Assessing racial bias in type 2 diabetes risk prediction algorithms

Héléne T. Cronjé, PhD[1,¶], Alexandros Katsiferis, MSc[1,¶], Leonie K. Elsenburg, PhD[1], Thea O. Andersen, MSc[1], Naja H. Rod, PhD[1], Tri-Long Nguyen, PhD[1], Tibor V. Varga, PhD[1,*]

Short title: *Racial bias in diabetes risk scores*

Affiliations:

1. Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

¶ These authors contributed equally

* Corresponding author:

Tibor V. Varga

ORCID: https://orcid.org/0000-0002-2383-699X

Postal address:

Bartholinsgade 6Q; DK-1356 Copenhagen K; Denmark

Tel: +45 35 32 77 39

E-mail: tibor.varga@sund.ku.dk

Abstract

Risk prediction models for type 2 diabetes can be useful for the early detection of individuals at high risk. However, models may also bias clinical decision-making processes, for instance by differential risk miscalibration across racial groups. We investigated whether the Prediabetes Risk Test (PRT) issued by the National Diabetes Prevention Program, and two prognostic models, the Framingham Offspring Risk Score, and the ARIC Model, demonstrate racial bias between non-Hispanic Whites and non-Hispanic Blacks. We used National Health and Nutrition Examination Survey (NHANES) data, sampled in six independent two-year batches between 1999 and 2010. A total of 9,987 adults without a prior diagnosis of diabetes and with fasting blood samples available were included. We calculated race- and year-specific average predicted risks of type 2 diabetes according to the risk models. We compared the predicted risks with observed ones extracted from the US Diabetes Surveillance System across racial groups (summary calibration). All investigated models were found to be miscalibrated with regards to race, consistently across the survey years. The Framingham Offspring Risk Score overestimated type 2 diabetes risk for non-Hispanic Whites and underestimated risk for non-Hispanic Blacks. The PRT and the ARIC models overestimated risk for both races, but more so for non-Hispanic Whites. The risk of type 2 diabetes extracted from these landmark models were more severely overestimated for non-Hispanic Whites compared to non-Hispanic Blacks, potentially resulting in a larger fraction of non-Hispanic Whites being prioritized for a preventive intervention, but also more likely to be overdiagnosed and overtreated, with a larger fraction of non-Hispanic Blacks being potentially underprioritized and undertreated.

Introduction

Early detection of individuals who are at high risk for developing type 2 diabetes is a powerful strategy to tackle the diabetes epidemic through targeted prevention[1]. Type 2 diabetes affects one in eight adults in the United States (US) and is the nation's seventh leading cause of death. A look beyond the averages reveals a disparity in the incidence and mortality rate of diabetes across racial groups, with non-Hispanic Whites being the only group wherein rates are below the national average[2, 3].

Despite their comparatively lower risk[4], non-Hispanic White groups remain overrepresented in diabetes risk prediction literature[5]. Consequently, the implementation of evidence-based risk prediction in primary care is vulnerable to limited generalizability to other racial groups. Biased prediction models may prioritize individuals of certain racial groups for preventive action at different rates, or at different stages in their disease progression[6].

Today, the current standards of care focus on identifying individuals with prevalent type 2 diabetes or prediabetes in asymptomatic adults and referring them to appropriate prevention or treatment routes [10]. Screening relies mainly on the Prediabetes Risk Test (PRT) [10,11], a diagnostic model developed by the National Diabetes Prevention Program (https://diabetes.org/diabetes/risk-test and https://www.cdc.gov/prediabetes/risktest). To the best of our knowledge, despite the availability of numerous other US-based *prognostic* risk prediction models [7-9], very few or none of them are currently implemented in routine.

In this project, we set out to investigate whether the PRT[7] and two key US-based prognostic prediction models for type 2 diabetes that informed the current guidelines, demonstrate consistent (mis)calibration across Black and White non-Hispanic US residents. We further

65 investigate to what extent the current clinical guidelines may address or perpetuate these biases
66 and discuss concrete action points for stakeholders to (further) promote fair prognostic disease
67 prediction.

Research Design and Methods

Study population

The Continuous National Health and Nutrition Examination Survey (NHANES) is a repeated cross-sectional survey that has been providing a vast amount of publicly available health data on a randomly selected nationally representative sample of the civilian, non-institutionalized US population since 1999[8]. Demographic, questionnaire, physical examination, and biochemical data are collected in two-year intervals. Reliability of sub-group data is ensured by the oversampling of minority racial groups and persons over the age of 60.

Our population of interest comprised non-Hispanic White and non-Hispanic Black adults at risk of type 2 diabetes. We included NHANES survey content between 1999 and 2010, sampled in six independent 2-year batches (1999-2000, 2001-2002, …, 2009-2010). A total of 62,160 participants were surveyed during this timeframe, of whom 22,661 provided fasting blood samples. Analyses were restricted to the latter group due to the utilization of fasting biochemical measures in the prediction algorithms tested. Upon excluding individuals with diagnosed type 2 diabetes (based on self-report and medication use), those younger than 18 years, and those who did not self-report their race as non-Hispanic White or non-Hispanic Black (n=12,674), we retained an analytical sample of 9,987 individuals (**Figure 1, Appendix Figure 1**). We used fasting subsample weights to correct for representativeness in our analyses and allow for inferences made from this subset to be representative of the larger US population of interest.

Prediction models

Our analyses focused on the PRT[7] issued by the National Diabetes Prevention Program, and two widely known North-American prognostic type 2 diabetes risk models: the Framingham Offspring Risk Score[9] and the Atherosclerosis Risk in Communities (ARIC) Risk Model[10] (**Appendix Table 1**). We used the NHANES database to extract the data needed to calculate model estimates according to the authors' guidelines (**Figure 1**). Continuous predictors included age (years), body mass index (BMI, kg/m^2), waist circumference (cm), height (cm), systolic and diastolic blood pressure (mmHg), fasting glucose (mmol/L), high-density lipoprotein cholesterol (HDL-C) (mmol/L), and triglycerides (mmol/L).

Binary predictors were sex (male/female), race (non-Hispanic Black and non-Hispanic White), anti-hypertensive medication use (yes/no), previously diagnosed hypertension (yes/no), physical activity (yes/no, coded yes for datasets between 1999-2006 if the participant responded "More active" to a single question "Compared with most men/women your age, would you say that you are…", and coded yes for datasets between 1999-2006 if the participant reported either occupational or leisure time vigorous physical activity), and family history of diabetes (yes/no; yes when participants answered in the affirmative to any of the specific family members questioned explicitly between the 1999 and 2004 surveys, and from 2005 when participants answered yes to the single question of whether a close relative has diabetes). For the Framingham Offspring Risk Score and ARIC Risk Model, we included the latter-described variable as a proxy for *parental history of diabetes* due to the unavailability of a better-defined variable in NHANES.

Statistical analysis

All analyses were undertaken using R 4.1.2. The NHANES data were extracted and prepared for analysis using the framework depicted in **Appendix Figure 1**. Missing data were imputed using multivariate imputation by chained equations[11]. For all variables, random forest was utilized (five iterations), and 15 imputed copies were generated. The imputation models included the sampling weights to account for the complex survey design[12]. Convergence was visually inspected for randomly selected imputed datasets. Race- and survey-specific estimates derived from the final analytical cohort (N=9,987) from the multiple imputation framework were pooled using Rubin's rules[13].

We used the Framingham Offspring Risk Score and the ARIC Risk Model to predict the risk of type 2 diabetes of participants. Given those predictions, we subsequently calculated race- and year-specific average predicted incidence proportions and 95% confidence intervals (CIs) for eight or nine years (depending on the model used) subsequent to each survey batch interval. Those race- and year- specific averages were weighted to account for the survey design of the NHANES data. We also determined the proportion of each survey batch that would be identified for further screening based on their PRT score. The latter assigns a score between -1 and 9, with values of 5 or higher indicating a high risk of either prediabetes or type 2 diabetes. Cumulative type 2 diabetes incidences (matched to the respective risk score time frames) by race were calculated as the sum of yearly reported incidences extracted from the US Diabetes Surveillance System database[14]. We subsequently compared the latter with the NHANES-derived weighted average predicted risks of the two prognostic models. For illustration purposes we also plotted the observed cumulative incidences against the predicted proportion of undiagnosed type 2 diabetes cases of the PRT test **(Figure 1).** Thus, we evaluated the predictive performance of the

133 model in terms of overall calibration within different racial subgroups, by calculating expected-
134 to-observed incidence risk ratios[15].

135

136 Ethics statement

137 The NHANES was subject to ethics review by the National Center for Health Statistics Research
138 Ethics Review Board and approval was obtained bi-annually between each full proposal review.
139 All participants provided informed consent.

140

141 Role of the funding source

142 This project was supported by the University of Copenhagen, the Novo Nordisk Foundation, and
143 the Independent Research Fund Denmark. The funders had no role in the study design; in the
144 collection, analysis, and interpretation of data; in the writing of the report; and in the decision to
145 submit the paper for publication.

Results

Race-stratified weighted, unimputed descriptive statistics of the study population are presented per survey year in **Table 1**. Imputed descriptive statistics are shown in **Appendix Table 2**. Compared to non-Hispanic White NHANES participants, non-Hispanic Black groups were younger, less likely to have resided in the US from birth or to have obtained a high school diploma, and more likely to be hypertensive or have a family history of diabetes. While non-Hispanic Blacks were also more likely to have higher BMIs than their White counterparts, they had more favorable triglyceride profiles, and did not differ from non-Hispanic White groups in terms of waist circumference or HDL-C. For survey years spanning 1999-2004, non-Hispanic Whites were proportionally less physically active than non-Hispanic Blacks, with the trend reversing from 2005 onward.

Appendix Table 3 presents the race-stratified age-adjusted type 2 diabetes incidence rates for non-Hispanic Whites and non-Hispanic Blacks from the US Diabetes Surveillance System[14]. Overall, type 2 diabetes incidence peaked in 2008 and has been decreasing since, although incidences remain higher than what they were in 2000. There continues to be considerable variation across racial groups, with non-Hispanic Blacks consistently having higher incidence rates compared to non-Hispanic Whites.

Predicted average risks for each examined type 2 diabetes model were compared with cumulative incidences calculated from the US Diabetes Surveillance System. As an example, we used the NHANES cohort from 1999 to calculate race-stratified 8-year average predicted type 2 diabetes risk, and thus, calculated that by 2007, 7% of non-Hispanic Whites, and 6% of non-Hispanic Blacks are expected to develop type 2 diabetes. We compared our calculated estimates with real-life race-stratified cumulative incidences from the matching 8-year period (2000-2007)

(cumulative incidences for this period are 5% for non-Hispanic Whites and 8% for non-Hispanic Blacks). We repeated this analysis for the three models, and each of the six cohorts from NHANES between 1999 and 2010, to obtain race-stratified predicted estimates until 2017 (**Appendix Table 4**). Results for the three models are shown in **Figure 2**.

While the Framingham Offspring Risk Score overestimated type 2 diabetes risk for non-Hispanic Whites, it underestimated risk for non-Hispanic Blacks. The PRT and the ARIC Model overestimated type 2 diabetes risk for both races, but more so for non-Hispanic Whites compared to non-Hispanic Blacks. The PRT showed the highest overestimation; based on the scores approximately half of the total populations of the cohorts were prioritized for screening.

We calculated the ratios of the calculated average predicted risks (incidence proportion) to the true cumulative incidences (**Figure 3**). All three models demonstrated overestimation of risk for non-Hispanic Whites. The ARIC model delivered a lower overestimation for non-Hispanic Blacks (average ratio = 1.22) compared to non-Hispanic Whites (average ratio = 2.31). Similarly, the PRT delivered a lower overestimation for non-Hispanic Blacks (average ratio = 5.05) compared to non-Hispanic Whites (average ratio = 9.51) when considering 8-year risk of diabetes.

We visualized the correspondence between the dichotomized PRT scores of individuals and the individual predicted probabilities from the two prognostic type 2 diabetes risk prediction algorithms (**Figure 4**). The PRT scores showed moderate correlation with the predicted probabilities by the Framingham (Kendall's tau=0.39) and the ARIC models (Kendall's tau=0.56). Above and below the threshold of scoring five on the PRT, non-Hispanic Whites demonstrated a slightly higher risk of type 2 diabetes compared to non-Hispanic Blacks using

191 both the Framingham and the ARIC models, however this difference did not reach statistical

192 significance (95% CIs overlapping 0).

193

194

Discussion

In recent decades, algorithmic decision making has become a routine part of healthcare. Simple risk scores are routinely used to evaluate an individual's risk of developing a disease for most cardiometabolic outcomes, including type 2 diabetes[16], and complex artificial intelligence-driven models are continuously developed to predict common complications of type 2 diabetes[17]. Nonetheless, regardless of how powerful artificial intelligence models can be in capturing complex interactions and patterns in data, the appropriateness of the available datasets in terms of representativeness and quality, remains crucial to algorithmic design.

Recent examples have shown that models developed within biased datasets, or not directly addressing data biases, will likely propagate inequalities into clinical decisions[6, 18, 19]. As a result of biased algorithmic decision making, those who are already marginalized and less represented, may encounter further obstacles in accessing optimal healthcare, generating a vicious cycle. In this report, we investigated whether the PRT—a nationally adopted screening algorithm for prediabetes and type 2 diabetes—and two landmark type 2 diabetes predictive models developed in the US were racially biased. We compared the proportion of the population identified as being at high risk for incident type 2 diabetes with corresponding reported national statistics on type 2 diabetes incidence by racial group, within the predicted time horizon.

We interpret the model-derived proportions, indicating the population at risk for type 2 diabetes within a certain timeframe, as the fraction of the racial groups that could be prioritized for preventive intervention in that timeframe. When models underestimate the proportion of the population at risk for type 2 diabetes compared with national statistics, that indicates that the models identify less individuals in need of preventive action than the actual needs based on national statistics. Conversely, in case of overestimation, larger fractions of the population could

be prioritized for preventive action and larger proportions would potentially receive support to achieve their health targets than would be necessary based on national statistics. Imbalances in over- and underestimation between racial groups reflect algorithmic bias, i.e., systematic differences in the performance of algorithms across specific groups. Overdiagnosis, and overtreatment pose significant challenges, such as the detection and unnecessary treatment of individuals who will ultimately remain asymptomatic[20]. However, in the case of type 2 diabetes, the consensus on preventive action largely involves non-invasive lifestyle interventions, which are less likely to pose harm to false positives compared to e.g., pharmaceutical treatment. Nonetheless, imposing individuals to unnecessary, even nonaggressive, procedures, can lead to negative psychosocial consequences, i.e., increased stress levels, reduced quality of life, or social stigma. Ultimately, and given the nature and guidelines of the specific chronic disease, we consider a potential underestimation of risk as a bigger threat compared to overestimation.

Our results show that PRT and the two examined prognostic type 2 diabetes risk prediction models consistently demonstrated a larger risk overestimation for non-Hispanic Whites than Blacks. Moreover, the Framingham Offspring Risk Score, developed in a 99% White and non-Hispanic population, underestimated type 2 diabetes risk for non-Hispanic Blacks.

Although the nationally adopted PRT appears to be a highly sensitive (i.e., very unlikely to misclassify diabetes cases as non-cases), it is also non-specific, and identifies approximately half of the total population for the screening of type 2 diabetes, very likely falsely identifying individuals who are normoglycemic. Despite the wide net the PRT casts, this risk scoring still identifies proportionally more non-Hispanic Whites for preventive action than non-Hispanic Blacks. When compared to the observed national incidence rates, the PRT fails to fully account

for the racial differences that we observe in real-life incidence rates. However, a certain amount of overestimation from PRT was expected, given that it also indicates cases of prediabetes.

Notably, all three examined algorithms are almost exclusively accounting for inherent risk (biological sex and family history of diabetes) or markers of metabolic health (biochemical profiles, blood pressure, and anthropometry). The ARIC Risk Model was the only algorithm to include race, a non-metabolic parameter; and the PRT was the only algorithm incorporating a lifestyle factor (physical activity). In this nationally representative dataset, the three examined models consistently predicted higher average risks for non-Hispanic Whites compared to non-Hispanic Blacks, contrary to the official national statistics. This observation indicates that the metabolic health variables implemented in the models failed to capture the observed lower risk of type 2 diabetes in non-Hispanic Whites compared to Blacks. On the contrary, we observed higher predicted risks for non-Hispanic Whites than for non-Hispanic Blacks. While our results are only suggestive, such phenomenon could be explained by the absence in the models of socio-economic related determinants describing health literacy status, access to healthcare, the latter being on average higher among non-Hispanic Whites, thus serving as compensatory, protective factors that would result in lower realized risks of type 2 diabetes compared to non-Hispanic Blacks. Our descriptive statistics of the analyzed data confirm these patterns, with non-Hispanic Blacks having lower educational attainment on average and a higher likelihood of being immigrants. As an acute solution to the observed algorithmic bias resulting from these health inequalities, the examined algorithms – and specifically the PRT that is already adopted by healthcare – would likely benefit from the explicit inclusion of (additional) markers related to education, health-literacy, and other socioeconomic determinants (that are expected to correlate with race)[21, 22]. The inclusion of such features in the models would likely shift the

distributions of predicted probabilities higher for non-Hispanic Blacks and thus prioritize a larger fraction of this population for preventive action.

Given the underrepresentation of minorities in studies where models were developed, limited algorithmic generalizability might be expected within some racial groups. When it comes to sample size considerations in developing novel algorithms, three major approaches are to be considered. The first approach is to develop models in nationally representative populations. This will result in predictive models that will perform more accurately for the majority. The cohort investigated by the developers of the ARIC Model[10] most closely represented this approach by comprising 85% non-Hispanic Whites and 15% non-Hispanic Blacks. In the original publication, the ARIC model did perform most accurately in the majority population, although the positive coefficient in the risk equation for Blacks increased the risk estimates for this group resulting in an accuracy nearing that of the majority population. When tested externally using NHANES data, however, we still see marked differences in model performance between non-Hispanic Whites and Blacks, reflecting algorithmic bias. The second approach is to develop models in cohorts with roughly equal sample sizes across groups, resulting in models that are expected to perform with the same confidence and precision across groups, but may perform less optimally for the majority. The third approach is to develop separate models in separate groups. This approach was taken when developing the Reynolds Risk Scores separately for men[23] and women[24] to predict sex-specific 10-year cardiovascular disease risk. While both total sample size and sample size of respective groups are important considerations in the development of prediction models, societal and data biases (e.g. differential access to healthcare) may also impact the development and subsequent performance of algorithms, as observed before[6].

Prediction models can perpetuate and reinforce data biases that lie in the core of the societies we live in. Explicit counteraction is warranted to address these biases when developing prediction models. In recent years, algorithmic fairness, as an emergent research area, has been receiving increasing attention and the development of toolkits for the development of fair models have taken flight[25]. It has been raised that without the consideration of sensitive attributes, it is cumbersome to correct for data biases, and develop fair models[26]. However, the inclusion of race as a feature in predictive models has been at the center of academic debate[27-29]. For instance, the vast majority of clinical prediction models for cardiovascular diseases do not include race in the assessment of individual risk[27]. Arguments for the omission of these sensitive attributes are that racial discrimination can potentially be reinforced with the addition of race in the models (i.e. racial profiling), and that there is only weak evidence for genetic or biological differences across races[30]. Additionally, researchers have expressed concern that race insertion might create risks of falsely interpreting racial inequities as immutable facts rather than disparities that require intervention[31, 32]. Nonetheless, variables encoding race are deemed to reflect a complex combination of factors related to biological, cultural, as well as socio-contextual aspects. Therefore, while the use of race as a predictor may suffer from crucial flaws, it can also pave the way for better algorithmic decision-making across various racial groups, for example by promoting personalized care, targeting those in the most need of intervention[27]. The exclusion of race from algorithms may lead to differential external validity and potentially harmful decision making, as models may fail to account for differential risk across groups[30]. Undeniably, we are facing a ‘data problem’: data currently used to generate models are likely to embed inherent racial imbalances, lack intersectionality, and/or important social-contextual features that potentially confound the association between race and various

clinical endpoints[29, 31]. As an example, Obermeyer *et al*. demonstrated significant racial bias in a predictive algorithm using healthcare expenditure as predictive features for prioritizing individuals for healthcare interventions[6]. Thus, it is crucial for future studies to investigate complex diseases through a 'structural racism lens'[33, 34] and disentangle causal relationships between various socioeconomic factors, human physiology, and outcomes, in pursuit of identifying the most appropriate candidate features for predictive modeling.

Based on these findings, we suggest that any published and/or candidate diagnostic or prognostic models should demonstrate algorithmic fairness before adoption in healthcare, e.g., via a systematic comparison of their performance in external samples stratified by sensitive attributes, such as race, ethnicity, and sex. Until fairness is a default consideration in algorithmic development, and models adopted by healthcare are thoroughly checked for fairness, it is likely that societal inequalities will keep propagating into clinical decisions. Thus, we call for medical journals to recommend reporting criteria related to the fairness of novel predictive algorithms. While numerous toolkits, checklists, and governance frameworks have been developed on the topic[35], we emphasize five key points that journals may want to require reporting on: (i) potential data biases; (ii) possible measurements of bias; (iii) utilized bias mitigation strategies; (iv) potential impacts of the biases; (v) and the generalizability of results across relevant sensitive attributes. We recommend that expert groups and healthcare policymakers responsible for the adoption of any risk models (e.g., cutoffs, simple risk scores, or complex algorithms) in clinical practice consider fairness a decisive factor. In the specific context of type 2 diabetes, there is a need to develop an accurate risk score for new-onset diabetes that helps identify groups with unmet needs within minorities.

Our study needs to be considered under some limitations. First, the predictive performance of the models could not be assessed in terms of discrimination as in their traditional reported form. The repeated cross-sectional design of the NHANES did not allow us to compare the individual predicted probabilities returned by the prediction models with the individual observed outcomes, since participants were not followed-up over time. While prediction models are often used for decision-making at the individual level, one needs to bear in mind that predicted probabilities are in fact drawn from *groups* of persons with comparable characteristics in terms of predictor values. To show disparities of predictive performance across racial groups, we reported ratios of average predicted to average observed type 2 diabetes incidences as an overall, summary measure of calibration within the different racial groups[15]. The interpretation of our findings thus remains at the group level. Second, we caution against an interpretation of the trend over time in the estimated predictive performance, as changes in observed incidence over time could be merely due to changes in case detection. Skewed case detection across racial groups (i.e., differential measurement error), could exaggerate or attenuate the reported disparities. This is, however, less of a concern due to the lack of observable differential detection in the available national diagnosed/undiagnosed type 2 diabetes data[36]. Third, the complex survey design of NHANES complicated the estimation procedure in the presence of missing item responses. To address this issue, we performed multiple imputation, while including the sampling weights in the imputation models[12], and using random forest models to allow for complex interactions and non-linearities to be captured[37]. Last, we acknowledge that a large number of type 2 diabetes models are available, and it would be possible to extend our list. Here we selected widely known models developed in the US, but future reports could possibly include a wider range of models for testing.

353 In summary, our study shows that the PRT currently adopted by US healthcare, and prognostic
354 type 2 diabetes prediction models available for adoption in US healthcare are likely attached with
355 some degree of racial bias, which in turn is likely to perpetuate inequalities by providing fewer
356 benefits to minorities, who already demonstrate higher risk for metabolic diseases. We provide
357 specific recommendations that we believe can improve the standards of publishing and the
358 adoption of algorithmic processes in healthcare.

359

Additional Information

Author contributions

Authors' contributions: The corresponding authors attest that all listed authors meet the ICMJE authorship criteria and that no others meeting the criteria have been omitted. AK, TOA, TN, and TVV performed the statistical analysis. HTC, AK, LKE, TOA, TN, and TVV drafted the manuscript. HTC, AK, and TVV created the visualizations. HTC, LKE performed a literature review for the project. TN, and TVV were responsible for the conceptualization of the project. TVV was responsible for project management. TVV is the guarantor of this manuscript. All listed authors interpreted the results, reviewed, and edited the manuscript, verified the data, and approved the final, submitted version.

Conflict-of-interest statement

The authors declare no conflict of interest.

Data availability statement

NHANES data is publicly available and can be accessed via https://www.cdc.gov/nchs/nhanes/index.htm

The authors are willing to share codes.

Funding

HTC and AK are supported by a grant from Novo Nordisk Foundation Challenge Programme for the project Harnessing the Power of Big Data to Address the Societal Challenge of Aging (NNF17OC0027812). TOA is supported by a grant from the Independent Research Fund Denmark (7025-00005B). TVV is supported by the "Data Science Investigator - Emerging 2022" grant from Novo Nordisk Foundation (NNF22OC0075284) and the Department of Public Health (University of Copenhagen).

References

1. Dunkley AJ, Bodicoat DH, Greaves CJ, Russell C, Yates T, Davies MJ, et al. Diabetes prevention in the real world: effectiveness of pragmatic lifestyle interventions for the prevention of type 2 diabetes and of the impact of adherence to guideline recommendations: a systematic review and meta-analysis. Diabetes care. 2014;37(4):922-33.
2. Control CfD, Prevention. National diabetes statistics report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services. 2020:12-5.
3. Heron M. Deaths: Leading Causes for 2019. Natl Vital Stat Rep. 2021;70(9):1-114. Epub 2021/09/15. PubMed PMID: 34520342.
4. Wang MC, Shah NS, Carnethon MR, O'Brien MJ, Khan SS. Age at diagnosis of diabetes by race and ethnicity in the United States from 2011 to 2018. JAMA internal medicine. 2021;181(11):1537-9.
5. Ayensa-Vazquez JA, Leiva A, Tauler P, López-González AA, Aguiló A, Tomás-Salvá M, et al. Agreement between Type 2 Diabetes Risk Scales in a Caucasian Population: A Systematic Review and Report. Journal of clinical medicine. 2020;9(5):1546.
6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.
7. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, et al. Development and validation of a patient self-assessment score for diabetes risk. Annals of internal medicine. 2009;151(11):775-83.
8. Statistics. CfDCaPNCfH. National Health and Nutrition Examination Survey. [https://www.cdc.gov/nchs/nhanes/index.htm]. Accessed: 26 Jan 2022.
9. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Archives of internal medicine. 2007;167(10):1068-74.
10. Schmidt MI, Duncan BB, Bang H, Pankow JS, Ballantyne CM, Golden SH, et al. Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study. Diabetes care. 2005;28(8):2013-8.
11. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45:1-67.
12. Kim JK, Michael Brick J, Fuller WA, Kalton G. On the bias of the multiple‐imputation variance estimator in survey sampling. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006;68(3):509-21.
13. Rubin D. Multiple Imputation for Nonresponse in SurveysWiley J&Sons New York. 1987.
14. US Diabetes Surveillance System. [https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html#]. Accessed: 26 Jan 2022.
15. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. Bmj. 2017;356.
16. Association AD. https://www.diabetes.org/risk-test.
17. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA : the journal of the American Medical

Association. 2016;316(22):2402-10. Epub 2016/11/30. doi: 10.1001/jama.2016.17216. PubMed PMID: 27898976.
18. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. New England Journal of Medicine. 2020;383(25):2477-8.
19. Chapman CH, Schechter CB, Cadham CJ, Trentham-Dietz A, Gangnon RE, Jagsi R, et al. Identifying equitable screening mammography strategies for Black women in the United States using simulation modeling. Annals of internal medicine. 2021;174(12):1637-46.
20. Vogt H, Green S, Ekstrøm CT, Brodersen J. How precision medicine and screening with big data could increase overdiagnosis. Bmj. 2019;366.
21. Laraia BA, Karter AJ, Warton EM, Schillinger D, Moffet HH, Adler N. Place matters: neighborhood deprivation and cardiometabolic risk factors in the Diabetes Study of Northern California (DISTANCE). Social science & medicine. 2012;74(7):1082-90.
22. Kimenai DM, Pirondini L, Gregson J, Prieto D, Pocock SJ, Perel P, et al. Socioeconomic Deprivation: An Important, Largely Unrecognized Risk Factor in Primary Prevention of Cardiovascular Disease. Circulation. 2022;146(3):240-8.
23. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. Circulation. 2008;118(22):2243-51.
24. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA : the journal of the American Medical Association. 2007;297(6):611-9.
25. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development. 2019;63(4/5):4: 1-4: 15.
26. Zliobaite I. Fairness-aware machine learning: a perspective. arXiv preprint arXiv:170800754. 2017.
27. Paulus JK, Kent DM. Race and ethnicity: a part of the equation for personalized clinical decision making? Circulation: Cardiovascular Quality and Outcomes. 2017;10(7):e003823.
28. Essien UR, Jackson LR. Race effects in CVD prediction models. Journal of general internal medicine. 2019;34(4):484-.
29. Waters EA, Colditz GA, Davis KL. Essentialism and Exclusion: Racism in Cancer Risk Prediction Models. JNCI: Journal of the National Cancer Institute. 2021.
30. Paulus JK, Wessler BS, Lundquist CM, Kent DM. Effects of race are rarely included in clinical prediction models for cardiovascular disease. Journal of general internal medicine. 2018;33(9):1429-30.
31. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. Mass Medical Soc; 2020. p. 874-82.
32. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences Research. Annals of family medicine.2792.
33. Adkins-Jackson PB, Chantarat T, Bailey ZD, Ponce NA. Measuring structural racism: a guide for epidemiologists and other health researchers. American journal of epidemiology. 2021.
34. Robinson WR, Renson A, Naimi AI. Teaching yourself about structural racism will improve your machine learning. Biostatistics. 2020;21(2):339-44.

35. Madaio MA, Stark L, Wortman Vaughan J, Wallach H, editors. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020.
36. Prevention CfDCa. Prevalence of Both Diagnosed and Undiagnosed Diabetes. [https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html]. Accessed: 26 Jan 2022.
37. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology. 2014;179(6):764-74.

Tables

Table 1. Descriptive statistics of the unimputed NHANES data (N=9,987).

	1999–2000		2001–2002		2003–2004		2005–2006		2007–2008		2009–2010	
	non-Hispanic White	non-Hispanic Black	non-Hispanic White	non-Hispanic Black	non-Hispanic White	non-Hispanic Black	non-Hispanic White	non-Hispanic Black	non-Hispanic White	non-Hispanic Black	non-Hispanic White	non-Hispanic Black
N	1,222	175	1,575	231	1,451	221	1,450	221	1,474	224	1,502	241
Age, years	44.9 (17.1)	40.2 (14.5)	45.4 (16.9)	39.9 (15.5)	45.6 (17.2)	40.1 (16.8)	46.4 (17.5)	40.7 (16.2)	46.5 (17.3)	40.8 (15.6)	47.1 (17.5)	41.1 (15.5)
Female, %	52.2	54.2	52.0	55.0	52.1	55.6	50.9	54.8	51.7	54.4	52.0	54.3
Highschool graduate, %	82.8	62.1	84	69.2	84.6	66.1	86.5	71.7	83.9	70.6	85.8	71.9
Missing, %	3.5	4.9	3.5	5.1	4.5	6.1	3.4	7.5	3.5	6.1	3.0	4.5
Born in the US, %	95.0	86.6	95.8	92.4	94.4	93.9	95.5	90.8	96.1	92.0	94.0	84.5
Physically active, %	37.5	40.7	35.7	37.9	34.0	38.9	38.2	37.1	43.2	42.5	40.1	37.2
Family history of diabetes, %	44.5	46.4	46.5	46.9	45.5	52.4	34.7	41.9	31.4	37.9	29.8	42.7
Missing, %	5.3	5.2	5.3	6.7	6.7	7.5	6.5	8.6	5.6	6.9	4.8	5.3
Body mass index, kg/m^2	26.9 (5.8)	29.1 (7.1)	27.7 (6.1)	28.8 (7.6)	27.8 (6.1)	29.9 (7.6)	28.2 (7.0)	30.3 (7.8)	27.6 (5.8)	29.1 (7.1)	28.2 (6.4)	30.5 (8.1)
Missing, %	0.5	1.8	6.6	3.6	1.4	2.0	1.7	2.2	1.9	2.1	1.2	0.4
Waist circumference, cm	93.5 (15.4)	94.9 (16.6)	95.5 (15.4)	93.9 (16.7)	96.9 (15.6)	97.1 (17.0)	97.0 (16.6)	98.0 (17.5)	96.3 (15.1)	96.2 (16.2)	97.8 (16.0)	98.1 (16.9)
Missing, %	1.7	4.1	5.2	4.2	4.6	6.7	4.5	4.9	6.6	5.4	4.4	5.2
Height, cm	170.3 (9.9)	169.9 (9.5)	170.1 (9.9)	169.6 (9.9)	170.8 (9.9)	169.7 (9.4)	170.4 (9.9)	169.6 (9.7)	170.8 (10.1)	170.5 (9.0)	170.5 (9.7)	169.0 (9.7)
Missing, %	0.5	1.8	3.6	1.2	1.3	2.0	1.6	2.0	1.8	2.1	0.9	0.4
Fasting glucose, mmol/L	5.1 (1.0)	5.2 (1.5)	5.1 (0.9)	5.1 (1.0)	5.2 (0.7)	5.2 (0.9)	5.2 (1.0)	5.1 (0.8)	5.1 (0.7)	5.1 (1.1)	5.2 (0.8)	5.2 (0.8)
Missing, %	5.6	10.4	4.4	13.5	5.5	6.1	5.3	9.4	4.1	14.7	5.1	9.7
HDL-C, mmol/L	1.3 (0.4)	1.4 (0.4)	1.3 (0.4)	1.4 (0.4)	1.4 (0.4)	1.5 (0.4)	1.5 (0.4)	1.5 (0.5)	1.4 (0.4)	1.6 (0.4)	1.4 (0.4)	1.4 (0.5)
Missing, %	5.7	10.8	4.5	14.1	5.2	5.9	5.1	8.8	4.2	14.1	4.9	9.5
Triglycerides, mmol/L	1.6 (1.1)	1.1 (0.7)	1.7 (2.1)	1.2 (0.7)	1.7 (1.4)	1.2 (0.8)	1.6 (1.2)	1.2 (1.2)	1.5 (1.1)	1.0 (0.6)	1.4 (1.1)	1.1 (0.7)
Triglycerides, ln(mmol/L)	0.3 (0.5)	0.0 (0.5)	0.3 (0.6)	0.0 (0.5)	0.3 (0.6)	0.0 (0.5)	0.3 (0.6)	0.0 (0.5)	0.2 (0.5)	-0.1 (0.6)	0.2 (0.5)	0.0 (0.5)
Missing, %	5.9	11.0	4.5	14.1	5.2	6.3	5.3	9.9	4.2	14.1	4.9	9.5
Systolic BP, mmHg	121.2 (17.7)	123.2 (17.8)	122.4 (18.8)	124.1 (19.4)	121.2 (17.9)	124.8 (20.7)	121.2 (17.2)	125.3 (18.0)	119.2 (16.0)	121.5 (17.3)	118.2 (15.5)	122.9 (16.7)
Diastolic BP, mmHg	72.0 (11.9)	74.1 (12.2)	72.6 (11.5)	73.4 (13.2)	70.4 (13.1)	71.2 (12.8)	68.8 (12.4)	69.8 (14.7)	69.3 (11.6)	70.4 (12.5)	67.8 (11.7)	71.1 (13.5)
Missing, %	2.4	6.5	3.5	5.9	5.1	5.5	3.8	8.5	4.6	6.2	3.6	7.6
Antihypertensive use, %	14.9	15.4	16.1	19.1	18.4	18.4	18.5	21.8	21.2	22.4	21.3	21.4
Missing, %	27.1	28.6	28.2	28.6	24.8	24.8	22.2	21.5	18.2	17.3	17.5	22.1
Hypertension diagnosis, %	41.4	44.3	43.7	46.9	41.4	41.4	39.7	44.2	37.1	41.5	37.0	44.4

Figure legends

Figure 1. Study design

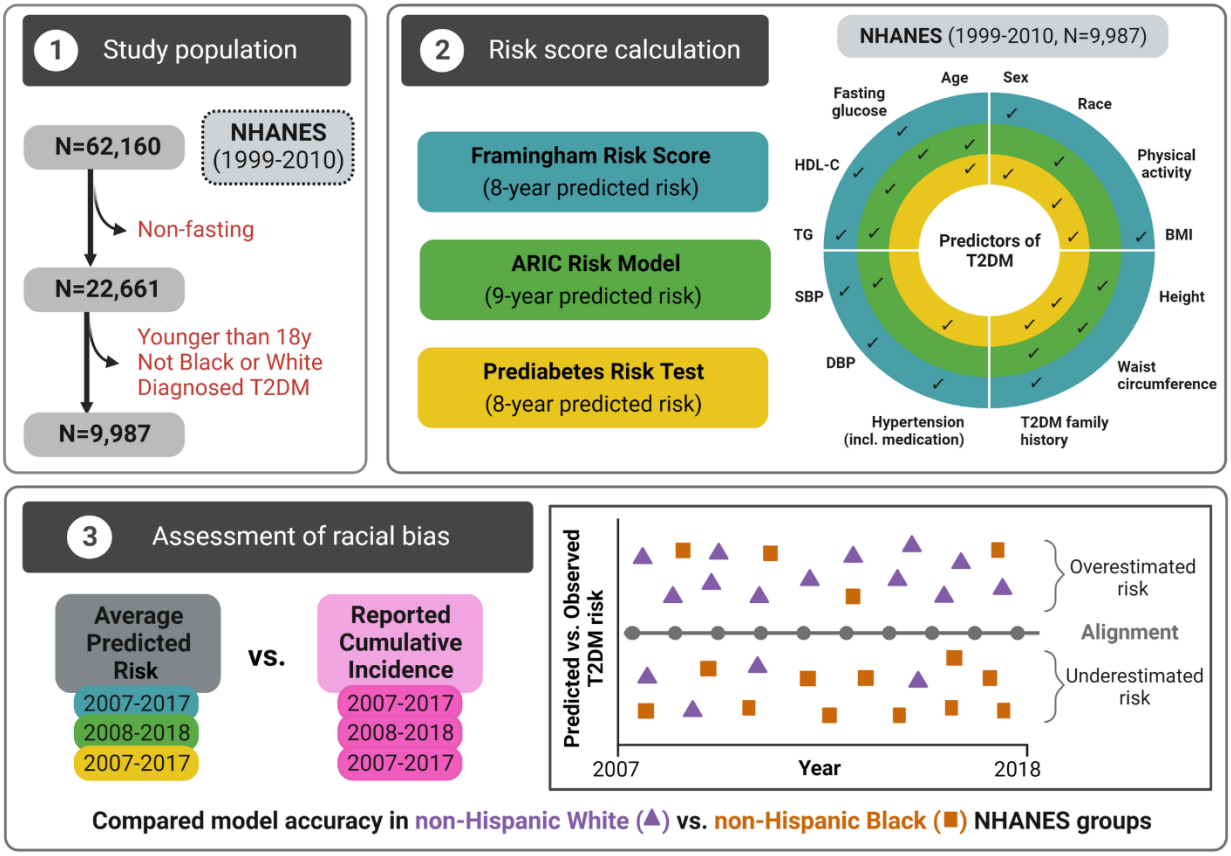


Figure 2. The first and third panel show the 8-year predicted type 2 diabetes risk for the Framingham Offspring Risk Score and the Prediabetes Risk Test and 8-year cumulative incidences of type 2 diabetes from the US Diabetes Surveillance System, combined and per racial group. The second panel shows the 9-year predicted type 2 diabetes risks for the ARIC Model and 9-year cumulative incidences of type 2 diabetes from the US Diabetes Surveillance System, combined and per racial group.

Abbreviations: API - average predicted incidence (proportion); CI - cumulative incidence; PRT – Prediabetes Risk Test.

Error bars reflect 95% confidence intervals.

502 **Figure 3**. Ratios between predicted Framingham Offspring Risk Score, ARIC model, and
503 Prediabetes Risk Test incidence proportions and reported cumulative incidences of type 2 diabetes
504 overall, and per racial group.

505

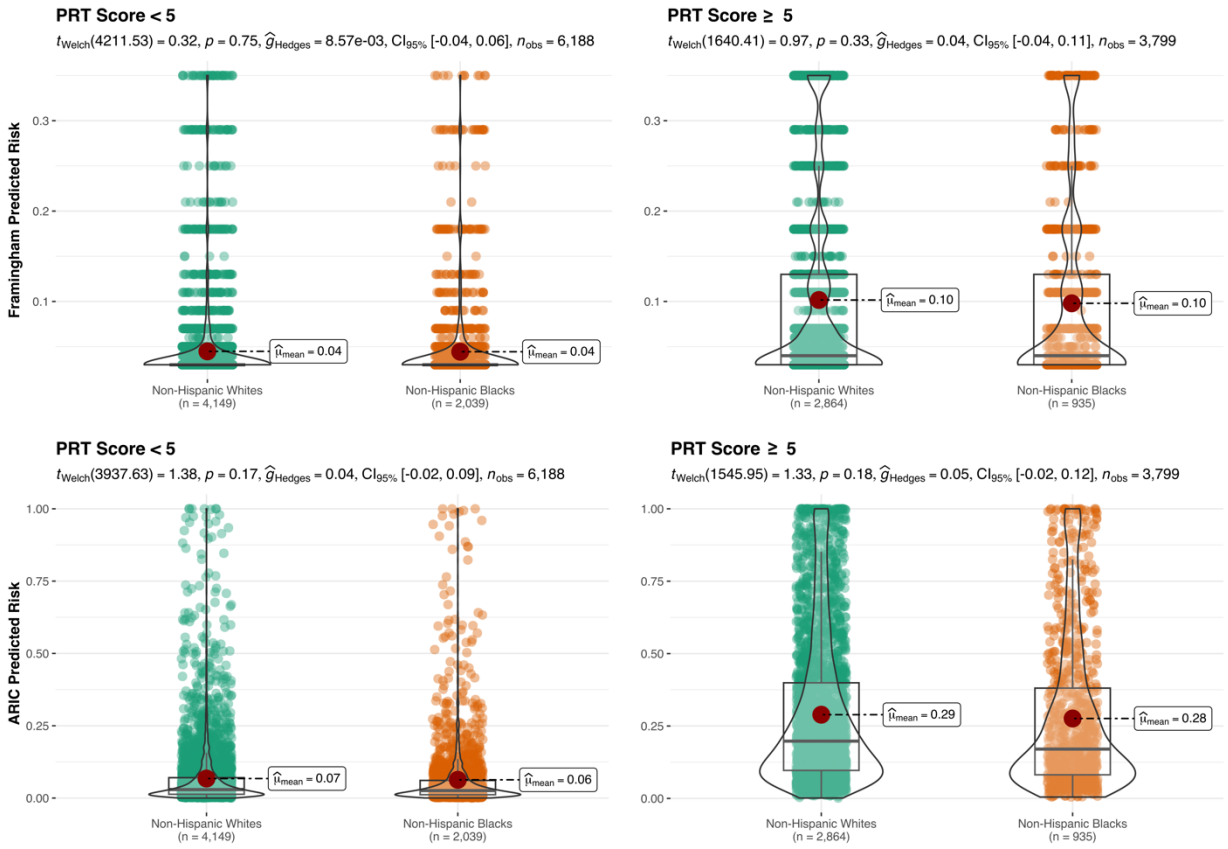
506 Abbreviations: API - average predicted incidence (proportion); CI - cumulative incidence.

507

508 **Figure 4**. Box-Violin Plots of Framingham Offspring Risk Score and ARIC Model individual

509 predicted risks in groups scoring less or greater than 5 in the Prediabetes Risk Test.

510



511

512 Group differences were estimated using Welch's t-tests.

513

Appendix headings

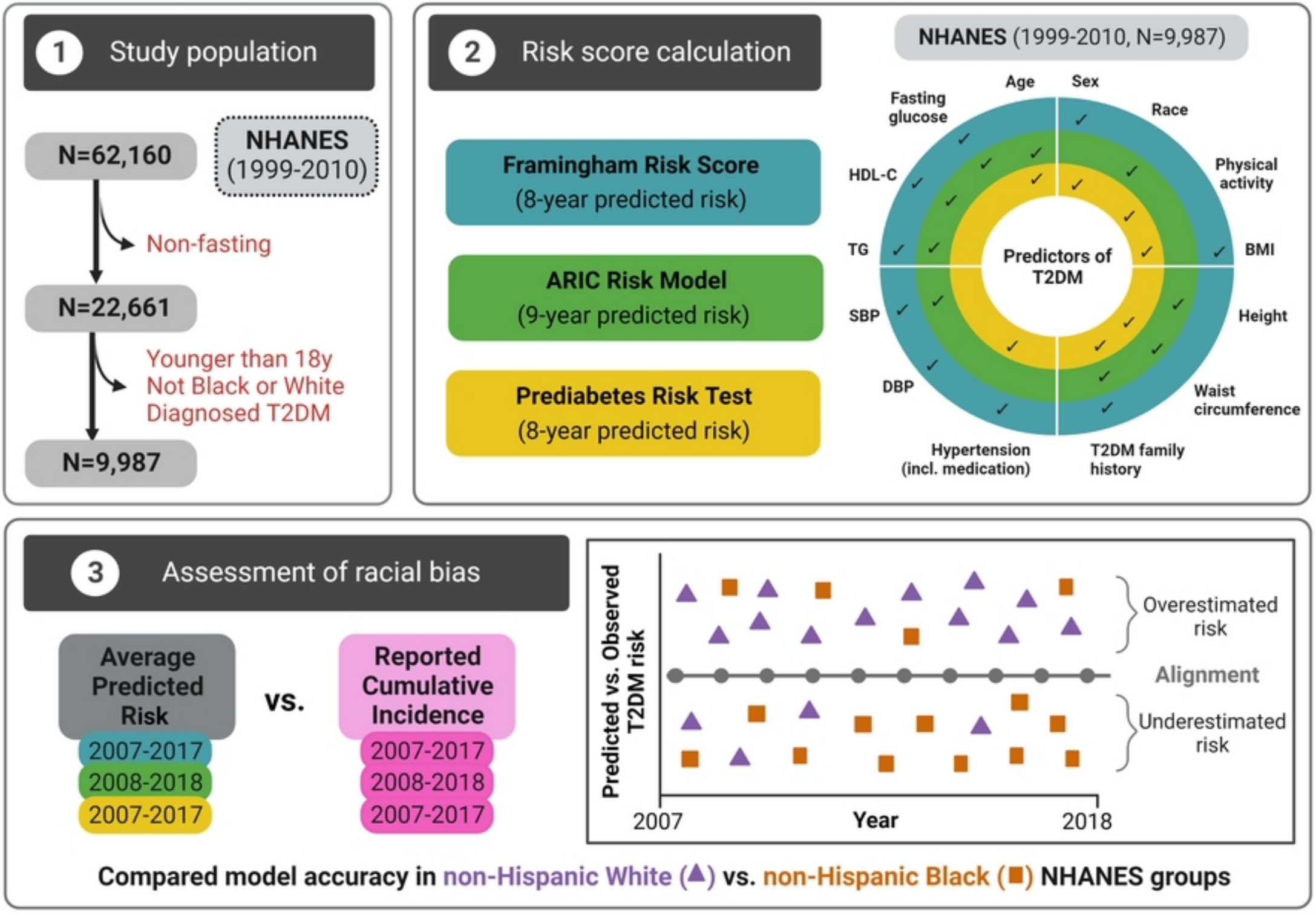

1 Study population
N=62,160
NHANES (1999-2010)
Non-fasting
N=22,661
Younger than 18y
Not Black or White
Diagnosed T2DM
N=9,987
2 Risk score calculation
Framingham Risk Score
(8-year predicted risk)
ARIC Risk Model
(9-year predicted risk)
Prediabetes Risk Test
(8-year predicted risk)
NHANES (1999-2010, N=9,987)
Age
Sex
Race
Physical activity
BMI
Height
Waist circumference
T2DM family history
Hypertension (incl. medication)
DBP
SBP
TG
HDL-C
Fasting glucose
Predictors of T2DM
3 Assessment of racial bias
Average Predicted Risk
2007-2017
2008-2018
2007-2017
vs.
Reported Cumulative Incidence
2007-2017
2008-2018
2007-2017
Predicted vs. Observed T2DM risk
Overestimated risk
Alignment
Underestimated risk
2007
Year
2018
Compared model accuracy in non-Hispanic White (▲) vs. non-Hispanic Black (■) NHANES groups

Figure 1

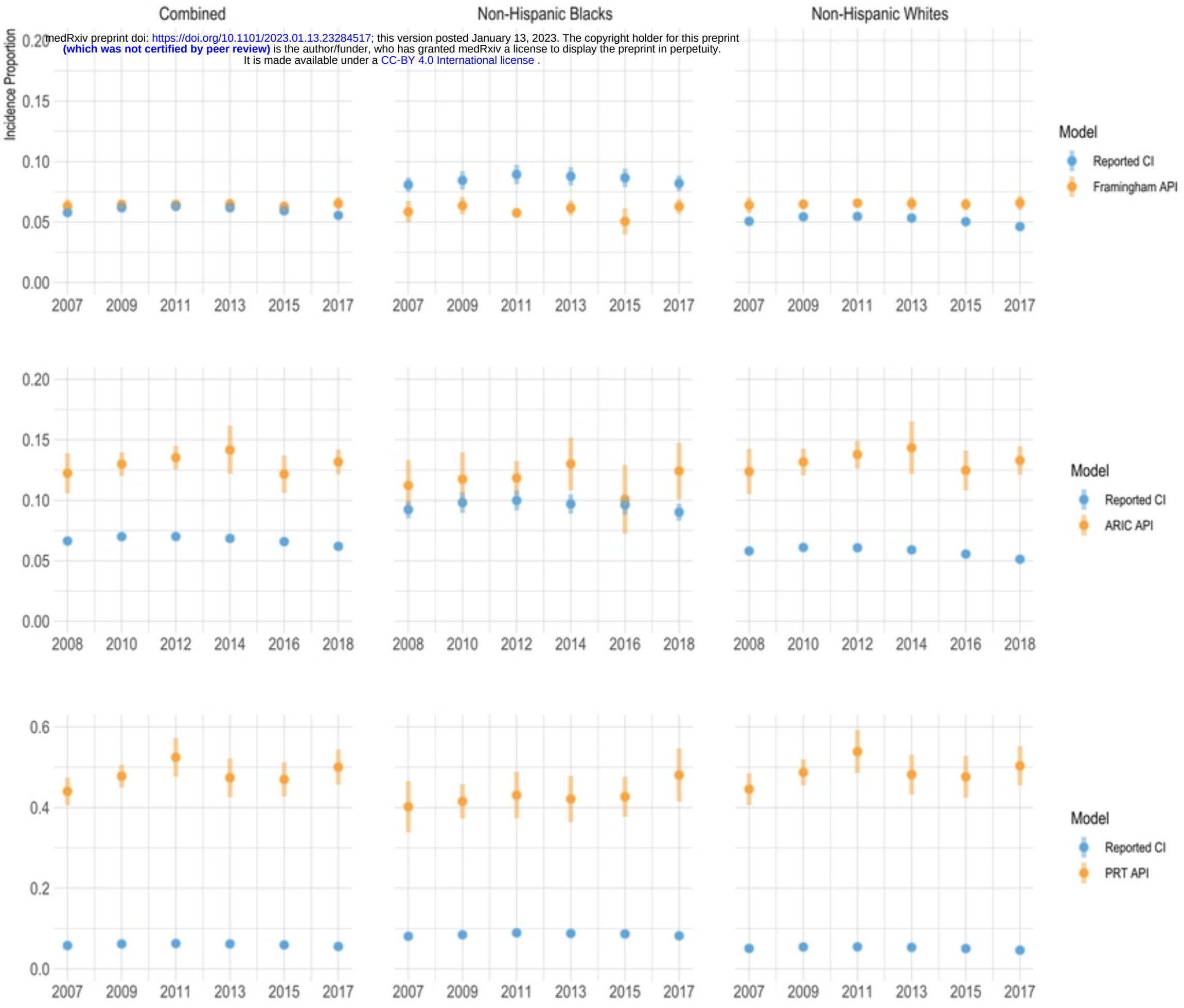


Figure 2

Figure 3

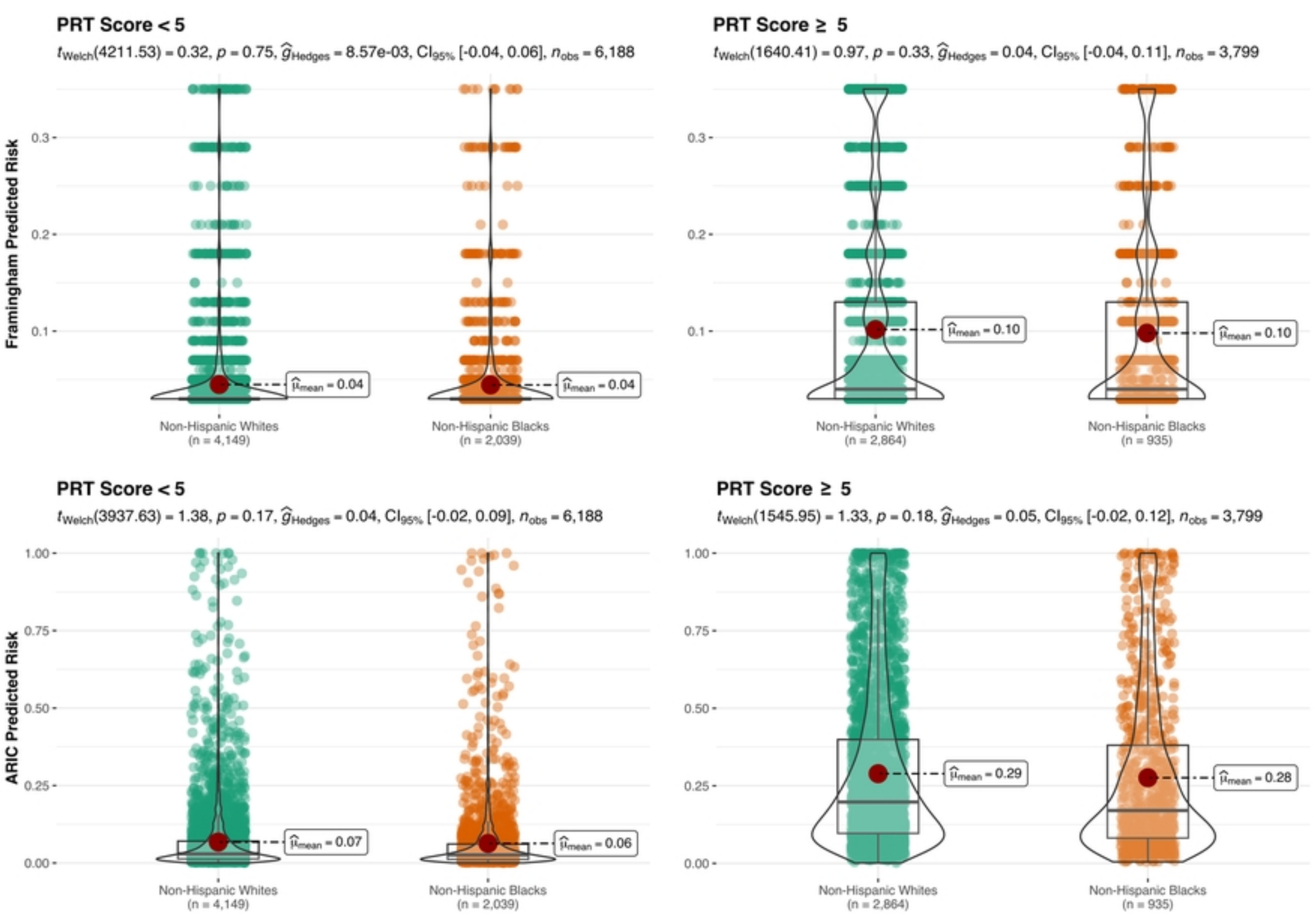

PRT Score < 5
$t_{Welch}(4211.53) = 0.32$, $p = 0.75$, $\hat{g}_{Hedges} = 8.57e\text{-}03$, $CI_{95\%}$ [-0.04, 0.06], $n_{obs} = 6,188$
Framingham Predicted Risk
0.3
0.2
0.1
$\hat{\mu}_{mean} = 0.04$
$\hat{\mu}_{mean} = 0.04$
Non-Hispanic Whites
(n = 4,149)
Non-Hispanic Blacks
(n = 2,039)
PRT Score ≥ 5
$t_{Welch}(1640.41) = 0.97$, $p = 0.33$, $\hat{g}_{Hedges} = 0.04$, $CI_{95\%}$ [-0.04, 0.11], $n_{obs} = 3,799$
0.3
0.2
0.1
$\hat{\mu}_{mean} = 0.10$
$\hat{\mu}_{mean} = 0.10$
Non-Hispanic Whites
(n = 2,864)
Non-Hispanic Blacks
(n = 935)
PRT Score < 5
$t_{Welch}(3937.63) = 1.38$, $p = 0.17$, $\hat{g}_{Hedges} = 0.04$, $CI_{95\%}$ [-0.02, 0.09], $n_{obs} = 6,188$
ARIC Predicted Risk
1.00
0.75
0.50
0.25
0.00
$\hat{\mu}_{mean} = 0.07$
$\hat{\mu}_{mean} = 0.06$
Non-Hispanic Whites
(n = 4,149)
Non-Hispanic Blacks
(n = 2,039)
PRT Score ≥ 5
$t_{Welch}(1545.95) = 1.33$, $p = 0.18$, $\hat{g}_{Hedges} = 0.05$, $CI_{95\%}$ [-0.02, 0.12], $n_{obs} = 3,799$
1.00
0.75
0.50
0.25
0.00
$\hat{\mu}_{mean} = 0.29$
$\hat{\mu}_{mean} = 0.28$
Non-Hispanic Whites
(n = 2,864)
Non-Hispanic Blacks
(n = 935)

Figure 4