HHS Public Access

Author manuscript

Circ Heart Fail. Author manuscript; available in PMC 2025 January 01.

Published in final edited form as:

Circ Heart Fail. 2024 January ; 17(1): e010879. doi:10.1161/CIRCHEARTFAILURE.123.010879.

Race, Gender and Age Disparities in the Performance of ECG Deep Learning Models Predicting Heart Failure

Dhamanpreet Kaur, BS[1], **John W. Hughes, BS**[1], **Albert J. Rogers, MD, MBA**[1], **Guson Kang, MD**[2], **Sanjiv M. Narayan, MD, PhD**[1], **Euan A. Ashley, FRCP, DPhil**[1], **Marco V. Perez, MD**[1]

[1.] Stanford University

[2.] VA Palo Alto Health Care System

Abstract

Background—Deep learning models may combat widening racial disparities in heart failure outcomes through early identification of individuals at high risk. However, demographic biases in the performance of these models have not been well studied.

Methods—This retrospective analysis used 12-lead ECGs taken between 2008 – 2018 from 326,518 patient encounters referred for standard clinical indications to Stanford Hospital. The primary model was a convolutional neural network model trained to predict incident heart failure within 5 years. Biases were evaluated on the testing set (160,312 ECGs) using area under the receiver operating curve (AUC), stratified across the protected attributes of race, ethnicity, age, and gender.

Results—59,817 incident cases of heart failure were observed within 5 years of ECG collection. The performance of the primary model declined with age. There were no significant differences observed between racial groups overall. However, the primary model performed significantly worse in Black patients aged 0 – 40 compared to all other racial groups in this age group, with differences most pronounced among young Black women. Disparities in model performance did not improve with integration of race, ethnicity, gender, and/or age into model architecture, by training separate models for each racial group, nor by providing the model with a dataset of equal racial representation. Using probability thresholds individualized for race, age, and gender offered substantial improvements in F1-scores.

Conclusion—The biases found in this study warrant caution against perpetuating disparities through the development of machine learning tools for the prognosis and management of heart failure. Customizing the application of these models by using probability thresholds individualized by race/ethnicity, age, and gender may offer an avenue to mitigate existing algorithmic disparities.

Keywords

Algorithmic biases; deep learning; electrocardiograms; heart failure

Corresponding Author, Dhamanpreet Kaur, dkaur24@stanford.edu, 291 Campus Drive, Stanford, CA 94305.

Introduction

Heart failure remains one of the leading causes of death in the US, currently affecting 6.2 million adults[1]. The burden of the disease varies greatly by age, race, and gender[2]. Despite advances in medical care that have allowed incidence to stabilize or decline[3], disparities in outcomes persist; in fact, the gap in age-adjusted death rates between Black patients and White patients has widened from 1999–2017, especially among younger patients[4]. Heart failure may be underdiagnosed at higher rates in Black patients and women in the outpatient setting[5]. Earlier detection and closer monitoring of high-risk individuals may aid in reducing occurrence and improving prognosis of the disease to ultimately combat these disparities.

Several studies have used machine learning algorithms to identify cardiovascular conditions from electrocardiograms (ECGs)[6–8]. Deep neural networks can outperform cardiologists in recognizing several abnormalities from 12-lead ECG recordings, achieving F1-scores above 80% and specificity above 99%[9]. They can also achieve superior performance when compared to commercial rule-based methods, such as GE Healthcare's MUSE ECG insights[10]. The application of artificial intelligence to ECG data has additionally been used to identify conditions that are not typically detected by the human eye. Deep learning models have been successfully developed to screen for asymptomatic left ventricular dysfunction (area under the receiver operator curve (AUC) of 0.93)[11], atrial fibrillation (AUC of 0.87)[12], aortic stenosis (AUC of 0.884)[13] and anemia (AUC of 0.923)[14], illustrating the broad potential of this non-invasive tool.

However, the performance and applicability of these models is highly contingent on the quality of the training data used, as well as the populations from which they are derived, and therefore may be prone to perpetuating implicit biases[15]. Several instances of this have been noted in the medical field: higher rates of underdiagnosis in chest radiographs among intersectional underserved populations[16], significantly worse AUC values reported for dermatology artificial intelligence algorithms when tested on diverse datasets[17], lower risk scores for Black patients equally as ill as White patients[18], and a 50% reduction in diagnostic accuracy of skin lesions among darker-skinned patients[19]. On the other hand, there have been cases of machine learning algorithms producing less biased results in comparison to other scoring methods[20]. Nonetheless, the presence of such biases remains largely understudied, particularly in the realm of machine learning applications for cardiovascular data.

In the case of ECG data specifically, disparities may be exacerbated by baseline differences due to age, race, and sex.[21,22] Several studies have particularly noted benign variations in ECG patterns among Black patients: higher QRS voltages among healthy young black males[22], ST segment flattening in young Black women[23], greater ST-elevation thresholds in Black patients than White patients[24], and biphasic T-waves in Black women[25]. The differences in ECG characteristics may also carry prognostic significance, showing inconsistent association with mortality independent of confounding risk factors, which holds implications for race-specific reference ranges and cardiovascular risk.[26] Few studies have evaluated the disparities present in machine learning models applied to ECG data. Noseworthy et al. noted that in spite of racial differences in ECG data, a convolutional

neural network trained on a homogenous population generalizes well to detecting low ejection fractions for several racial subgroups[27]. However, the effects of the intersectionality of age, gender, and race/ethnicity on model performance have not been well-studied.

This work aims to holistically investigate the existence of algorithmic biases as they pertain to age, race, ethnicity, and gender in a deep learning model trained to predict heart failure from ECG data and further explore how various modifications to the training and application of the model affect its performance.

Methods

Study Population and Data Sources

Because of the sensitive nature of the data collected for this study, requests to access the dataset from qualified researchers trained in human subject confidentiality protocols may be sent to the corresponding author. This study was approved by the Stanford University Institutional Review Board. There were 326,518 12-lead ECGs used in this retrospective analysis derived from patients referred for standard clinical indications to Stanford University Medical Center. A total of 954,817 ECGs taken between March 2008 and May 2018 were extracted from the Phillips TraceMaster system. All ECGs were saved as 10 second signals from all 12 leads of the ECG, sampled at 500Hz. Band pass and wandering baseline filters were applied to the signals, normalized on a per-lead basis, and down sampled to 250Hz.

Race, ethnicity, and gender were derived from self-report by the patients at the time of hospital enrollment. Follow-up heart failure from March of 2008 to February of 2022 was queried from STARR-OMOP[28] (Stanford Medicine Research Data Repository - Observational Medical Outcomes Partnership), a common data model for accessing electronic health records. Our primary outcome of interest was incident heart failure, defined as a first instance of heart failure within five years of the ECG. Heart failure was defined following prior work to include SNOMED (Systematized Nomenclature of Medicine) code 84114007 (heart failure) and all descendants, excluding 82523003 (congestive rheumatic heart failure)[29]. Patients with prior heart failure were excluded. During training and testing, positive cases were defined as patients who developed heart failure within five years of the ECG; negative cases were defined as patients with no heart failure within five years of ECG, given at least five years of follow-up evidenced by measurement, admission, or mortality.

Model development and training

The primary model was a convolutional neural network trained to predict occurrence of heart failure within 5 years of the ECG recording from solely the input of 12-lead ECG data. A detailed model architecture can be found in Figure S1. Model development was performed using Python 3.9 and PyTorch 1.11, and models were trained on single Nvidia Titan Xp Graphics Processing Units using Stanford's Sherlock computing cluster. Hyperparameters were tuned by learning from the training set and evaluating on the validation set, which comprised 164,630 ECGs (Figure S2). The parameters were set using a batch size of 128, a weight decay hyperparameter of 10^{-4}, and the Adam optimizer. The learning rate was

initialized to 10^{-3} and reduced by a factor of 10 each time the validation loss plateaued for more than five epochs, limited at a lower bound of 10^{-6}. After establishing hyperparameters, four more models were trained using cross-validation; the dataset was partitioned into five different subsets such that one of the five would be excluded from training and used as validation in each model. The outputs of these five models were averaged to generate predictions on the testing data (160,312 ECGs), which was then used for the analyses below.

Statistical Analysis

A fair model would be expected to have equal predictive capacity across demographic divisions; the existence of group-based disparities in model performance constitutes bias. Age, race, ethnicity, and gender were chosen as the protected attributes across which to evaluate intersectional biases in model performance. Analyses focused on the four largest racial/ethnic groups: Non-Hispanic White, Hispanic, Black, and Asian patients. Bias was defined as significant differences in area under the receiver operating curve (AUC) across demographic subgroups. Age was discretized into four groups: 0–40 years, 40–60 years, 60–80 years, and >80 (80+) years. AUC was computed in a stratified manner for each demographic division. The receiver operating characteristic curves, AUC values and 95% confidence intervals were computed using the pROC R package[30]. Unless otherwise noted, AUC values were computed at a five-year time horizon, comparing all examples with an event of incident heart failure within five years against all examples with follow-up data at five years.

To assess differences in overdiagnosis versus underdiagnosis, calibration curves were graphed for each demographic subgroup using the Python Scikit-learn package[31] as follows: the predicted probabilities were binned and plotted against the proportion of individuals in that bin that were observed to develop heart failure. The ideal model would show a 1:1 correlation; a slope of less than one indicates overdiagnosis and a slope of greater than one indicates underdiagnosis.

The optimal threshold used in this model was based on F1-score using data from the entire population; precision, recall, and negative predictive value were computed within subgroups of race and gender. Moreover, the distribution of positive and negative ground truth labels for each subgroup was plotted against the probabilities assigned to the cases by the model.

Reducing Model Biases

Three primary avenues in the design and application of the algorithm were investigated to reduce model biases: optimizing training data (pre-processing), modifying the architecture of the model, or tailoring application of the model (post-processing) (Figure S3).

In optimizing training data, a series of experiments were performed to modify the subset of training data presented to the model; the efficacy of these variations in reducing disparities in model performance was assessed through stratified AUC values. The first approach involved training a separate model for each racial subgroup and each age subgroup. Four individual models were trained and tested on datasets consisting solely of Non-Hispanic White patients, Black patients, Asian patients, and Hispanic patients (e.g., the AUC value reported for Asian patients reflects performance of a model trained only on ECGs from

Asian patients). Similarly, four individual models were trained and tested on datasets consisting solely of patients aged 0 – 40, 40 – 60, 60 – 80, and > 80. The second approach entailed providing the model with a dataset that has equal representation from each racial subgroup. Stratified sampling was applied to the full set of ECGs to take the same number of patients from each of the predominant four racial groups: using the size of the smallest group (Black patients) as the sampling quantity, this produced a test set of 27,176 ECGs.

In modifying the architecture of the model, demographic variables were incorporated through early and late fusion. In early fusion, the model is passed demographic data as an additional channel of input data. For example, a model trained on ECG and gender would be passed 13 channels of data, where the first 12 correspond to the 12-lead ECG and the 13^{th} is 0 where the patient is male and 1 if the patient is female (Figure S4a). In late fusion, demographic data is appended to an intermediate layer of the model, after the convolutional layers and before the fully connected layers, allowing for some mingling of information without passing the demographic data through the convolutional backbone (Figure S4b). AUC values were compared across the demographic subgroups to validate the effectiveness of these approaches in reducing model bias.

In tailoring the application of the model, the probability threshold for classification of a case as a positive label was optimized separately for each demographic subgroup based on F1-score. F1-score is a metric that can be used to assess precision and recall (the percentage of heart failure cases identified by the algorithm) once the model's outputs of risk scores are translated to binary predictions of whether a patient develops heart failure; it is computed as follows: $F1 = \frac{precision * recall}{precision + recall}$. For each demographic subgroup, the differences were computed between metrics (precision, recall, negative predictive value) that resulted from using the threshold value individualized for that group versus the overall optimal threshold value.

Results

Study Population

There were 326,518 patient ECGs derived from patients who were followed for an average of 6.8 years. The baseline characteristics stratified by incident heart failure are described in Table 1. The patients were on average 59.3 years old and comprised 49.7% women. The racial and ethnic composition consisted of Non-Hispanic White (56.5%), Asian (14.2%), Black (4.0%), Hispanic (12.3%), Native Hawaiian/Pacific Islander (1.2%), and American Indian or Alaskan Native (0.2%).

Incidence of Heart Failure

There were 59,817 incident cases of heart failure within 5 years of ECG collection. The fraction of patients who developed heart failure within 5 years increased from 9.0% in the youngest age group of 0–40 year-olds to 36.6% in those over 80 years of age. The fraction of women (16.4%) developing heart failure within 5 years was lower than that of men (20.2%). The incidence of 5-year heart failure was higher among Black patients (23.5%)

than White patients (19.0%), Hispanic patients (17.5%), or Asian patients (18.6%) (Table 1). Breakdowns of incidence rates by race/ethnicity, age, and gender can be found in Table S1.

Comparisons of Model Performance

The performance of the primary model declined significantly with age. The model performed significantly better in those 0 – 40 years old (AUC 0.80 [0.79 – 0.81]) compared to those who were >80 years old (AUC 0.66 [0.65 – 0.66]) (Figure 1a). The model performed slightly worse in men (AUC 0.77 [0.77 – 0.77]) compared to women (AUC 0.78 [0.78 – 0.79]) (Figure 1b). There were no significant differences observed in model performance between racial groups (Hispanic patients AUC 0.79 [0.79 – 0.80], Asian patients 0.78 [0.77 – 0.79], Non-Hispanic White patients 0.77 [0.77 – 0.78], Black patients 0.78 [0.77 – 0.79]) (Figure 1c).

However, the trends in race- and gender-based disparities differed when broken down by age group. The primary model performed significantly worse in Black patients aged 0 – 40 years old (AUC 0.69 [95% CI 0.64 – 0.75]) compared to all other racial groups in the same age group (Non-Hispanic White AUC 0.80 [0.78 – 0.81]; Hispanic AUC 0.81 [0.79 – 0.83]; and Asian AUC 0.82 [0.79 – 0.85]) (Figure 2). The racial differences in AUC were much less pronounced among men, whereas the AUC for Black women (AUC 0.69 [0.62 – 0.77]) was substantially lower than women of all other racial groups aged 0 – 40 years (Figure S5). Gender-based differences varied by age group (Figure S6).

The distributions of predicted probabilities for positive and negative labels vary between race and age subgroups (Figure S7). The overlap between healthy patients and heart failure patients increased substantially with age across all races. Among Black patients aged 0 – 40 years, a lower predicted probability was more likely to correlate with incident heart failure compared to other race and gender subgroups.

The calibration curves indicate that the model is best calibrated for Asian and Non-Hispanic White patients. The observed fraction of cases with heart failure exceeds the probability predicted by the model among Black patients, indicating greater underdiagnosis in comparison to other racial groups, especially among Black women (Figure 3).

Incorporating Race and Ethnicity in Model Building

The model was provided with different subsets of training data to determine if the disparities may have resulted from differences in racial and ethnic group sample sizes. Using a dataset with equal racial representation did not eliminate disparities between Black patients and patients of other racial groups in the 0 – 40 age group (Figure S8). The AUC values did not improve from the primary model. Similarly, there was no improvement in performance amongst the different race and ethnic subgroups when compared using models that were trained on the same race and ethnicity as the test set (Figure S9). Moreover, there was no improvement in age-related disparities when the models were trained and tested on data from separate age groups (Figure S10).

The incorporation of race/ethnicity, age, gender, or the combination of those three demographic variables into training did not significantly improve performance in any

subgroup of race and age. Nonetheless, there was a reduction in racial disparities among young patients from the combined effect of improving performance for Black patients and diminishing performance for other racial groups (Figure 4).

When selecting the probability threshold based on optimization of F1-score, the optimal threshold varied between subgroups of race/ethnicity, age, and gender (Figure S11). Black men and women in the 0 – 40 and 40 – 60 year age groups had a lower optimal probability score threshold than all other groups. The optimal probability threshold for the entire population was 0.20; the highest individualized optimal threshold was 0.30 for Asian men aged 0 – 40 years and the lowest was 0.10 for Black men aged >80 years. All racial and gender groups aged 0 – 40 years showed increases in F1-scores when using individualized thresholds (Figure 5). Black women in this age group showed the greatest improvement using an individualized threshold of 0.15 with an 11%-point increase in F1-score – a 3% increase in PPV and a 21% increase in recall. Black men aged >80 years showed a 10% increase in F1-score using an individualized threshold of 0.10 – a 1%-point decrease in PPV and a 27%-point increase in recall (Table S2).

Discussion

The utility of deep learning models for ECG data arises from a reduction in the demand for cardiologist labor, comparable or increased accuracy of the automated output, and the ability to provide insights above and beyond those typically detected by the human eye. They provide clinicians with a powerful tool in advising treatment and predicting prognosis.

In this study, we assessed the algorithmic biases present in a model designed to detect 5-year occurrence of heart failure. We used a range of approaches previously proposed to assess algorithmic equity: predictive parity, predictive equality (false positive error rate balance), equal opportunity (false negative error rate balance), equalized odds, conditional use accuracy quality, and Well-calibration[32]. Some of the proposed standards are mathematically incompatible (i.e., fairness in one metric may preclude fairness in another)[33], underscoring the need for a holistic evaluation and consideration of factors that influence utility of the model's application.

The differences in AUC across demographic subgroups confirms the hypothesis that disparities exist in model performance, and further indicates that these disparities reflect the intersectionality of race/ethnicity, gender, and age in cardiovascular disease. Using the metric of AUC, we found that the model performed slightly better for women than men, and significantly worse among the older population and the Black population aged 0 – 40 than all other races. The small observed difference between men and women is notable given that there was a higher percentage of positive labels among men and consequently, less class imbalance in the data, which would be expected to enhance model performance[34].

Upon further analyzing subgroups of race and gender, we found that the disparity observed among young Black patients stems primarily from the model's diminished performance among Black women in comparison to women of other racial groups aged 0 – 40; significant differences do not exist between races among male patients or between Black men and

Black women aged 0 – 40. There have been several cases of ECG patterns observed among healthy Black women that would typically be considered malignant in the White population[25]. This may be a contributing factor given the training data was predominantly comprised of White patients. Selected ECG examples in Figure S12 illustrate a few cases that may provide insight into the trends observed in poor model performance, including inaccurate model predictions for young Black women with T-wave inversion and for patients over 80 years old with typically low voltages and flat T-waves.

We found that model performance was worse in older patients compared to younger patients. Given that the composition of the training data was biased toward older patients and that older patients had a larger fraction developing heart failure, the most likely explanation for this trend is that there were more obvious ECG differences between healthy and diseased patients in the younger age groups. The age distribution varied across racial groups; Black patients had a lower average age than Asian patients or White patients (Table S3). Given that the model performed better for younger patients, this may be a confounding factor in the insignificant AUC difference observed between racial groups overall.

Given the notable reduction in model performance pertaining to race and gender, we investigated multiple avenues of modifying the training and application of the model to improve the existing disparities. First and foremost, we acknowledge the widespread debate surrounding the use of race as a variable in algorithms for medical decision making.[35–37] One study noted that race-specific models in the prediction of heart failure risk demonstrated superior performance to non-race-specific and traditional risk models.[38] In this study, we do not seek to draw conclusions on whether race should be an input into prognosis of cardiovascular disease, but rather to investigate this as a method of improving predictions for minority groups. We did not find that the inclusion of race, gender, or age as variables in model training improved overall performance of the model. We did observe some reduction in racial disparities among the 0–40 age group, but this was due to the combined effects of improved performance in Black patients and diminished performance in the other racial groups. Models trained and tested on a dataset with equal racial representation or on datasets with individual race and age groups did not eliminate disparities in performance. However, it is difficult to discern the effects of the separation by race or age and stratified sampling from the decreased quantity of training data. Nevertheless, we believe that these findings warrant caution in using machine learning models as a tool for ECG interpretation and heart failure prognosis in older patients or in young, Black patients.

Although modifications to model training did not produce the desired reductions in disparities, we sought insight into tailoring the downstream application of the model to ameliorate the inherent biases. In doing so, we considered F1-score to evaluate the utility of the model where a single binary cutoff is chosen for classification. The variance in thresholds at which F1-score peaks for different demographic subgroups suggests that individualized applications of the model may be optimal. The range in optimal threshold cutoffs varied from 0.10 in Black and Hispanic men aged >80 to 0.30 in Asian men aged 0–40. These threshold choices are supported by the distributions of positive case labels: in Black and Hispanic patients aged >80, positive cases follow a similar distribution to negative cases, so a lower threshold is necessary to capture more of those developing heart

failure, whereas in young Asian men, the distribution of positive case labels is skewed more strongly toward higher probabilities. Moreover, the calibration curves showed overdiagnosis being most prominent among Black and Hispanic men, further concurring with a higher optimal threshold for these subgroups. In this study, we used F1-Score as the criteria for optimization, which gives equal weight to precision and recall. In doing so, we found a 21.0% increase in recall for Black women. When applying such a model to the healthcare setting, we recognize that there may be greater value in recall than precision and recommend that the user assign an appropriate utility function for optimization. Thus, calibrating the algorithm for race, age, and gender after training may offer a promising avenue to reduce the disparities otherwise present in model performance.

Limitations

The primary limitations of the study lie around the parameters of the outcome and cohort used to develop the model. We only considered patients with at least five years of follow-up data within the Stanford health system from the time of ECG collection. The exclusion of cases due to lack of follow-up or prior heart failure may be biased by demographic group; however, adjustment for these variables did not preclude demographics as a significant predictor of model performance (Table S4). However, one might expect that limiting the cohort to patients who sought longitudinal care within the Stanford system would create greater homogeneity in the data, thus rendering the findings of disparity in model performance even more striking. Furthermore, we assumed that any misclassification of cases would be non-differential based on demographic status. In the case that the misclassification is biased, we would expect higher rates of underdiagnosis for Black patients and women[5], which only amplifies the finding of model underdiagnosis observed particularly for that group. To support the codes used for heart failure diagnosis, we analyze the fraction of diagnosed patients with high BNP (B-type natriuretic peptide) (Table S5) and perform a sensitivity analysis using BNP as part of the diagnostic criteria (Figure S13). The etiology of heart failure was not available for analysis, and we cannot rule out that disparities in model performance are not due to differences in heart failure etiology among patients of different demographic groups; nonetheless, the differences in performance of the model remain clinically valid. We considered the number of days a patient interacted with the medical system to account for likelihood of diagnosis following ECG; significant differences were not observed across race and age (Table S6). Some patients may have had undiagnosed heart failure at the time of ECG; analyses of AUC differences using a blanking period of three days shows consistent results (Figure S14). Lastly, it is important to note that although differential performance was observed among various demographic subgroups, the black box nature of such models renders it difficult to infer underlying biological variability as the root cause of these disparities. We did find that the primary model showed differential changes in risk scores by race, age, and gender compared to a linear regression model trained on clinical factors and demographic variables (Table S7). The scope of this work lies in highlighting disparities that may arise in the development of models on ECG data, alongside a series of approaches to mitigate the biases.

Conclusion

We analyzed demographic disparities on the basis of race, ethnicity, age, and gender present in a machine learning model trained to predict occurrence of 5-year heart failure from ECG data and found that performance suffered significantly for young Black women. We explored the mechanisms underlying these disparities through statistical covariates, calibration curves, and distributions of predicted probabilities. We investigated methods of improving these disparities – manipulation of the training data, modifications to model architecture, and optimization of threshold choice. This study serves to highlight the need to consider differential performance of machine learning models among demographic subgroups, and offers a framework for understanding, investigating, and improving the disparities that may otherwise be perpetuated by algorithms used for medical decision making.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We do not have any acknowledgements to make.

Sources of Funding

Ms. Kaur is funded by the Stanford Medical Scholars Research fellowship. Mr. Hughes is an NSF Graduate Research Fellow (DGE-1656518). Dr. Perez reports funding from NIH/NHLBI and Apple Inc.

Disclosures

Dr. Narayan reports research grants from NIH (R01 HL149134 “Machine Learning in Atrial Fibrillation”, and R01 HL83359 “Dynamics of Atrial Fibrillation”), consulting compensation from Abbott Inc., Up to Date, and LifeSignals.ai, and intellectual property rights from University of California Regents and Stanford University. Dr. Ashley reports consulting fees from Apple Inc. Dr. Perez reports consulting fees from Apple Inc., Boston Scientific, Biotronik Inc., Bristol Myers Squibb, QALY Inc., Johnson & Johnson, and has an equity interest in QALY Inc.

Non-standard Abbreviations and Acronyms

AUC	area under the curve
BNP	B-type natriuretic peptide
ECG	electrocardiogram

References

1. Virani SS, Alonso A, Benjamin EJ, et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation. 2020;141(9):e139–e596. doi:10.1161/CIR.0000000000000757 [PubMed: 31992061]
2. Bosch L, Assmann P, de Grauw WJC, Schalk BWM, Biermans MCJ. Heart failure in primary care: prevalence related to age and comorbidity. Prim Health Care Res Dev. 2019;20:e79. doi:10.1017/S1463423618000889 [PubMed: 31868152]
3. Groenewegen A, Rutten FH, Mosterd A, Hoes AW. Epidemiology of heart failure. European Journal of Heart Failure. 2020;22(8):1342–1356. doi:10.1002/ejhf.1858 [PubMed: 32483830]

4. Glynn P, Lloyd-Jones DM, Feinstein MJ, Carnethon M, Khan SS. Disparities in Cardiovascular Mortality Related to Heart Failure in the United States. Journal of the American College of Cardiology. 2019;73(18):2354–2355. doi:10.1016/j.jacc.2019.02.042 [PubMed: 31072580]
5. Sandhu AT, Tisdale RL, Rodriguez F, et al. Disparity in the Setting of Incident Heart Failure Diagnosis. Circulation: Heart Failure. 2021;14(8):e008538. doi:10.1161/CIRCHEARTFAILURE.121.008538
6. Alfaras M, Soriano MC, Ortín S. A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection. Frontiers in Physics. 2019;7. Accessed August 24, 2022. https://www.frontiersin.org/articles/10.3389/fphy.2019.00103
7. Aziz S, Ahmed S, Alouini MS. ECG-based machine-learning algorithms for heartbeat classification. Sci Rep. 2021;11(1):18738. doi:10.1038/s41598-021-97118-5 [PubMed: 34548508]
8. Darmawahyuni A, Nurmaini S, Rachmatullah MN, et al. Deep learning-based electrocardiogram rhythm and beat features for heart abnormality classification. PeerJ Comput Sci. 2022;8:e825. doi:10.7717/peerj-cs.825
9. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun. 2020;11(1):1760. doi:10.1038/s41467-020-15432-4 [PubMed: 32273514]
10. Hughes JW, Olgin JE, Avram R, et al. Performance of a Convolutional Neural Network and Explainability Technique for 12-Lead Electrocardiogram Interpretation. JAMA Cardiology. 2021;6(11):1285–1295. doi:10.1001/jamacardio.2021.2746 [PubMed: 34347007]
11. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med. 2019;25(1):70–74. doi:10.1038/s41591-018-0240-2 [PubMed: 30617318]
12. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet. 2019;394(10201):861–867. doi:10.1016/S0140-6736(19)31721-0
13. Kwon J, Lee SY, Jeon K, et al. Deep Learning–Based Algorithm for Detecting Aortic Stenosis Using Electrocardiography. Journal of the American Heart Association. 2020;9(7):e014717. doi:10.1161/JAHA.119.014717 [PubMed: 32200712]
14. Kwon J myoung, Cho Y, Jeon KH, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. The Lancet Digital Health. 2020;2(7):e358–e367. doi:10.1016/S2589-7500(20)30108-4 [PubMed: 33328095]
15. Tat E, Bhatt DL, Rabbat MG. Addressing bias: artificial intelligence in cardiovascular medicine. The Lancet Digital Health. 2020;2(12):e635–e636. doi:10.1016/S2589-7500(20)30249-1 [PubMed: 33328028]
16. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176–2182. doi:10.1038/s41591-021-01595-0 [PubMed: 34893776]
17. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances. 2022;8(32):eabq6147. doi:10.1126/sciadv.abq6147
18. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453. doi:10.1126/science.aax2342 [PubMed: 31649194]
19. Kamulegeya LH, Okello M, Bwanika JM, et al. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. Published online October 31, 2019:826057. doi:10.1101/826057
20. Allen A, Mataraso S, Siefkas A, et al. A Racially Unbiased, Machine Learning Approach to Prediction of Mortality: Algorithm Development Study. JMIR Public Health Surveill. 2020;6(4):e22400. doi:10.2196/22400 [PubMed: 33090117]
21. Macfarlane PW. The Influence of Age and Sex on the Electrocardiogram. In: Kerkhof PLM, Miller VM, eds. Sex-Specific Analysis of Cardiovascular Function. Springer International Publishing; 2018:93–106. doi:10.1007/978-3-319-77932-4_6

22. Macfarlane PW, Katibi IA, Hamde ST, et al. Racial differences in the ECG — selected aspects. Journal of Electrocardiology. 2014;47(6):809–814. doi:10.1016/j.jelectrocard.2014.08.003 [PubMed: 25193321]
23. Kornberg B. S-T segment variants in young, adult, black women. Journal of the American Osteopathic Association. 1983;83(2):125–127. [PubMed: 6643146]
24. Kashou AH, Basit H, Malik A. ST Segment. In: StatPearls. StatPearls Publishing; 2022. Accessed October 31, 2022. http://www.ncbi.nlm.nih.gov/books/NBK459364/
25. Walsh B, Macfarlane PW, Prutkin JM, Smith SW. Distinctive ECG patterns in healthy black adults. J Electrocardiol. 2019;56:15–23. doi:10.1016/j.jelectrocard.2019.06.007 [PubMed: 31229678]
26. Santhanakrishnan R, Wang N, Larson MG, et al. Racial Differences in Electrocardiographic Characteristics and Prognostic Significance in Whites Versus Asians. J Am Heart Assoc. 2016;5(3):e002956. doi:10.1161/JAHA.115.002956
27. Noseworthy PA, Attia ZI, Brewer LC, et al. Assessing and Mitigating Bias in Medical Artificial Intelligence. Circulation: Arrhythmia and Electrophysiology. 2020;13(3):e007988. doi:10.1161/CIRCEP.119.007988
28. Datta S, Posada J, Olson G, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. Published online March 17, 2020. doi:10.48550/arXiv.2003.10534
29. Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. The Lancet. 2019;394(10211):1816–1826. doi:10.1016/S0140-6736(19)32317-7
30. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77. doi:10.1186/1471-2105-12-77 [PubMed: 21414208]
31. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–2830.
32. Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. ACM; 2018:1–7. doi:10.1145/3194770.3194776
33. Kleinberg J. Inherent Trade-Offs in Algorithmic Fairness. In: Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems. SIGMETRICS ‘18. Association for Computing Machinery; 2018:40. doi:10.1145/3219617.3219634
34. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data. 2019;6(1):27. doi:10.1186/s40537-019-0192-5
35. Bonham VL, Green ED, Pérez-Stable EJ. Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. JAMA. 2018;320(15):1533–1534. doi:10.1001/jama.2018.13609 [PubMed: 30264136]
36. Cooper RS, Nadkarni GN, Ogedegbe G. Race, Ancestry, and Reporting in Medical Journals. JAMA. 2018;320(15):1531–1532. doi:10.1001/jama.2018.10960 [PubMed: 30264132]
37. Eneanya ND, Yang W, Reese PP. Reconsidering the Consequences of Using Race to Estimate Kidney Function. JAMA. 2019;322(2):113–114. doi:10.1001/jama.2019.5774 [PubMed: 31169890]
38. Segar MW, Jaeger BC, Patel KV, et al. Development and Validation of Machine Learning–Based Race-Specific Models to Predict 10-Year Risk of Heart Failure: A Multicohort Analysis. Circulation. 2021;143(24):2370–2383. doi:10.1161/CIRCULATIONAHA.120.053134 [PubMed: 33845593]

COMMENTARY

What is new?

- This study is among the first to conduct an analysis of intersectional biases in deep learning models for ECG data with respect to race, ethnicity, gender, and age.
- The primary model was trained to predict the incidence of heart failure within 5 years solely from 12-lead ECG data. The results indicate that model performance suffers for older patients, as well as for young Black patients, particularly women.
- We explore several avenues of approaches to ameliorating algorithmic biases that may serve as a generalized framework for improving model disparities.

What are the clinical implications?

- Our findings warrant caution in using this ECG deep learning model for heart failure prognosis among certain demographic subgroups.
- Analyses of intersectional biases are key to ensuring deep learning models do not perpetuate existing disparities in health outcomes.
- Finetuning a model's application in the clinical setting by calibrating model outputs to assigned risk scores may ameliorate existing model biases.

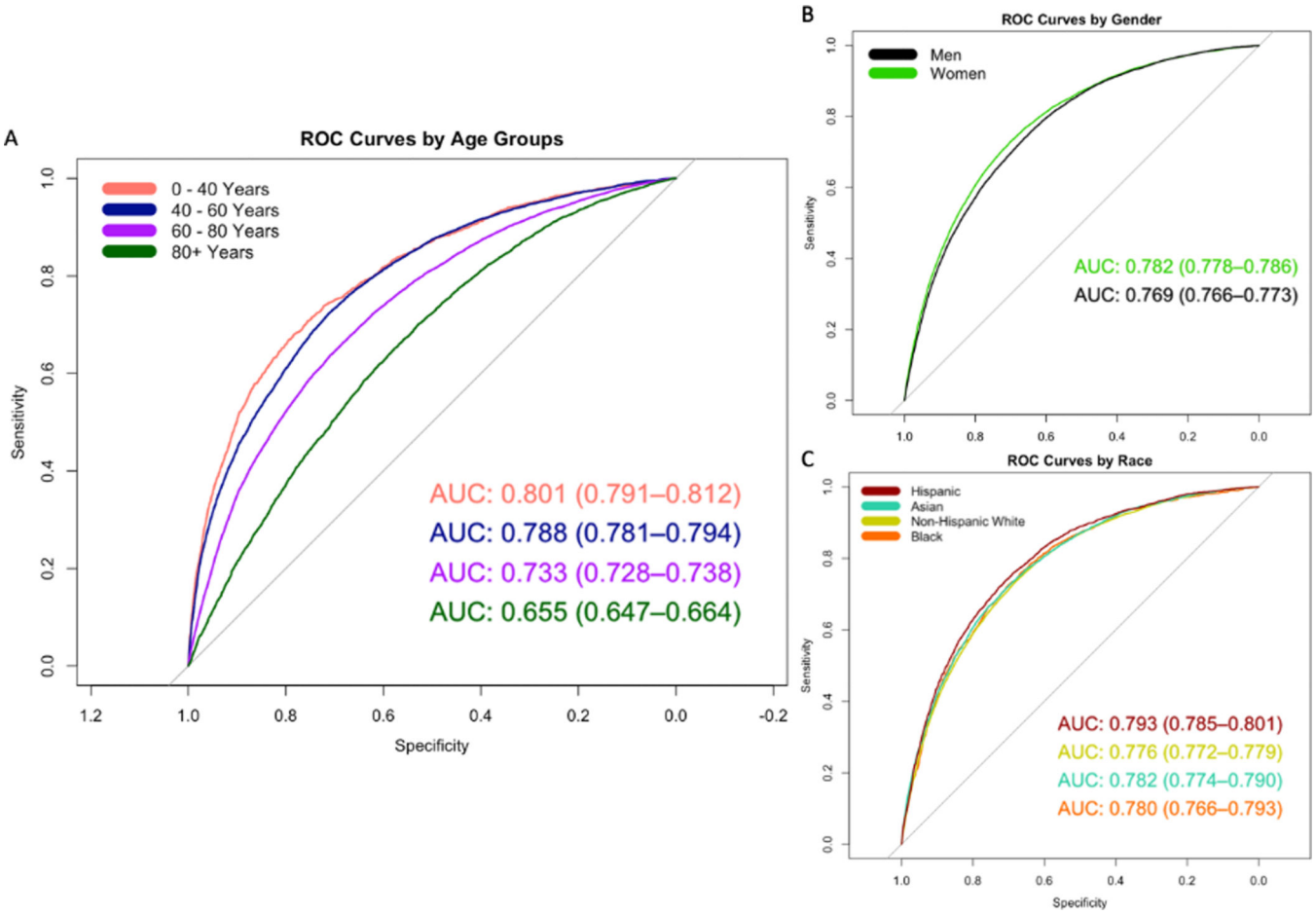

Figure 1.

Receiver operator characteristic (ROC) curves and area under the curve (AUC) with 95% confidence intervals stratified by (a) age, (b) gender, (c) race

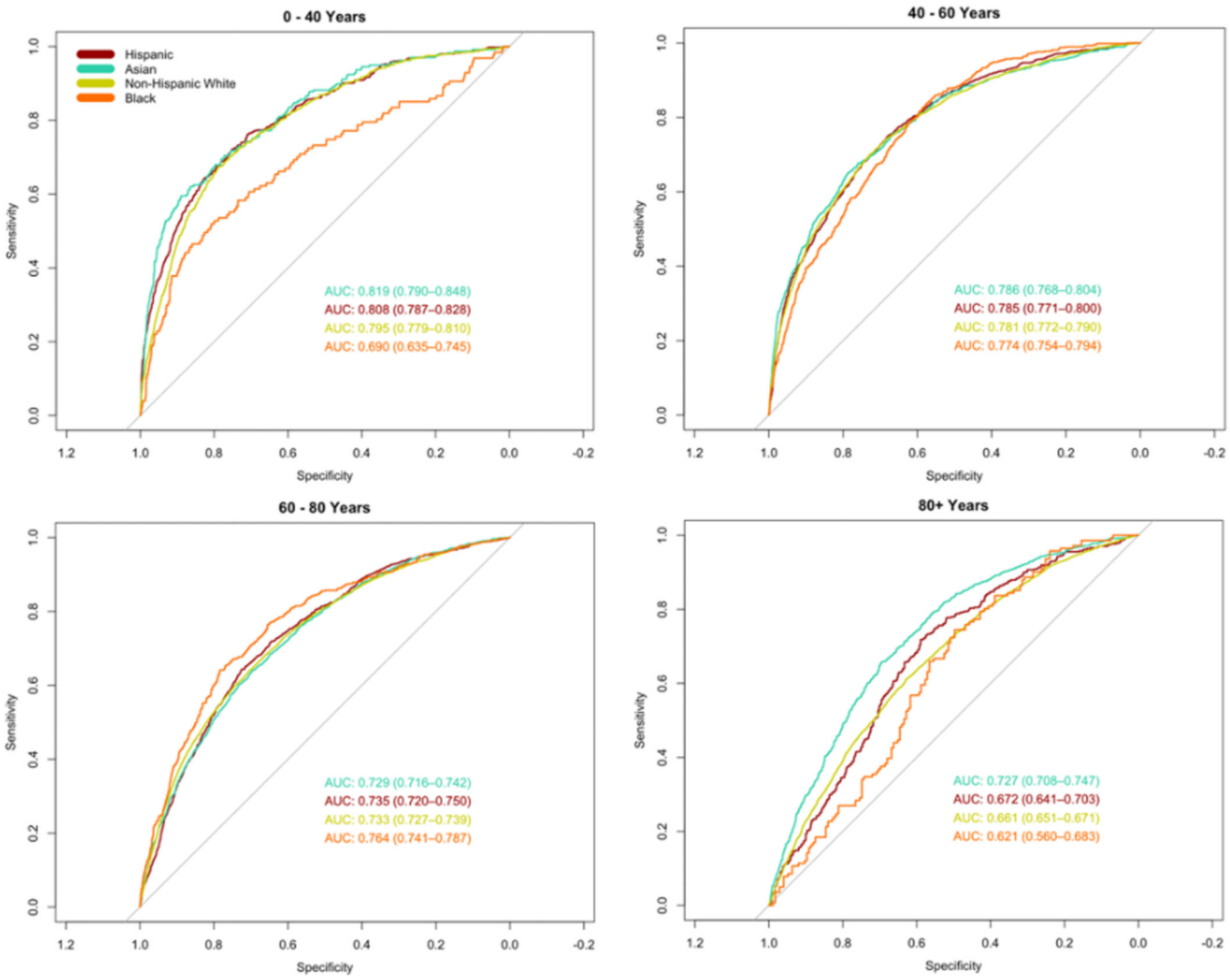

Figure 2.

Receiver operator characteristic (ROC) curves and area under the curve (AUC) stratified by race across age groups

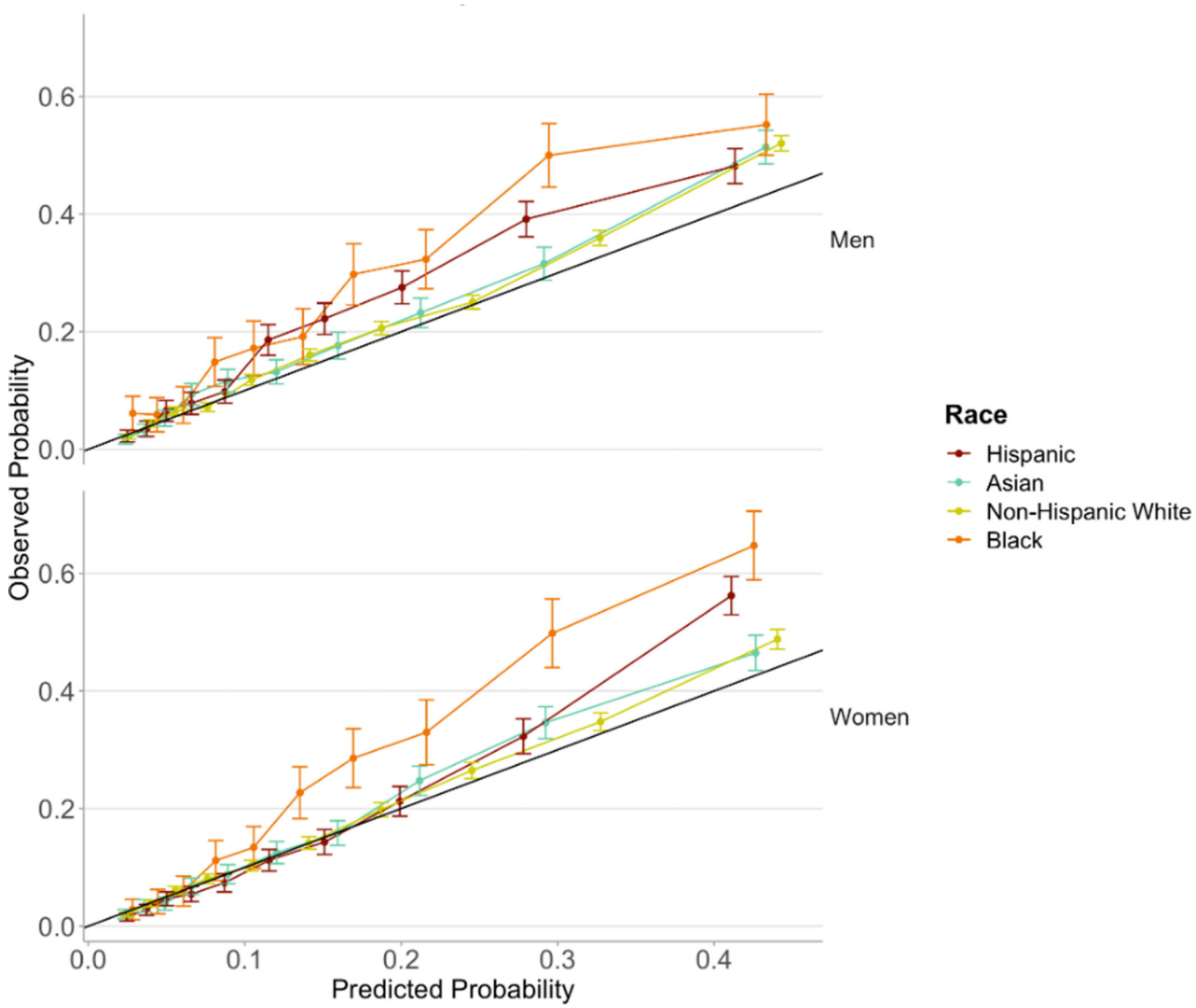

Figure 3.
Calibration curves by race and gender

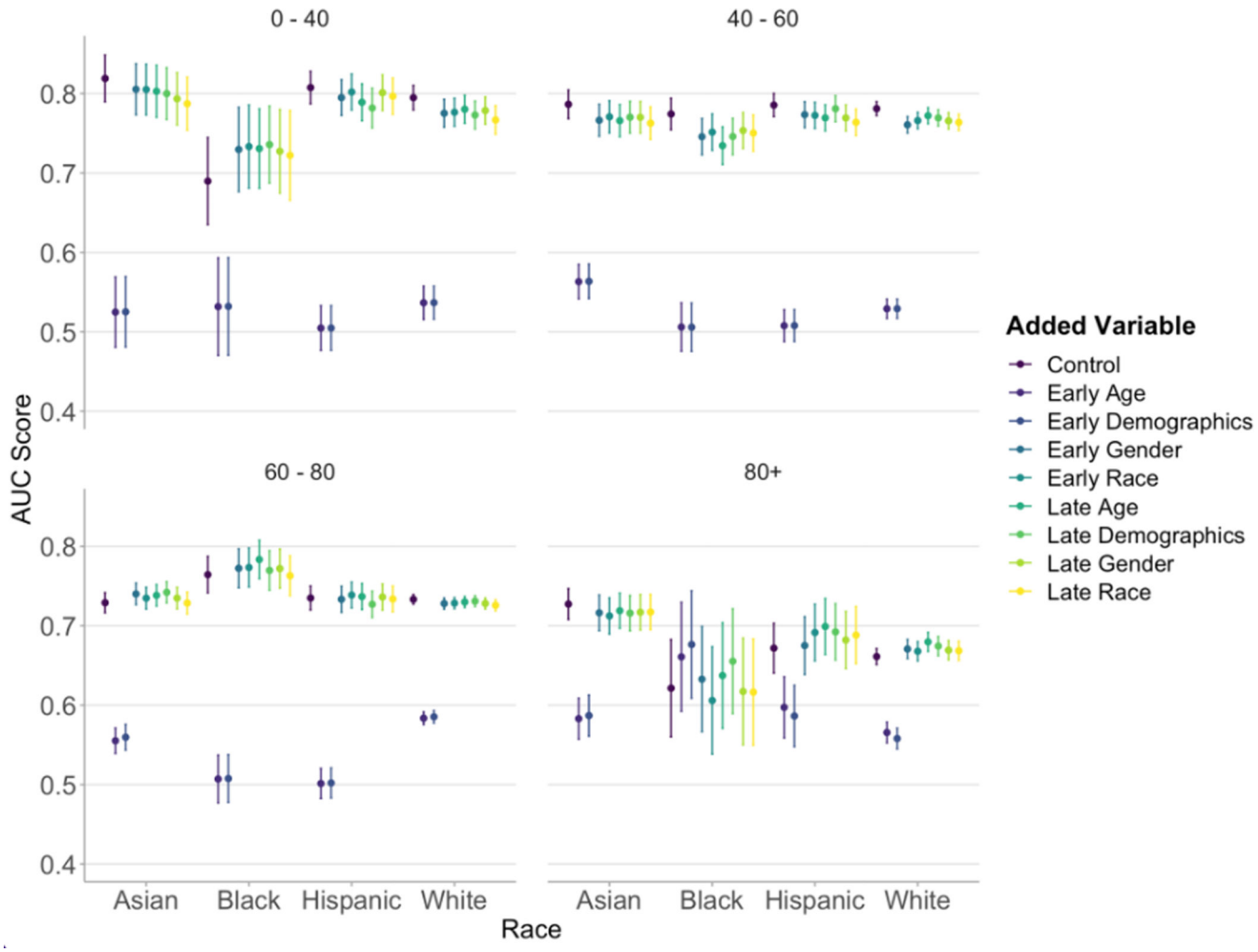

Figure 4.

AUC (area under the curve) values with error bars representing 95% confidence intervals by race and age groups with integration of age, race, and gender variables at earlier and later points in model architecture

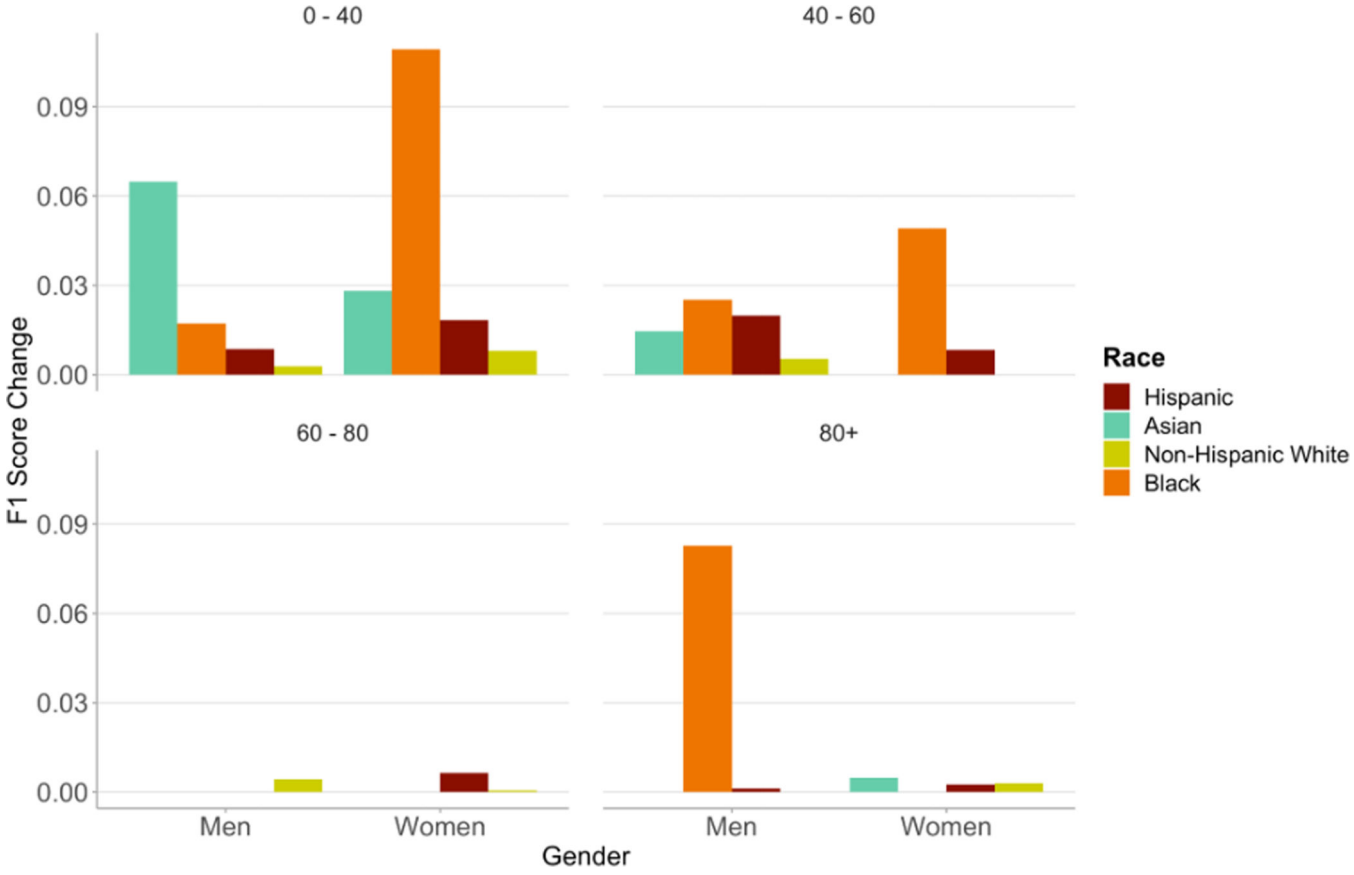

Figure 5.
Changes in F1-Score when using probability thresholds individualized for race, age, and gender subgroups

Table 1.

Baseline characteristics (Hypertension defined as diastolic >80 or systolic >130; hyperlipidemia defined as low-density lipoprotein > 160; abbreviated CAD - coronary artery disease, MI – myocardial infarction, BMI – body mass index)

Demographic	Subgroup	All (% of Total)	No Heart Failure (% of Subgroup)	Heart Failure (% of Subgroup)
Age	0 – 40	47657 (14.6%)	43389 (91.0%)	4268 (9.0%)
	40 – 60	106811 (32.7%)	93434 (87.5%)	13377 (12.5%)
	60 – 80	134669 (41.2%)	106235 (78.9%)	28434 (21.1%)
	≥ 80	37526 (11.5%)	23789 (63.4%)	13737 (36.6%)
Race	Asian	46393 (14.2%)	37771 (81.4%)	8622 (18.6%)
	Hispanic	40301 (12.3%)	33259 (82.5%)	7042 (17.5%)
	Non-Hispanic White	184410 (56.4%)	149234 (81.0%)	35086 (19.0%)
	Black or African American	13063 (4.0%)	9997 (76.5%)	3066 (23.5%)
	Native Hawaiian/ Pacific Islander	4080 (1.3%)	2987 (73.2%)	1093 (26.8%)
	American Indian/ Alaskan Native	700 (0.2%)	576 (82.3%)	124 (17.7%)
Gender	Male	164483 (50.3%)	131331 (79.8%)	33152 (20.2%)
	Female	162190 (49.6%)	135526 (83.6%)	26664 (16.4%)
BMI	< 18.5	5539 (1.7%)	4431 (80.0%)	1108 (20.0%)
	18.5 – 24.9	66660 (20.4%)	53418 (80.1%)	13242 (19.9%)
	25.0 – 29.9	68788 (21.1%)	54095 (78.6%)	14693 (21.4%)
	≥ 30.0	57996 (17.8%)	43889 (75.7%)	14107 (24.3%)
Hypertension	Positive History	109941 (33.7%)	86713 (78.9%)	23228 (21.1%)
Diabetes	Positive History	44453 (13.6%)	30198 (67.9%)	14255 (32.1%)
CAD or MI	Positive History	43845 (13.4%)	27730 (63.2%)	16115 (36.8%)
Stroke	Positive History	14577 (4.5%)	10928 (75.0%)	3649 (25.0%)
Hyperlipidemia	Positive History	1891 (0.6%)	1618 (85.6%)	273 (14.4%)
Smoking	Positive History	2267 (0.7%)	1753 (77.3%)	514 (22.7%)