Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis

Peter A. Noseworthy, MD[1,2], **Zachi I. Attia, MSc**[1], **LaPrincess C. Brewer, MD, MPH**[1], **Sharonne N. Hayes, MD**[1,3], **Xiaoxi Yao, PhD**[2,4], **Suraj Kapa, MD**[1], **Paul A. Friedman, MD**[1], **Francisco Lopez-Jimenez, MD, MSc**[1]

[1]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN

[2]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN

[3]Office of Diversity and Inclusion, Mayo Clinic, Rochester, MN

[4]Division of Health Care Policy and Research, Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract

Background -—Deep learning algorithms derived in homogeneous populations may be poorly generalizable and have the potential to reflect, perpetuate, and even exacerbate racial/ethnic disparities in health and healthcare. In this study we aimed to (1) assess if the performance of a deep learning algorithm designed to detect low left ventricular ejection fraction (LVEF) using the 12-lead electrocardiogram (ECG) varies by race/ethnicity, and to (2) determine whether its performance is determined by the derivation population, or by racial variation in the ECG.

Methods -—We performed a retrospective cohort analysis that included 97,829 patients with paired ECGs and echocardiograms. We tested the model performance by race/ethnicity for convolutional neural network (CNN) designed to identify patients with a LVEF ≤ 35% from the 12-lead ECG.

Results -—The CNN which was previously derived in a homogeneous population (derivation cohort N=44,959; 96.2% non-Hispanic White) demonstrated consistent performance to detect low LVEF across a range of racial/ethnic subgroups in a separate testing cohort (N=52,870): Non-Hispanic white (N= 44,524, AUC 0.931), Asian (N=557, AUC 0.961), black/African American (N=651, AUC 0.937), Hispanic/Latino (N=331, AUC 0.937), and American Indian/Native Alaskan (N=223, AUC 0.938). In secondary analyses, a separate neural network was able to discern racial subgroup category (Black/African-American [AUC of 0.84], and white, non-Hispanic [AUC 0.76] in a five-class classifier), and a network trained only in non-Hispanic whites from the original

Correspondence: Francisco Lopez-Jimenez, MD, MSc, Professor of Medicine, Mayo Clinic, Department of Cardiovascular Medicine, 200 1st St SW, Rochester, MN 55905, Tel: 507-284-4375, Fax: 507-266-7929, lopez@mayo.edu.

derivation cohort performed similarly well across a range of racial/ethnic subgroups in the testing cohort with an AUC of at least 0.930 in all racial/ethnic subgroups.

Conclusions -—Our study demonstrates that while ECG characteristics vary by race, this did not impact the ability of a CNN to predict low LVEF from the ECG. We recommend reporting of performance amongst diverse ethnic, racial, age and gender groups for all new AI tools to ensure responsible use of AI in medicine.

Graphical Abstract

Convolutional Neural Network (CNN) able to predict from ECG low LVEF across races
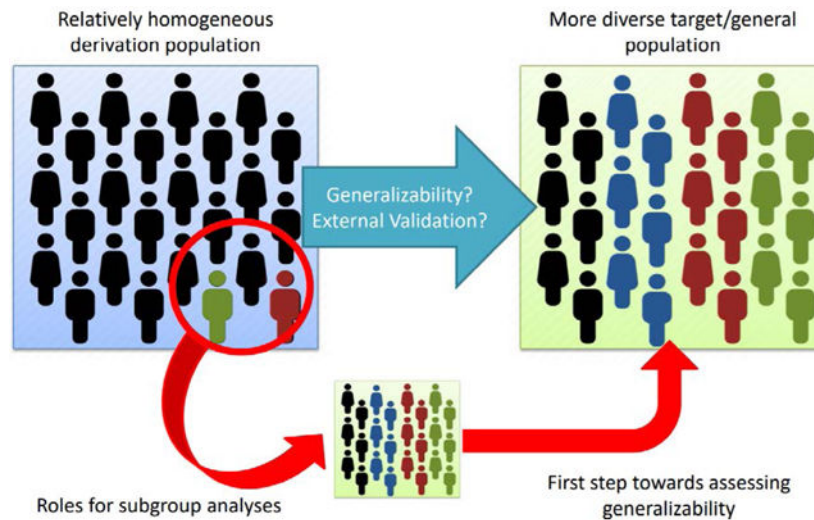
• The CNN which was previously derived in a homogeneous population (N=44,959; 96.2% non-Hispanic whites)

• The CNN was tested in a separate cohort (N=52,870):

• Non-Hispanic white (N= 44,524, AUC 0.931),

• Asian (N=557, AUC 0.961)

• black/African American 45 (N=651, AUC 0.937)

• Hispanic/Latino (N=331, AUC 0.937)

• American Indian/Native Alaskan 46 (N=223, AUC 0.938)

Keywords

electrocardiography; race and ethnicity; machine learning; artificial intelligence

Journal Subject Terms:

Race and Ethnicity; Contractile function; Electrocardiology (ECG)

Introduction

Though momentum for the application of artificial intelligence (AI) to medicine continues to build,[1–3] there are several examples of initially promising findings that struggle when applied to diverse populations. For instance, Google's algorithm for diagnosis of diabetic retinopathy performed poorly in populations in India outside where the model had been developed.[4] Other dramatic examples outside clinical medicine include failure of facial recognition software to recognize non-White faces,[5] and, more ominously, examples from the criminal justice system in which structural flaws in the data create prediction networks that perpetuate bias and inequity.[6, 7]

These problems can arise when datasets used to train the algorithms lack demographic diversity or when the datasets themselves reflect disparities in outcomes or bias in human behavior. For instance, if a network used to predict outcomes after a particular intervention reflects racial disparities unrelated to either the patient's intrinsic risk profile or the intervention (uninsured status, for example), then application of the model for patient selection could perpetuate the disparity. As such, these technologies have the potential to reflect, perpetuate, and even exacerbate disparities in health and healthcare. Addressing generalizability in AI, and the contribution of racial and ethnic categories to algorithm performance, will therefore remain key barriers towards its implementation into clinical practice.

We have previously developed a convolutional neural network (CNN) that can determine the presence of left ventricular (LV) systolic dysfunction from an electrocardiogram (ECG). The algorithm demonstrated very strong overall performance with an area under the curve (AUC) of 0.93,[8, 9] and has the potential to offer a scalable and low-cost screening test for LV systolic dysfunction. The model performance was consistent across age and sex subgroups. However, the derivation population was overwhelmingly non-Hispanic White (>95%) and, like many biometric variables, ECG features are known to vary by race.[10–12] In our initial report, we did not assess the performance of the CNN by race, leaving a key issue of generalizability unexplored.

In this analysis, we aimed to determine the impact of race on the performance of our CNN to detect the presence of LV systolic dysfunction from an ECG. Furthermore, to explain the generalizability (or lack thereof) of the network, we also evaluated whether a similar algorithm, trained on a completely homogenous population, would achieve the same outcomes. Finally, we assessed whether a CNN could discern race using the ECG itself.

Methods

Data Sources and Study Population

As previously described,[8, 9] we identified all adult patients (18 years or older) with at least one digital, standard 10-second 12-lead ECG and a standard transthoracic echocardiogram obtained within two weeks of each other, from January 1994 and February 2017. All ECGs were acquired at a sampling rate of 500 Hz using a GE-Marquette ECG machine (Marquette, WI) and stored using the MUSE data management system for later retrieval. All ECGs were

recorded by a trained technician while the patients were supine and still for 10 seconds. We did not perform any preprocessing of the signals. The ECG recording system used 0.15–100 Hz filters that resulted in consistent recordings with minimal noise. LVEF was routinely determined using a hierarchical priority from the most to the least valid method (3-D, volumetrics, biplane, M-Mode, and visual estimation).[13] Race and ethnicity were self-reported according the U.S. Census Bureau Office of Management and Budget standards on Race and Ethnicity categories[14] and recorded in the medical record (non-Hispanic White, Asian, Black/African American, Hispanic/Latino, and American Indian/Native Alaskan). Patients who did not report race or ethnicity or who reported multiple categories were not included in the analysis. No other exclusions were applied. The Mayo Clinic Internal Review Board approved waiver of the requirement to obtain informed consent in accordance with 45 CFR 46.116, and waiver of HIPAA authorization in accordance with applicable HIPAA regulations. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Primary Analysis

The primary outcome was the ability of the deep learning network to identify patients with a low LVEF of 35% using the ECG signal alone in patients with prior echocardiographic assessments of LV function performed within each racial/ethnic subgroup: non-Hispanic White, Asian, Black/African American, Hispanic /Latino, and American Indian/Native Alaskan. Details of the previously-described CNN identification of patients with a low LVEF of 35% using the ECG signal are included in Appendix 1. We used the previously published model architecture. The model input was the 12 lead ECG as a matrix of 5000×12 values representing 10 seconds of ECG sampled at 500Hz. The last layer of the model was a 1×12 convolution filter that combines the data from the different leads. The weights for the lead-specific combining function were learned in the training process. We did not use regularization.

Based on the network output (the probability of the patient having an LVEF 35%), we created a receiver-operating curve (ROC) and AUC as the primary assessment of network strength to diagnose low LVEF in each of the racial subgroups. The threshold probability to define a positive screen for low LVEF was determined from the prior publication on the internal validation set. [9]

Secondary Analyses

We performed two secondary analyses. First, we retrained a network using a population restricted to non-Hispanic White patients and tested the algorithm on each of the racial subgroups. Similar to the primary analysis, we created a receiver-operating curve (ROC) and measured its AUC as primary assessment of network strength to diagnose low LVEF in each of the racial/ethnic subgroups.

Second, a separate neural network was trained on the self-reported race/ethnicity categories in order to detect ECG features suggestive of an individual's racial/ethnic subgroup. The network structure of the multiclass network was similar to the LVD classification network, however, the output layer was replaced by a 5-neuron layer, each reprinting a different racial/

ethnic subgroup. During training, the network with the highest overall validation accuracy was selected and testing in the testing dataset (the testing dataset was not used for any model training and was 'unseen' by the model until model testing phase).

Statistical Considerations

Continuous data are presented as mean +/− standard deviation (SD) and median and interquartile range (IQR) if highly skewed. For measures of diagnostic performance (AUC, ROC, sensitivity, specificity), 95% confidence intervals were computed under the assumption of independence. Since the current study was aimed at validating the previously derived model, the threshold for determining accuracy, sensitivity and specificity was selected based on the original work and alerted for any ECG that had an output greater than 25.6% (i.e. the ECG indicates at least a 25.6% probability of EF 35%). These routine statistical analyses were conducted using R (version 3.4.2).

The original testing cohort was used for the primary analyses (testing the performance of the CNN across racial/ethnic subgroups) and the original derivation cohort was used in the two sensitivity analyses to retrain two networks (for race classification and for low LVEF detection using a homogeneous sample) which were then applied to the "unseen" testing cohort. The testing cohort was a holdout dataset and was never used or shown to the model during any stage of the development.

Results

Study Population

The overall sample included 97,829 patients with paired ECG and echocardiographic data (Figure 1). During the previously published model derivation and testing,[9] the group was divided into a derivation cohort (N=44,959) and a testing cohort (N=52,870). The breakdown of self-reported race/ethnicity in the testing cohort are shown in Figure 1. Self-reported race was not available in 6,587 patients. The majority of patients in the sample were non-Hispanic Whites (44,524/46,283, 96.1%).

The baseline characteristics, stratified by self-reported race and ethnicity are shown in Table 1. As shown in Figure 2, the distribution of age and the probability of low LVEF varied by race/ethnicity. The non-Hispanic white patients were slightly older but had a lower proportion of patients with LVEF 35%.

Primary Analysis: Model Performance across Racial and Ethnic Subgroups

The previously derived CNN demonstrated consistent performance to detect low LVEF across a range of racial subgroups in the testing cohort (Figure 3; N=52,870): non-Hispanic White (N=44,524, AUC 0.931), Asian (N=557, AUC 0.961), Black/African American (N=651, AUC 0.937), Hispanic/Latino (N=331, AUC 0.937), and American Indian/Native Alaskan (N=223, AUC 0.938).

Secondary Analysis: Detection of Racial Subgroup using ECG Features

A multi class neural network (with more balanced representation of the different racial/ethnic groups) was able to discern racial/ethnicity subgroup in the testing cohort with an accuracy of 56.2% (Figure 4). The performance was best of classification of Black/African-American (AUC of 0.84), and White, non-Hispanic(AUC 0.76) race/ethnicity, but less robust for the other subgroups (AUC was between 0.65–0.70).

Secondary Analysis: Performance of a CNN trained on a Completely Homogeneous Group

A network trained only on White non-Hispanic individuals from the derivation cohort performed similarly well across a range of racial/ethnic subgroups in the testing cohort with an AUC of at least 0.930 in all racial/ethnic subgroups.

Discussion

The primary finding of this study is that a previously derived CNN demonstrated consistent performance for the detection of low LVEF using the ECG across a range of racial/ethnic subgroups. Though this and other studies have demonstrated that ECG features vary by race[10–12] and the population used to train the network was quite homogeneous, our network was able to classify tracings with high accuracy. Taken together, these findings illustrate the importance of validating the applicability of AI across racial and ethnic categories, while also highlighting the unpredictable relationship between race, biophysical signals, and algorithm performance.

LV systolic dysfunction is associated with adverse outcomes,[15] but progression to overt heart failure is preventable.[16, 17] There is currently no established approach to population-level screening for asymptomatic LV systolic dysfunction,[15, 18] but this algorithm could potentially be incorporated into existing ECG interpretation work flow in order to provide front-line clinicians with a tool to identify such patients. Implementation of the algorithm would allow prospective collection of data on model performance and provide data for further tuning and refinement of the network.

Although it is highly preferable to derive models in heterogeneous populations, our study demonstrates that even when diverse populations are not available for training, some networks may be (at least in this particular case) widely applicable. It is important, however, not to make this assumption but instead to examine rigorously the performance of any given algorithm across a range of demographic subgroups. In order to avoid potentially hidden bias in AI tools used for healthcare, subgroup reporting and AI test assessment in different populations is critical. In light of our findings, the features used by this specific neural network for the detection of ventricular dysfunction appear to be race invariant. However, it is imperative that future algorithms intended for broad application similarly pursue validation according to the diversity reflected in the target population.

Need for consistent reporting of racial/ethnic subgroup effects

Although the current algorithm was demonstrated to be race-invariant in its performance, other algorithms may not be. It will be critical as new algorithms are introduced, that the

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

performance across a range of demographic subgroups is tested and reported. Only with consistent and rigorous reporting will we be able to determine which algorithms may be less generalizable and require further training before they can be integrated into clinical practice. It is also critical that race and ethnicity data be consistently collected and reported to allow consistency of these algorithms.

We recommend four potential strategies to minimize bias in AI. We suggest that (1) investigators remain mindful of the potential for racial/ ethnic bias in AI, (2) investigators report performance across a range of demographic subgroups, particularly race/ethnicity in all initial derivation/validation studies, (3) when models are poorly generalizable, investigators perform additional model training on diverse populations (seeking collaborations when needed), and (4) investigators perform external validation in populations that are representative of the population for intended clinical application.

Limitations

While the number of racial/ethnic minorities in our study was sufficient to calculate an AUC for the model in each group, the sample sizes were still small. Furthermore, the sample was geographically limited to Mayo Clinic, Rochester MN, which in addition to being a relatively homogenous population of predominantly non-Hispanic Whites, could have other issues related to geographic, socioeconomic, or healthcare setting generalizability. The algorithm was derived and validated on ECG signals obtained using the GE system and other healthcare organizations that obtain ECG data through other means may also not be able to apply the algorithm. Lastly, as is true of all deep learning exercises, the specific ECG characteristics used by our unsupervised CNN to classify individuals are not known. We are conducting ongoing studies to create interpretable read outs from the network, but this is not required to apply the current algorithm.

Conclusion

A previously-derived CNN had consistent performance for the detection of low LVEF using the ECG across a range of racial/ethnic subgroups. This is not to say that the ECG is race-invariant (indeed a separate CNN was able to discern race using ECG features), rather it suggests that the ECG features associated with ventricular dysfunction are race invariant. Other applications of the ECG to risk stratification could be highly sensitive to race. As such, we need to be aware that AI has the potential to exacerbate racial bias and health and healthcare disparities. We recommend vigilance, maintenance of diverse datasets, consistent subgroup reporting, and external validation in order to ensure responsible use of AI in medicine.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Nonstandard Abbreviations and Acronyms

AI	artificial intelligence
AUC	area under the curve
CNN	convolutional neural network
ECG	electrocardiogram
IQR	interquartile range
LVEF	left ventricular ejection fraction
ROC	receiver-operating curve
SD	standard deviation

References:

1. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316:2402–2410. [PubMed: 27898976]

2. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, Helbich TH, Chevalier M, Tan T, Mertelmeier T, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111:916–922. doi: 10.1093/jnci/djy222. [PubMed: 30834436]

3. Nakagawa M, Nakaura T, Namimoto T, Kitajima M, Uetani H, Tateishi M, Oda S, Utsunomiya D, Makino K, Nakamura H, et al. Machine learning based on multi-parametric magnetic resonance imaging to differentiate glioblastoma multiforme from primary cerebral nervous system lymphoma. Eur J Radiology. 2018;108:147–154.

4. Abrams C, Google's Effort to Prevent Blindness Shows AI Challenges. The Wall Street Journal. 1 26, 2019.

5. Hardesty L. Study finds gender and skin-type bias in commercial artificial-intelligence systems. MIT News Office 2018.

6. Budds D. Biased AI Is A Threat To Civil Liberties. The ACLU Has A Plan To Fix It. https://www.fastcompany.com/90134278/biased-ai-is-a-threat-to-civil-liberty-the-aclu-has-a-plan-to-fix-it. 2017;2019.

7. Tashea J. Courts are using AI to sentence criminals. That must stop now. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/ 2017;2019.

8. Attia ZI, Kapa S, Yao X, Lopez-Jimenez F, Mohan TL, Pellikka PA, Carter RE, Shah ND, Friedman PA, Noseworthy PA. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. J Cardiovascular Electrophysiology. 2019;30:668–674.

9. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med. 2019;25:70–74. [PubMed: 30617318]

10. Macfarlane PW, Katibi IA, Hamde ST, Singh D, Clark E, Devine B, Francq BG, Lloyd S, Kumar V. Racial differences in the ECG--selected aspects. J Electrocardiology. 2014;47:809–14.

11. Macfarlane PW, McLaughlin SC, Devine B, Yang TF. Effects of age, sex, and race on ECG interval measurements. J Electrocardiology. 1994;27 Suppl:14–9.

12. Rautaharju PM, Park LP, Gottdiener JS, Siscovick D, Boineau R, Smith V, Powe NR. Race- and sex-specific ECG models for left ventricular mass in older populations. Factors influencing overestimation of left ventricular hypertrophy prevalence by ECG criteria in African-Americans. J Electrocardiology. 2000;33:205–18.

13. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, Flachskampf FA, Foster E, Goldstein SA, Kuznetsova T, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiography. 2015;28:1–39 e14.

14. Bureau USC. U.S. Census Bureau Office of Management and Budget (OMB) standards on Race and Ethnicity. https://www.census.gov/topics/population/race/about.html2018;2019.

15. McDonagh TA, McDonald K, Maisel AS. Screening for asymptomatic left ventricular dysfunction using B-type natriuretic Peptide. Congest Heart Fail. 2008;14:5–8. [PubMed: 18772638]

16. Al-Khatib SM, Stevenson WG, Ackerman MJ, Bryant WJ, Callans DJ, Curtis AB, Deal BJ, Dickfeld T, Field ME, Fonarow GC, et al. 2017 AHA/ACC/HRS Guideline for Management of Patients With Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. Circulation. 2018;138:e272–e391. [PubMed: 29084731]

17. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr., Drazner MH, Fonarow GC, Geraci SA, Horwich T, Januzzi JL, et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. Circulation. 2013;128:e240–327. [PubMed: 23741058]

18. Redfield MM, Rodeheffer RJ, Jacobsen SJ, Mahoney DW, Bailey KR, Burnett JC Jr. Plasma brain natriuretic peptide to detect preclinical ventricular systolic or diastolic dysfunction: a community-based study. Circulation. 2004;109:3176–81. [PubMed: 15184280]

What is Known:

- ECG features vary by race which could affect the generalizability of deep learning approaches to ECG interpretation.

- Many examples of poorly generalizable AI algorithms exist indicating a potential for these technologies to reflect, perpetuate, and even exacerbate racial/ethnic disparities in healthcare.

What this study adds:

- A previously developed convolutional neural network that can determine the presence of LV systolic dysfunction from the ECG performs well across a wide range of racial/ethnic subgroups.
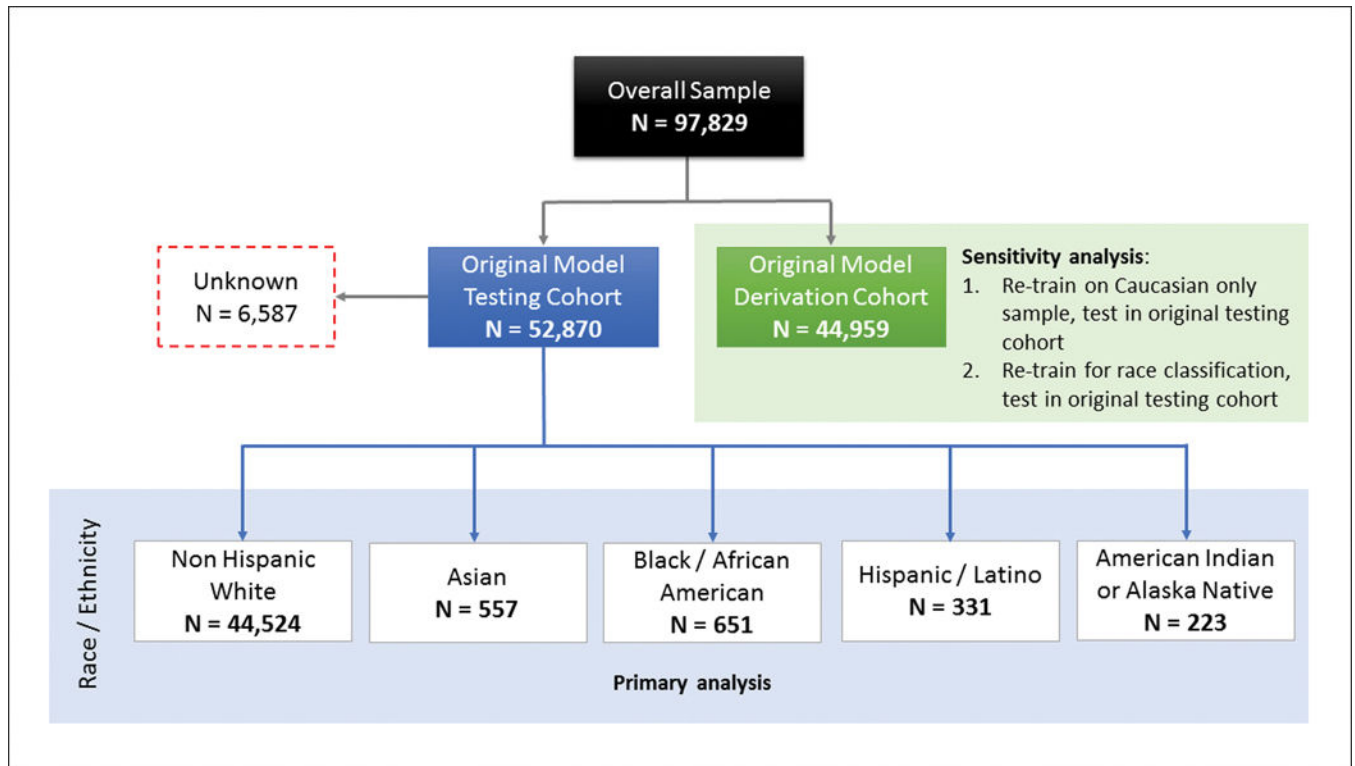
Figure 1.
Breakdown of race and ethnicity in the testing cohort (N=52,870, these previously 'unseen' patients were not part of the derivation cohort). The population was divided into four groups based on self-reported race (non-Hispanic White, Asian, Black/African American, and Native American/Alaskan).
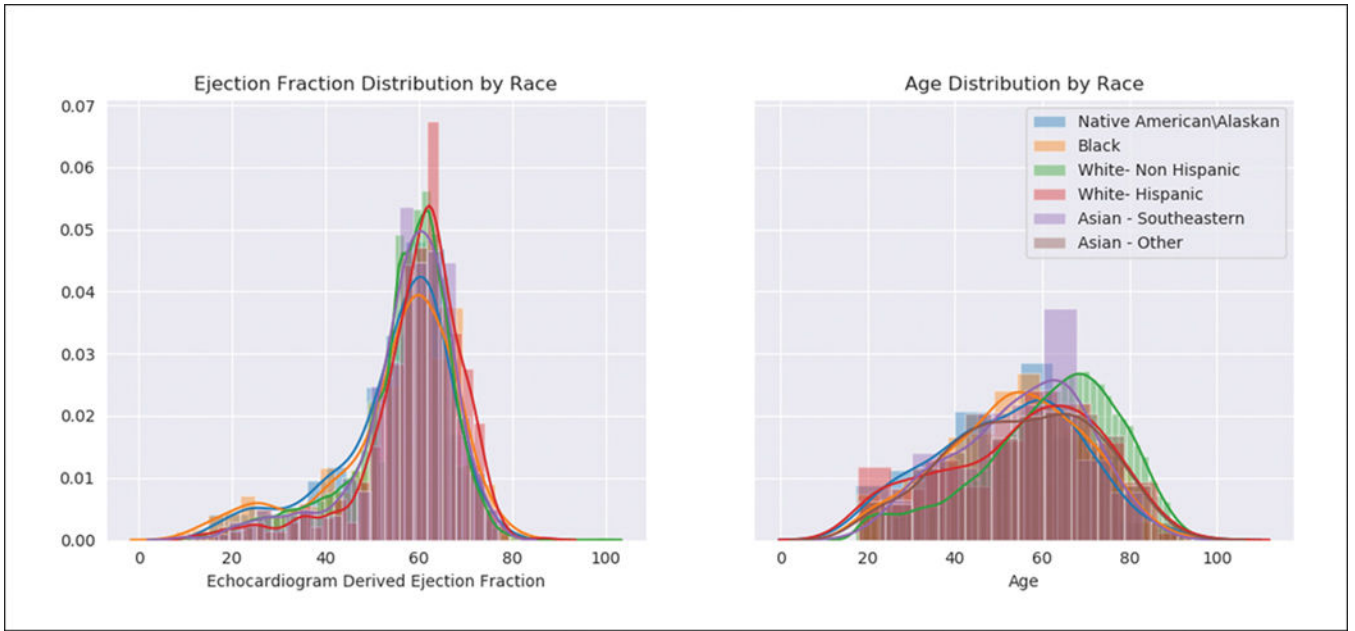
Figure 2.
Histograms demonstrating the distribution of (A) age and (B) ejection fraction, stratified by race/ethnicity group.
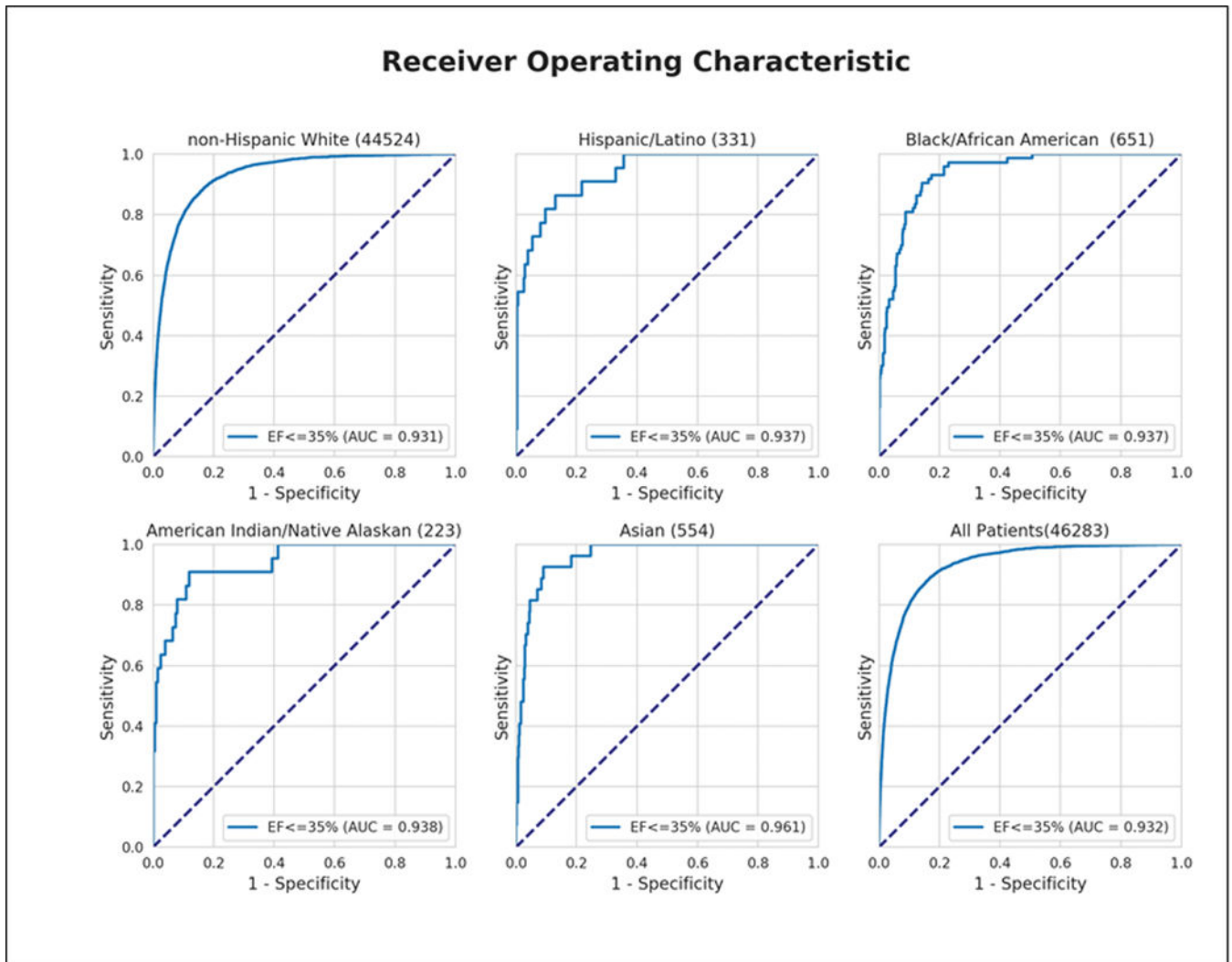
Figure 3.
The performance of the CNN across each of the six race/ethnic subgroups for the detection
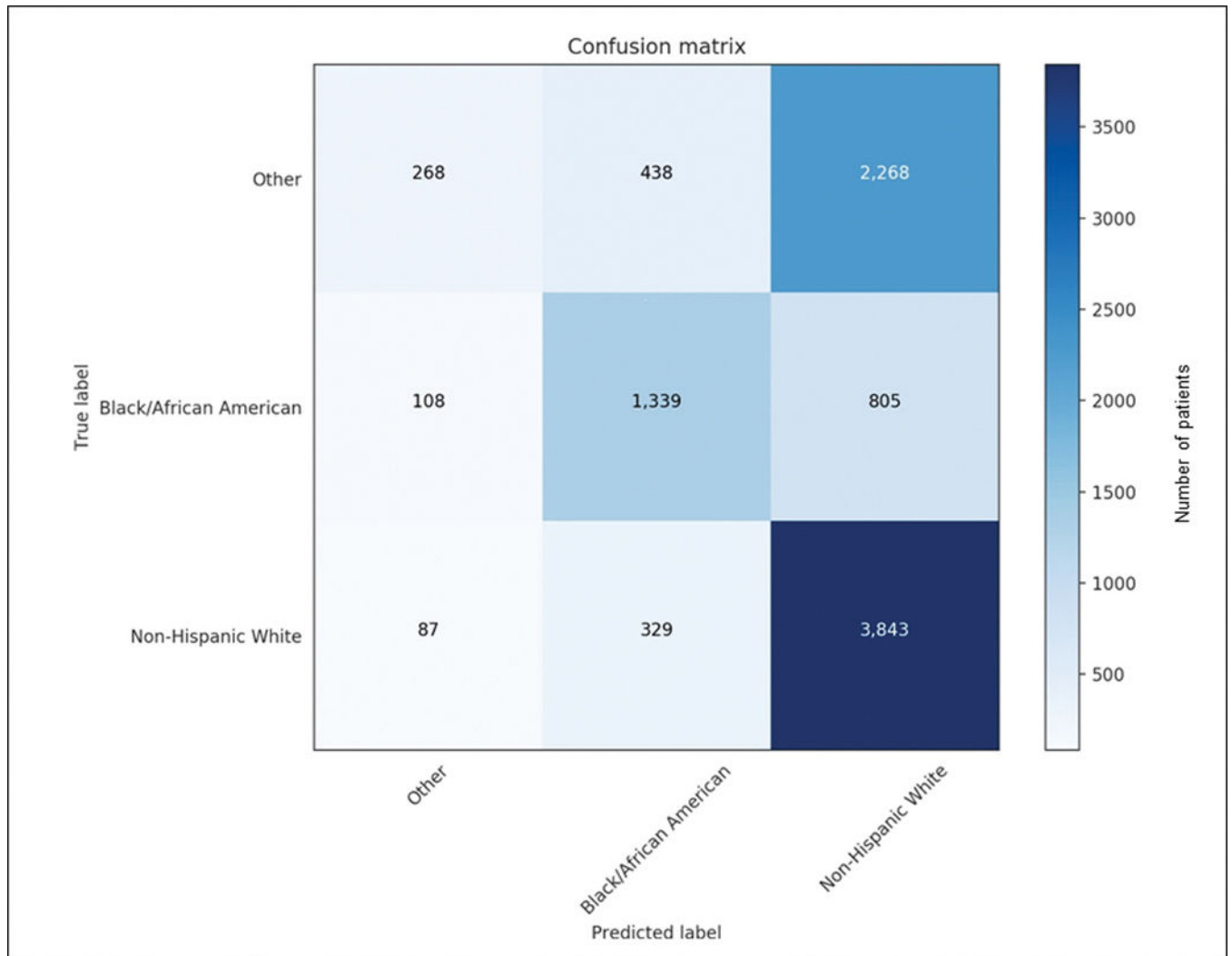of an ejection fraction 35% and 40%.

Figure 4.
Confusion matrix demonstrating the accuracy of the multiclass classifier (trained on the initial derivation cohort) to identify race based on the ECG (testing in the 'untouched' testing cohort). The classifier was most accurate for Black/African American and non-Hispanic White racial/ethnic groups but less accurate for the other subgroups.

Table 1.

Patient demographics, stratified by racial and ethnic subgroup

	Non Hispanic White	Black / African American	Hispanic / Latino	Asian	American Indian / Alaskan Native	Unknown
Patients, N	44,524	651	331	554	223	6,587
% male	52.8%	57.8%	40.8%	49.8%	50.5%	57.2%
Age, years	53.5	55.4	59.6	51.9	55.2	62.4
Ejection fraction, %	54.8	59.2	55.9	54.5	57.1	56.2
Medical history						
Diabetes mellitus	29.6%	24.4%	20.2%	23.8%	24.8%	23.9%
Dyslipidemia	33.5%	35.9%	35.1%	34.5%	37.2%	45.2%
Hypertension	47.5%	37.9%	40.0%	37.7%	34.4%	47.8%
Coronary artery disease	24.3%	29.6%	33.7%	30.0%	28.4%	38.9%
Angina	9.2%	11.6%	11.1%	9.9%	8.2%	14.2%
History of myocardial infarction	7.5%	6.7%	11.0%	11.2%	7.9%	13.3%
Coronary revascularization	6.9%	11.6%	10.5%	13.0%	7.9%	14.9%
Cardiomyopathy	26.4%	17.1%	15.1%	17.9%	16.9%	16.9%
Chronic kidney disease	16.6%	8.8%	7.2%	12.6%	6.6%	10.2%
Cerebrovascular Disease	10.4%	9.0%	9.7%	7.6%	8.8%	13.4%
Ischemic stroke	5.4%	5.2%	6.4%	4.0%	5.4%	8.9%