# Characterizing and Quantifying Performance Heterogeneity in Cardiovascular Risk Prediction Models — A Step Towards Improved Disease Risk Assessment

**Uri Kartoun** ( uri.kartoun@ibm.com )
IBM Research

**Shaan Khurshid**
Broad Institute of the Massachusetts Institute of Technology and Harvard University

**Bum Chul Kwon**
IBM Research https://orcid.org/0000-0002-9391-6274

**Aniruddh Patel**
Broad Institute of the Massachusetts Institute of Technology and Harvard University

**Puneet Batra**
Broad Institute https://orcid.org/0000-0001-6822-0593

**Anthony Philippakis**
Broad Institute of MIT and Harvard

**Amit Khera**
Massachusetts General Hospital

**Patrick Ellinor**
The Broad Institute of MIT and Harvard https://orcid.org/0000-0002-2067-0533

**Steve Lubitz**
MGH https://orcid.org/0000-0002-9599-4866

**Kenney Ng**
IBM Research https://orcid.org/0000-0003-0792-070X

**Article**

**Keywords:** Atrial fibrillation, Atherosclerotic cardiovascular disease, Risk prediction, Discrimination, Calibration, Fairness

**Posted Date:** October 8th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-936366/v1

# Characterizing and Quantifying Performance Heterogeneity in Cardiovascular Risk Prediction Models — A Step Towards Improved Disease Risk Assessment

**Uri Kartoun PhD[1†], Shaan Khurshid MD MPH[2,3†], Bum Chul Kwon PhD[1], Aniruddh P Patel[2,4], Puneet Batra PhD[5], Anthony Philippakis MD PhD[2], Amit V Khera MD MSc[2,4], Patrick T Ellinor MD PhD[2,3], Steven A Lubitz MD MPH[2,3], Kenney Ng PhD[1*]**

**1. Center for Computational Health, IBM Research, Cambridge, Massachusetts, USA; 2. Cardiovascular Disease Initiative, Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA; 3. Demoulas Center for Cardiac Arrhythmias, Massachusetts General Hospital, Boston, Massachusetts, USA; 4. Division of Cardiology, Massachusetts General Hospital, Boston, Massachusetts, USA; 5. Data Sciences Platform, Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA.**

† Authors contributed equally.

* Corresponding Author:

Kenney Ng

IBM Research, 314 Main St., Cambridge MA, 02142, USA

+1-617-504-4665

kenney.ng@us.ibm.com

**Short Title:** Heterogeneity in cardiovascular risk prediction

**Word Count**: 4,051

## Abstract

Prediction models are commonly used to estimate risk for cardiovascular diseases; however, performance may vary substantially across relevant subgroups of the population. Here we investigated the variability of performance and fairness across a variety of subgroups for risk prediction of two common diseases, atherosclerotic cardiovascular disease (ASCVD) and atrial fibrillation (AF). We calculated the Cohorts for Heart and Aging in Genomic Epidemiology Atrial Fibrillation (CHARGE-AF) for AF and the Pooled Cohort Equations (PCE) score for ASCVD in three large data sets: 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 of established cardiovascular risk scores across subpopulations defined by age, sex, and presence of preexisting disease. For example, in CHARGE-AF, discrimination declined with increasing age, with concordance index of 0.72 [95% CI, 0.72–0.73] for the youngest (45–54y) subgroup to 0.57 [0.56–0.58], for the oldest (85–90y) subgroup in Explorys. The statistical parity difference (i.e., likelihood of being classified as high risk) was considerable between males and females within the 65–74y subgroup with a value of -0.33 [95% CI, -0.33–-0.33]. We observed also that large segments of the population suffered from both decreased discrimination (i.e., <0.7) and poor calibration (i.e., calibration slope outside of 0.7–1.3); for example, all individuals 75 or older in Explorys (17.4%). Our findings highlight the need to characterize and quantify how clinical risk models behave and perform within specific subpopulations so they can be used appropriately to facilitate more accurate and equitable assessment of disease risk.

**Keywords**: Atrial fibrillation, Atherosclerotic cardiovascular disease, Risk prediction, Discrimination, Calibration, Fairness

## Abbreviations

1K PY – Per 1,000 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

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 standard performance metrics to assess cardiovascular disease (CVD) risk has frequently been reported[6,11] with findings highlighting that performance varies depending on the type of the groups, for example, sex groups[1], racial groups (in the US[2,3,4] and out of the US[5,8,9]), and groups with specific clinical factors[7,10]. With the continued growth of large collections of electronic health records accessible for research purposes it is now possible to more thoroughly explore and better understand performance heterogeneity, considering 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[13,44]. The Cohorts for Heart and Aging Research in Genomic Epidemiology AF (CHARGE-AF) score[14,15] has consistently demonstrated good predictive performance for incident AF risk across multiple community cohorts[16,17] and electronic health record (EHR)-based repositories[18].

Leveraging three large and distinct datasets, one from a prospective cohort and two from electronic health records, covering millions of individuals, we aimed to robustly characterize CVD risk score performance heterogeneity across multiple subpopulations defined by clinically relevant strata (e.g., age, sex, and presence of relevant diseases at baseline). Specifically, we deployed the CHARGE-AF and PCE scores within

subpopulations across each dataset and quantified model performance, including discrimination, calibration, and fairness metrics, assessing for important and consistent patterns of heterogeneity[19].

## Methods

## Data sources

A high-level summary of our methodology is illustrated in **SUPPLEMENTARY FIGURE 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[20,18,21]. Data were statistically de-identified[22], standardized, and 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[23], an EHR dataset comprising over 520,000 individuals who received care at any of the 7 academic and community hospitals in MGB.

The UKBB is a prospective cohort of over 500,000 participants enrolled during 2006–2010[24]. 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.

## Cohort construction

To ensure adequate data ascertainment and follow-up, we included in Explorys individuals with at least two outpatient encounters greater than or equal to 2 years apart[25]. 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. We excluded all enrolled individuals who decided at a later point to withdraw 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, as an enrollment-based resource, start of follow-up was the date of the initial assessment visit. In each dataset, 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 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 measured relative to baseline in all 3 datasets[18,45]. 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. Patients who indicated themselves as Black (possibly with one or more other race types) were considered Black for risk calculations, and White otherwise. 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 (UK Biobank). All covariates were used in accordance with the CHARGE-AF and PCE definitions[12,15,31]. Clinical factor definitions of all covariates appear in **SUPPLEMENTARY TABLE VII**.

## Follow-up and outcome definitions

The primary outcomes were 5-year incident AF (for the AF Subsets), and 10-year incident ASCVD (for the ASCVD Subsets). Incident AF was defined using a modified version of a previously validated EHR-based AF ascertainment algorithm (positive predictive value 92%), in which electrocardiographic criteria were not used given the absence of electrocardiogram reports in the Explorys Dataset[26]. Incident ASCVD was defined as a composite of myocardial infarction (MI) and stroke, each defined using previously published sets of diagnosis codes[27]. Outcome definitions are shown in **SUPPLEMENTARY TABLE VII**.

All models were censored at the earliest of death, 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, and date of last available linked hospital data in UKBB.

## 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 stratified further 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 1,000 patient years (1K 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. 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 c-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[28].

To assess further potential biases in performance, we calculated fairness measures, including differences in statistical parity, true positive rates, and true negative rates[29]. These analyses focused on subgroups most likely to be affected by potential bias, including age, sex (female and male) and race (Black and White). For these analyses, the CHARGE-AF and PCE scores were converted to event probabilities using their published equations[14,12]. 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 $\geq$ 5.0% using CHARGE-AF[30,18] and high ASCVD risk as estimated 10-year ASCVD risk $\geq$ 7.5%[31,1,2,6].

All analyses were performed using R version 3.6, including the “survival,” “rms,” “data.table,” and “prodlim” packages[32].

## Results

A summary of baseline characteristics for the three data sets and their associated two distinct outcomes is shown in **TABLE 1**, including mean (SD) for continuous measurements, percentage for binary attributes, and follow-up durations for each of the six scenarios (i.e., two scores applied to three distinct datasets). For brevity, only the PCE model with the largest cohort (female-White; n = 1,603,450) is described in the sections below; results for all four PCE models are presented in **SUPPLEMENTARY TABLE VIII** and **SUPPLEMENTARY FIGURE 2**.

### Association between age and incidence of AF and ASCVD

As shown in **FIGURE 1A** (AF) and **FIGURE 1B** (ASCVD) incidence rate increased with age in each dataset. Explorys and MGB showed similar incidence rates in each age group, whereas UKBB patients had substantially lower AF incidence. Similarly, the ASCVD incidence rate increased with age. 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 poor discrimination, poor calibration, or both (**SUPPLEMENTARY TABLE X and XI**); for example, patients 75 or older had discrimination lower than 0.7 and calibration slope out of the 0.7–1.3 range (17.4% in Explorys, 10.6% in MGB). All patients with prevalent HF had the two measures out of boundaries as well (3.7% in Explorys, 1.9% in MGB).

**FIGURE 2** summarizes performance measures for the CHARGE-AF score.

Discrimination consistently decreased with increased age (**FIGURE 2A**); for example, discrimination declined with increasing age from concordance index of 0.721 [95% CI, 0.716–0.726] for the youngest (45–54y) subgroup to 0.566 [0.556–0.577], for the oldest (85–90y) subgroup in Explorys. Discrimination was higher for females than for males, consistent with prior findings[15,18,6,30], 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 (**FIGURE 2B**); for example, in Explorys, values declined from 1.222 [95% CI, 1.198–1.246] for the youngest (45–54y) subgroup to 0.422 [0.371–0.474] for the oldest (85–90y) subgroup.

The strength of association between the CHARGE-AF score and incident AF (as measured using SHRs) decreased with older age (**FIGURE 2C**); for example, SHR declined from 3.395 [95% CI, 3.315–3.477] for the youngest (45–54y) subgroup to 1.526 [1.449–1.606] for the oldest (85–90y) 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.

**Biased behaviors for CHARGE-AF**

As shown in **FIGURE 3A**, 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 most biased

subgroup was 65–74y with a statistical parity difference of -0.331 [95% CI, -0.333–-0.329] in Explorys. As shown in **FIGURE 3B**, consistent across each dataset, sensitivity was lower for females, particularly in intermediate age groups (65–74 and 75–84); for example, the most biased subgroup was 65–74y with sensitivity difference of -0.311 [95% CI, -0.319–-0.304] in Explorys. As shown in **FIGURE 3C**, specificity was higher for females in intermediate age groups (65–74 and 75–84); for example, the most biased subgroup was 65–74y with specificity difference of 0.328 [95% CI, 0.326–0.330] in Explorys.

Similar to the bias pattens for sex, biases for race were notable in intermediate age groups (65–74 and 75–84). As shown in **FIGURE 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 75–84y subgroup had statistical parity difference of -0.228 [95% CI, -0.232–-0.225] in Explorys. Likely as a result of systematically lower predicted risk estimates, CHARGE-AF exhibited lower sensitivity (**FIGURE 3E)** and greater specificity (**FIGURE 3F**) among Black individuals; as an example, sensitivity difference was -0.168 [95% CI, -0.180–-0.157], and specificity difference was 0.231 [0.228–0.235] for the 75–84y subgroup in Explorys. For both sex and race, biased behavior 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 poor discrimination, poor calibration, or both (**SUPPLEMENTARY TABLE 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 0.661 [95% CI, 0.650–0.672] for the 40–49y subgroup to 0.569 [0.564–0.574] for the 70–79y subgroup in Explorys (**FIGURE 4A)**. This behavior was consistent across all 3 datasets. Discrimination among individuals with prevalent HF was similar to the overall 70–79y subgroup.

We also observed miscalibration 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 (**FIGURE 4B**); for example, in Explorys, values were the lowest for the 40–49y subgroup with a slope of 0.582 [95% CI, 0.549–0.615], and 0.396 [0.368–0.423] for the 70–79y subgroup, in comparison to values above 0.7 for the intermediate age subgroups. Similar to CHARGE-AF, calibration was poor among individuals with prevalent HF, again with a general tendency to overestimate risk.

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 (**FIGURE 4C**); for example, highest SHR was 2.083 [95% CI, 2.025–2.142] for the 50–59 subgroup and 1.485 [1.446–1.526] for the 70–79 subgroup, in Explorys.

**Biased behaviors for PCE**

As shown in **FIGURE 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 60–69y subgroup had the most bias with a statistical parity difference of -0.424 [95% CI, -0.425–-0.422]. As shown in **FIGURE 5B,** across all datasets, sensitivity was lower for females, especially at intermediate age groups (50–59 and 60–69); for example, the subgroup with the most bias was 50–59y with sensitivity difference of -0.354 [95% CI, -0.367–-0.341] in Explorys. Specificity was higher among females (**FIGURE 5C**), especially at intermediate age groups (50–59 and 60–69); for example, the subgroup with the most bias was 60–69y with specificity difference of 0.428 [95% CI, 0.426–0.429] in Explorys. Overall, patterns observed on the basis of sex using the PCE were similar to those observed using CHARGE-AF.

As shown in **FIGURE 5D**, unlike CHARGE-AF, risk estimates using the PCE were higher in Black individuals in all datasets; this effect was especially noticeable at intermediate age groups (50–59 and 60–69); for example, statistical parity difference for the 50–59y subgroup was the largest compared to the other subgroups in Explorys at 0.249 [95% CI, 0.246–0.252]. Again unlike CHARGE-AF, PCE had increased sensitivity among Black individuals versus White individuals (**FIGURE 5E**); for example, sensitivity difference for the 40–49y and 50–59y subgroups were the largest compared to the other subgroups in Explorys at 0.256 [95% CI, 0.235–0.278] and 0.268 [0.254–0.283], respectively. Differences in sensitivity on the basis of race decreased with increased age in all 3 datasets, with very little difference observed in the oldest age group (70–79).

As shown in **FIGURE 5F** and again unlike CHARGE-AF, across specific age ranges, specificity was lower for Black individuals than for White individuals; this effect was especially noticeable at intermediate age groups (50–59 and 60–69); for example, specificity difference for the 50–59y subgroup was the greatest compared to the other subgroups in Explorys at -0.246 [95% CI, -0.249–-0.243].

## 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 the nature of AF and of ASCVD risk in several keyways. First, we assessed the scores on very large databases, allowing us to perform more granular subgroup analyses. Second, we provide results applicable to 3 resources, allowing us to assess consistency in results across independent datasets. Third, our results provide analyses focused on 2 distinct outcomes, which allows a comparison of performance measures not only using different resources, but also different conditions. Fourth, our results highlight the magnitude of poor performance affecting a large proportion of the population (discrimination, calibration, or both), especially 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 (to predict 5-year incident AF) and PCE (to predict 10-year incident ASCVD) scores to assess possible biases considering sex and race differences.

Patterns of variability were fairly consistent across the CHARGE-AF and PCE models. Importantly, we observed that discrimination and calibration were consistently lower at extremes of age, as well as for individuals with certain prevalent conditions (e.g., HF). Furthermore, we observed evidence of potentially biased performance, with important differences in fairness metrics for sex and race in both scores; for instance, sensitivity was much lower for females than males for both scores in intermediate

subgroups, a finding that was consistent in all datasets. 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 bias 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[16,17] and the PCE[33,7] 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 and increasing miscalibration for higher age groups. Since risk of the majority of incident CVD occurs among older individuals, our findings suggest that 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 biased performance across strata of sex and race. Use of the CHARGE-AF score led to lower sensitivity and greater specificity among women, as well as for 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 existed despite the fact that the PCE has dedicated models stratified on the basis of race and sex (i.e., it is based on 4 distinct equations). Since PCE model predictions were generally better calibrated among White individuals (as shown in **SUPPLEMENTARY FIGURE 2B**), 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 bias.

Of the 3 databases we analyzed, 2 were EHR-based (Explorys and MGB) and the other (UKBB) was a prospective cohort study. While we did identify a strong consistency between MGB and Explorys, patterns identified in the UKBB were not as consistent in all scenarios with the EHR databases. To make more accurate comparisons, additional studies are required to account for differences in EHR resources compared to enrollment-based resources. Individuals appearing in EHR resources are typically associated with higher prevalence of comorbid conditions (as highlighted in **TABLE 1**). Furthermore, EHR resources contain data entries collected for archiving and retrieval purposes; differently, prospective resources are based on systematic data collection mechanisms and are thus susceptible to selection biases.

Our study has several limitations. First, mirroring definitions of race for CHARGE-AF and the PCE, we classified race as White and Black, which limits our ability to assess for more granular effects of race on model behavior and performance. Second, we were unable to assess the effects of socioeconomic deprivation[37,38,39] given the lack of available data in Explorys and MGB. Third, although we analyzed data from large

datasets representing very different settings (i.e., two EHR-based datasets and a prospective cohort study), the majority of individuals across the datasets were White. Inclusion of data sources comprising larger proportions of Black individuals may have allowed us to examine heterogeneity with greater precision. Fourth, cause of death was not available in any of the 3 datasets, affecting calculations of incident ASCVD and AF measures (we considered in our analyses all death causes, not just CVD-related). Fifth, although our findings provide important evidence of performance heterogeneity and potential bias in commonly used risk estimators, we did not explore methods to mitigate these biases. Sixth, we have not applied recently proposed fairness metrics that assess individual fairness (rather than assessing bias at the population level)[42,43].

There are several potential strategies to mitigate the important heterogeneity in performance we characterized and quantified in the current study. One strategy is to adjust models according to empirically observed patterns of bias, such as a recalibration methodology, which have been previously proposed as a potential method to reduce bias and minimize, in particular, decisions related to the overtreatment of healthy individuals[5,34]. Another potential approach is to reweight existing models[36,40,41] within each subgroup of the population, resulting in distinct weights for each subgroup of interest. Yet another strategy is to create new larger models that include certain variables (e.g., socioeconomic deprivation)[35,5] that may offer more consistent prognostic value across subgroups, as well as variables defined to greater precision (e.g., more precise quantification of self-reported race(s)). Applying mitigation as well as individual-level fairness assessment techniques are outside the scope of the current study and are the subject of planned future work.

In summary, we identified evidence of important performance heterogeneity and bias in two cardiovascular risk scores, CHARGE-AF and the PCE. We observed consistent patterns across three large and contrasting populations totaling millions of individuals, including consistently worse risk discrimination among older individuals and substantial miscalibration at extremes of age. We also observed that use of common score thresholds may lead to notable biases on the basis of sex and race. Our study – characterizing and quantifying the performance heterogeneity and bias in clinical risk models – is just an initial step toward improved disease assessment. These results can help inform clinicians on when it may be appropriate to use and not use a particular risk model for an individual patient. They can also inform the important next step: the development of risk models that are more robust to differences across clinical settings and patient characteristics, to facilitate more accurate and equitable risk estimation to guide improved clinical decisions. A major challenge, however, may still remain – even if much more robust models will be developed, care systems that extensively rely on existing simple models must be convinced that not only the new models are significantly much more robust, but are also easy to use and interpretable.

## 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 contain 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 biological samples and 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).

## Funding Sources

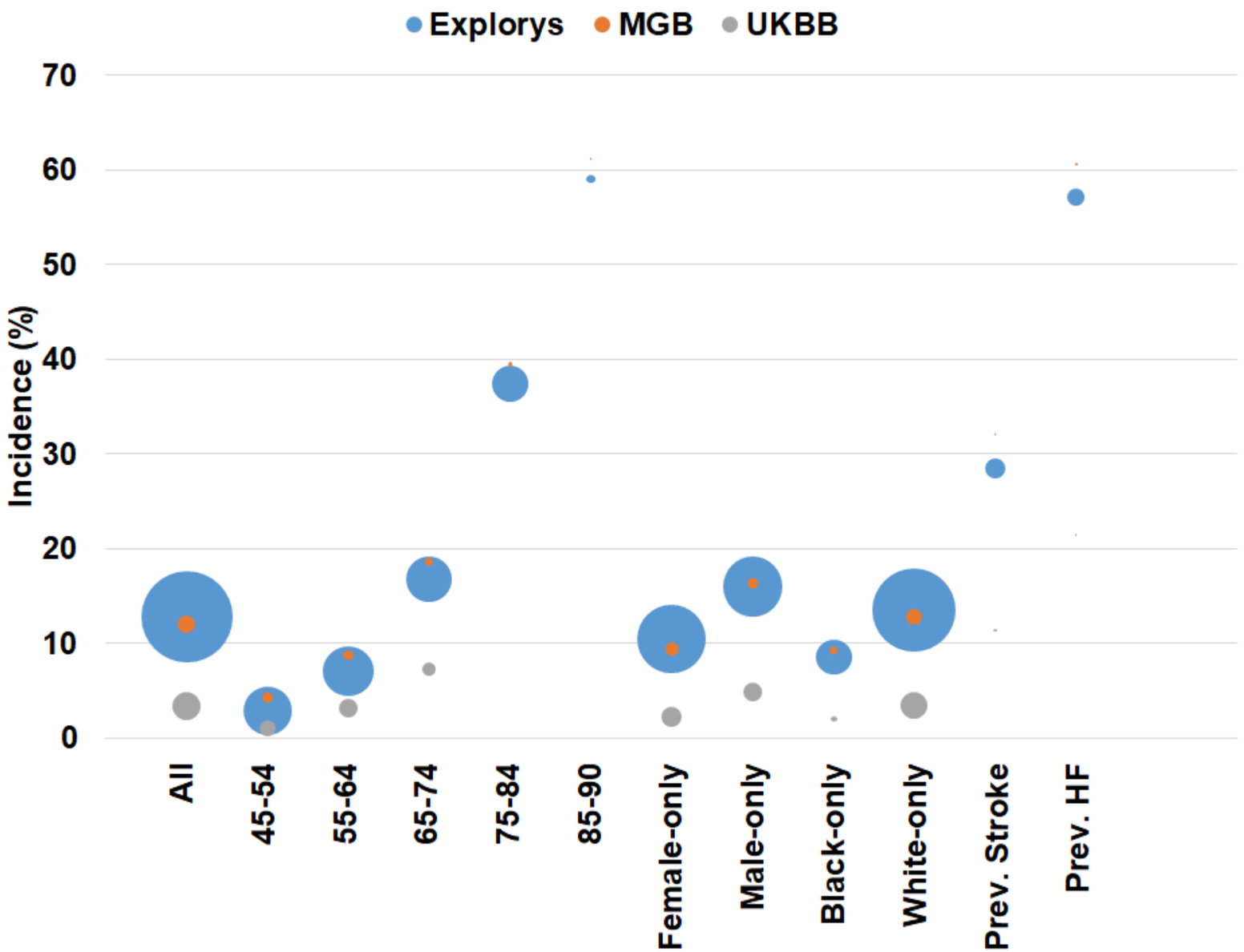
This work was supported by a collaboration between IBM and the Broad Institute and by NIH grants 1K08HG010155 (Khera), R01HL139731 (Lubitz), 2R01HL092577 (Ellinor), and K24HL105780 (Ellinor), T32HL007208 (Khurshid, Patel); American Heart Association (Dallas, Texas) 18SFRN34250007 (Lubitz); a Doris Duke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

## Disclosures

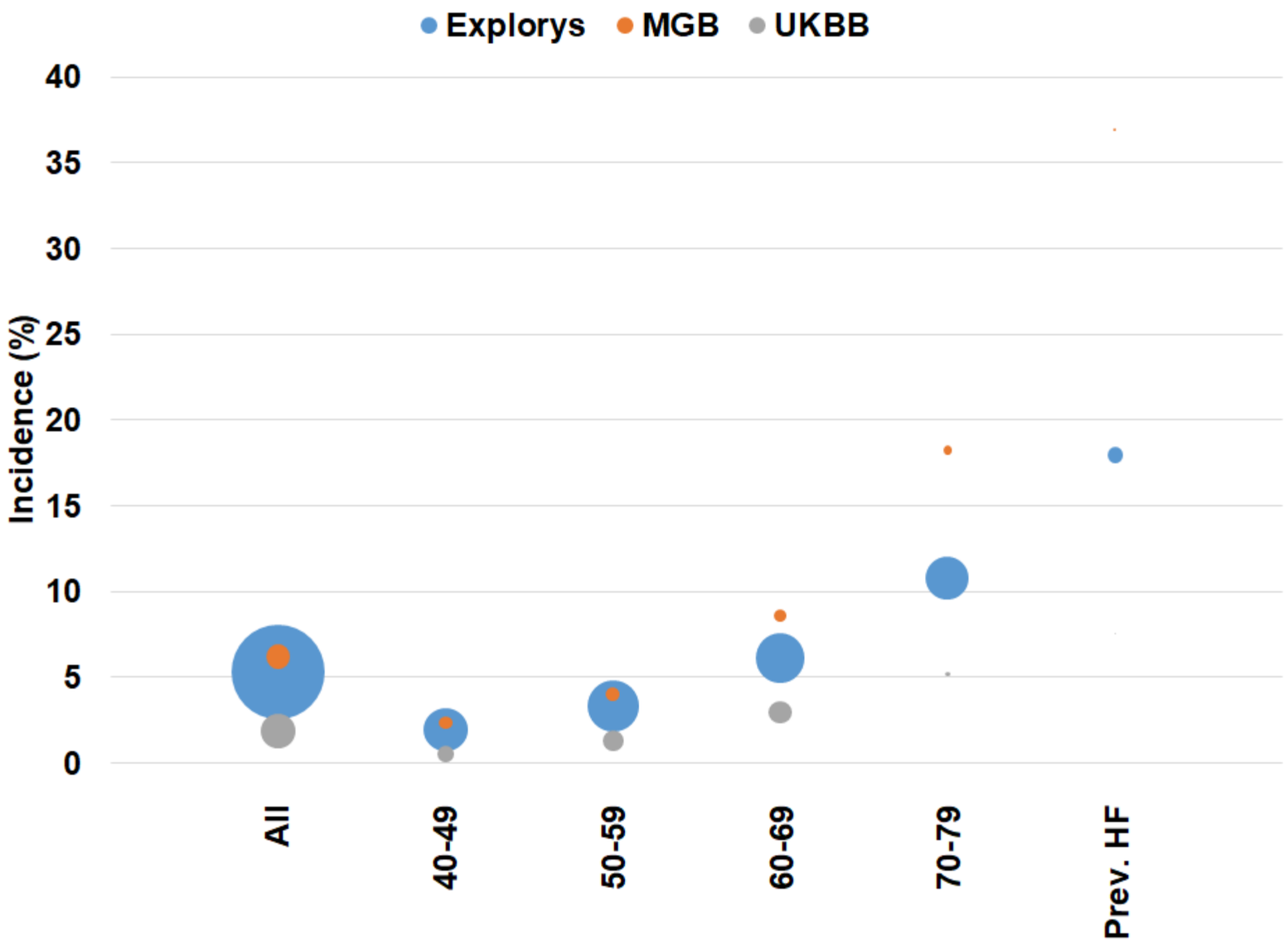
Dr. Lubitz receives sponsored research support from Bristol-Myers Squibb / Pfizer, Bayer AG, Biotronik, and Boehringer Ingelheim, and has consulted for Bristol-Myers Squibb and Bayer AG. Dr. Ellinor receives sponsored research support from Bayer AG and IBM, and he has consulted for Bayer AG, Novartis, and MyoKardia. Drs. Kartoun, Kwon, and Ng are employees of IBM. The remaining authors have no disclosures.

**Table 1. Baseline characteristics**

| | Incident AF (5Y) | | | Incident ASCVD (10Y) | | |
|---|---|---|---|---|---|---|
| | **Explorys (N = 4,750,660)** | **UKBB (N = 445,329)** | **MGB (N = 174,644)** | **Explorys (N = 3,328,992)** | **UKBB (N = 408,154)** | **MGB (N = 198,184)** |
| **N events:** | 196,252 | 7,404 | 7,877 | 97,883 | 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) | 4.1 (2.0, 6.9) | 8.9 (8.2, 9.7) | 6.8 (2.6, 10.0) |
| **Characteristics** | **% or Mean (SD)** | | | | | |
| Female (%) | 56.7 | 55.0 | 60.9 | 56.0 | 54.8 | 58.8 |
| Age (years) | 62.6 (10.8) | 58.4 (7.0) | 60.9 (10.0) | 59.0 (10.6) | 56.9 (8.1) | 57.0 (10.3) |
| White race (%) | 84.2 | 94.7 | 79.6 | 87.3 | 98.4 | 78.1 |
| Smoking (%) | 17.3 | 10.7 | 8.0 | 19.2 | 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 (kg) | 168.5 (10.9) | 168.2 (9.2) | 166.6 (10.4) | | | |
| Weight (cm) | 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. | | | 52 (17) | 1.46 (0.4) | 57 (18) |
| TC (US: mg/dL; UK: mmol/L) | | | | 189 (43) | 5.7 (1.1) | 195 (39) |
| Hypertensive therapy (%) | 50.1 | 30.5 | 44.8 | 54.2 | 27.9 | 39.3 |
| Diabetes (%) | 21.3 | 2.5 | 16.0 | 22.0 | 5.0 | 14.8 |
| Heart failure (%) | 3.7 | 0.4 | 1.9 | 3.4 | 0.3 | 1.6 |




**(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 1K PY and population sizes.** All population and subpopulation sizes and exact incidence rates are provided in **SUPPLEMENTARY TABLE IX**.

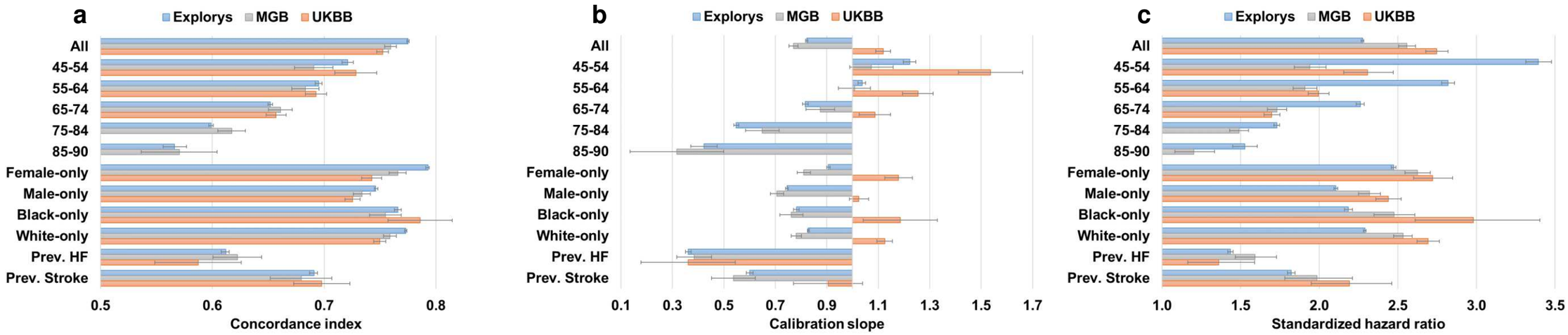
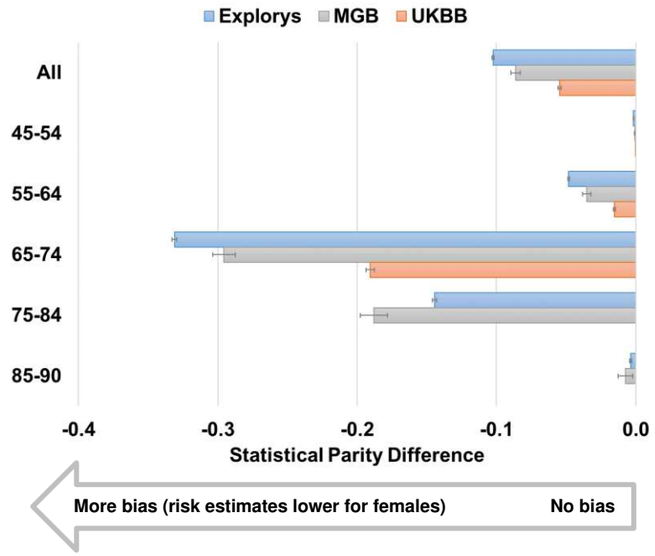
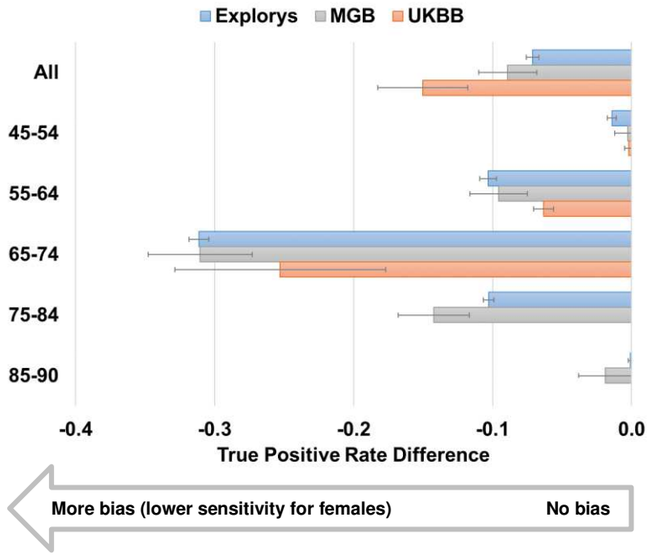


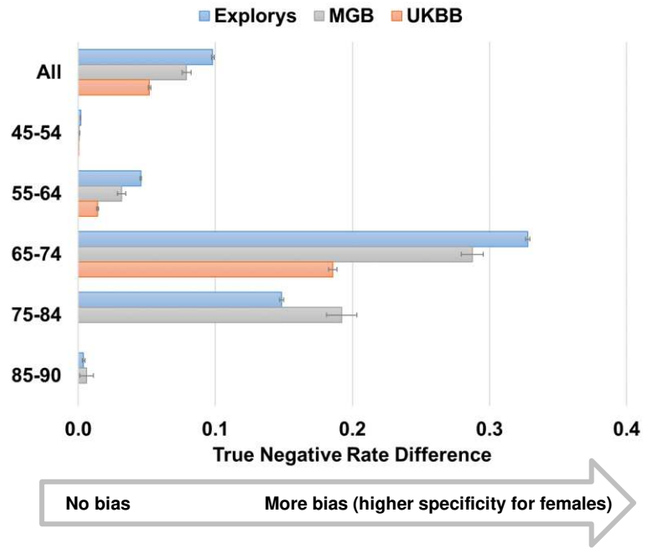

**Figure 2. Performance measures for CHARGE-AF.** Prev. = Prevalence; HF = Heart failure.



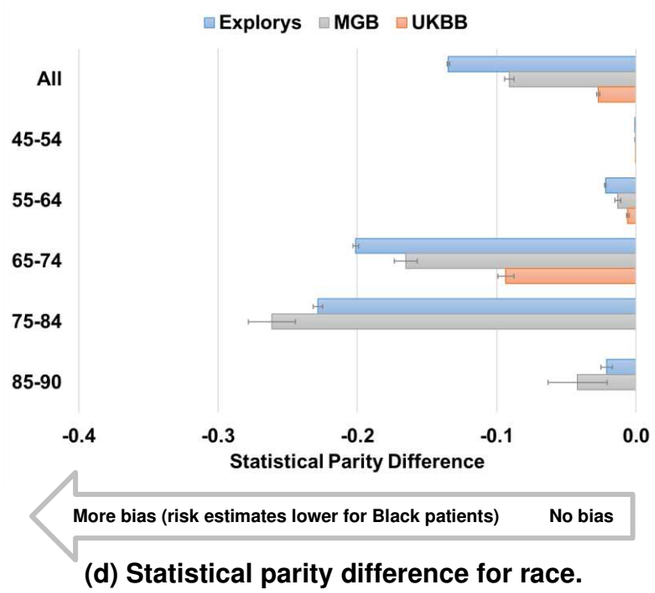

(a) Statistical parity difference for sex.



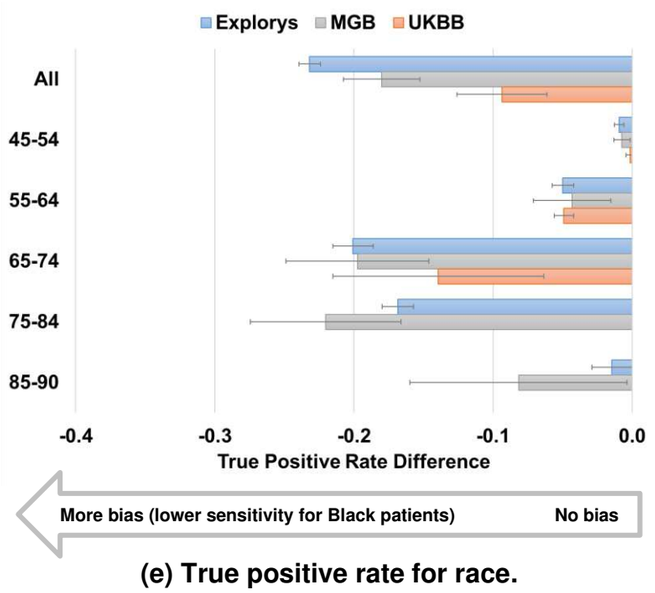

(b) True positive rate for sex.



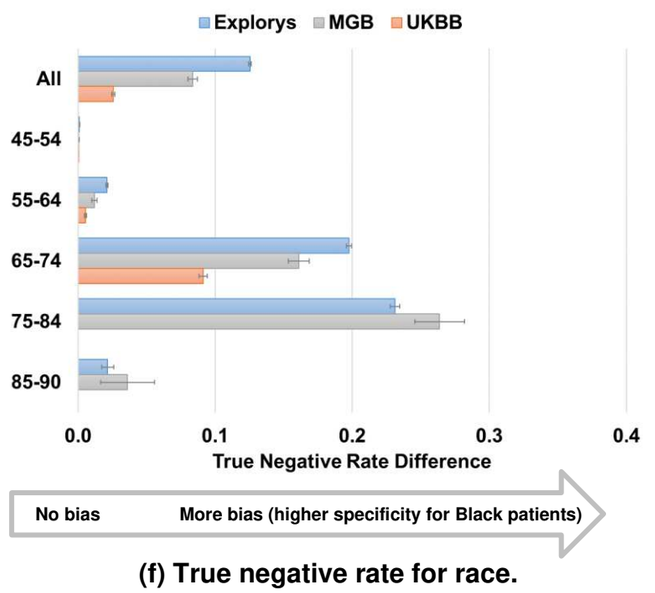

(c) True negative rate for sex.




(d) Statistical parity difference for race.




(e) True positive rate for race.




(f) True negative rate for race.

**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.

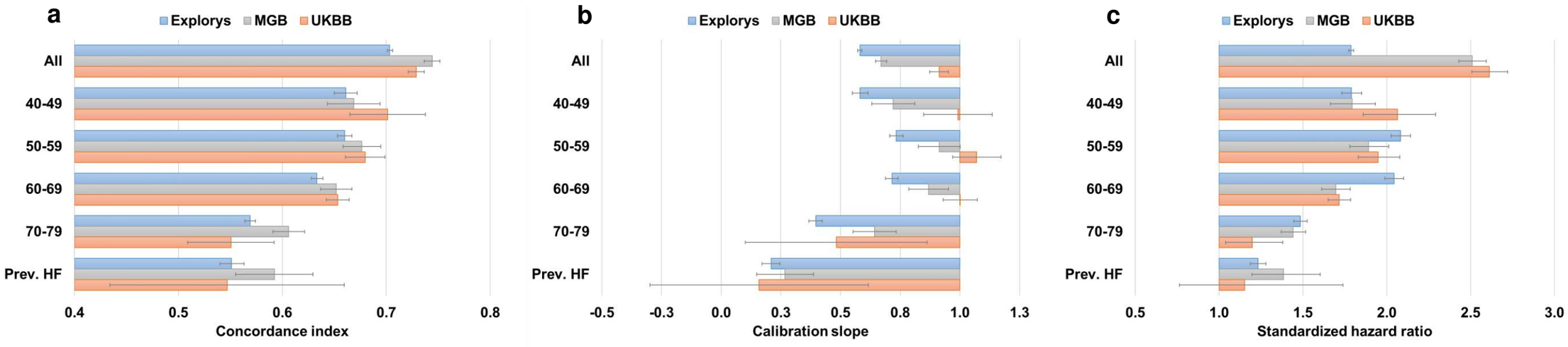
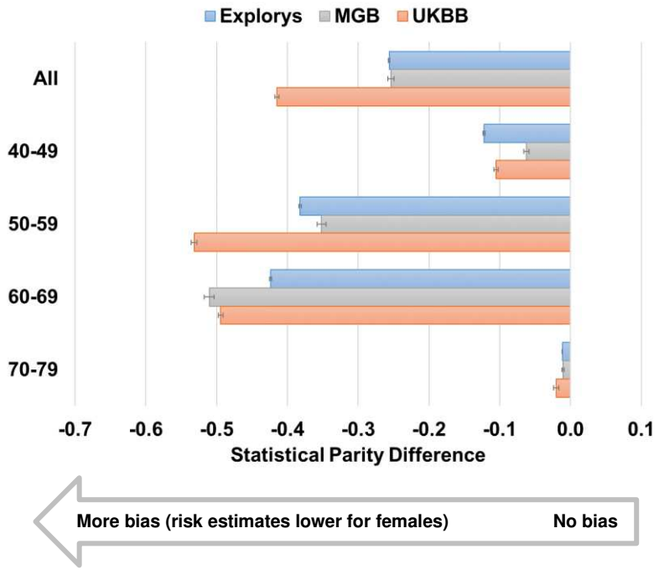
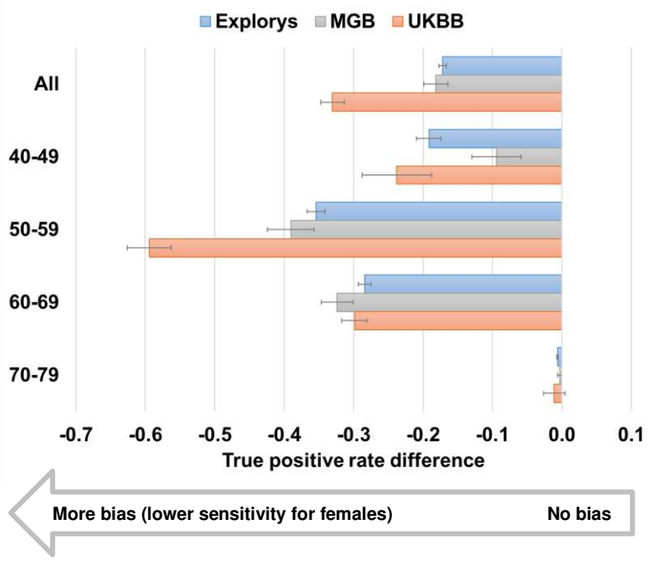


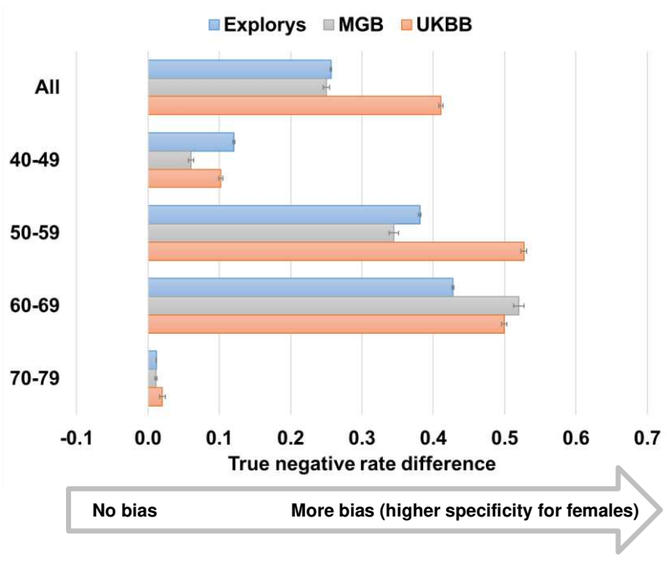

**Figure 4. Performance measures for PCE (Female-White).** Prev. = Prevalence; HF = Heart failure. Refer to **SUPPLEMENTARY TABLE VII** for additional PCE models.



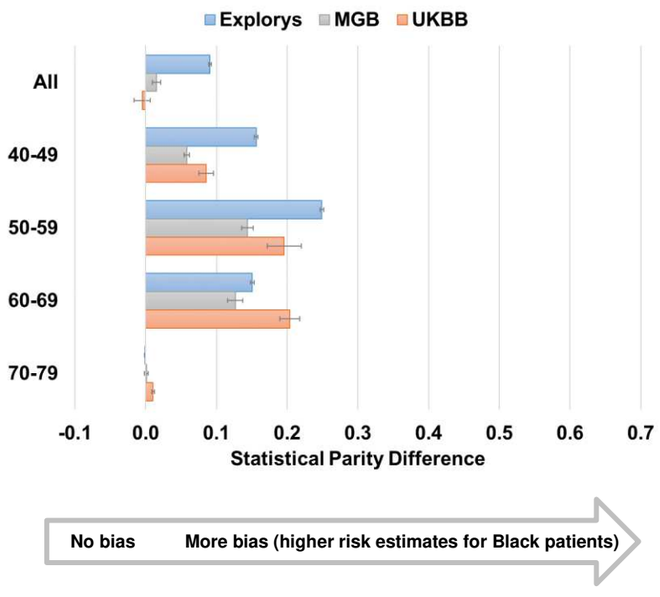

(a) Statistical parity difference for sex.



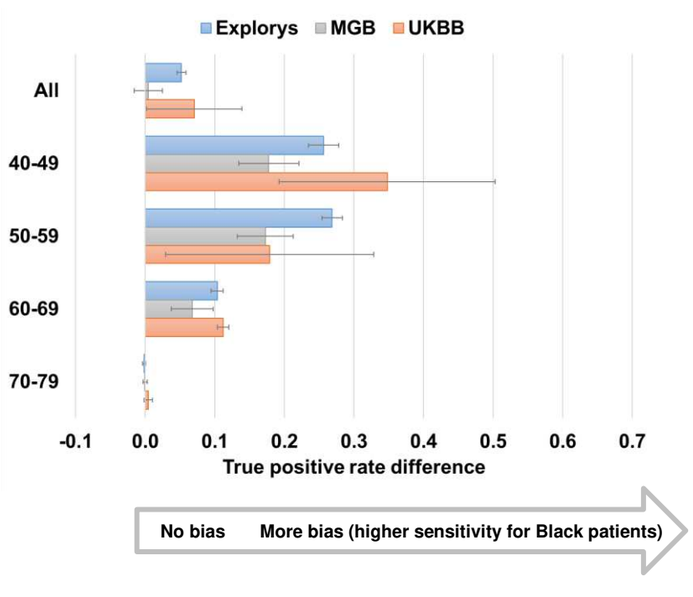

(b) True positive rate for sex.



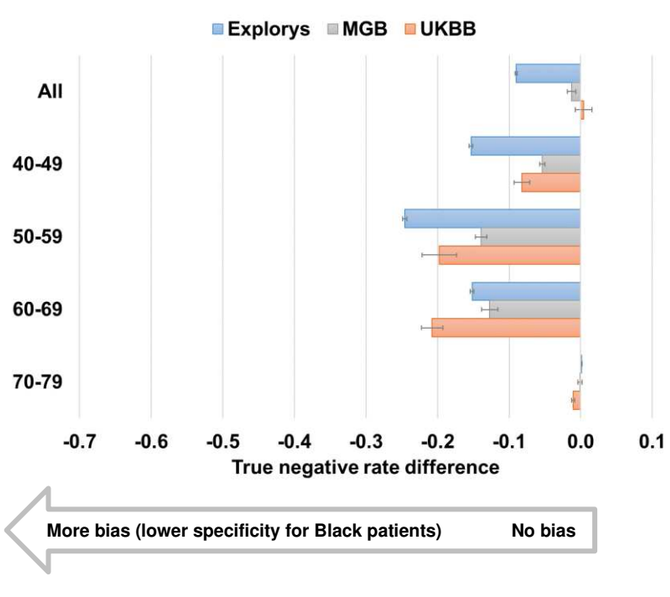

(c) True negative rate for sex.




(d) Statistical parity difference for race.




(e) True positive rate for race.




(f) True negative rate for race.

**Figure 5. Fairness analysis for PCE.**

## References


1. Kavousi M, Leening MJ, Nanchen D, Greenland P, Graham IM, Steyerberg EW, Ikram MA, Stricker BH, Hofman A, Franco OH. 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. 2014;311(14):1416–23.

2. DeFilippis AP, Young R, Carrubba CJ, et al. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. Ann Intern Med. 2015;162(4):266–75. doi:10.7326/M14–1281.

3. Rana JS, Tabada GH, Solomon MD, et al. Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic population. J Am Coll Cardiol 2016;67:2118–30.

4. DeFilippis AP, Young R, McEvoy JW, 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 2017;38:598–608.

5. Pylypchuk R, Wells S, Kerr A, et al. Cardiovascular disease risk prediction equations in 400,000 primary care patients in New Zealand: a derivation and validation study. Lancet 2018;391:1897–901.

6. Damen JA, Pajouheshnia R, Heus P, Moons KGM, Reitsma JB, Scholten RJPM, Hooft L, Debray TPA. 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. 2019 Jun 13;17(1):109. doi: 10.1186/s12916-019-1340-7. PMID: 31189462; PMCID: PMC6563379.

7. Khera R, Pandey A, Ayers CR, Carnethon MR, Greenland P, Ndumele CE, Nambi V, Seliger SL, Chaves PHM, Safford MM, Cushman M, Xanthakis V, Vasan RS, Mentz RJ, Correa A, Lloyd-Jones DM, Berry JD, de Lemos JA, Neeland IJ. Performance of the Pooled Cohort Equations to estimate atherosclerotic cardiovascular disease risk by body mass index. JAMA Netw Open. 2020 Oct 1;3(10):e2023242. doi: 10.1001/jamanetworkopen.2020.23242. Erratum in: JAMA Netw Open. 2020 Dec 1;3(12):e2030880. PMID: 33119108; PMCID: PMC7596579.

8. Lee CH, Woo YC, Lam JK, Fong CH, Cheung BM, Lam KS, Tan KC. Validation of the Pooled Cohort equations in a long-term cohort study of Hong Kong Chinese. J Clin Lipidol. 2015 Sep-Oct;9(5):640-6.e2. doi: 10.1016/j.jacl.2015.06.005. Epub 2015 Jun 16. PMID: 26350809.

9. Jung KJ, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, Seung KB, Kim HK, Yun YD, Choi SH, Sung J, Lee TY, Kim SH, Koh SB, Kim MC, Chang Kim H, Kimm H, Nam C, Park S, Jee SH. The ACC/AHA 2013 pooled cohort equations compared to a Korean

Risk Prediction Model for atherosclerotic cardiovascular disease. Atherosclerosis. 2015 Sep;242(1):367–75. doi: 10.1016/j.atherosclerosis.2015.07.033. Epub 2015 Jul 22. PMID: 26255683.

10. Nguyen QD, Odden MC, Peralta CA, Kim DH. Predicting risk of atherosclerotic cardiovascular disease using Pooled Cohort Equations in older adults with frailty, multimorbidity, and competing risks. J Am Heart Assoc. 2020 Sep 15;9(18):e016003. doi: 10.1161/JAHA.119.016003. Epub 2020 Sep 2. PMID: 32875939; PMCID: PMC7727000.

11. Muntner P, Colantonio LD, Cushman M, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. JAMA 2014;311:1406–15.

12. Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC Jr, Sorlie P, Stone NJ, Wilson PW, Jordan HS, Nevo L, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK, Smith SC Jr, Tomaselli GF; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 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. 2014 Jun 24;129(25 Suppl 2):S49–73. doi: 10.1161/01.cir.0000437741.48606.98. Epub 2013 Nov 12. Erratum in: Circulation. 2014 Jun 24;129(25 Suppl 2):S74–5. PMID: 24222018.

13. Kemp Gudmundsdottir K, Fredriksson T, Svennberg E, Al-Khalili F, Friberg L, Frykman V, Hijazi Z, Rosenqvist M, Engdahl J. Stepwise mass screening for atrial fibrillation using N-terminal B-type natriuretic peptide: the STROKESTOP II study. Europace. 2020 Jan 1;22(1):24-32. doi: 10.1093/europace/euz255. PMID: 31790147; PMCID: PMC6945054.

14. Alonso A, Krijthe BP, Aspelund T, Stepas KA, Pencina MJ, Moser CB, Sinner MF, Sotoodehnia N, Fontes JD, Janssens AC, Kronmal RA, Magnani JW, Witteman JC, Chamberlain AM, Lubitz SA, Schnabel RB, Agarwal SK, McManus DD, Ellinor PT, Larson MG, Burke GL, Launer LJ, Hofman A, Levy D, Gottdiener JS, Kääb S, Couper D, Harris TB, Soliman EZ, Stricker BH, Gudnason V, Heckbert SR, Benjamin EJ. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. J Am Heart Assoc. 2013 Mar 18;2(2):e000102. doi: 10.1161/JAHA.112.000102. PMID: 23537808; PMCID: PMC3647274.

15. Alonso A, Roetker NS, Soliman EZ, Chen LY, Greenland P, Heckbert SR. Prediction of atrial fibrillation in a racially diverse cohort: The Multi-Ethnic Study of Atherosclerosis (MESA). J Am Heart Assoc. 2016;5.

16. Shulman E, Kargoli F, Aagaard P, Hoch E, Di Biase L, Fisher J, Gross J, Kim S, Krumerman A, Ferrick KJ. 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. 2016 Jan 1;117(1):76–83. doi: 10.1016/j.amjcard.2015.10.009. Epub 2015 Oct 19. PMID: 26589820.

17. Christophersen IE, Yin X, Larson MG, Lubitz SA, Magnani JW, McManus DD, Ellinor PT, Benjamin EJ. 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. 2016;178:45–54.

18. Khurshid S, Kartoun U, Ashburner JM, Trinquart L, Philippakis A, Khera AV, Ellinor PT, Ng K, Lubitz SA. Performance of atrial fibrillation risk prediction models in over 4 million individuals. Circ Arrhythm Electrophysiol. 2021 Jan;14(1):e008997. doi: 10.1161/CIRCEP.120.008997. Epub 2020 Dec 9. PMID: 33295794; PMCID: PMC7856013.

19. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. BMC Med Ethics. 2017;18(1):19. Published 2017 Mar 1. doi:10.1186/s12910-017-0179-8

20. Kartoun U, Corey KE, Simon TG, Zheng H, Aggarwal R, Ng K, Shaw SY. The MELD-Plus: A generalizable prediction risk score in cirrhosis. PLoS ONE. 2017;12:e0186301.

21. Dron JS, Wang M, Patel AP, Kartoun U, Ng K, Hegele RA, Khera AV. Genetic predictor to identify individuals with high Lipoprotein(a) concentrations. Circ Genom Precis Med. 2021 Feb;14(1):e003182. doi: 10.1161/CIRCGEN.120.003182. Epub 2021 Feb 1. PMID: 33522245; PMCID: PMC7887018.

22. 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. Available from: https://www.ncbi.nlm.nih.gov/books/NBK285994/

23. Khurshid S, Reeder C, Harrington LX, et al. Cohort design and natural language processing to reduce bias in electronic health records research: The Community Care Cohort Project. medRxiv 2021.05.26.21257872; doi: https://doi.org/10.1101/2021.05.26.21257872

24. Sudlow C, Gallacher J, Allen N, 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 2015;12:e1001779.

25. Hulme OL, Khurshid S, Weng L-C, Anderson CD, Wang EY, Ashburner JM, Ko D, McManus DD, Benjamin EJ, Ellinor PT, Trinquart L, Lubitz SA. Development and validation of a prediction model for atrial fibrillation using electronic health records. JACC Clin Electrophysiol. 2019;5:1331–41.

26. Khurshid S, Keaney J, Ellinor PT, Lubitz SA. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. Am J Cardiol. 2016;117:221–5.

27. Wang EY, Hulme OL, Khurshid S, Weng LC, Choi SH, Walkey AJ, Ashburner JM, McManus DD, Singer DE, Atlas SJ, Benjamin EJ, Ellinor PT, Trinquart L, Lubitz SA. Initial precipitants and recurrence of atrial fibrillation. Circ Arrhythm Electrophysiol. 2020 Mar;13(3):e007716. doi: 10.1161/CIRCEP.119.007716. Epub 2020 Feb 12. PMID: 32078361; PMCID: PMC7141776.

28. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45:562–5.

29. Bellamy R, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S, Ramamurthy K, Richards JT, 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.

30. Himmelreich JCL, Lucassen WAM, Harskamp RE, Aussems C, van Weert HCPM, Nielen MMJ. 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. 2021 Jan;8(1):e001459. doi: 10.1136/openhrt-2020-001459. PMID: 33462107; PMCID: PMC7816907.

31. Stone NJ, Robinson JG, Lichtenstein AH, Bairey Merz CN, Blum CB, Eckel RH, Goldberg AC, Gordon D, Levy D, Lloyd-Jones DM, McBride P, Schwartz JS, Shero ST, Smith SC Jr, Watson K, Wilson PW; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 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. 2014 Jul 1;63(25 Pt B):2889–934. doi: 10.1016/j.jacc.2013.11.002. 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.

32. 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/.

33. Muntner P, Colantonio LD, Cushman M, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. JAMA 2014;311:1406–15.

34. Pennells L, Kaptoge S, Wood A, Sweeting M, Zhao X, White I, Burgess S, Willeit P, Bolton T, Moons KGM, van der Schouw YT, Selmer R, Khaw KT, Gudnason V, Assmann G, Amouyel P, Salomaa V, Kivimaki M, Nordestgaard BG, Blaha MJ, Kuller LH, Brenner H, Gillum RF, Meisinger C, Ford I, Knuiman MW, Rosengren A, Lawlor DA, Völzke H, Cooper C, Marín Ibañez A, Casiglia E, Kauhanen J, Cooper JA,

Rodriguez B, Sundström J, Barrett-Connor E, Dankner R, Nietert PJ, Davidson KW, Wallace RB, Blazer DG, Björkelund C, Donfrancesco C, Krumholz HM, Nissinen A, Davis BR, Coady S, Whincup PH, Jørgensen T, Ducimetiere P, Trevisan M, Engström G, Crespo CJ, Meade TW, Visser M, Kromhout D, Kiechl S, Daimon M, Price JF, Gómez de la Cámara A, Wouter Jukema J, Lamarche B, Onat A, Simons LA, Kavousi M, Ben-Shlomo Y, Gallacher J, Dekker JM, Arima H, Shara N, Tipping RW, Roussel R, Brunner EJ, Koenig W, Sakurai M, Pavlovic J, Gansevoort RT, Nagel D, Goldbourt U, Barr ELM, Palmieri L, Njølstad I, Sato S, Monique Verschuren WM, Varghese CV, Graham I, Onuma O, Greenland P, Woodward M, Ezzati M, Psaty BM, Sattar N, Jackson R, Ridker PM, Cook NR, D'Agostino RB, Thompson SG, Danesh J, Di Angelantonio E; Emerging Risk Factors Collaboration. Equalization of four cardiovascular risk algorithms after systematic recalibration: individual-participant meta-analysis of 86 prospective studies. Eur Heart J. 2019 Feb 14;40(7):621-631. doi: 10.1093/eurheartj/ehy653. PMID: 30476079; PMCID: PMC6374687.

35. Dalton JE, Perzynski AT, Zidar DA, Rothberg MB, Coulton CJ, Milinovich AT, Einstadter D, Karichu JK, Dawson NV. Accuracy of cardiovascular risk prediction varies by neighborhood socioeconomic position: A retrospective cohort study. Ann Intern Med. 2017 Oct 3;167(7):456–64. doi: 10.7326/M16-2543. Epub 2017 Aug 29. PMID: 28847012; PMCID: PMC6435027.

36. Park Y, Hu J, Singh M, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open. 2021;4(4):e213909. doi:10.1001/jamanetworkopen.2021.3909

37. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018 Dec 18;169(12):866-872. doi: 10.7326/M18-1990. Epub 2018 Dec 4. PMID: 30508424; PMCID: PMC6594166.

38. Townsend P, Phillimore P, Beattie A. Health and Deprivation: Inequality and the North. Routledge, London. 1988.

39. Foster HME, Celis-Morales CA, Nicholl BI, Petermann-Rocha F, Pell JP, Gill JMR, O'Donnell CA, Mair FS. 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. 2018 Dec;3(12):e576-e585. doi: 10.1016/S2468-2667(18)30200-7. Epub 2018 Nov 20. PMID: 30467019.

40. Calders T, Kamiran F, Pechenizkiy M. Building classifiers with independency constraints. ICDM Workshops - IEEE International Conference on Data Mining. 2009:13–8. August 6–9, 2009; Miami, Florida.

41. Hainmueller J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1), 25–46. 2012. doi:10.1093/pan/mpr025

42. Yurochkin M, Bowery A, Sun Y. Training individually fair ML models with sensitive subspace robustness. ICLR 2020.

43. Maity S, Xue S, Yurochkin M, Sun Y. Statistical inference for individual fairness. ICLR 2021.

44. Khurshid S, Mars N, Haggerty CM, Huang Q, Weng LC, Hartzel DN, Lunetta KL, Ashburner JM, Anderson CD, Benjamin EJ, Salomaa V, Ellinor PT, Fornwalt BK, Ripatti S, Trinquart L, Lubitz SA; Regeneron Genetics Center. Predictive accuracy of a clinical and genetic risk model for atrial fibrillation. Circ Genom Precis Med. 2021 Aug 31:CIRCGEN121003355. doi: 10.1161/CIRCGEN.121.003355. Epub ahead of print. PMID: 34463125.

45. Patel AP, Wang M, Kartoun U, Ng K, Khera AV. Quantifying and understanding the higher risk of atherosclerotic cardiovascular disease among South Asian individuals: results from the UK Biobank prospective cohort study. Circulation. 2021 Aug 10;144(6):410-422. doi: 10.1161/CIRCULATIONAHA.120.052430. Epub 2021 Jul 12. PMID: 34247495; PMCID: PMC8355171.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- KartounetalNatMedSupplementaryFinal.pdf
- flatKartounepc.pdf
- flatKartounrs.pdf