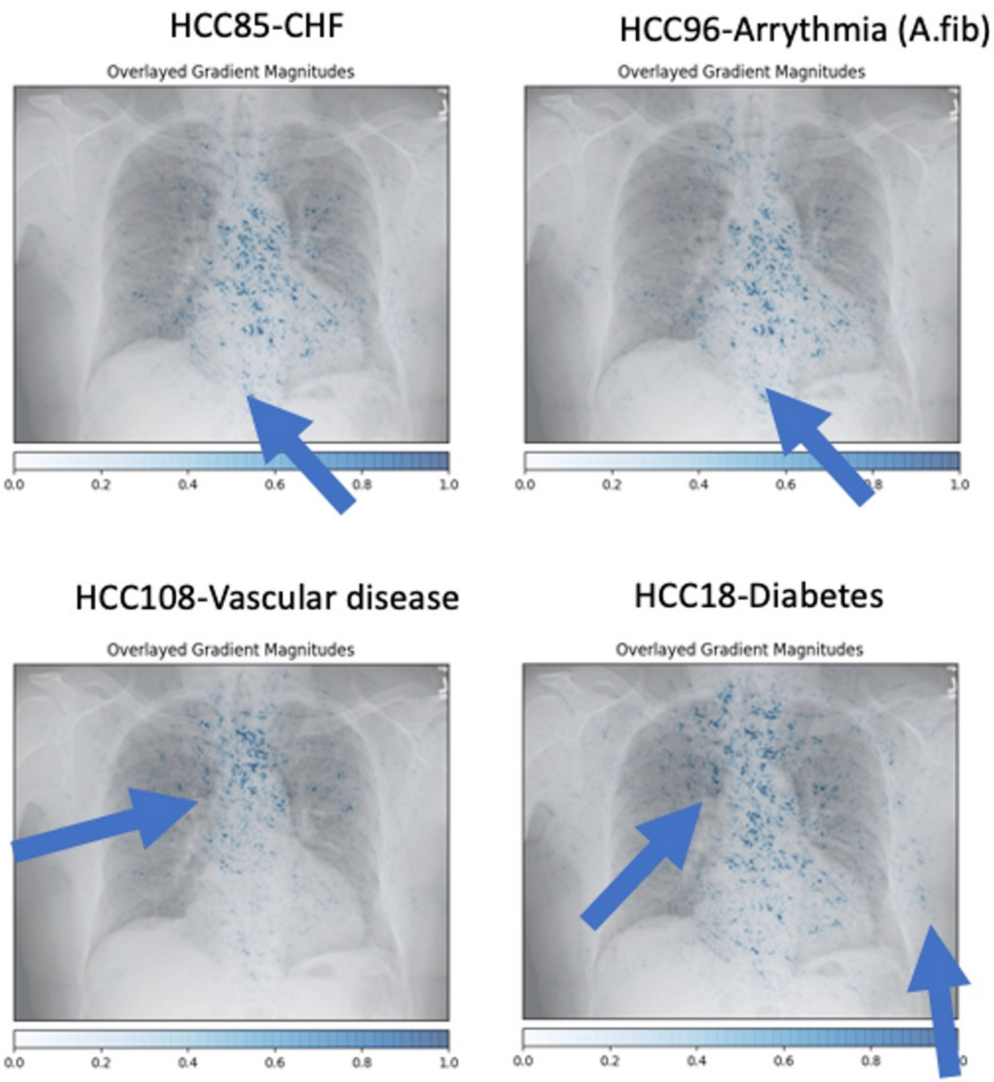


ORIGINAL ARTICLE DATA SCIENCE | VOLUME 19, ISSUE 1, P184-191, JANUARY 01, 2022

Detecting Racial/Ethnic Health Disparities Using Deep Learning From Frontal Chest Radiography

Ayis Pyrros, MD • Jorge Mario Rodríguez-Fernández, MD • Stephen M. Borstelmann, MD • Judy Wawira Gichoya, MD • Jeanne M. Horowitz, MD • Brian Fornelli, MS • Nasir Siddiqui, MD • Yury Velichko, PhD • Oluwasanmi Koyejo, PhD • William Galanter, MD, PhD • Show less

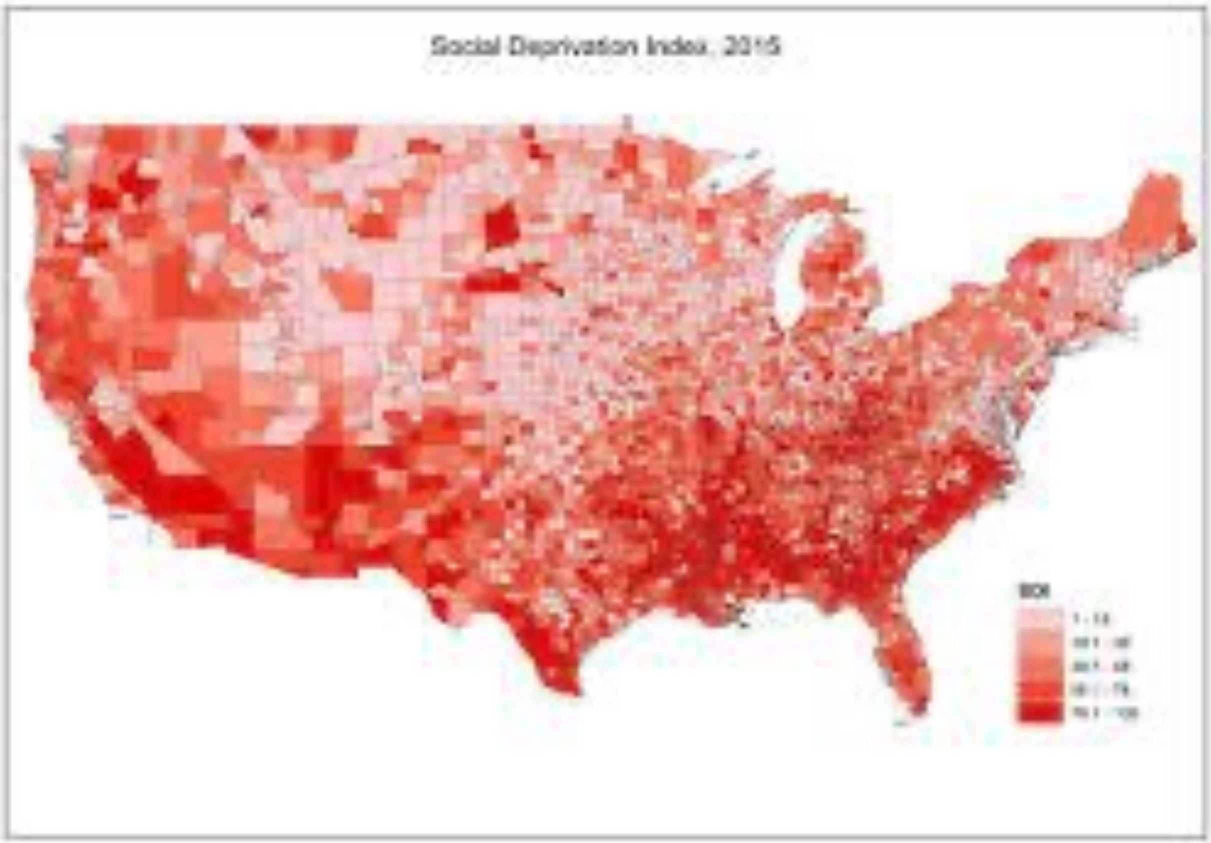
DOI: https://doi.org/10.1016/j.jacr.2021.09.010 • Check for updates



One CXR generates six HCC probabilities

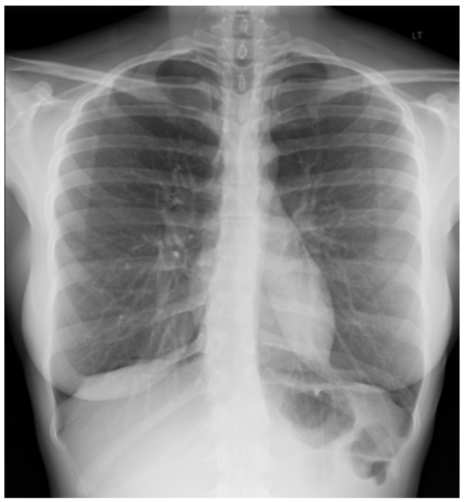
- Cardiac arrythmia maps demonstrate subtle increased activation in the atrial regions
- CHF maps shows enlarged heart
- Vascular disease activation at the aortic knob
- COPD shows increased activation of the lungs
- Diabetes activation in the axillary soft-tissues and aorta

Social Deprivation Index



NONWHITE: ASIAN,BLACK, HISPANIC	**AVEARGE AGE 48**	
	SDI average 45, median 40, std 28	**Social Deprivation Index**
	252 (30%)	114 MALE
		138 FEMALE
WHITE	**AVERAGE AGE 52**	
	SDI average 27, median 21, std 23	**Social Deprivation Index**
	562 (70%)	275 MALE
		287 FEMALE

RECAP



ELSEVIER

The Lancet Digital Health

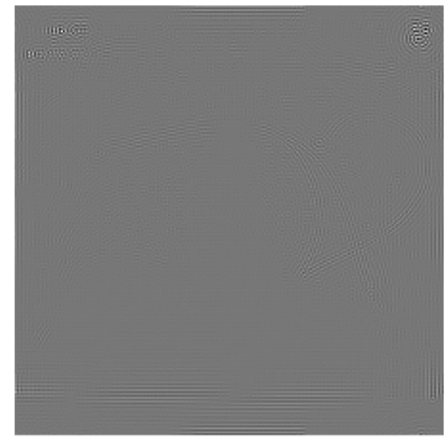
Available online 11 May 2022

In Press, Corrected Proof

Articles

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya MD [a], Imon Banerjee PhD [c], Ananth Reddy Bhimireddy MS [a], John L Burns MS [d], Leo Anthony Celi MD [e, g], Li-Ching Chen BS [h], Ramon Correa BS [c], Natalie Dullerud MS [i], Marzyeh Ghassemi PhD [e, f], Shih-Cheng Huang [j], Po-Chih Kuo PhD [h], Matthew P Lungren MD [j], Lyle J Palmer PhD [k, l], Brandon J Price MD [m], Saptarshi Purkayastha PhD [d], Ayis T Pyrros MD [n], Lauren Oakden-Rayner MD [k], Chima Okechukwu MS [o] … Haoran Zhang MS [i]



1) Performance of deep learning models to detect race from medical images across modalities and external datasets
2) Assessment of possible anatomic and phenotype confounders such as body habitus and disease distribution
3) Investigation into underlying mechanisms by which AI models can recognize race.

- judywawira@emory.edu
- @judywawira

Healthcare Innovation and Translational Informatics
(HITI) Lab at Emory