npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Confounders mediate AI prediction of demographics in medical imaging

Grant Duffy[1], Shoa L. Clarke[2], Matthew Christensen[1], Bryan He[3], Neal Yuan[4], Susan Cheng[1] and David Ouyang[1,5]✉

Deep learning has been shown to accurately assess "hidden" phenotypes from medical imaging beyond traditional clinician interpretation. Using large echocardiography datasets from two healthcare systems, we test whether it is possible to predict age, race, and sex from cardiac ultrasound images using deep learning algorithms and assess the impact of varying confounding variables. Using a total of 433,469 videos from Cedars-Sinai Medical Center and 99,909 videos from Stanford Medical Center, we trained video-based convolutional neural networks to predict age, sex, and race. We found that deep learning models were able to identify age and sex, while unable to reliably predict race. Without considering confounding differences between categories, the AI model predicted sex with an AUC of 0.85 (95% CI 0.84–0.86), age with a mean absolute error of 9.12 years (95% CI 9.00–9.25), and race with AUCs ranging from 0.63 to 0.71. When predicting race, we show that tuning the proportion of confounding variables (age or sex) in the training data significantly impacts model AUC (ranging from 0.53 to 0.85), while sex and age prediction was not particularly impacted by adjusting race proportion in the training dataset AUC of 0.81–0.83 and 0.80–0.84, respectively. This suggests significant proportion of AI's performance on predicting race could come from confounding features being detected. Further work remains to identify the particular imaging features that associate with demographic information and to better understand the risks of demographic identification in medical AI as it pertains to potentially perpetuating bias and disparities.

npj Digital Medicine (2022)5:188 ; https://doi.org/10.1038/s41746-022-00720-8

INTRODUCTION

Recent advances in deep learning have resulted in leaps in performance in analyzing and assessing image data, both with natural images as well as diagnostic medical images[1–3]. While traditional computer vision datasets are used by algorithms to perform common perceptive visual tasks that are achievable by most humans (for example, recognizing a cat when an animal is in the image)[2,4,5], deep learning in medicine has extended to tasks of prognosis and detection beyond the normal abilities of human clinicians. From evaluating blood pressure in fundoscopic images[6] to predicting prognosis and biomarkers from videos[7,8], convolutional neural networks are being applied to tasks not traditionally performed by clinicians.

Recent work from a variety of researchers on many medical imaging modalities have suggested deep learning can identify demographic features from medical waveforms, images, or videos[6,8–12]. The ability of deep learning algorithms to identify age, sex, and race from medical imaging raises challenging questions about whether using artificial intelligence (AI) black box models can be a vector for perpetuating biases and disparities[13–15]. The current regulatory environment does not require the demonstration of standard performance across different populations[16], however it has been shown that even when race is not directly used as an input, complicated decision support systems can learn patterns that reinforce disparities in access or treatment[17].

In this analysis, we sought to systematically evaluate whether demographic information can be captured from echocardiography, cardiac ultrasound, videos using deep learning. Using video-based deep learning architectures known to be able to capture quantitative traits commonly assessed by clinicians as well as textual patterns associated with disease[18,19], we evaluate whether echocardiogram videos can be used to predict age, sex, and race, and whether these models are robust across varying confounding variables and generalize across institutions with different geographic and demographic characteristics.

RESULTS

Study cohort characteristics

This study used cohorts of patients from two geographically distinct independent health care systems with different patient demographics. The Cedars-Sinai Medical Center (CSMC) study cohort consists of 30,762 patients who underwent 51,640 echocardiogram studies, individual instances when imaging were obtained, between 2011 and 2021. The same patient can undergo multiple studies over time as clinicians assess for change over time or for disease surveillance. A total of 433,469 videos from 51,640 studies were used representing apical-4-chamber, apical-2-chamber, parasternal long axis, and subcostal view videos. The mean age at the time of echocardiogram study was 66.5 ± 16.4 years, 44.8% were women, and 68.8% self-identified as White. The Stanford Healthcare (SHC) study cohort consists of 99,909 patients who underwent 99,909 echocardiogram studies between 2000 and 2019. The mean age at the time of echocardiogram study was 59.9 ± 17.7 years, 44.3% were women, and 56.5% self-identified as White. Demographic characteristics are shown in Table 1.

When initially trained on 99,909 apical 4 chamber videos from SHC without balancing confounding covariables, video-based deep learning models successfully learned features of age, sex,

[1]Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [2]Division of Cardiovascular Medicine, Department of Medicine, Stanford University, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. [4]San Francisco Veteran Affairs Medical Center, University of California San Francisco, San Francisco, CA, USA. [5]Division of Artificial Intelligence in Medicine, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA.
✉email: David.Ouyang@cshs.org

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.

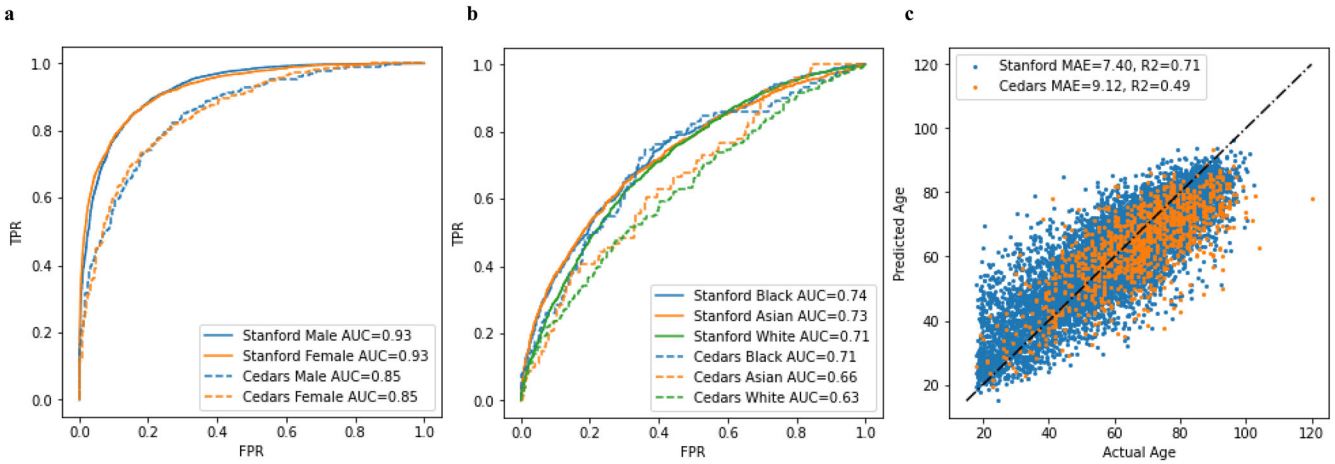

Fig. 1 AI model performance in predicting demographics in with unadjusted training and test datasets. A Performance in Predicting Sex of Model trained at Stanford on Internal Held-out Test Set and External (Cedars) Held-out Test Set. **B** Performance in Predicting Race of Model trained at Stanford on Internal Held-out Test Set and External (Cedars) Held-out Test Set. **C** Performance in Predicting Age of Model trained at Stanford on Internal Held-out Test Set and External (Cedars) Held-out Test Set.

and race. The deep learning model accurately predicted sex with AUC of 0.93 on the hold-out test set and 0.85 on the external test set of apical 4 chamber views from CSMC. Predicting Black, Asian, and White races, the AI model achieved an AUC of 0.74, 0.73, and 0.71 respectively on the hold-out test set and 0.71, 0.66, and 0.63 on the external test set. This performance was similar in both inpatient as well as outpatient echocardiograms (Supplementary Table 1) when trained and evaluated on CSMC data. The AI model predicted age with a MAE of 7.40 years on the hold-out test set and 9.29 years on the external test set. In general, there was a slight drop off in model performance with external validation as shown in Fig. 1; however, the model learned generalizable features that allowed for high accuracy in external validation datasets with sufficient training examples. When trained on echocardiogram videos from CSMC without balancing confounding covariables, video-based deep learning models similarly successfully learned features of age and sex and, to a lesser extent, race. With 150,913 CSMC apical 4 chamber video training examples, the deep learning model accurately predicted sex with an AUC of 0.84, age with a MAE of 9.66 years, and race with an AUC ranging from 0.54 to 0.60 on the held-out test dataset. A similar trend was seen when training models using different echocardiographic views as input (Table 2). Ensembling the information from all views modestly improved model performance, but continues to show limited predictive ability for race compared to age and sex.

We hypothesized that the modest ability to predict race from echocardiograms may depend on biased distributions of predictable features in race-stratified cohorts. For example, in the CSMC data, the proportion of males among the White subset is slightly higher (58.5%) compared to the proportion of males in the Black subset (54.1%). To test the impact of bias in a single predictable covariate, we artificially created subset datasets where race was confounded by sex. When sex is matched (bias = 0.5) in the training set, the model performance decreased for predicting White race (AUC 0.57) compared to the model trained without matching (AUC 0.59). We then created datasets with artificially introduced bias by selecting patients from subgroups unevenly. For example, a model for predicting race was trained with a dataset with an 80% sex bias means that 80% of the White patients are male, 20% are female while 80% of non-White patients are female while 20% are male. For predicting binary race

Table 2. Model performance when trained on different views.

	Sex (AUC)	Age (MAE)	White vs Rest (AUC)	Black vs Rest (AUC)	Asian vs Rest (AUC)
Cedars-Sinai					
Apical 4 chamber	0.84 (0.84–0.85)	9.66 (9.55–9.77)	0.59 (0.58–0.60)	0.60 (0.59–0.61)	0.54 (0.52–0.55)
Apical 2 chamber	0.80 (0.79–0.81)	10.82 (10.61–11.03)	0.58 (0.57–0.60)	0.60 (0.58–0.61)	0.55 (0.53–0.58)
Parasternal long axis	0.84 (0.83–0.85)	9.11 (8.97–9.25)	0.63 (0.61–0.64)	0.62 (0.61–0.64)	0.58 (0.56–0.60)
Subcostal	0.74 (0.73–0.75)	11.73 (11.62–11.81)	0.55 (0.54–0.57)	0.55 (0.53–0.57)	0.55 (0.53–0.58)
Ensemble of all views	0.93 (0.92–0.94)	7.78 (7.55–8.0)	0.71 (0.69–0.74)	0.72 (0.69–0.74)	0.60 (0.56–0.64)
Stanford					
Apical 4 chamber	0.93 (0.92–0.93)	7.40 (7.28–7.53)	0.71 (0.70–0.73)	0.74 (0.71–0.76)	0.73 (0.71–0.74)

classification biased by both age (binarized at 65 years old) and sex, model performance is poor but increases as bias increases. As bias in the training dataset approaches 100%, model performance approaches the performance of the confounding task, 0.85 and 0.84 for age and sex classification respectively. For predicting binary age and sex, performance is mostly unaffected when confounded by binary race suggesting that race adds little to no signal as a confounder for these tasks (Fig. 2). These results demonstrate how even a single confounder may be used by a deep learning model to predict race.

We expect that in most real-world datasets of clinical data, there are many relevant confounders that vary across race categories. Some of these confounders are comorbidities as shown in Table 3. To evaluate the predictive power of these confounders, we used logistic regression to predict White from non-White from comorbidity data alone. This method resulted in an AUC of 0.62, comparable to the AI computer vision model trained on CSMC data. Many of these cardiovascular comorbidities have known cardiac imaging findings. If an AI is able to detect imaging findings of these confounders, then it would be able to "shortcut" predict race without learning any additional information.

DISCUSSION

In this study, we systemically evaluated whether deep learning models can learn features of age, sex, and race from large datasets of echocardiogram videos. Consistent with prior applications of AI to medical imaging, we show that echocardiogram videos can be used to accurately predict age and sex and the models generalize across institutions. In contrast, we were not able to reproduce similarly accurate predictions of race and the model performance did not generalize well across institutions. Our experiments suggest race prediction results from shortcutting through predictions of known confounders commonly seen when stratifying large population cohorts by race.

Consistent with our deep learning model's high accuracy in predicting age and sex, there are well known age-associated changes[20,21] and sexual dimorphism[22,23] in cardiac structures visualized by ultrasound. While clinicians do not routinely use echocardiography to assess age, established age-dependent references for echocardiography and cardiac MRI highlights the recognized changes seen with normal cardiac remodeling of aging[24]. Conversely, conventionally measured metrics in echocardiography do not significantly vary across race, often within measurement error across cohorts[21,24]. Others have shown that AI models can predict race accurately using chest X-rays, thoracic CT scans, and breast mammograms[12]. These results are surprising given racial categories are imprecise, lack objective definitions, and have varied over time and by geography[25]. Nonetheless, identifying *if* and *when* AI may predict race is important for the thoughtful development of equitable applications of AI to medicine[26]. We believe it is equally important to understand *how* and *why* AI may predict race. Without careful analyses and discussion of such predictions, we risk sending the dangerous and false message that race reflects distinct biological features. This falsehood has been the basis of several missteps in medicine[27].

Our results suggest that AI models may predict race by leveraging the non-random distribution of predictable features in race-stratified cohorts. We use sex and age, two highly predictable features previously shown to be predictable by AI on medical imaging[11], to demonstrate that the degree of bias in a predictable confounding feature can arbitrarily impact prediction of a downstream task, such as predicting race even when there is limited information. Consequentially, we expect that there are many imaging-relevant features with varying predictability that contribute to the performance of AI race prediction from medical imaging. Importantly, some features, particularly related to health and disease, reflect the impacts of social determinant of health, and thus systemic differences in these features by racial category in population cohorts is a marker of the structural racism in medicine and society[28].

There are a few limitations in this study worth mentioning. First, the echocardiogram videos used in this study were only of two institutions, although geographically distinct and with different population demographics. The predominant ultrasound machine make and model was the same in both institution, which could standardize input video information and facilitate external validity but through imaging characteristics that are particular to that particular ultrasound machine. Second, the medical imaging used in this study were obtained in the course of routine clinical care, thus are enriched for individuals with access to healthcare and comorbidities that might have particular relationships with age, sex, and race. While this work was motivated to understand if AI models can shortcut prediction tasks through predicting demographics, there are biases regarding who is able to access healthcare, at what stage of disease, and for whom imaging is obtained. Third, we are unable to remove all confounders in the experimental set-up as there are many other measured and unmeasured differential covariates among populations. However, even in a setting with likely residual remaining confounding, models to predict race in echocardiography performed poorly and without generalizability. Without release of demographic information from other studies, we cannot generalize these findings across other medical imaging modalities, although we took particular care in adjusting for confounders. Finally, different views in medical imaging encode different information. In echocardiography, we think the A4c view is one of the most informative views and our experiments suggest the A4c view has the highest performance in predicting demographics, but additional future experiments might better detail how demographics are encoded in medical imaging and be a source of bias. Ethical considerations must be considered carefully, as fair application of AI is required to avoid perpetuating or exacerbating current biases in the healthcare system.

In summary, echocardiogram videos contain information that is detectable by AI models and is predictive of demographic

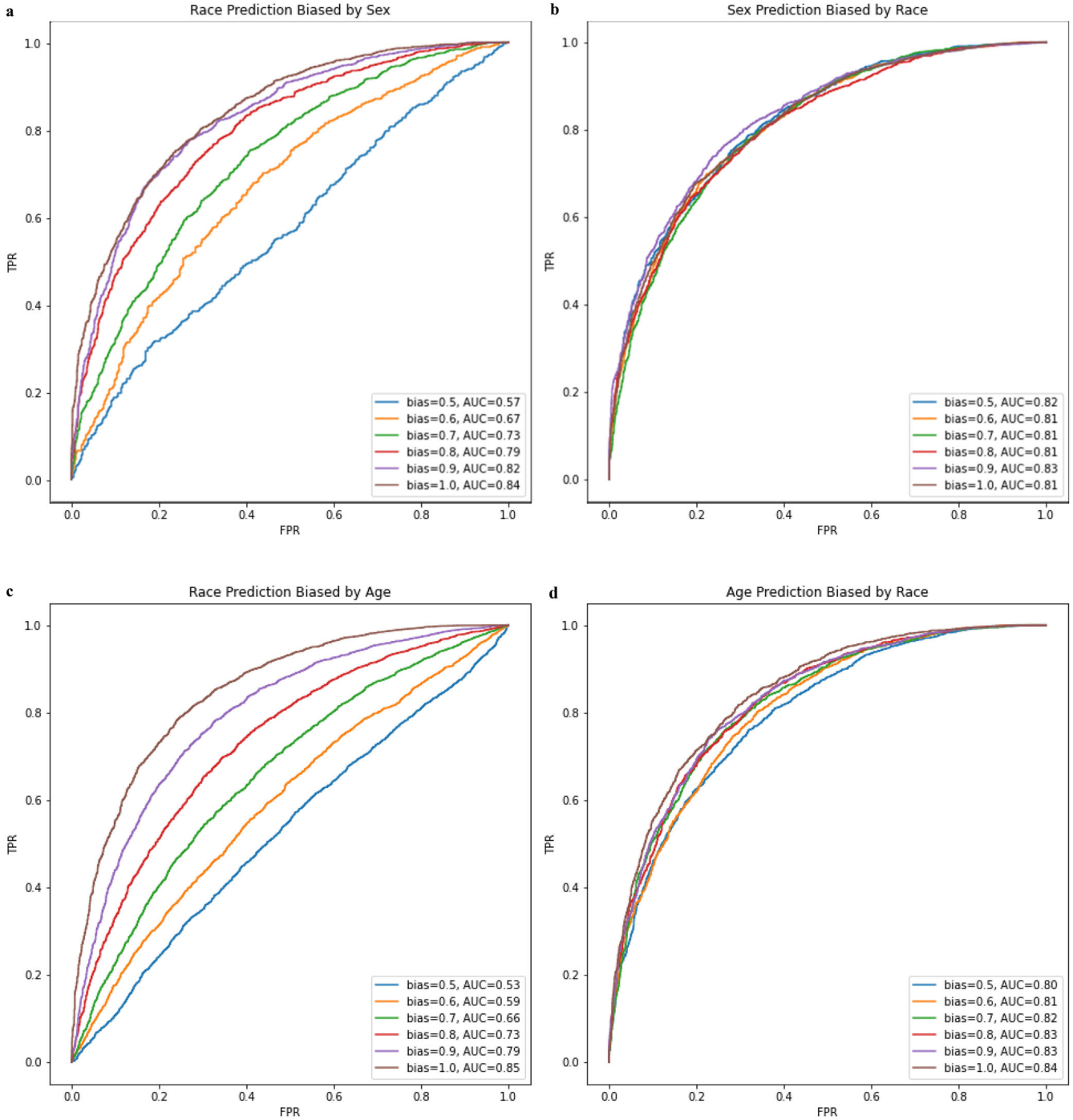

Fig. 2 Model performance mediated by varying training dataset ratio of confounding variable. a Model performance in predicting race with varying levels of sex bias in the training set. **b** Model performance in predicting sex with varying levels of race bias in the training set. **c** Model performance in predicting race with varying levels of age bias in the training set. **d** Model performance in predicting age with varying levels of race bias in the training set.

features. Age and sex features appear to be recognizable and generalize across geography and institution, while race prediction has significant dependence on the construction of the training dataset and performance could be mostly attributable to bias in training data. Further work remains to identify the particular imaging features that associate with demographic information and to better understand the risks of demographic identification in medical AI as it pertains to potentially perpetuating bias and disparities.

METHODS

Datasets

We used echocardiogram video data from two large academic medical centers, CSMC and SHC. Originally stored as DICOM videos after acquisition by GE or Philips ultrasound machines, we used a standard pre-processing workflow to remove information outside of the ultrasound sector, identify metadata[29], and save files in AVI format. Videos were stored as 112×112 pixel video files and view classified into four standard echocardiographic

Table 3. Comorbidities for Cedars-Sinai Medical Center cohort, separated by race.

Prevalent diagnoses at the time of the study	White	Black or African American	Asian	Total	*P* value
Atrial fibrillation	33,883 (29.83%)	6132 (24.83%)	3577 (27.46%)	43,592 (28.81%)	<0.001
Heart failure	49,411 (43.50%)	13,084 (52.99%)	5990 (45.98%)	68,485 (45.26%)	<0.001
Hypertension	68,114 (59.96%)	17,561 (71.12%)	8014 (61.51%)	93,689 (61.92%)	<0.001
Diabetes	26,511 (23.34%)	8292 (33.58%)	4392 (33.71%)	39,195 (25.90%)	<0.001
Ischemic stroke	12,017 (10.58%)	4190 (16.97%)	1405 (10.78%)	17,612 (11.64%)	<0.001
Transient ischemic attack	7431 (6.54%)	1457 (5.90%)	410 (3.15%)	9298 (6.14%)	<0.001
Systemic embolism	1076 (0.95%)	346 (1.40%)	113 (0.87%)	1535 (1.01%)	<0.001
Pulmonary embolism	4981 (4.38%)	2146 (8.69%)	340 (2.61%)	7467 (4.93%)	<0.001
Prior myocardial infarction	14,573 (12.83%)	4148 (16.80%)	1930 (14.81%)	20,651 (13.65%)	<0.001
Stroke/transient ischemic attack/thromboembolism	32,387 (28.51%)	9380 (37.99%)	3529 (27.09%)	45,296 (29.94%)	<0.001
Peripheral arterial disease	18,318 (16.13%)	3471 (14.06%)	1861 (14.28%)	23,650 (15.63%)	<0.001
Vascular disease	29,029 (25.56%)	6721 (27.22%)	3421 (26.26%)	39,171 (25.89%)	<0.001
Coronary artery disease	40,848 (35.96%)	6977 (28.26%)	4664 (35.80%)	52,489 (34.69%)	<0.001
Chronic kidney disease	28,346 (24.95%)	10,148 (41.10%)	4044 (31.04%)	42,538 (28.11%)	<0.001
Liver disease	7566 (6.66%)	1177 (4.77%)	1011 (7.76%)	9754 (6.45%)	<0.001
Chronic obstructive pulmonary disease	7329 (6.45%)	2474 (10.02%)	387 (2.97%)	10,190 (6.73%)	<0.001
Prior smoker	7077 (6.23%)	2089 (8.46%)	678 (5.20%)	9844 (6.51%)	<0.001

P values are from Chi-square test.

views (apical-4-chamber, apical-2-chamber, parasternal long axis, and subcostal views). At time of training, a random 32 frame clip was selected and every other frame (16 frames per clip) were inputted into the model.

Echocardiogram videos were split into training, validation, and test datasets by patient to prevent data leakage across splits. Demographic information was obtained from the electronic health record. Age was calculated from time from the echocardiogram study to date of birth. Information about sex and race were obtained from the electronic health record based on self-report from the clinical record. Comorbidities were extracted from the electronic health record by International Classification of Disease Ninth or Tenth Revision codes present in problem lists or visit associated diagnoses within 1 year of the echocardiogram imaging study. This analysis did not independently re-survey or use other instruments to evaluate data labels.

For experiments sweeping model performance with various degrees of biased training datasets, we subsetted the whole dataset to form simplified cohorts that binarized demographic categories and maintained the same training set size across experiments. The breakdown of these subgroups can be found in Supplementary Tables 2 and 3. Additionally, Supplementary Table 4 shows the distribution of ultrasound machines and transducer models used in the study. For predicting race, we focused on white/non-white binary classification and varying the confounding variable of sex and binary classification of age, binarized at 65 years old. For example, a bias of 0.5 corresponds to a dataset where 50% of both white and non-white examples are male and female. A bias of 1.0 corresponds to a dataset where all of the white patients are male and all of the non-white patients are female. Other than training dataset construction, all other model training details (architectures, loss, learning rate, number of epochs, etc) were held the same. The parallel set of experiments predicting sex used the same format only with race proportion in the training set varied as the confounding variable.

AI model architecture

Spatiotemporal relationships were captured by our deep learning model using 3D convolutions using standard ResNet architecture (R2 + 1D)[18]. This model interprets 3D videos by using 2D special convolutions and 1D temporal convolutions at every convolutional layer. Models were trained to minimize L2 loss for predicting age (a regression task), binary cross entropy for predicting sex (a binary classification task), and cross entropy for predicting race (a multi-class classification task). Models were trained using stochastic gradient descent with an ADAM optimizer using an initial learning rate of 0.01, momentum of 0.9, and learning rate decay. Models were trained using an array of NVIDIA 2080, 3090, and A6000 graphical processing units. Deep learning models were trained with input videos of each individual view and an ensemble model was constructed by logistical regression with inputs of the inference prediction from models of each view.

Analysis

The performance of deep learning models was assessed on internal held out datasets or external datasets from another institution. The performance of predicting age was evaluated by the mean absolute difference between the model prediction and actual age at time of echocardiogram study. The prediction of sex and race was evaluated by area under receiver operating curve (AUROC). Confidence intervals were computed using 10,000 bootstrapped samples of the test datasets. To benchmark with the predictive ability of differences in population level comorbidity rates, we developed a logistic regression model using all comorbidities in Table 3 as well as age and sex as independent input variables to predict race. The continuous output of the logistic regression model was used to assess AUROC. This research was approved by the Stanford University and Cedars-Sinai Medical Center Institutional Review Boards.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

A representative subset of the data is publicly available at https://echonet.github.io/dynamic/. The full dataset is available with a data use agreement with the respective institutions after IRB approval by emailing David.Ouyang@cshs.org.

Received: 27 July 2022; Accepted: 2 November 2022;
Published online: 22 December 2022

REFERENCES

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
2. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
3. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. Preprint at *arXiv* https://arxiv.org/abs/1512.00567 (2015).
4. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. Preprint at *arXiv* https://arxiv.org/abs/1512.03385 (2015).
5. Lin, T.-Y. et al. Microsoft COCO: common objects in context. Preprint at *arXiv* https://arxiv.org/abs/1405.0312 (2014).
6. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
7. Weston Hughes, J. et al. Deep learning prediction of biomarkers from echocardiogram videos. *EBioMedicine* **73**, 103613 (2021).
8. Ghorbani, A. et al. Deep learning interpretation of echocardiograms. *NPJ Digit Med* **3**, 10 (2020).
9. Yi, P. H. et al. Radiology "forensics": determination of age and sex from chest radiographs using deep learning. *Emerg. Radiol*. https://doi.org/10.1007/s10140-021-01953-y (2021).
10. Raghu, V. K., Weiss, J., Hoffmann, U., Aerts, H. J. W. L. & Lu, M. T. Deep learning to estimate biological age from chest radiographs. *JACC Cardiovasc. Imaging* **14**, 2226–2236 (2021).
11. Attia, Z. I. et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ. Arrhythm. Electrophysiol.* **12**, e007284 (2019).
12. Banerjee, I. et al. Reading race: AI recognises patient's racial identity in medical images. Preprint at *arXiv* https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00063-2/fulltext. 10356 (2021).
13. Zou, J. & Schiebinger, L. AI can be sexist and racist — it's time to make it fair. *Nature* **559**, 324–326 (2018).
14. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Preprint at *arXiv* https://arxiv.org/abs/2003.00827 (2020).
15. Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S. & Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat. Med.* **27**, 136–140 (2021).
16. Wu, E. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).
17. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
18. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
19. Duffy, G. et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA Cardiol.* **7**, 386–395 (2022).
20. Cheng, S. et al. Age-related left ventricular remodeling and associated risk for cardiovascular outcomes: the Multi-Ethnic Study of Atherosclerosis. *Circ. Cardiovasc. Imaging* **2**, 191–198 (2009).
21. Asch, F. M. et al. Similarities and differences in left ventricular size and function among races and nationalities: results of the World Alliance Societies of Echocardiography Normal Values Study. *J. Am. Soc. Echocardiogr.* **32**, 1396.e2–1406.e2 (2019).
22. Cain, P. A. et al. Age and gender specific normal values of left ventricular mass, volume and function for gradient echo magnetic resonance imaging: a cross sectional study. *BMC Med. Imaging* **9**, 2 (2009).
23. Kawel-Boehm, N. et al. Reference ranges ("normal values") for cardiovascular magnetic resonance (CMR) in adults and children: 2020 update. *J. Cardiovasc. Magn. Reson.* **22**, 1–63 (2020).
24. Miyoshi, T. et al. Left ventricular diastolic function in healthy adult individuals: results of the World Alliance Societies of Echocardiography Normal Values Study. *J. Am. Soc. Echocardiogr.* **33**, 1223–1233 (2020).
25. Deyrup, A. & Graves, J. L. Jr. Racial biology and medical misconceptions. *N. Engl. J. Med.* **386**, 501–503 (2022).
26. Ganapathi, S. et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat. Med*. https://doi.org/10.1038/s41591-022-01987-w (2022).
27. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. *N. Engl. J. Med.* **383**, 874–882 (2020).
28. Bailey, Z. D., Feldman, J. M. & Bassett, M. T. How structural racism works - racist policies as a root cause of U.s. racial health inequities. *N. Engl. J. Med.* **384**, 768–773 (2021).
29. Ouyang, D. ConvertDICOMToAVI.ipynb at master · echonet/dynamic. Github. https://github.com/echonet/dynamic/blob/master/scripts/ConvertDICOMToAVI.ipynb. Accessed Jan 1, 2022.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grants NIH K99 HL157421-01., R01-HL077477, R01-HL131532, R01HL134168, R01-DK080739, R01-HL126136, R01-HL080124, R01-HL077477, R01-HL070100, and U54 AG065141.

AUTHOR CONTRIBUTIONS

G.D., M.C., and B.H. performed the experiments. G.D., S.L.C., N.Y., S.C., and D.O. wrote the manuscript and provided experimental feedback. All authors reviewed the experiments, analyses, and manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00720-8.

Correspondence and requests for materials should be addressed to David Ouyang.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.