scientific reports

Check for updates

OPEN

Prediction performance and fairness heterogeneity in cardiovascular risk models

Uri Kartoun[1,6], Shaan Khurshid[2,3,6], Bum Chul Kwon[1], Aniruddh P. Patel[2,4], Puneet Batra[5], Anthony Philippakis[2], Amit V. Khera[2,4], Patrick T. Ellinor[2,3], Steven A. Lubitz[2,3] & Kenney Ng[1]✉

Prediction models are commonly used to estimate risk for cardiovascular diseases, to inform diagnosis and management. However, performance may vary substantially across relevant subgroups of the population. Here we investigated heterogeneity of accuracy and fairness metrics across a variety of subgroups for risk prediction of two common diseases: atrial fibrillation (AF) and atherosclerotic cardiovascular disease (ASCVD). We calculated the Cohorts for Heart and Aging in Genomic Epidemiology Atrial Fibrillation (CHARGE-AF) score for AF and the Pooled Cohort Equations (PCE) score for ASCVD in three large datasets: Explorys Life Sciences Dataset (Explorys, n = 21,809,334), Mass General Brigham (MGB, n = 520,868), and the UK Biobank (UKBB, n = 502,521). Our results demonstrate important performance heterogeneity across subpopulations defined by age, sex, and presence of preexisting disease, with fairly consistent patterns across both scores. For example, using CHARGE-AF, discrimination declined with increasing age, with a concordance index of 0.72 [95% CI 0.72–0.73] for the youngest (45–54 years) subgroup to 0.57 [0.56–0.58] for the oldest (85–90 years) subgroup in Explorys. Even though sex is not included in CHARGE-AF, the statistical parity difference (i.e., likelihood of being classified as high risk) was considerable between males and females within the 65–74 years subgroup with a value of − 0.33 [95% CI − 0.33 to − 0.33]. We also observed weak discrimination (i.e., < 0.7) and suboptimal calibration (i.e., calibration slope outside of 0.7–1.3) in large subsets of the population; for example, all individuals aged 75 years or older in Explorys (17.4%). Our findings highlight the need to characterize and quantify the behavior of clinical risk models within specific subpopulations so they can be used appropriately to facilitate more accurate, consistent, and equitable assessment of disease risk.

Abbreviations

1 K PY	1000 Patient years
AF	Atrial fibrillation
ASCVD	Atherosclerotic cardiovascular disease
CHARGE-AF	Cohorts for Heart and Aging Research in Genomic Epidemiology atrial fibrillation
CPT	Current Procedural Terminology
DBP	Diastolic blood pressure
EHR	Electronic health record
HDL	High-density lipoprotein
HF	Heart failure
ICD	International Classification of Diseases
Inc	Incidence
MGB	Mass General Brigham
MI	Myocardial infarction
PCE	Pooled Cohort Equations

[1]Center for Computational Health, IBM Research, 314 Main St., Cambridge, MA 02142, USA. [2]Cardiovascular Disease Initiative, Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA. [3]Demoulas Center for Cardiac Arrhythmias, Massachusetts General Hospital, Boston, MA, USA. [4]Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA. [5]Data Sciences Platform, Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA. [6]These authors contributed equally: Uri Kartoun and Shaan Khurshid. ✉email: kenney.ng@us.ibm.com

SBP	Systolic blood pressure
SD	Standard deviation
SHR	Standardized hazard ratio
T2DM	Type 2 diabetes mellitus
TC	Total cholesterol
TIA	Transient ischemic attack
UKBB	United Kingdom Biobank

Variability in the accuracy of models used to classify cardiovascular disease (CVD) risk has frequently been reported[1,2], with findings highlighting that performance appears to vary on the basis of sex[3], race (in the US[4–6] and out of the US[7–9]), and the presence of specific clinical factors[10,11]. With the continued growth of large collections of electronic health records (EHRs) accessible for research purposes, it is now possible to more thoroughly explore and better understand the performance heterogeneity of risk estimators, including within more refined subgroups.

CVD risk models are commonly used to prioritize individuals for preventive counseling (e.g., weight loss, alcohol cessation) and therapies (e.g., cholesterol-lowering medication). For atherosclerotic CVD (ASCVD), risk estimation using the Pooled Cohort Equations (PCE) is recommended by U.S. guidelines for determining whether individuals without established ASCVD should be considered for cholesterol-lowering therapy[12]. For atrial fibrillation (AF), in which the presence of arrhythmia is associated with an increased risk of stroke and heart failure (HF), risk estimation may also prioritize individuals for screening to detect asymptomatic disease[13,14]. The Cohorts for Heart and Aging Research in Genomic Epidemiology AF (CHARGE-AF) score[15,16] has consistently demonstrated good predictive performance for incident AF risk across multiple community cohorts[17,18] and EHR-based repositories[19].

Leveraging three large and distinct datasets, one from a prospective cohort and two from electronic health records, in total covering millions of individuals, we aimed to quantify the robustness of established models used to predict risk for AF and ASCVD. Specifically, we deployed the CHARGE-AF and PCE scores within subpopulations defined by clinically relevant strata (e.g., age, sex, and presence of relevant diseases at baseline), and quantified model performance, including discrimination, calibration, and fairness metrics, assessing for important and consistent patterns of heterogeneity[20].

Methods

Data sources. A high-level summary of our methodology is illustrated in Supplementary Fig. 1. We analyzed 3 independent data sources: the Explorys Dataset, Mass General Brigham (MGB), and the UK Biobank (UKBB).

The Explorys Dataset is comprised of the healthcare data of over 21 million individuals, pooled from different healthcare systems with distinct EHRs that have been previously used for medical research[19,21,22]. Data were statistically de-identified[23], standardized, normalized using common ontologies, and made searchable after being uploaded to a Health Insurance Portability and Accountability Act-enabled platform. The data included EHR entries for all patients who were seen between January 1, 1999, and December 31, 2020.

MGB is a large healthcare network serving the New England region of the US. We utilized the Community Care Cohort Project[24], an EHR dataset comprising over 520,000 individuals who received longitudinal primary within the MGB system, which includes 7 academic and community hospitals with associated outpatient clinics.

The UKBB is a prospective cohort of over 500,000 participants enrolled during 2006–2010[25]. Briefly, approximately 9.2 million individuals aged 40–69 years living within 25 miles of 22 assessment centers in the UK were invited, and 5.4% participated in the baseline assessment. Questionnaires and physical measures were collected at recruitment, and all participants are followed for outcomes through linkage to national health-related datasets provided by the Health & Social Care Information Centre, the Patient Episode Database for Wales, and by Scottish Morbidity Records[26]. We confirm that all methods were performed in accordance with the relevant guidelines and regulations.

Cohort construction. To ensure adequate data ascertainment and follow-up, we included individuals in Explorys with at least two outpatient encounters greater than or equal to 2 years apart[27]. Individuals in the MGB dataset had at least one pair of primary care office visits 1–3 years apart. We included all individuals who enrolled in the UKBB study, excluding those who subsequently withdrew consent.

In Explorys, the start of follow-up was defined as the first encounter following the second qualifying outpatient encounter. In MGB, the start of follow-up was defined as the second office visit of the earliest qualifying pair. In UKBB, the start of follow-up was the initial assessment visit. In each dataset, baseline variables were defined at or before the start of follow-up. Individuals with missing data for AF risk estimation at baseline were excluded. We refer to the AF analysis sets as the "AF Subsets". We defined the ASCVD analysis set analogously, with the exclusion of individuals with missing data needed to calculate the PCE score ("ASCVD Subsets"). Full details of the cohort construction for the 3 datasets are shown in Supplementary Tables I–VI.

Clinical factors. Age, sex, race, and smoking status were defined using EHR fields in Explorys and MGB and were self-reported at the initial assessment visit in UKBB. Height, weight, blood pressure, total cholesterol, and high-density lipoprotein cholesterol values were similarly extracted from the EHR in MGB and Explorys and measured at the baseline assessment in UKBB[19,28]. For patients with multiple eligible values in the baseline period, only the most recent was used. Smoking status was classified as present or absent, and race was classified as White or Black. Since dedicated PCE models are available only for White and Black individuals, as performed

previously[29] the models developed for Black individuals were utilized for individuals identifying as Black, while the models developed for White individuals were utilized for individuals of all other races. The presence of clinical comorbidities was ascertained using diagnostic (International Classification of Diseases-9th [ICD-9] and -10th [ICD-10] revisions) and procedural (Current Procedural Terminology, CPT) codes, either extracted from the EHR (Explorys and MGB), or from linked national health record data (UKBB). All covariates were used in accordance with the CHARGE-AF and PCE definitions[12,16,30]. Clinical factor definitions for all outcomes and covariates appear in Supplementary Table VII.

Follow-up and outcome definitions. The primary outcomes were the 5-year incident AF (for the AF Subsets), and the 10-year incident ASCVD (for the ASCVD Subsets). In the EHR samples, incident AF was defined using a previously validated EHR-based AF ascertainment algorithm (positive predictive value 92%), with the exception that electrocardiographic criteria were not used in Explorys given absence of electrocardiogram reports[31]. In the UKBB, AF was defined using a previously published set of self-reported data and diagnostic and procedural codes, which had been previously validated in an external dataset with a positive predictive value of 92%[32]. Incident ASCVD was defined as a composite of myocardial infarction (MI) and stroke, each defined using diagnosis codes[33]. The codes used to define ASCVD in UKBB and Explorys have been previously published[19,32], and those used in MGB have been previously validated with positive predictive value of ≥ 85%[27]. Outcome definitions are shown in Supplementary Table VII.

All models were censored at last follow-up or the end of the relevant prediction window (i.e., 5 years for CHARGE-AF and 10 years for the PCE). Last follow-up was defined as the last office visit or hospital encounter in Explorys, last EHR encounter in MGB (or administrative censoring date of August 31, 2019), and date of last available linked hospital data in UKBB. Since date of death is known in UKBB and MGB, follow-up was also censored at death in these analyses. However, since the precise date of death was not available in Explorys, we did not attempt to censor death (i.e., death was presumed to occur after the last office visit or hospital encounter).

Subgroup types. Per the original design of the PCE, we assessed the 4 sex- and race-specific models within their respective populations (Black women, Black men, White women, White men). All populations were further stratified into 10-year age ranges. These age-based analyses included 6 age strata for CHARGE-AF (45–54, 55–64, 65–74, 75–84, 85–90, and all) and 5 age strata for PCE (40–49, 50–59, 60–69, 70–79, and all). In the AF analyses, we evaluated the following additional subgroups: females, males, Black race, White race, prevalent HF, and prevalent stroke. In the PCE analyses, we also evaluated prevalent HF.

Quantification of model performance. We computed incidence rates for each outcome, reported per 1000 patient years (1 K PY). For each risk score and subgroup, we assessed the association between the risk score and its respective outcome using Cox proportional hazards regression, with 5-year AF as the outcome of interest for CHARGE-AF and 10-year ASCVD as the outcome of interest for PCE. Since the CHARGE-AF and PCE models did not account for death as a competing risk, date of death is not available in Explorys, and the proportion of individuals who died prior to the end of follow-up was low in both UKBB (AF 1.6%, PCE 3.1%) and MGB (AF 0.3%, PCE 0.4%), we did not model the competing risk of death. Hazard ratios were scaled by the within-sample standard deviation (SD) of the linear predictor of each score for comparability (Standardized Hazard Ratio [SHR]). Therefore, the SHR reflects the relative increase in event hazard observed with a 1-SD increase in the respective linear predictor. We also assessed the discrimination of each score by calculating Harrell's concordance index. We compared calibration slopes, defined as the beta coefficient of a univariable Cox proportional hazards model with the prediction target as the outcome and the linear predictor of the respective risk score as the sole covariate, where an optimally calibrated slope has a value of one[34]. To calculate 95% confidence intervals, we applied bootstrap resampling with 100 replicates.

For the purposes of identifying subgroups in which performance was particularly suboptimal, we utilized a concordance index of < 0.7. For calibration, in the absence of a consensus definition of a poor calibration slope, we utilized arbitrary calibration slope thresholds of < 0.7 (general tendency to overestimate) or > 1.3 (general tendency to underestimate) to define suboptimal calibration.

To assess performance heterogeneity beyond traditional model metrics, we calculated fairness measures, including statistical parity difference, true positive rate difference, and true negative rate difference[35]. Such measures assess fairness within the context of a protected attribute (e.g., sex, race). Statistical parity difference represents differences in the predicted risk according to the score. True positive and negative rates represent differences in sensitivity and specificity. These analyses focused on subgroups most likely to be affected by potential unfairness, including age, sex (female and male) and race (Black and White). A score is considered potentially unfair if it exhibits unexplained performance variation across different subpopulations. Fairness measures may be independent of traditional model metrics for accuracy (e.g., a score may provide very good discrimination within a subpopulation but could still be unfair).

For these analyses, the CHARGE-AF and PCE scores were converted to event probabilities using their published equations[12,15]. Where fairness metrics required application of binary risk cutoffs (i.e., true positive rate difference and false positive rate difference), we defined high AF risk as estimated 5-year AF risk ≥ 5.0% using CHARGE-AF[19,36] and high ASCVD risk as estimated 10-year ASCVD risk ≥ 7.5%[1,3,4,30].

All analyses were performed using R version 3.6, including the "survival," "rms," "data.table," and "prodlim" packages[37].

	Incident AF (5 years)			Incident ASCVD (10 years)		
	Explorys (N = 4,750,660)	UKBB (N = 445,329)	MGB (N = 174,644)	Explorys (N = 3,656,680)	UKBB (N = 408,154)	MGB (N = 198,184)
N events	196,252	7404	7877	346,159	10,906	10,201
Median follow-up, years (Q1, Q3)	3.6 (1.6, 5.0)	5.0 (5.0, 5.0)	5.0 (2.3, 5.0)	3.8 (1.8, 6.6)	8.9 (8.2, 9.7)	6.8 (2.6, 10.0)
Characteristics	**% or mean (SD)**					
Female (%)	56.7	55.0	60.9	55.9	54.8	58.8
Age (years)	62.6 (10.8)	58.4 (7.0)	60.9 (10.0)	59.0 (10.7)	56.9 (8.1)	57.0 (10.3)
White race (%)	84.2	94.7	79.6	87.4	98.4	78.1
Smoking (%)	17.3	10.7	8.0	18.7	10.4	7.4
SBP (mmHg)	131 (18)	139 (19)	128 (17)	129 (17)	139 (20)	126 (17)
DBP (mmHg)	77 (11)	83 (10)	76 (10)	DBP, Height, and Weight were not necessary to calculate PCE scores		
Height (cm)	168.5 (10.9)	168.2 (9.2)	166.6 (10.4)			
Weight (kg)	86.1 (22.1)	77.9 (15.8)	79.4 (19.5)			
HDL (US: mg/dL; UK: mmol/L)	HDL and TC were not necessary to calculate CHARGE-AF scores			51 (17)	1.46 (0.4)	57 (18)
TC (US: mg/dL; UK: mmol/L)				189 (42)	5.7 (1.1)	195 (39)
Hypertensive therapy (%)	50.1	30.5	44.8	52.8	27.9	39.3
Diabetes (%)	21.3	2.5	16.0	21.4	5.0	14.8
Heart failure (%)	3.7	0.4	1.9	3.5	0.3	1.6

Table 1. Baseline characteristics.

Results

A summary of baseline characteristics for the three datasets and their associated two outcomes is shown in Table 1, including mean (SD) for continuous measurements, percentage for binary attributes, and follow-up durations. For brevity, only the PCE model with the largest sample size (female-White; n = 1,763,103) is described in the sections below; results for all four PCE models are presented in Supplementary Table VIII and Supplementary Fig. 2.

Association between age and incidence of AF and ASCVD. As shown in Fig. 1A (AF) and B (ASCVD), incidence rate increased with age in each dataset. Explorys and MGB showed similar incidence rates in each age group, whereas UKBB participants had substantially lower AF incidence. Similarly, ASCVD incidence rate increased with age, but higher in Explorys compared to MGB and the UKBB. The effect of age on ASCVD within each of the four PCE groups is shown in Supplementary Table VIII.

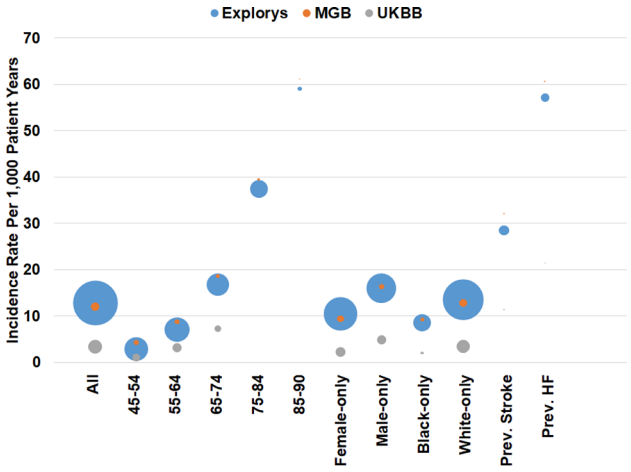
Performance heterogeneity of CHARGE-AF. We observed that a variety of subgroups were affected by limited discrimination, suboptimal calibration, or both (Supplementary Tables X and XI); for example, discrimination was lower than 0.7 and calibration slope was out of the 0.7–1.3 range among individuals aged 75 years or older (17.4% in Explorys, 10.6% in MGB). Discrimination and calibration also met criteria for poor performance among patients with prevalent HF (3.7% in Explorys, 1.9% in MGB).

Figure 2 summarizes performance measures for the CHARGE-AF score. Discrimination consistently decreased with increased age (Fig. 2A); for example, discrimination declined with increasing age from concordance index of 0.721 [95% CI 0.716–0.726] for the youngest (45–54 years) subgroup to 0.566 [0.556–0.577], for the oldest (85–90 years) subgroup in Explorys. Discrimination was higher for females than for males, consistent with prior findings[1,16,19,36], whereas differences across White versus Black race were minor. Discrimination was substantially lower among individuals with prevalent HF and stroke.

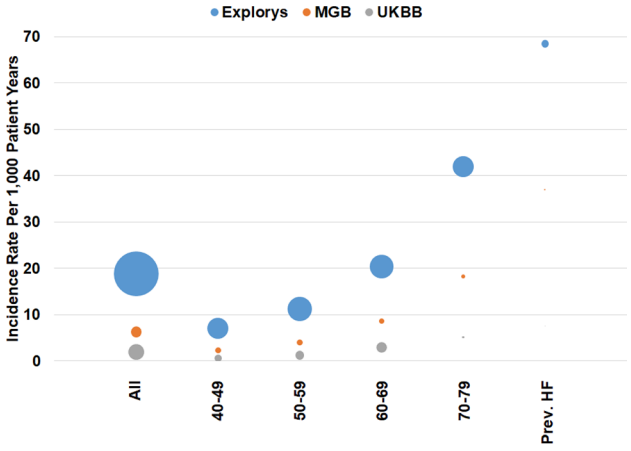
We also observed miscalibration within subgroups of age. For all 3 datasets calibration slopes decreased with increasing age, reflecting a general tendency toward underestimation at younger ages and overestimation at older ages (Fig. 2B); for example, in Explorys, values declined from 1.222 [95% CI 1.198–1.246] for the youngest (45–54 years) subgroup to 0.422 [0.371–0.474] for the oldest (85–90 years) subgroup.

The strength of association between the CHARGE-AF score and incident AF (as measured using SHRs) decreased with older age (Fig. 2C); for example, SHR declined from 3.395 [95% CI 3.315–3.477] for the youngest (45–54 years) subgroup to 1.526 [1.449–1.606] for the oldest (85–90 years) subgroup in Explorys. Within strata defined by sex and race, SHRs were highest in the UKBB, followed by MGB and Explorys. SHRs were substantially lower among individuals with prevalent HF and stroke.

Unfair behaviors for CHARGE-AF. As shown in Fig. 3A, even though sex is not included in CHARGE-AF, risk estimates using the CHARGE-AF model were much lower for females than for males, with regard to the population as a whole and particularly in the age groups 65–74 and 75–84; for example, the 65–74 years subgroup had a statistical parity difference of − 0.331 [95% CI − 0.333 to − 0.329] in Explorys. As shown in Fig. 3B, consistent across each dataset, sensitivity was lower for females, particularly in intermediate age groups (65–74



(a) AF. Zoom-in to better view details for the prevalent stroke and HF subgroups. Note that data for patients 75 or older was not available in the UKBB.



(b) ASCVD (female-White). Zoom-in to better view details for the prevalent HF subgroup.

Figure 1. Incidence rates per 1 K PY and population sizes. All population and subpopulation sizes and exact incidence rates are provided in Supplementary Table IX.

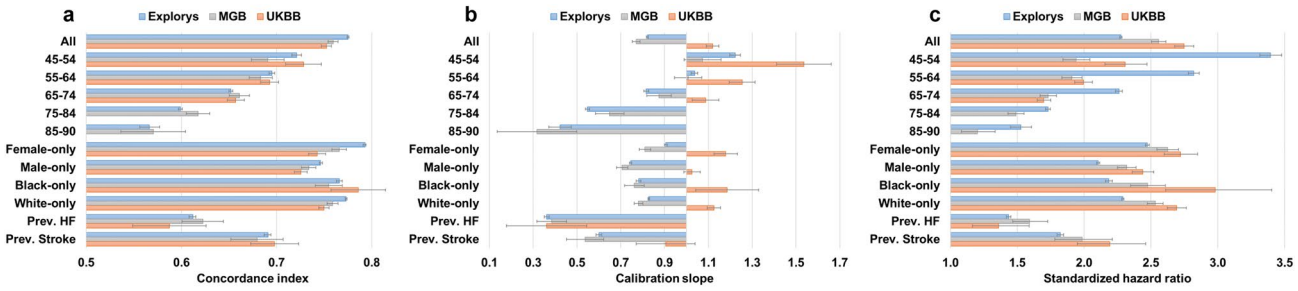


Figure 2. Performance measures for CHARGE-AF. Prev. = Prevalence; HF = Heart failure.

and 75–84); for example, the 65–74 years subgroup had a sensitivity difference of − 0.311 [95% CI − 0.319 to − 0.304] in Explorys. As shown in Fig. 3C, specificity was higher for females in intermediate age groups (65–74 and 75–84); for example, the 65–74 years subgroup had a specificity difference of 0.328 [95% CI 0.326–0.330] in Explorys.

Similar to the unfairness of pattens for sex, unfairness for race was notable in intermediate age groups (65–74 and 75–84). As shown in Fig. 3D, risk estimates using the CHARGE-AF model were much lower for Black individuals than for White individuals, as expected since White race is a risk enhancing factor in the CHARGE-AF model; for example, the 75–84 years subgroup had statistical parity difference of − 0.228 [95% CI − 0.232 to − 0.225] in Explorys. Likely as a result of systematically lower predicted risk estimates, CHARGE-AF exhibited

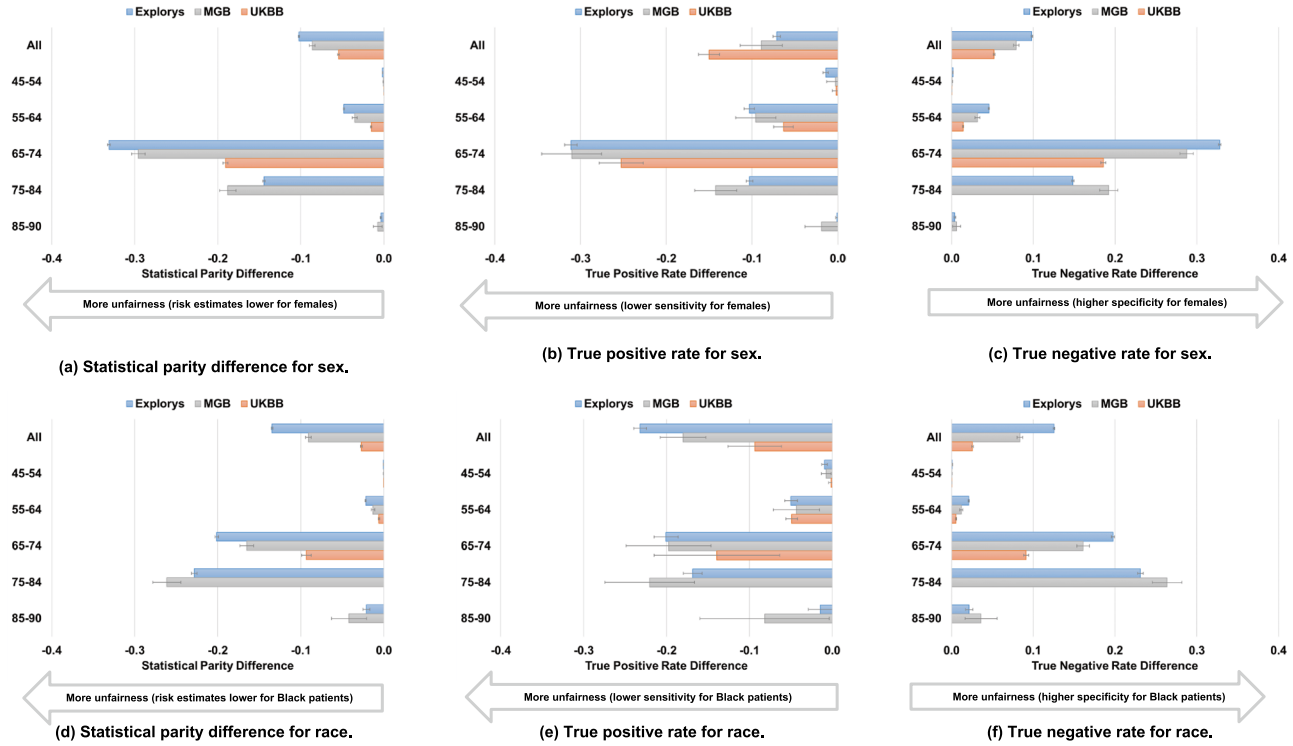


Figure 3. Fairness analysis for CHARGE-AF. Note that data was not available in the UKBB for the 75–84 and 85–90 age subpopulations.

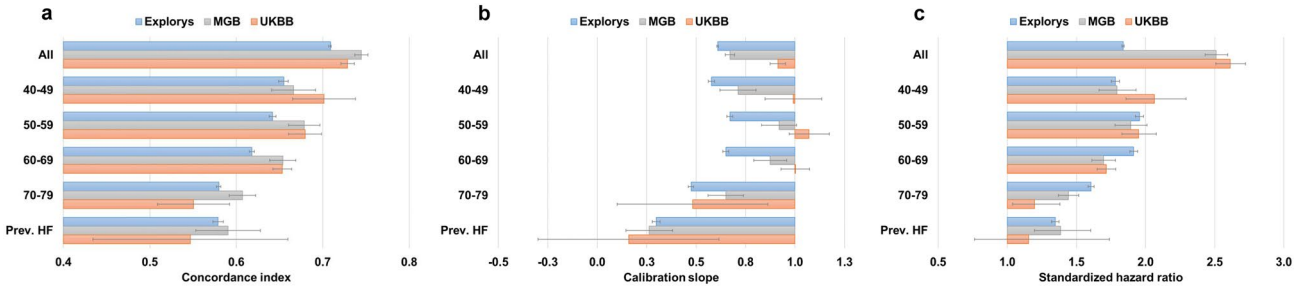


Figure 4. Performance measures for PCE (Female-White). Prev. = Prevalence; HF = Heart failure. Refer to Supplementary Table VIII for additional PCE models.

lower sensitivity (Fig. 3E) and greater specificity (Fig. 3F) among Black individuals; as an example, sensitivity difference was −0.168 [95% CI −0.180 to −0.157], and specificity difference was 0.231 [0.228–0.235] for the 75–84 years subgroup in Explorys. For both sex and race, behavior indicating unfairness was similar between Explorys and MGB but less prominent in the UKBB.

Performance heterogeneity of PCE. As with CHARGE-AF, we observed that a variety of subgroups were affected by limited discrimination, limited calibration, or both (Supplementary Tables XII and XIII). Only a few of the subgroups across the 3 datasets were associated with both good discrimination and calibration (e.g., female-White 40–49 in the UKBB with a percentage of 21.9% of the total patients in this subgroup).

Consistent with CHARGE-AF, discrimination using the PCE decreased with older age from a concordance index of 0.655 [95% CI 0.649–0.660] for the 40–49 years subgroup to 0.580 [0.577–0.582] for the 70–79 years subgroup in Explorys (Fig. 4A). This behavior was consistent across all 3 datasets. Discrimination among individuals with prevalent HF was similar to the overall 70–79 years subgroup.

We also observed suboptimal calibration using the PCE within subgroups of age, with consistently lower calibration slopes in the youngest and oldest groups, indicating an overall tendency to overestimate risk at extremes of age (Fig. 4B); for example, in Explorys, values were the lowest for the 40–49 years subgroup with a slope of 0.577 [95% CI 0.561–0.594], and 0.474 [0.460–0.487] for the 70–79 years subgroup, in comparison to values above 0.7 for the intermediate age subgroups. Similar to CHARGE-AF, calibration performance was limited among individuals with prevalent HF, again with a general tendency to overestimate risk.

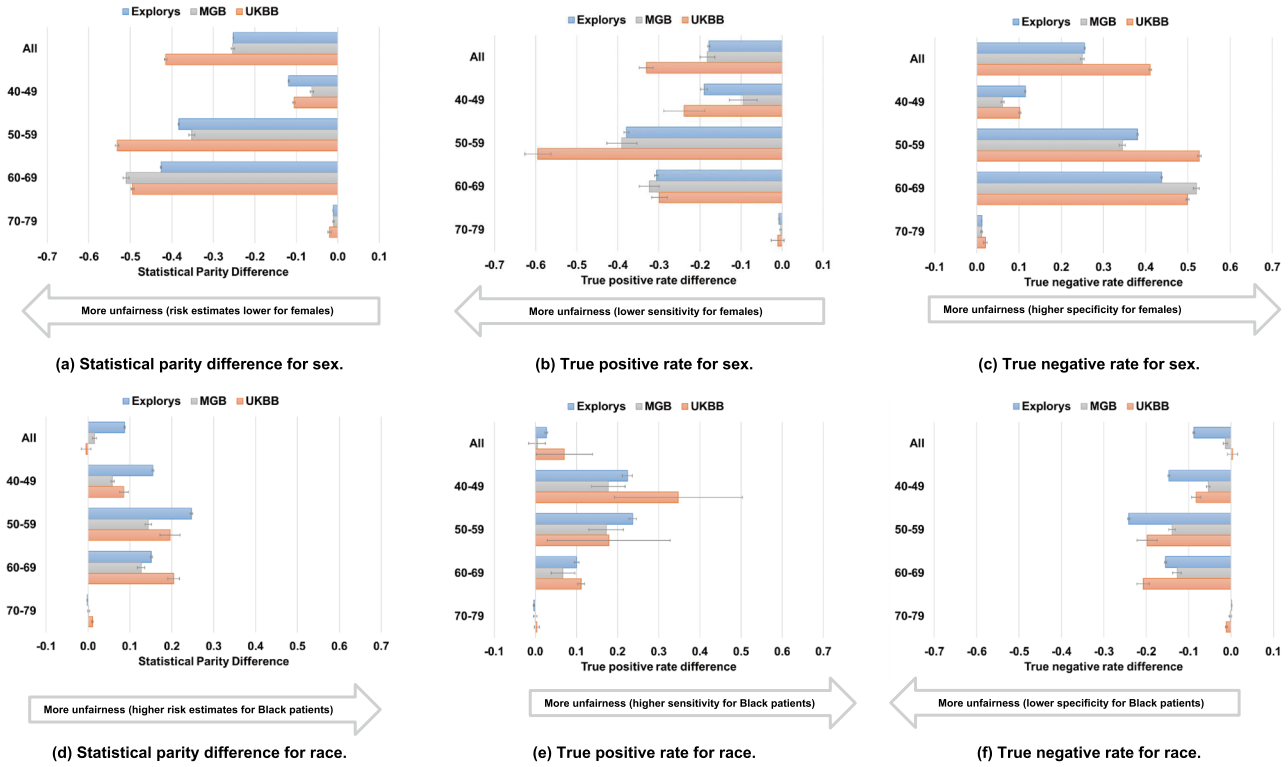


Figure 5. Fairness analysis for PCE.

The strength of association between the PCE score on incident ASCVD (as measured using SHRs) was highest in intermediate age groups (50–59 and 60–69) compared to the younger (40–49) and older (70–79) age groups (Fig. 4C); for example, highest SHR was 1.956 [95% CI 1.927–1.985] for the 50–59 subgroup and 1.606 [1.585–1.628] for the 70–79 subgroup, in Explorys.

Unfair behaviors for PCE. As shown in Fig. 5A, risk estimates using the PCE were much lower for females than for males in the overall population as well as within the intermediate age groups (50–59 and 60–69); for example, in Explorys, the 60–69 years subgroup had a statistical parity difference of − 0.426 [95% CI − 0.427 to − 0.424]. As shown in Fig. 5B, across all datasets, sensitivity was lower for females, especially in intermediate age groups (50–59 and 60–69); for example, the 50–59 years subgroup had a sensitivity difference of − 0.379 [95% CI − 0.386 to − 0.373] in Explorys. Specificity was higher among females (Fig. 5C), especially in intermediate age groups (50–59 and 60–69); for example, the 60–69 years subgroup had a specificity difference of 0.438 [95% CI 0.436–0.439] in Explorys. Overall, patterns observed on the basis of sex using the PCE were similar to those observed using CHARGE-AF.

As shown in Fig. 5D, unlike CHARGE-AF, risk estimates using the PCE were higher in Black individuals in all datasets; this effect was especially noticeable in intermediate age groups (50–59 and 60–69); for example, statistical parity difference between the 50–59 years subgroup was the largest compared to the other subgroups in Explorys at 0.247 [95% CI 0.244–0.250]. In contrast to CHARGE-AF, greater risk estimates led to increased sensitivity among Black individuals versus White individuals (Fig. 5E); for example, sensitivity difference between the 40–49 years and 50–59 years subgroups were the largest compared to the other subgroups in Explorys at 0.224 [95% CI 0.211–0.237] and 0.237 [0.228–0.246], respectively. Differences in sensitivity on the basis of race decreased with increasing age in all 3 datasets, with very little difference observed in the oldest age group (70–79). As shown in Fig. 5F, across specific age ranges, specificity was lower for Black individuals than for White individuals; this effect was especially noticeable in intermediate age groups (50–59 and 60–69); for example, specificity difference between the 50–59 years subgroup was the greatest compared to the other subgroups in Explorys at − 0.241 [95% CI − 0.244 to − 0.239].

Discussion

We analyzed three large independent datasets including millions of individuals and identified important patterns of performance heterogeneity across clinically relevant subgroups as indicated by standard performance measures including discrimination, calibration, SHRs, and fairness metrics. Our results build on previous efforts to understand estimation of AF and ASCVD risk in several key ways. First, we assessed the scores on very large databases, allowing us to quantify performance within granular subgroups. Second, we provide results applicable to 3 resources, allowing us to assess consistency in results across independent samples. Third, we perform analyses of two distinct outcomes, which allows for identification of potential patterns of heterogeneity that may be shared across risk estimators for different conditions. Fourth, our results highlight the magnitude of important

limitations in performance affecting sizeable portions of the population, in particular patients at older ages and with prevalent conditions. Fifth, to our knowledge, our study is the first to report on fairness-related measures for the CHARGE-AF and PCE scores in relation to sex and race.

Patterns of variability were fairly consistent across the CHARGE-AF and PCE models. Importantly, we observed that discrimination and calibration were consistently worse at extremes of age, as well as for individuals with certain prevalent conditions (e.g., HF). Furthermore, we observed evidence of potentially unfair performance, with significant differences in fairness metrics for sex and race in both scores. For instance, the sensitivity difference of both scores was much lower for females than males in the intermediate-age subgroups, suggesting that current scores may miss more women at high risk for events, potentially worsening existing sex-related treatment gaps[38]. Overall, our findings underscore the importance of evaluating prognostic models across the many specific subpopulations in which risk prediction is intended, in order to better understand the accuracy and potential unfairness of the prognostic information used to drive clinical decisions at the point of care.

Our findings suggest that clinicians utilizing prognostic models should not assume that a given level of performance in the overall population will translate to similar accuracy within a subgroup of the population to which their patient belongs. Consistent with prior findings suggesting good overall performance of CHARGE-AF[17,18] and the PCE[2,10] across multiple populations, we observed moderate or greater discrimination using each score in our datasets. However, we observed that multiple standard metrics (e.g., discrimination and calibration) vary substantially within subpopulations. Specifically, we observed a consistent pattern of decreasing discrimination for higher age groups, a finding which may be attributable to less variability in event risk among older individuals. Furthermore, since assessing discrimination within a subgroup defined by a certain feature precludes classification of risk on the basis of that feature (i.e., discrimination is adjusted), stratification by variables with substantial effects on event risk will decrease discrimination. Similar to discrimination, we also observed increasing miscalibration in higher age groups, which may be related to greater average event risk. In addition to age, miscalibration related to baseline event risk may also be impacted by varying treatment patterns across different settings and over time. Ultimately, since the majority of incidents CVD occur among older individuals, more accurate models for an older population remains a critical unmet need. Future work is needed to assess whether models derived within specific subgroups of clinical importance may lead to better and more consistent model performance across important subsets of the population.

In addition to variation across standard model metrics, our findings also suggest that common prognostic models may have performance indicating unfairness across strata of sex and race. As discussed above, CHARGE-AF had lower sensitivity and greater specificity among women. A similar pattern was observed among Black individuals. Although use of the PCE also led to lower sensitivity and greater specificity among women, it demonstrated the opposite pattern (greater sensitivity and lower specificity) among Black individuals. It is notable that these differences exist despite the fact that the PCE has dedicated models specific to race and sex (i.e., there are 4 distinct equations). Since PCE model predictions were generally better calibrated among White individuals, our findings suggest that model derivation in populations having greater representation of women and Black individuals may lead to more accurate and generalizable models with less unfairness.

There are several potential strategies to mitigate the significant heterogeneity in performance we characterized and quantified in the current study. One strategy is to adjust models according to empirically observed patterns of unfairness, which has been previously proposed as a method to reduce unfairness and minimize overtreatment of healthy individuals[7,39]. Another approach is to reweight existing models[40–42] within each subgroup of the population, resulting in distinct weights for each subgroup of interest. Yet another strategy is to create new higher capacity models that include additional (e.g., socioeconomic deprivation)[7,43] or more precisely defined predictors (e.g., granular race definitions), which may offer more consistent prognostic value across subgroups. Any chosen strategy should consider both calibration and discrimination not only separately but also jointly; for example, even if a mitigation strategy could handle limited calibration performance in a certain subgroup, effects may not translate to other subgroups. Furthermore, certain strategies may result in a tradeoff in which one measure is improved (e.g., discrimination), while another is worsened (e.g., fairness-related).

Our study has several limitations. First, despite analysis of three large datasets, the majority of individuals included were White, limiting the precision of subgroup-based estimates in Black individuals. Second, since dedicated PCE models are available only for White and Black individuals, as performed previously[29], the models for Black individuals were utilized for individuals identifying as Black, and the models for White individuals were utilized for individuals of all other races. Evidence suggests that cardiovascular risk and outcomes[5,29] may differ importantly on account of more granular classification of race and ethnicity, and therefore we acknowledge that our race classification may have contributed to observed heterogeneity in PCE performance. We submit that future work is warranted to develop more accurate methods of risk ASCVD risk stratification in these populations. Third, we were unable to assess the effects of socioeconomic deprivation[44–46] given the lack of available data in Explorys and MGB. Fourth, given that the CHARGE-AF and PCE scores did not model death as a competing risk, and death data are not available in the Explorys, we did not adjust for the competing risk of death (note that death rates within the windows of interest in the UKBB and MGB datasets were low). Fifth, as with any EHR-based study, misclassification of exposures and outcomes is possible. Additionally, cause of death data is available only in UKBB, and therefore fatal ASCVD events not resulting in hospitalization may have been missed in the EHR samples. To mitigate misclassification, we utilized previously published disease definitions and constructed our EHR samples to include individuals receiving longitudinal ambulatory care. Furthermore, predictive utility was similar to expectations for both scores in all 3 datasets compared to values observed from prior prospective cohort studies[12,15]. Sixth, we have not applied recently proposed fairness metrics that assess individual fairness (rather than assessment at the population level)[47,48]. Sixth, although our findings provide important evidence of performance heterogeneity and potential unfairness in commonly used risk estimators, we did not explore mitigation methods.

In summary, we evaluated the CHARGE-AF and the PCE scores in three independent datasets totaling over 5 million individuals, identifying important performance heterogeneity and unfairness. The patterns we observed were consistent, including worse discrimination of risk among older individuals and substantial miscalibration at extremes of age. We also observed that use of common score thresholds may lead to unfairness on the basis of sex and race, which may worsen existing treatment gaps. Overall, users of current clinical risk stratification methods should exercise caution when interpreting risk estimates obtained in certain subgroups (e.g., extremes of age), and there is a critical need to develop more robust risk estimators that display more consistent accuracy and fairness.

Data availability

The institutional review boards of Mass General Brigham (MGB) and IBM approved this study and its methods, including the EHR cohort assembly using the Explorys Dataset, data extraction, and analyses. MGB data contains potentially identifying information and may not be shared publicly. Explorys data can be made available through a commercial license (for details see: https://www.ibm.com/downloads/cas/4P0QB9JN). We are indebted to the UKBB and its participants who provided data for this analysis (UKBB Applications #7089 and #50658). All UKBB participants provided written informed consent. The UK Biobank was approved by the UK Biobank Research Ethics Committee (reference# 11/NW/0382). Source data are provided with this paper.

Received: 18 March 2022; Accepted: 12 July 2022
Published online: 22 July 2022

References

1. Damen, J. A. *et al.* Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: A systematic review and meta-analysis. *BMC Med.* **17**(1), 109. https://doi.org/10.1186/s12916-019-1340-7 (2019) (**PMID: 31189462; PMCID: PMC6563379**).
2. Muntner, P. *et al.* Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA* **311**, 1406–1415 (2014).
3. Kavousi, M. *et al.* Comparison of application of the ACC/AHA guidelines, Adult Treatment Panel III guidelines, and European Society of Cardiology guidelines for cardiovascular disease prevention in a European cohort. *JAMA* **311**(14), 1416–1423 (2014).
4. DeFilippis, A. P. *et al.* An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Ann. Intern. Med.* **162**(4), 266–275. https://doi.org/10.7326/M14-1281 (2015).
5. Rana, J. S. *et al.* Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic population. *J. Am. Coll. Cardiol.* **67**, 2118–2130 (2016).
6. DeFilippis, A. P. *et al.* Risk score overestimation: The impact of individual cardiovascular risk factors and preventive therapies on the performance of the American Heart Association-American College of Cardiology-Atherosclerotic Cardiovascular Disease risk score in a modern multi-ethnic cohort. *Eur. Heart J.* **38**, 598–608 (2017).
7. Pylypchuk, R. *et al.* Cardiovascular disease risk prediction equations in 400,000 primary care patients in New Zealand: A derivation and validation study. *Lancet* **391**, 1897–1901 (2018).
8. Lee, C. H. *et al.* Validation of the Pooled Cohort equations in a long-term cohort study of Hong Kong Chinese. *J. Clin. Lipidol.* **9**(5), 640-646.e2. https://doi.org/10.1016/j.jacl.2015.06.005 (2015) (**Epub 2015 Jun 16. PMID: 26350809**).
9. Jung, K. J. *et al.* The ACC/AHA 2013 pooled cohort equations compared to a Korean Risk Prediction Model for atherosclerotic cardiovascular disease. *Atherosclerosis* **242**(1), 367–375. https://doi.org/10.1016/j.atherosclerosis.2015.07.033 (2015) (**Epub 2015 Jul 22. PMID: 26255683**).
10. Khera, R. *et al.* Performance of the Pooled Cohort Equations to estimate atherosclerotic cardiovascular disease risk by body mass index. *JAMA Netw. Open.* **3**(10), e2023242. https://doi.org/10.1001/jamanetworkopen.2020.23242 (2020) (**Erratum in: JAMA Netw Open. 2020 Dec 1;3(12):e2030880. PMID: 33119108; PMCID: PMC7596579**).
11. Nguyen, Q. D., Odden, M. C., Peralta, C. A. & Kim, D. H. Predicting risk of atherosclerotic cardiovascular disease using Pooled Cohort Equations in older adults with frailty, multimorbidity, and competing risks. *J. Am. Heart Assoc.* **9**(18), e016003. https://doi.org/10.1161/JAHA.119.016003 (2020) (**Epub 2020 Sep 2. PMID: 32875939; PMCID: PMC7727000**).
12. Goff, D. C. Jr. *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* **129**(25 Suppl 2), S49–S73. https://doi.org/10.1161/01.cir.0000437741.48606.98 (2014) (**Epub 2013 Nov 12. Erratum in: Circulation. 2014 Jun 24;129(25 Suppl 2):S74–5. PMID: 24222018**).
13. Kemp Gudmundsdottir, K. *et al.* Stepwise mass screening for atrial fibrillation using N-terminal B-type natriuretic peptide: The STROKESTOP II study. *Europace* **22**(1), 24–32. https://doi.org/10.1093/europace/euz255 (2020) (**PMID: 31790147; PMCID: PMC6945054**).
14. Khurshid, S. *et al.* Predictive accuracy of a clinical and genetic risk model for atrial fibrillation. *Circ. Genom. Precis Med.* https://doi.org/10.1161/CIRCGEN.121.003355 (2021) (**Epub ahead of print. PMID: 34463125**).
15. Alonso, A. *et al.* Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: The CHARGE-AF consortium. *J. Am. Heart Assoc.* **2**(2), e000102. https://doi.org/10.1161/JAHA.112.000102 (2013) (**PMID: 23537808; PMCID: PMC3647274**).
16. Alonso, A., Roetker, N.S., Soliman, E.Z., Chen, L.Y., Greenland, P., Heckbert, S.R. Prediction of atrial fibrillation in a racially diverse cohort: The Multi-Ethnic Study of Atherosclerosis (MESA). *J. Am. Heart Assoc.* **5** (2016).
17. Shulman, E. *et al.* Validation of the Framingham Heart Study and CHARGE-AF risk scores for atrial fibrillation in Hispanics, African–Americans, and Non-Hispanic Whites. *Am. J. Cardiol.* **117**(1), 76–83. https://doi.org/10.1016/j.amjcard.2015.10.009 (2016) (**Epub 2015 Oct 19. PMID: 26589820**).
18. Christophersen, I. E. *et al.* A comparison of the CHARGE-AF and the CHA_2DS_2-VASc risk scores for prediction of atrial fibrillation in the Framingham Heart Study. *Am. Heart J.* **178**, 45–54 (2016).
19. Khurshid, S. *et al.* Performance of atrial fibrillation risk prediction models in over 4 million individuals. *Circ. Arrhythm. Electrophysiol.* **14**(1), e008997. https://doi.org/10.1161/CIRCEP.120.008997 (2021) (**Epub 2020 Dec 9. PMID: 33295794; PMCID: PMC7856013**).
20. FitzGerald, C. & Hurst, S. Implicit bias in healthcare professionals: A systematic review. *BMC Med. Ethics.* **18**(1), 19. https://doi.org/10.1186/s12910-017-0179-8 (2017) (**Published 2017 Mar 1**).
21. Kartoun, U. *et al.* The MELD-Plus: A generalizable prediction risk score in cirrhosis. *PLoS ONE* **12**, e0186301 (2017).
22. Dron, J. S. *et al.* Genetic predictor to identify individuals with high Lipoprotein(a) concentrations. *Circ. Genom. Precis. Med.* **14**(1), e003182. https://doi.org/10.1161/CIRCGEN.120.003182 (2021) (**Epub 2021 Feb 1. PMID: 33522245; PMCID: PMC7887018**).

23. Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington (DC): National Academies Press (US); 2015 Apr 20. Appendix B, Concepts and Methods for De-identifying Clinical Trial Data. https://www.ncbi.nlm.nih.gov/books/NBK285994/
24. Khurshid, S. *et al.* Cohort design and natural language processing to reduce bias in electronic health records research: The Community Care Cohort Project. *medRxiv.* https://doi.org/10.1101/2021.05.26.21257872 (2021).
25. Sudlow, C. *et al.* UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med* **12**, e1001779 (2015).
26. UK Biobank. Integrating Electronic Health Records into the UK Biobank Resource. http://biobank.ctsu.ox.ac.uk/showcase/showcase/docs/DataLinkageProcess.pdf (2014).
27. Hulme, O. L. *et al.* Development and validation of a prediction model for atrial fibrillation using electronic health records. *JACC Clin. Electrophysiol.* **5**, 1331–1341 (2019).
28. Patel, A. P., Wang, M., Kartoun, U., Ng, K. & Khera, A. V. Quantifying and understanding the higher risk of atherosclerotic cardiovascular disease among South Asian individuals: Results from the UK Biobank prospective cohort study. *Circulation* **144**(6), 410–422. https://doi.org/10.1161/CIRCULATIONAHA.120.052430 (2021) (**Epub 2021 Jul 12. PMID: 34247495; PMCID: PMC8355171**).
29. Rodriguez, F. *et al.* Atherosclerotic cardiovascular disease risk prediction in disaggregated Asian and Hispanic subgroups using electronic health records. *J Am Heart Assoc.* **8**(14), e011874. https://doi.org/10.1161/JAHA.118.011874 (2019) (**Epub 2019 Jul 11. PMID: 31291803; PMCID: PMC6662141**).
30. Stone, N. J. *et al.* 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* **63**(25 Pt B), 2889–2934. https://doi.org/10.1016/j.jacc.2013.11.002 (2014) (**Epub 2013 Nov 12. Erratum in: J Am Coll Cardiol. 2014 Jul 1;63(25 Pt B):3024–25. Erratum in: J Am Coll Cardiol. 2015 Dec 22;66(24):2812. PMID: 24239923**).
31. Khurshid, S., Keaney, J., Ellinor, P. T. & Lubitz, S. A. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. *Am. J. Cardiol.* **117**, 221–225 (2016).
32. Khurshid, S. *et al.* Frequency of cardiac rhythm abnormalities in a half million adults. *Circ. Arrhythm. Electrophysiol.* **11**(7), e006273. https://doi.org/10.1161/CIRCEP.118.006273 (2018) (**PMID: 29954742; PMCID: PMC6051725**).
33. Wang, E. Y. *et al.* Initial precipitants and recurrence of atrial fibrillation. *Circ. Arrhythm. Electrophysiol.* **13**(3), e007716. https://doi.org/10.1161/CIRCEP.119.007716 (2020) (**Epub 2020 Feb 12. PMID: 32078361; PMCID: PMC7141776**).
34. Cox, D. R. Two further applications of a model for binary regression. *Biometrika* **45**, 562–565 (1958).
35. Bellamy, R., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K., Richards, J.T., Saha, D., Sattigeri, P., Singh, M., Varshney, K., Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. ArXiv, abs/1810.01943. 2018.
36. Himmelreich, J. C. L. *et al.* CHARGE-AF in a national routine primary care electronic health records database in the Netherlands: Validation for 5-year risk of atrial fibrillation and implications for patient selection in atrial fibrillation screening. *Open Heart.* **8**(1), e001459. https://doi.org/10.1136/openhrt-2020-001459 (2021) (**PMID: 33462107; PMCID PMC7816907**).
37. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. URL https://www.R-project.org/.
38. Mehran, R., Vogel, B., Ortega, R., Cooney, R. & Horton, R. The Lancet Commission on women and cardiovascular disease: Time for a shift in women's health. *Lancet* **393**(10175), 967–968. https://doi.org/10.1016/S0140-6736(19)30315-0 (2019) (**Epub 2019 Feb 11. PMID: 30765122**).
39. Pennells, L. *et al.* Equalization of four cardiovascular risk algorithms after systematic recalibration: Individual-participant meta-analysis of 86 prospective studies. *Eur. Heart J.* **40**(7), 621–631. https://doi.org/10.1093/eurheartj/ehy653 (2019) (**PMID: 30476079; PMCID: PMC6374687**).
40. Park, Y. *et al.* Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw. Open.* **4**(4), e213909. https://doi.org/10.1001/jamanetworkopen.2021.3909 (2021).
41. Calders, T., Kamiran, F., Pechenizkiy, M. Building classifiers with independency constraints. in *ICDM Workshops—IEEE International Conference on Data Mining*. 2009:13–8. August 6–9, 2009; Miami, Florida.
42. Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20**(1), 25–46. https://doi.org/10.1093/pan/mpr025 (2012).
43. Dalton, J. E. *et al.* Accuracy of cardiovascular risk prediction varies by neighborhood socioeconomic position: A retrospective cohort study. *Ann. Intern. Med.* **167**(7), 456–464. https://doi.org/10.7326/M16-2543 (2017) (**Epub 2017 Aug 29. PMID: 28847012; PMCID: PMC6435027**).
44. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**(12), 866–872. https://doi.org/10.7326/M18-1990 (2018) (**Epub 2018 Dec 4. PMID: 30508424; PMCID: PMC6594166**).
45. Townsend, P., Phillimore, P. & Beattie, A. *Health and Deprivation: Inequality and The North* (Routledge, 1988).
46. Foster, H. M. E. *et al.* The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy lifestyle factors and health outcomes: A prospective analysis of the UK Biobank cohort. *Lancet Public Health.* **3**(12), e576–e585. https://doi.org/10.1016/S2468-2667(18)30200-7 (2018) (**Epub 2018 Nov 20 PMID: 30467019**).
47. Yurochkin, M., Bowery, A., Sun, Y. Training individually fair ML models with sensitive subspace robustness. ICLR 2020.
48. Maity, S., Xue, S., Yurochkin, M., Sun, Y. Statistical inference for individual fairness. ICLR 2021.

Author contributions

U.K., S.K., S.A.L., K.N.: conception and design of the study; interpreting data; writing the manuscript. B.C.K., A.P.P., P.B., A.P., A.V.K., P.T.E.: interpreting data; writing the manuscript.

Funding

This work was supported by a collaboration between IBM and the Broad Institute and by NIH Grants 1K08HG010155 (Khera), R01HL139731 (Lubitz), 2R01HL092577 (Ellinor), and K24HL105780 (Ellinor), T32HL007208 (Khurshid, Patel); American Heart Association (Dallas, Texas) 18SFRN34250007 (Lubitz); a Doris Duke Charitable Foundation Clinical Scientist Development Award 2014105 (Lubitz); and by the Fondation Leducq 14CVD01 (Ellinor).

Competing interests

Dr. Lubitz receives sponsored research support from Bristol-Myers Squibb/Pfizer, Bayer AG, Biotronik, and Boehringer Ingelheim, and has consulted for Bristol-Myers Squibb and Bayer AG. Dr. Ellinor receives sponsored research support from Bayer AG and IBM, and he has consulted for Bayer AG, Novartis, and MyoKardia, and Quest Diagnostic. Dr. Philippakis is also employed as a Venture Partner at GV and consulted for Novartis; and

has received funding from Intel, Verily and MSFT. Dr. Batra serves as a consultant for Novartis. Dr. Khera has served as a scientific advisor to Sanofi, Amgen, Maze Therapeutics, Navitor Pharmaceuticals, Sarepta Therapeutics, Novartis, Verve Therapeutics, Silence Therapeutics, Veritas International, Color Health, Third Rock Ventures, and Columbia University (NIH); received speaking fees from Illumina, MedGenome, Amgen, and the Novartis Institute for Biomedical Research; and received a sponsored research agreement from the Novartis Institute for Biomedical Research. Drs. Kartoun, Kwon, and Ng are employees of IBM. The remaining authors have no disclosures.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-16615-3.

Correspondence and requests for materials should be addressed to K.N.

Reprints and permissions information is available at www.nature.com/reprints.

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