

## RESEARCH ARTICLE

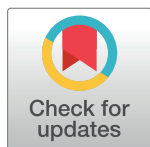
# Predicting incident cardiovascular disease among African-American adults: A deep learning approach to evaluate social determinants of health in the Jackson heart study

Matthew C. Morris<sup>1,2\*</sup>, Hamidreza Moradi<sup>3,4</sup>, Maryam Aslani<sup>5</sup>, Mario Sims<sup>6</sup>, David Schlundt<sup>7</sup>, Chrystyna D. Kouros<sup>8</sup>, Burel Goodin<sup>9,10</sup>, Crystal Lim<sup>11</sup>, Kerry Kinney<sup>7</sup>

**1** Department of Anesthesiology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **2** Department of Psychiatry and Human Behavior, University of Mississippi Medical Center, Jackson, Mississippi, United States of America, **3** Department of Data Science, University of Mississippi Medical Center, Jackson, Mississippi, United States of America, **4** Department of Computer Science, University of North Carolina Agricultural and Technical State University, Greensboro, North Carolina, United States of America, **5** Department of Data Analytics, University of North Texas, Denton, Texas, United States of America, **6** Department of Social Medicine, Population, and Public Health, University of California, Riverside, California, United States of America, **7** Department of Psychology, Vanderbilt University, Nashville, Tennessee, United States of America, **8** Department of Psychology, Southern Methodist University, Dallas, Texas, United States of America, **9** Department of Psychology, University of Alabama at Birmingham, Birmingham, Alabama, Texas, United States of America, **10** Department of Anesthesiology, Washington University in St. Louis, St. Louis, Missouri, United States of America, **11** Department of Health Psychology, University of Missouri, Columbia, Missouri, Texas, United States of America

\* These authors contributed equally to this work.

\* [matthew.c.morris@vumc.org](mailto:matthew.c.morris@vumc.org)



## OPEN ACCESS

**Citation:** Morris MC, Moradi H, Aslani M, Sims M, Schlundt D, Kouros CD, et al. (2023) Predicting incident cardiovascular disease among African-American adults: A deep learning approach to evaluate social determinants of health in the Jackson heart study. PLoS ONE 18(11): e0294050. <https://doi.org/10.1371/journal.pone.0294050>

**Editor:** Jonathan Turner, Luxembourg Institute of Health, LUXEMBOURG

**Received:** January 12, 2023

**Accepted:** October 24, 2023

**Published:** November 10, 2023

**Copyright:** © 2023 Morris et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** This dataset contains patient level health records with intellectual property rights held by The Jackson Heart Study, which is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National

## Abstract

The present study sought to leverage machine learning approaches to determine whether social determinants of health improve prediction of incident cardiovascular disease (CVD). Participants in the Jackson Heart study with no history of CVD at baseline were followed over a 10-year period to determine first CVD events (i.e., coronary heart disease, stroke, heart failure). Three modeling algorithms (i.e., Deep Neural Network, Random Survival Forest, Penalized Cox Proportional Hazards) were used to evaluate three feature sets (i.e., demographics and standard/biobehavioral CVD risk factors [FS1], FS1 combined with psychosocial and socioeconomic CVD risk factors [FS2], and FS2 combined with environmental features [FS3]) as predictors of 10-year CVD risk. Contrary to hypothesis, overall predictive accuracy did not improve when adding social determinants of health. However, social determinants of health comprised eight of the top 15 predictors of first CVD events. The social determinates of health indicators included four socioeconomic factors (insurance status and types), one psychosocial factor (discrimination burden), and three environmental factors (density of outdoor physical activity resources, including instructional and water activities; modified retail food environment index excluding alcohol; and favorable food stores). Findings suggest that whereas understanding biological determinants may identify who is currently at risk for developing CVD and in need of secondary prevention,

Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). Jackson Heart Study data may be requested from the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) repository and the database of Genotypes and Phenotypes (dbGaP). Additionally, investigators with a manuscript proposal or ancillary study proposal that has been approved by study committees may request data directly from the Jackson Heart Study Coordinating Center at the University of Mississippi Medical Center. BioLINCC: <https://biolincc.nhlbi.nih.gov/studies/jhs/> dbGaP: <https://www.ncbi.nlm.nih.gov/gap/> Jackson Heart Study Coordinating Center: <https://www.jacksonheartstudy.org/Research/Study-Data/Data-Access>.

**Funding:** The Jackson Heart Study is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C, and grant R01 HL117285 from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities.

**Competing interests:** The authors have declared that no competing interests exist.

**Abbreviations:** CVD, cardiovascular disease; CHD, coronary heart disease; MI, myocardial infarction; NHB, Non-Hispanic Black; NHW, Non-Hispanic White; SES, socioeconomic status; JHS, Jackson Heart Study; HF, heart failure; ML, machine learning; DNN, Deep Neural Network; RSF, Random Survival Forest; CPH, Cox Proportional Hazards; SHAP, Shapley Additive Explanation; SDOH, social determinants of health.

understanding upstream social determinants of CVD risk could guide primary prevention efforts by identifying where and how policy and community-level interventions could be targeted to facilitate changes in individual health behaviors.

## Introduction

Cardiovascular disease (CVD), which includes fatal and nonfatal coronary heart disease (CHD), myocardial infarction (MI), and stroke, is the leading cause of death in the United States, accounting for 1 in 4 deaths [1,2]. Staggering racial disparities exist in CVD morbidity and mortality, with Non-Hispanic Black (NHB) adults exhibiting the highest CVD mortality across all ages compared to other racial and ethnic groups [3]. In 2017, CHD death rates (per 100,000) for adults ages 35 and older were 204 and 182 for NHB and Non-Hispanic White (NHW) adults, respectively [4]. Overall, mortality rates for NHB as compared to NHW adults are estimated to be 30% higher for CHD and 45% higher for stroke [5].

There is now strong empirical support for the association between social determinants of health (SDOH) and CVD risk [6]. Social determinants refer to the environments in which people are born, live, and age, and include socioeconomic status (SES), social support, neighborhood and housing conditions, exposure to stressors and discrimination, and access to quality education, food, and health care. Socioeconomic and environmental risk factors for CVD lie upstream relative to more downstream, individual-level behavioral and biological risk factors. Increased risk for CVD morbidity and mortality has been linked to a host of social determinants, including lower per capita household income [7], higher rates of neighborhood poverty, higher levels of neighborhood violence and crime exposure [8], unemployment, greater percentages of single family households, overcrowding, greater racial segregation [8], lower levels of perceived social support [9], reduced access to medical care [6,8,10,11], limited neighborhood walkability and access to public open spaces [12], and the presence of food deserts [13]. The added value of social determinants as predictors of CVD above and beyond traditional biobehavioral risk factors (e.g., smoking, systolic/diastolic blood pressure, eating behaviors) has yet to be determined.

Psychosocial factors such as stress and discrimination are also considered social determinants of risk for CVD and typically lie downstream relative to socioeconomic and environmental factors. Adults who report experiencing four or more early life stressors are 2.2 times more likely to develop CHD and 2.4 times more likely to develop stroke [14]. Meta-analytic findings also suggest a medium effect size for the association between early adversity and CVD [15]. Stressors, such as discrimination and being the victim of an assault, are associated with increased risk for CVD (e.g., elevated blood pressure, atherosclerosis) and mortality [16–20]. Conversely, growing up in an environment with lower stress levels and higher SES is associated with lower CVD risk in adulthood [21]. According to minority stress theory [22], racial disparities in CVD are driven, in part, by disproportionately high exposure to stressors across the lifespan [6,23–28]. Most cross-sectional studies find higher rates of stressful events in NHB compared to NHW adults, though few carefully control for SES [29]. Moreover, relations between discrimination and CVD appear stronger for NHB than non-minority adults [16,30]. In the Jackson Heart Study (JHS), greater risk for CVD has been associated with higher negative affect (i.e., depressive symptoms, cynicism, anger) and stress levels (i.e., weekly stress levels, major life events) [31–36].

Racial disparities in CVD are complex and multi-faceted [37], yet studies have primarily examined risk and protective factors separately. One significant barrier to this research is that

traditional analytic approaches do not include the wide array of cross-domain (e.g., environment, behavior, biology) and cross-level (geographic, socioeconomic, interpersonal, individual) exposures implicated in CVD risk, which exert statistically significant but weak individual effects. Data-driven techniques based on machine learning (ML) are well-suited for CVD risk prediction because they can overcome the restrictive modeling assumptions and limitations on number of predictors that characterize traditional multivariable regression approaches. Despite the promise of SDOH for improving CVD risk prediction beyond traditional risk scores (e.g., Framingham) that ignore SES [38], few studies have determined whether these predictors add predictive value to standard/biobehavioral CVD risk factors. One study using ML showed that greater area-level social service resources (i.e., food, employment, and nutrition) were associated with lower CVD risk (i.e., lower body mass index [BMI]) [39]. The present study addressed key gaps in the extant research by adopting ML to evaluate the predictive performance of social determinants for first CVD events among NHB adults in the JHS across multiple levels of their social ecology; these predictors included experiences of discrimination, low SES, neighborhood violence, and nearby physical activity facilities [26,40–44]. We hypothesized that models including social determinants (i.e., psychosocial, socioeconomic, and environmental factors) would exhibit superior predictive accuracy for incident CVD as compared to models including only standard/biobehavioral CVD risk factors.

## Materials and methods

### Participants

The JHS is a longitudinal cohort study focused on understanding the emergence of CVD [45]. During the baseline assessment, which occurred between March 2000 and September 2004, NHB adults, ages 21 to 94, were recruited from the tri-county area (Hinds, Madison, and Rankin) of the Jackson, Mississippi metropolitan area. Participants were excluded if they had a history of CVD at baseline as evident in any of the following conditions: self-reported history of MI; self-reported history of cardiac procedure; self-reported history of physician-diagnosed stroke; history of CHD from electrocardiogram and self-report; and self-reported history of carotid angioplasty. All JHS participants provided written informed consent and JHS was approved by the Institutional Review Boards of The University of Mississippi Medical Center, Jackson State University, and Tougaloo College.

**CVD outcomes.** CVD events, which included CHD (i.e., definite or probable MI, definite fatal CHD, cardiac procedures), stroke (definite or probable), and heart failure (HF; JHS surveillance and event adjudication started on January 1, 2005), were carefully documented and verified through data linkage hospital discharge lists and National Death Index and review of medical records of CVD-related hospitalizations and death certificates to adjudicate CVD events and deaths [46]. First CVD events were determined over a 10-year follow-up period.

### CVD risk factors

**Standard/Biobehavioral risk factors.** Standard/biobehavioral CVD risk factors assessed at baseline included BMI, systolic and diastolic blood pressure, ankle brachial index, blood pressure medication status, total cholesterol, low density lipoprotein (LDL) cholesterol, high density lipoprotein (HDL) cholesterol, triglycerides, fasting glucose, hemoglobin A1C (HbA1c), alcohol drinking, smoking, physical activity, diet, age, sex, and waist circumference. BMI was determined as weight (kilograms) divided by height (meters squared). Hypertension status was derived from the Joint National Committee on Prevention (JNC-7) and defined as systolic blood pressure  $\geq 140$ , diastolic blood pressure  $\geq 90$ , or use of antihypertensive medications [47]. Diabetes status was derived from the American Diabetes Association (ADA)

criteria and defined as fasting glucose  $\geq 126$  mg/dL, HbA1c  $\geq 6.5\%$ , or use of diabetic medication within 2 weeks of clinic visit [48]. Cholesterol measures included fasting LDL level (mg/dL), fasting HDL level (mg/dL), and total fasting cholesterol level (mg/dL). Additional CVD biospecimens included fasting triglyceride level (mg/dL), HbA1c (National Glycohemoglobin Standardization Program units [%]), and fasting plasma glucose level (mg/dL). Medications (anti-hypertensive, anti-diabetic) were determined by self-report. Alcohol use was assessed by self-report (i.e., frequency of use in the past 12 months). Smoking was assessed by self-report tobacco use forms (27-items) and used to determine current and history of cigarette smoking; prior work suggests CVD risk remains elevated in former-smokers for 3–15 years compared to those with no history of smoking [49,50]. Physical activity was determined by the Physical Activity (PA) scale from the Active Living Index (30-item [51]), and computed as a categorical variable (0 = *poor health* (0 min/week of physical activity); 1 = *intermediate health* (1–49 min/week of moderate activity or 1–74 min/week of vigorous activity or 1–149 min/week of moderate+vigorous activity); 2 = *ideal health* ( $\geq 150$  min/week of moderate activity or  $\geq 75$  min/week of vigorous activity or  $\geq 150$  min/week of moderate+vigorous activity). Diet was determined by the 158-item Food Frequency Questionnaire (FFQ). The following components of a 2000-kcal diet were used to determine nutrition categories (i.e., *poor health* = 0–1 components; *intermediate health* = 2–3 components; *ideal health* = 4–5 components) based on American Heart Association guidelines [52]: (1)  $\geq 4.5$  cups/day of fruits and vegetables; (2)  $> 3.5$  ounces twice/week of fish; (3)  $< 1500$  mg/day of sodium; (4)  $< 450$  kcal/week of sugary beverages; (5)  $\geq 3$  servings/day of whole grains).

**Social determinants: Psychosocial factors.** Psychosocial factors assessed at baseline included perceived daily discrimination, lifetime discrimination, burden of lifetime discrimination, perceived depressive symptoms, and perceived stress levels. Daily discrimination was assessed with a 9-item measure based on the scale developed by Williams and colleagues [53] (good internal consistency: alpha = 0.88) [54]. Lifetime discrimination was determined through a self-report measure assessing the occurrence of unfair treatment across 9 domains (adequate internal consistency: alpha = 0.78) [31,54]. Burden of lifetime discrimination (i.e., interference related to discrimination) was assessed by a measure exhibiting adequate internal consistency (alpha = 0.63). Depressive symptoms were determined by the 20-item [55] version of the Center for Epidemiological Studies Depression Scale (CES-D), with higher scores reflecting greater depression severity in the prior week (alpha = 0.82). Perceived stress was determined by the Global Perceived Stress Scale (GPSS; 8 items), adapted for the JHS from other validated stress measures [56] with adequate psychometric properties (alpha = 0.72) [34]. Higher scores reflected greater perceived stress levels over a 12-month period across multiple domains (e.g., employment, relationships, neighborhood, basic needs). In addition, the Weekly Stress Inventory (WSI; 87 items [57]) was used to capture minor stressors (e.g., work tasks, finances, household tasks, relationships) experienced by participants. Higher WSI-impact scores reflected greater stress ratings for events occurring in the past week. The WSI has excellent psychometric properties (alpha = 0.98) [34].

**Social determinants: Socioeconomic factors.** Socioeconomic factors assessed at baseline included family income (categories: less than \$5,000; \$5,000–7,999; \$8,000–11,999; \$12,000–15,999; \$16,000–19,999; \$20,000–24,999; \$25,000–34,999; \$35,000–49,999; \$50,000–74,999; \$75,000–99,999; \$100,000 or more), occupation (U.S. Department of Labor Standard Occupational Classifications: management/professional, service, sales, farming, construction, production, military, sick, unemployed, homemaker, retired, student, other), education (0 = less than high school; 1 = high school graduate/GED; 2 = attended vocational school, trade school, or college), and insurance status (any insurance; insurance type [uninsured, public only, private only, private & public]).

**Social determinants: Environmental factors.** A complete list of the environmental factors (census tract-level indicators) that comprise SDOH is included in the [S1 Table](#) [8,58]. Measures varied according to spatial function (i.e., simple or kernel density) and area (i.e., ½ mile, 1 mile, or 3 mile radius). Environmental data reflecting densities of physical activity resources and “favorable” food stores were obtained from the National Establishment Time-Series (NET-S) and Nieslen/TDLinx Service Supermarket Retail Category databases for the years 2000 to 2010, and linked to JHS baseline data using geocoded participant addresses as described elsewhere [59,60]. Favorable food store density was based on the density of groceries, supermarket chains and non-chain stores, and fruit and vegetable markets. Unfavorable food store density was based on the density of convenience stores, bakeries, candy/nut shops, ice cream stores, liquor stores, alcoholic drinking places, and fast food stores [59]. Physical activity facility density was based on the following resources: biking, bowling, dance, golf, indoor conditioning, physical activity instruction, swimming, team and racquet sports, and water activities. Environmental factors in the present study included Census-derived measures of median household income, percentage living below poverty, percentage black non-Hispanic residents, percentage white non-Hispanic residents, favorable food stores within 3 miles, physical activity facilities within 3 miles, percentage residential land use per square mile, and population density. Self-report measures of environmental characteristics assessed neighborhood problems (i.e., age- and sex-adjusted scale including participant reports of excessive noise, heavy traffic or speeding cars, lack of access to adequate food and/or shopping, lack of parks and playgrounds, trash and litter, and lacking or poorly maintained sidewalks in their neighborhoods), neighborhood social cohesion (i.e., age- and sex-adjusted participant reports of living in a close knit neighborhood, people willing to help neighbors, neighbors generally getting along, neighbors who can be trusted, neighbors who share the same values, neighborhood safety from crime), and neighborhood violence (age- and sex-adjusted scale including items assessing how often participant reported fights with weapons, violent arguments, gang fights, sexual assaults or rapes, and/or robbery or muggings in their neighborhoods).

## Data curation

We used a light-touch approach to participant exclusion and feature imputation to avoid affecting subsequent model interpretations with any prior assumptions. Data manipulation was minimized in the process of training and testing models. Accordingly, only participants with a history of CVD at baseline were excluded. Missing features imputation were conducted with constant values (medians) to preserve the informativeness of each feature in the dataset [61]. ML-based imputation strategies were eschewed due to concerns that learning from the existing correlational structure could affect the determination of feature importance. Preliminary analyses using Python’s Scikit-learn iterative random forest imputer [62] did not improve model accuracy. As a result, all missing values were imputed using the median and no missing features resulted in participant removal from analyses.

## Feature sets

To evaluate the effect of different sets of features on the model’s prediction accuracy, three feature sets were considered. In the first Feature Set (FS1), demographics and standard/biobehavioral CVD risk factors were considered as predictive features; the second Feature Set (FS2) combined psychosocial and socioeconomic features with FS1, and the third Feature Set (FS3) utilized FS2 along with environmental features.



## Modeling algorithms and evaluation

We evaluated three modeling algorithms for predicting 10-year CVD risk among JHS patients. For each algorithm, different feature sets as inputs were used to model the survival function. Comparisons between the models' estimated risks were conducted using Antolini time-dependent Concordance Index (CI) to account for non-proportional hazard model used in this study [63,64]. Please note that Antolini CI is equivalent to 'Harrell's C' for survival models with proportional hazards [65,66]. HyperOpt [67]—an open-source Bayesian optimization library—was used to address models' sensitivity to hyper-parameters, increase the models' accuracy, and to facilitate study replication. Hyper-parameter tuning was performed on random train, validation, and test splits of 60%, 20%, and 20%, respectively. Following hyper-parameter tuning, model evaluation was conducted by 10-fold cross-validation to report the model's average CI. All implementations were conducted in Python 3.8 using PyTorch 1.10, PyCox 0.2.3, and Scikit-Survival 0.17.2. The following three modeling algorithms were used in the present study:

**Deep neural network.** Neural Network (NN)-based models have been shown to improve prediction accuracy [68]. The present study implemented Deep Neural Network (DNN) models based on DeepHit [65]. DeepHit learns the distribution of survival time directly from the data without any prior assumption(s) about the underlying stochastic process. As a result, predictions depend directly on features in the dataset. The loss function of DeepHit is designed to handle censored data for survival analysis. For DeepHit, the number of layers, number of neurons in each layer, dropout rate, optimization algorithm, learning rate, activation function, and batch size are all considered as hyper-parameters and optimized.

**Random survival forest (RSF).** As an ensemble of tree-based learners, this algorithm ensures individual trees are de-correlated. Each tree is built on a bootstrap of the original training dataset and split criteria in each node include a random subset of features [69]. Final prediction results comprise the combined predictions from all trained trees. For this algorithm, we used scikit-survival implementation with number of estimators and max depth as hyper-parameters [70].

**Penalized cox proportional hazards (CPH).** This algorithm was included as a comparison model due to its ease of implementation and low computational requirement. Penalized CPH models were implemented in scikit-survival with regularization parameter and convergence criteria as hyper-parameters.

## Model interpretability

ML models are often viewed as black-box procedures yielding little insight or interpretability except for predictions of outcomes. However, recent improvements have been made in the generation of robust and interpretable insights from complex ML models [71]. Shapley Additive Explanation (SHAP [72]) values have gained attention because they can facilitate interpretation of complex ML models with high accuracy and robustness. By comparing SHAP values generated for input features, it is possible to assess the extent to which changes in the inputs influence the final model's prediction and, hence, to evaluate feature importance for complex models. The present study evaluated standard/biobehavioral risk factors and social determinants as features in a complex DNN model for CVD risk prediction. After optimizing model hyperparameters, training, and evaluating the model accuracy, the model was subsequently retrained on the full dataset using the same parameters; in this manner, the model could learn all existing interactions in the dataset. As recommended by SHAP best practice guidelines for calculation of the required background samples, we applied the K-Nearest Neighbor clustering algorithm ( $K = 100$ ) to the dataset; this provided a total of 100 cluster centroids to be used for

Table 1. Descriptive characteristics for patients with and without incident CVD.

Baseline Characteristic	Incident CVD (n = 382)	No Incident CVD (n = 3,598)	All (n = 3,980)	Missing values
Sex				0
Male (%)	154 (40)	1,265 (35)	1,419 (36)	
Female (%)	228 (60)	2,333 (65)	2,561 (64)	
Age (SD)	62.6 (11)	52.8 (12)	53.8 (12)	0
Education				12
Less than high school (%)	117 (30)	510 (14)	627 (15)	
Greater than High school (%)	265 (70)	3,076 (86)	3,341 (85)	
Occupation				8
Employed (%)	381 (>99)	3,569 (>99)	3,950 (>99)	
Not Employed (%)	1 (<1)	21 (<1)	22 (<1)	
Smoking status				
Current smoker (%)	61 (16)	395 (11)	456 (11)	36
Ever smoker (%)	149 (39)	1,019 (28)	1,168 (29)	9
Insurance Status				17
Yes (%)	348 (91%)	3,081 (86%)	3429 (86%)	
No (%)	33 (9%)	501 (14%)	534 (14%)	
BMI (SD)	31.9 (7)	31.6 (7)	31.6 (7)	7

<https://doi.org/10.1371/journal.pone.0294050.t001>

SHAP value calculation. SHAP values were then generated for input features based on the trained model and the background samples, providing insight into feature importance. Feature importance was then reported as the absolute mean value of the effect on the final model prediction.

## Results

### Study population

During the 10-year follow-up period of NHB adults with no history of CVD at baseline (n = 3,980), 382 participants experienced at least one CVD incident: there were 139 cases of incident CHD, 221 cases of incident heart failure, and 104 cases of incident stroke. Event was defined as the first adjudication of any CVD incident. Descriptive characteristics for NHB adults with and without CVD are presented in Table 1. Participants exhibited a mean age of 53.8 years and 64% were female.

**Model accuracy.** Penalized CPH, RSF, and DNN models were evaluated using CI with three sets of features as inputs (FS1, FS2, FS3); mean CI of 10-fold cross-validation and corresponding standard deviations for models are presented in Table 2. The DNN model exhibited

Table 2. Concordance indices for predictive models.

Feature sets	Deep Neural Network	Random Survival Forest	Penalized CPH Regression
	Mean (SD)	Mean (SD)	Mean (SD)
Standard risk factors (FS1)	0.76 (0.03)	0.75 (0.07)	0.74 (0.06)
Psychosocial/socioeconomic + standard risk factors (FS2)	0.76 (0.03)	0.75 (0.06)	0.74 (0.06)
Environmental + Psychosocial/socioeconomic + standard risk factors (FS3)	0.76 (0.03)	0.72 (0.05)	0.72 (0.06)

Note: CPH = Cox proportional hazards.

<https://doi.org/10.1371/journal.pone.0294050.t002>

the highest overall accuracy across feature sets. With higher numbers of input features, the accuracy of the RSF and CPH models decreased. In contrast, the DNN model exhibited consistent performance regardless of the number of input features, and detected the best set of features predictive of the outcome of interest in a high dimensional space.

### Feature importance

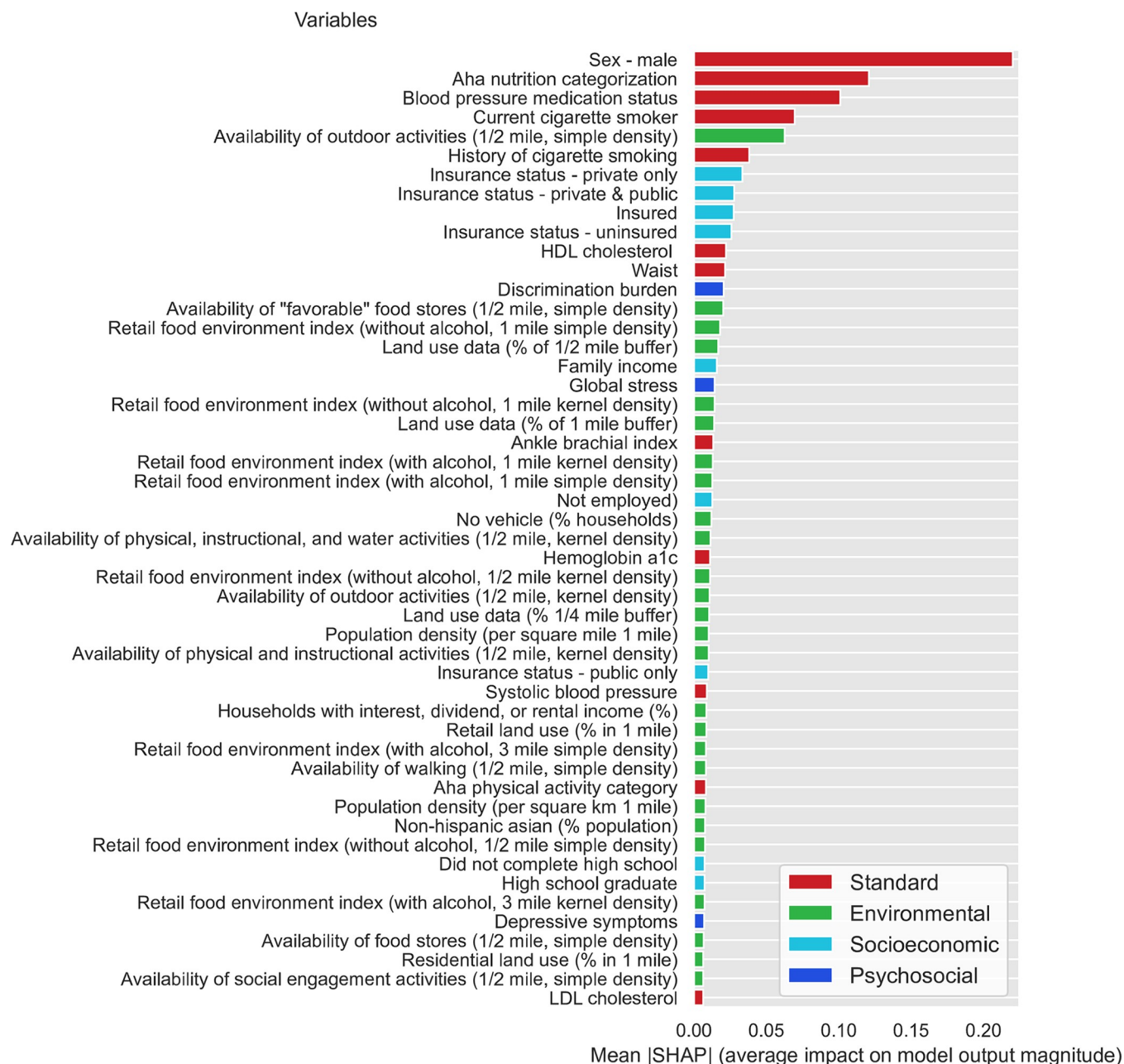
To evaluate feature importance, we calculated the SHAP values for all the features in the dataset using the DNN model trained on the FS3 feature set (i.e., all standard/biobehavioral, psychosocial/socioeconomic, and environmental factors). [Fig 1](#) presents the relative importance (average impact on final model output) for the top 50 features investigated in this study, sorted by their absolute mean SHAP value (relative importance for all features is presented in the [S1 Table](#)). Seven of the top 15 features—including five of the top 10—were standard/biobehavioral CVD risk factors, including sex (male), nutrition, blood pressure medication status, cigarette smoking status (current smoker, history of smoking), HDL cholesterol, and waist circumference. Notably, 8 of the top 15 features were SDOH. These included four socioeconomic factors (insurance status and type), one psychosocial factor (discrimination burden), and three environmental factors. The latter factors included area-level composite variables that reflect density of outdoor physical activity resources (including instructional and water activities) and favorable food stores (i.e., grocery stores, supermarket chains and non-chain stores, fruit and vegetable markets). Relative importance is also presented separately for standard CVD risk ([Fig 2](#)), socioeconomic ([Fig 3](#)), psychosocial ([Fig 4](#)), and environmental ([Fig 5](#)) features.

### Discussion

Based on prior work, the extent to which SDOH—including psychosocial, socioeconomic, and environmental factors—can improve predictive accuracy for incident CVD beyond standard/biobehavioral risk factors was unclear. To address this gap, the present study used ML models to determine overall predictive performance for incident CVD events (i.e., CHD, stroke, and/or HF) among NHB adults followed over time in the JHS, and to assess the relative importance of standard/biobehavioral and social determinant features in these models. The DNN model provided more accurate predictions regarding incident CVD than RSF or CPH models. Contrary to our hypothesis, overall predictive accuracy for DNN models did not improve when adding SDOH. Whereas accuracy was stable across feature sets for DNN models, decreases in accuracy were observed for RSF and CPH models with higher numbers of inputs. This was likely due to high dimensionality of the input datasets and inability of RSF or CPH models to accurately detect important features required to maintain or increase accuracy. Taken together, these results highlight the promise of DNN over RSF and CPH models for predicting incident CVD, but suggest that the psychosocial, socioeconomic, and environmental factors did not appreciably improve predictive accuracy for first CVD events beyond biobehavioral risk factors.

Recent work demonstrates improved prediction of first fatal or non-fatal CVD events using ML algorithms as compared to conventional statistical approaches focused on standard risk factors or scores typically derived from routinely collected clinical data [73–78]. In addition, ML approaches have shown that neighborhood-level predictors (e.g., prevalence of obesity, rates of binge drinking and leisure-time physical activity) were associated with higher rates of CHD and stroke [79]. Predictive accuracy for DNN in the present study was comparable to another study using Neural Networks to predict first CVD events in 423,604 participants in the UK Biobank (AUC-ROC: 0.755, 95% CI: 0.750–0.760); the latter study included 4,801 incident CVD cases within 5 years of baseline assessment and a host of features that overlapped





**Fig 1. Relative importance for the top 50 study features sorted by mean absolute SHAP value.**

<https://doi.org/10.1371/journal.pone.0294050.g001>

with the present study (e.g., diet, physical activity, sociodemographics, lipid profile, body composition, depressive symptoms) but notably did not include environmental factors [73]. The following sections explain the reasons the inclusion of psychosocial (e.g., stress levels and depressive symptoms), socioeconomic (e.g., family income and educational attainment), and environmental (e.g., neighborhood poverty) risk factors did not improve overall predictive performance for first CVD events beyond biobehavioral risk factors (e.g., diet, blood pressure, smoking).

First, it is important to note that even though overall predictive accuracy was not improved by adding social determinants as input features, this does not imply that these features are not

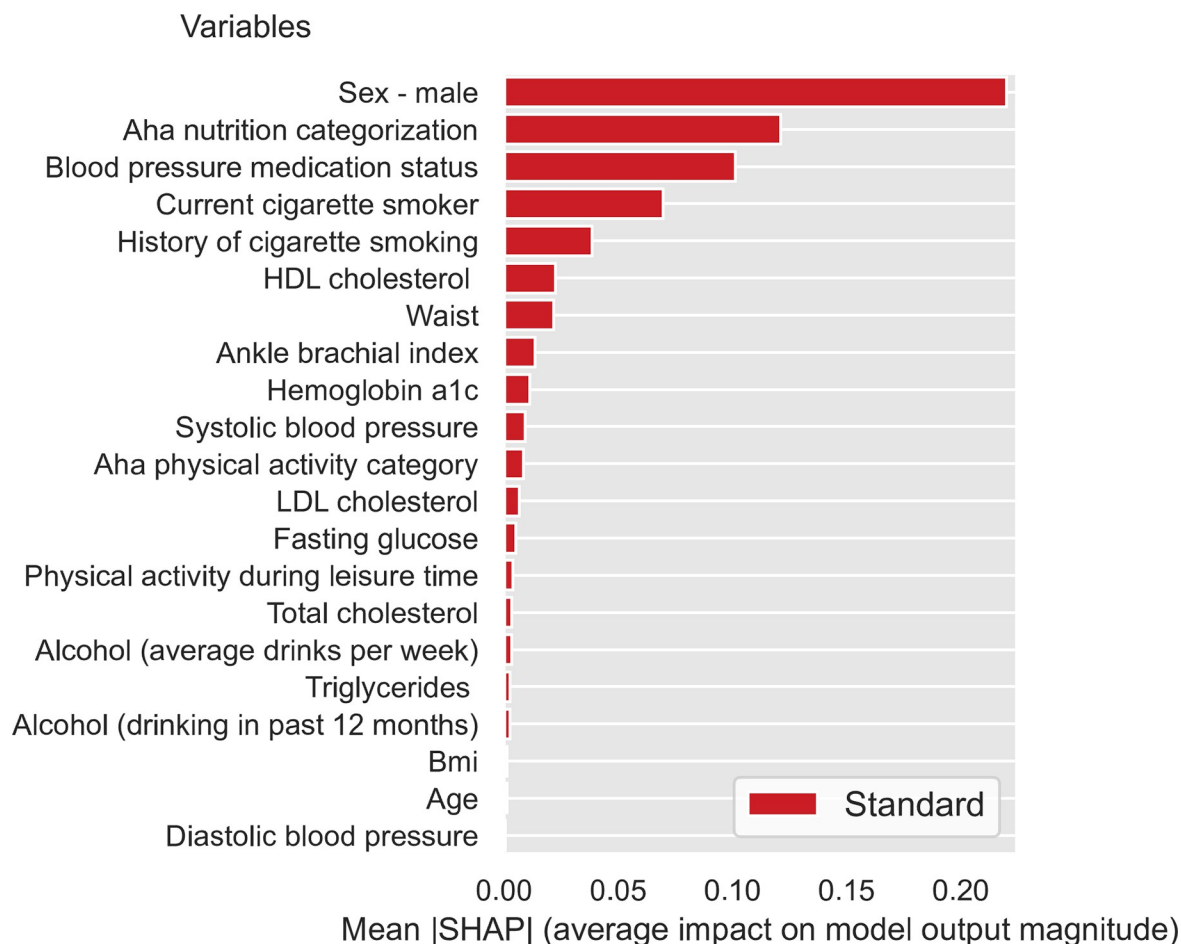


Fig 2. Relative importance for standard CVD risk features sorted by mean absolute SHAP value.

<https://doi.org/10.1371/journal.pone.0294050.g002>

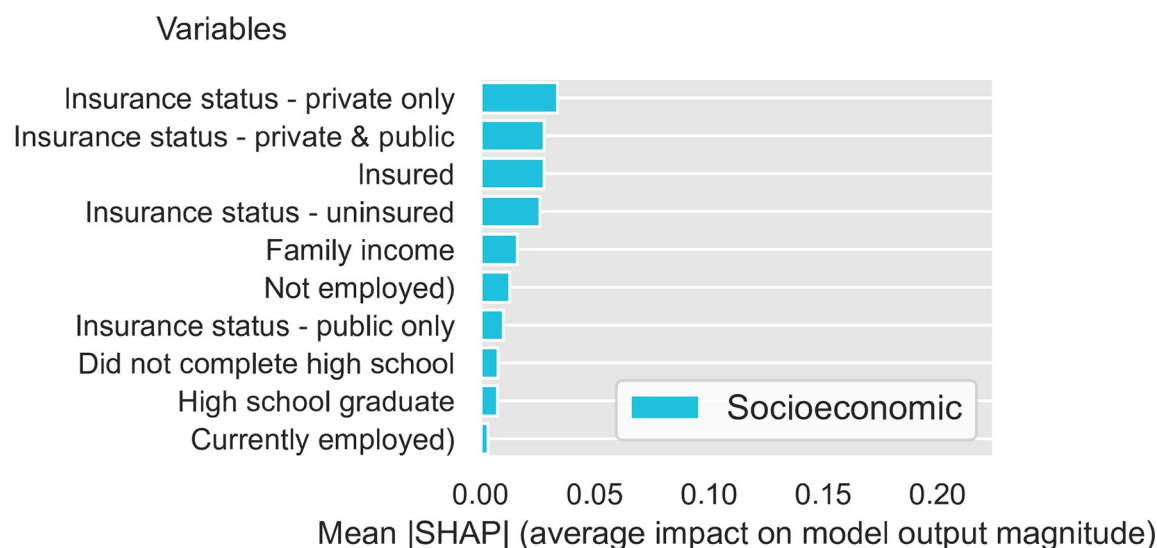
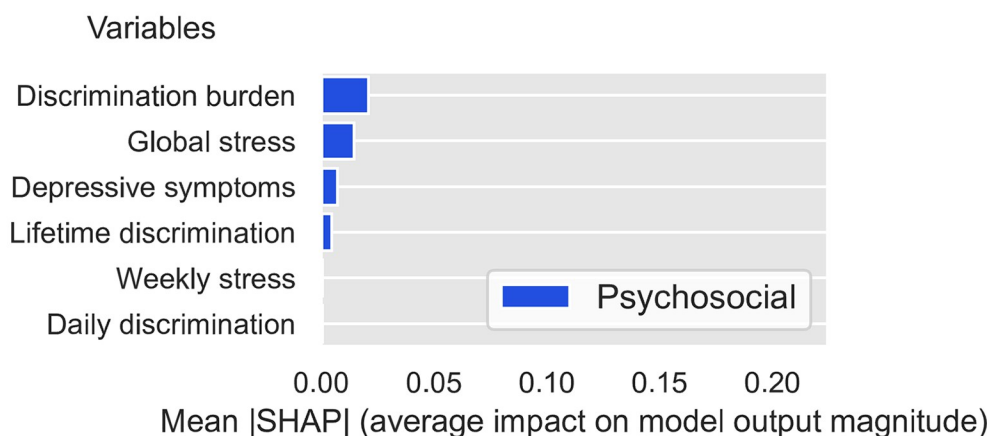


Fig 3. Relative importance for socioeconomic features sorted by mean absolute SHAP value.

<https://doi.org/10.1371/journal.pone.0294050.g003>



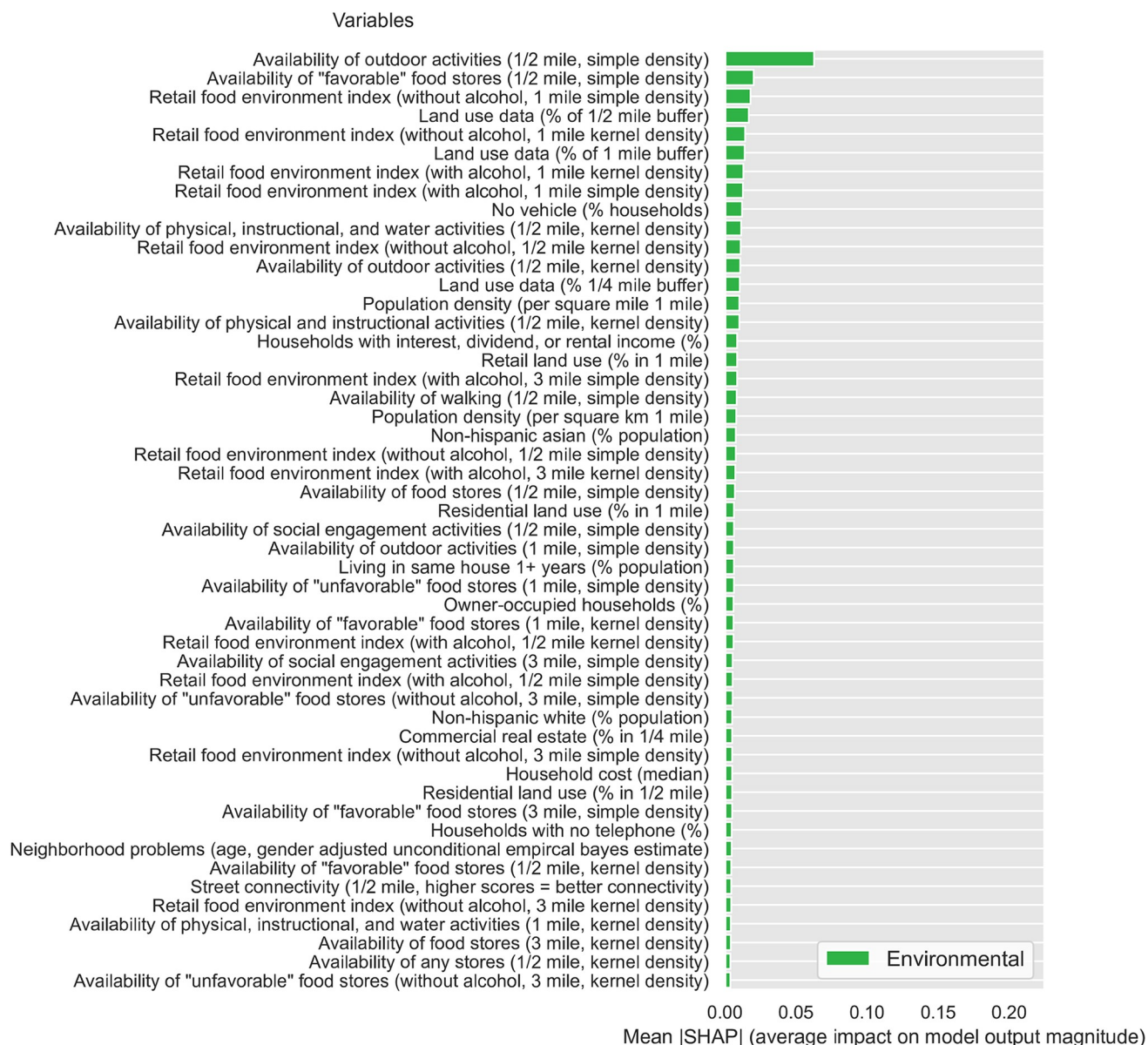
**Fig 4. Relative importance for psychosocial features sorted by mean absolute SHAP value.**

<https://doi.org/10.1371/journal.pone.0294050.g004>

important predictors of incident CVD. Analyses of feature importance (i.e., Shapley Additive Explanation [SHAP] values) can aid interpretation of complex DNN models and complement overall indicators of predictive accuracy (concordance index [CI]) by providing information on the relative importance of input features. SHAP values showed that standard/biobehavioral risk factors comprised seven of the top 15 predictors of first CVD events. However, the relative importance of psychosocial, socioeconomic, and environmental factors cannot be discounted. Discrimination burden, insurance status, and outdoor physical activity resources and ranked higher in importance than well-established standard/biobehavioral risk factors such as HbA1C, systolic blood pressure, physical activity levels, and LDL cholesterol [80].

Second, conceptual models depict pathways linking upstream (e.g., economic stability, neighborhood environment, structural discrimination, and access to education, health care, and healthy food) and midstream (e.g., exposure to stressors, experiences with discrimination, health behaviors, diet) social determinants to downstream (e.g., hypertension, obesity, lipid profiles, HbA1c) risk factors for CVD [81,82]. These models highlight “trickle-down effects” of socio-contextual factors on social position, lived experiences, biobehavioral responses, and, ultimately, CVD development and progression for marginalized groups [82]. One interpretation of the present findings is that this ‘trickle’ takes time: whereas biological determinants tell us who is currently at risk for developing CVD and in need of secondary prevention, social determinants tell us who may be at risk for developing CVD and could benefit from primary prevention. If social determinants exert their influence through biological determinants, then they may be less useful for forecasting incident CVD and more useful as targets for preventive policy, community, and individual interventions.

Third, a priori distinctions made in the present study between standard/biobehavioral, psychosocial, socioeconomic, and environmental factors may ignore other important dimensions and sources of cross-category overlap. For example, social determinants are likely to differ according to their timing and duration, with features such as education and experiences with discrimination exerting a cumulative effect on CVD risk over a lifetime as compared to more proximal features (e.g., stress levels, current neighborhood violence, depressive symptoms) that influence CVD risk factors on a day-to-day basis. It is also unclear whether and to what extent correlations among input features can influence accuracy and measures of feature



**Fig 5. Relative importance for environmental features sorted by mean absolute SHAP value.**

<https://doi.org/10.1371/journal.pone.0294050.g005>

importance in Neural Network models. Standard/biobehavioral factors such as physical activity and diet are likely to be correlated with environmental features such as walkability and density of favorable food stores, respectively. In addition, the environmental factors included identical resource measures that differed only by spatial function (i.e., simple or kernel density) and/or area (i.e., 1/2 mile, 1 mile, or 3 mile radius) as well as resource measures that differed only slightly by content (e.g., food stores with and without alcohol). To our knowledge, potential problems with multicollinearity in Neural Network models—including the extent to which correlations among inputs influence non-linear activation functions and advanced regularization methods—have yet to be evaluated. Notably, strong correlations among similar variables in the present study (e.g., current smoker and history of smoking) did not prevent them from emerging simultaneously as important features in SHAP analyses.

Fourth, it should be noted that DNN models typically require large datasets to be trained effectively. Therefore, given the relatively small size of the dataset, the DNN model may not outperform less complex models such as CPH or RSF. This could be a plausible reason why the addition of SDOH did not improve the predictive accuracy of our model, despite being among the top predictors of the first CVD event. Another possible explanation is that SDOH may not have a significant impact on predicting first CVD event beyond the predictive contribution of other risk factors. We hypothesize that this may be due to the complex and multifactorial nature of CVD risk. Finally, it is possible that the attainable accuracy for the provided dataset has been reached. In this regard, we acknowledge that the accuracy we achieved is similar to other studies utilizing full cohort [83].

The present findings have important implications for efforts to prevent the onset of CVD. First, more upstream SDOH are unlikely to improve accuracy of ML models that seek to distinguish NHB adults in terms of their risk for first CVD events over a 10-year follow-up period. While it is possible that assessing the cumulative impact of sociopolitical and economic factors on CVD risk at higher levels of the social ecology would improve predictive accuracy over a longer time frame, our results suggest that careful assessment of SDOH may not yield dividends for more immediate risk stratification over and above standard/biobehavioral measures. Second, our results suggest that identifying connections between upstream (e.g., neighborhood resources for physical activity) and midstream (e.g., stress levels) determinants [81] could prove useful for primary prevention of CVD. Neighborhoods identified as having higher rates of upstream social determinants of CVD risk could be prioritized for delivery of preventive interventions. Neighborhood features identified as important for incident CVD prediction could then be targeted by policy and community-level interventions to facilitate changes in—and remove barriers to—individual health and wellness behaviors.

## Limitations

Limitations of the present study may provide directions for future research. First, as noted above, the degree to which strong correlations among feature inputs influences SHAP values for relative importance remains unclear. Second, the relative importance of features with higher levels of missing data may be underestimated in the present study despite the use of median imputation. The top three features based on missingness among JHS participants were depressive symptoms ( $n = 1,309$ ), weekly stress levels ( $n = 1,663$ ), and percent retail land use within a 1/4 mile radius ( $n = 580$ ). Hence, the importance of these features for predicted CVD risk should be interpreted with caution. Third, the adjudication of HF in the JHS was initiated in 2005 despite enrollment beginning in 2000. Hence, there may have been a small number of participants for whom HF events were not captured and who may have been misclassified. Fourth, focusing primarily on middle-to-older aged adults and including a relatively short (10 year) follow-up period may have favored biobehavioral over other risk factors. Studies following younger individuals over a longer time frame may be better-suited to capturing more gradual effects of SDOH on incident CVD.

## Conclusion

The present findings highlight important features associated with risk for incident CVD across multiple levels of the social ecology and argue in favor of a biopsychosocial approach to CVD risk and prevention. Although SDOH did not augment predictive accuracy for CVD events over a 10-year period, they emerged as important features of predictive models that ranked above even well-established behavioral and biological risk factors. Future studies should leverage ML approaches to feature importance in order to guide prevention efforts by identifying



salient points along the stream where an individual's risk for developing subsequent CVD may be diverted.

## Supporting information

**S1 Table. Relative importance of all features sorted by absolute mean SHAP values.**  
(DOCX)

## Author Contributions

**Conceptualization:** Matthew C. Morris, Hamidreza Moradi, Mario Sims.

**Data curation:** Hamidreza Moradi.

**Formal analysis:** Hamidreza Moradi.

**Methodology:** Matthew C. Morris, Hamidreza Moradi, Mario Sims, David Schlundt.

**Project administration:** Matthew C. Morris.

**Visualization:** Hamidreza Moradi, Maryam Aslani.

**Writing – original draft:** Matthew C. Morris, Hamidreza Moradi, Mario Sims, David Schlundt, Chrystyna D. Kouros, Crystal Lim, Kerry Kinney.

**Writing – review & editing:** Matthew C. Morris, Hamidreza Moradi, Maryam Aslani, Mario Sims, David Schlundt, Chrystyna D. Kouros, Burel Goodin, Crystal Lim, Kerry Kinney.

## References

1. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2020 update: a report from the American Heart Association. *Circulation*. 2020; 141(9):e139–e596. <https://doi.org/10.1161/CIR.0000000000000757> PMID: 31992061
2. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Blaha MJ, et al. Executive Summary: Heart Disease and Stroke Statistics-2014 Update A Report From the American Heart Association. *Circulation*. 2014; 129(3):399–410. <https://doi.org/10.1161/01.cir.0000442015.53336.12> PMID: 24446411
3. Mensah GA, Mokdad AH, Ford ES, Greenlund KJ, Croft JB. State of disparities in cardiovascular health in the United States. *Circulation*. 2005; 111(10):1233–41. <https://doi.org/10.1161/01.CIR.0000158136.76824.04> PMID: 15769763
4. Vaughan AS, Schieb L, Casper M. Historic and recent trends in county-level coronary heart disease death rates by race, gender, and age group, United States, 1979–2017. *PloS one*. 2020; 15(7): e0235839. <https://doi.org/10.1371/journal.pone.0235839> PMID: 32634156
5. Underlying cause of death 1999–2019 on CDC WONDER Online Database [Internet]. Centers for Disease Control and Prevention,. 2020.
6. Havranek EP, Mujahid MS, Barr DA, Blair IV, Cohen MS, Cruz-Flores S, et al. Social Determinants of Risk and Outcomes for Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation*. 2015; 132(9):873–98. <https://doi.org/10.1161/CIR.0000000000000228> PMID: 26240271
7. Centers for Disease Control and Prevention (CDC). Racial/ethnic and socioeconomic disparities in multiple risk factors for heart disease and stroke: United States, 2003. *Mmwr-Morbidity and Mortality Weekly Report*. 2005; 54:113–7. PMID: 15703691
8. Barber S, Hickson DA, Wang X, Sims M, Nelson C, Diez-Roux AV. Neighborhood Disadvantage, Poor Social Conditions, and Cardiovascular Disease Incidence Among African American Adults in the Jackson Heart Study. *American journal of public health*. 2016; 106(12):2219–26. <https://doi.org/10.2105/AJPH.2016.303471> PMID: 27736207
9. Barth J, Schneider S, von Kanel R. Lack of Social Support in the Etiology and the Prognosis of Coronary Heart Disease: A Systematic Review and Meta-Analysis. *Psychosomatic Medicine*. 2010; 72(3):229–38. <https://doi.org/10.1097/PSY.0b013e3181d01611> PMID: 20223926

10. Daniel M, Moore S, Kestens Y. Framing the biosocial pathways underlying associations between place and cardiometabolic disease. *Health Place*. 2008; 14(2):117–32. <https://doi.org/10.1016/j.healthplace.2007.05.003> PMID: 17590377
11. O'Rand AM, Hamil-Luker J. Processes of cumulative adversity: Childhood disadvantage and increased risk of heart attack across the life course. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences*. 2005; 60:117–24. [https://doi.org/10.1093/geronb/60.special\\_issue\\_2.s117](https://doi.org/10.1093/geronb/60.special_issue_2.s117) PMID: 16251582
12. Paquet C, Coffee NT, Haren MT, Howard NJ, Adams RJ, Taylor AW, et al. Food environment, walkability, and public open spaces are associated with incident development of cardio-metabolic risk factors in a biomedical cohort. *Health & Place*. 2014; 28:173–6.
13. Larson C, Haushalter A, Buck T, Campbell D, Henderson T, Schlundt D. Development of a Community-Sensitive Strategy to Increase Availability of Fresh Fruits and Vegetables in Nashville's Urban Food Deserts, 2010–2012. *Preventing Chronic Disease*. 2013; 10:12. <https://doi.org/10.5888/pcd10.130008> PMID: 23886044
14. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults—The adverse childhood experiences (ACE) study. *American Journal of Preventive Medicine*. 1998; 14(4):245–58.
15. Wegman HL, Stetler C. A meta-analytic review of the effects of childhood abuse on medical outcomes in adulthood. *Psychosom Med*. 2009; 71(8):805–12. <https://doi.org/10.1097/PSY.0b013e3181bb2b46> PMID: 19779142
16. Lewis TT, Barnes LL, Bienias JL, Lackland DT, Evans DA, de Leon CFM. Perceived Discrimination and Blood Pressure in Older African American and White Adults. *Journals of Gerontology Series a-Biological Sciences and Medical Sciences*. 2009; 64(9):1002–8. <https://doi.org/10.1093/gerona/glp062> PMID: 19429703
17. Troxel WM, Matthews KA, Bromberger JT, Sutton-Tyrrell K. Chronic stress burden, discrimination, and subclinical carotid artery disease in African American and Caucasian women. *Health Psychology*. 2003; 22(3):300–9. <https://doi.org/10.1037/0278-6133.22.3.300> PMID: 12790258
18. Barnes LL, de Leon CFM, Lewis TT, Bienias JL, Wilson RS, Evans DA. Perceived discrimination and mortality in a population-based study of older adults. *American Journal of Public Health*. 2008; 98(7):1241–7. <https://doi.org/10.2105/AJPH.2007.114397> PMID: 18511732
19. Davidson K, Jonas BS, Dixon KE, Markovitz JH. Do depression symptoms predict early hypertension incidence in young adults in the CARDIA study? *Archives of Internal Medicine*. 2000; 160(10):1495–500.
20. Mezuk B, Eaton WW, Albrecht S, Golden SH. Depression and Type 2 Diabetes Over the Lifespan A meta-analysis. *Diabetes Care*. 2008; 31(12):2383–90. <https://doi.org/10.2337/dc08-0985> PMID: 19033418
21. Juonala M, Pulkki-Råback L, Elovainio M, Hakulinen C, Magnussen CG, Sabin MA, et al. Childhood Psychosocial Factors and Coronary Artery Calcification in Adulthood. *JAMA Pediatrics*. 2016; 170(5):466.
22. Meyer IH. Prejudice as stress: Conceptual and measurement problems. *American Journal of Public Health*. 2003; 93(2):262–5. <https://doi.org/10.2105/ajph.93.2.262> PMID: 12554580
23. Gibbons FX, Stock ML. Perceived racial discrimination and health behavior: Mediation and moderation. *The Oxford handbook of stigma, discrimination, and health*. 2018:355–77.
24. Kuzawa CW, Sweet E. Epigenetics and the embodiment of race: developmental origins of US racial disparities in cardiovascular health. *Am J Hum Biol*. 2009; 21(1):2–15. <https://doi.org/10.1002/ajhb.20822> PMID: 18925573
25. Wickrama KA, O'Neal CW, Lott RE. Early community contexts, race/ethnicity and young adult CVD risk factors: the protective role of education. *J Community Health*. 2012; 37(4):781–90. <https://doi.org/10.1007/s10900-011-9511-2> PMID: 22101680
26. Myers HF. Ethnicity- and socio-economic status-related stresses in context: an integrative review and conceptual model. *Journal of behavioral medicine*. 2009; 32(1):9–19. <https://doi.org/10.1007/s10865-008-9181-4> PMID: 18989769
27. West CM. Black women and intimate partner violence—New directions for research. *Journal of interpersonal violence*. 2004; 19(12):1487–93. <https://doi.org/10.1177/0886260504269700> PMID: 15492062
28. Turner RJ, Lloyd DA. Stress burden and the lifetime incidence of psychiatric disorder in young adults—Racial and ethnic contrasts. *Arch Gen Psychiatry*. 2004; 61(5):481–8. <https://doi.org/10.1001/archpsyc.61.5.481> PMID: 15123493

29. Hatch SL, Dohrenwend BP. Distribution of traumatic and other stressful life events by race/ethnicity, gender, SES and age: A review of the research. *American Journal of Community Psychology*. 2007; 40(3–4):313–32. <https://doi.org/10.1007/s10464-007-9134-z> PMID: 17906927
30. Guyll M, Matthews KA, Bromberger JT. Discrimination and unfair treatment: Relationship to cardiovascular reactivity among African American and European American women. *Health Psychology*. 2001; 20(5):315–25. <https://doi.org/10.1037/0278-6133.20.5.315> PMID: 11570645
31. Sims M, Diez-Roux AV, Dudley A, Gebreab S, Wyatt SB, Bruce MA, et al. Perceived discrimination and hypertension among African Americans in the Jackson Heart Study. *American journal of public health*. 2012; 102(S2):S258–S65. <https://doi.org/10.2105/AJPH.2011.300523> PMID: 22401510
32. Sims M, Glover LM, Norwood AF, Jordan C, Min Y-I, Brewer LC, et al. Optimism and cardiovascular health among African Americans in the Jackson Heart Study. *Preventive medicine*. 2019; 129:105826. <https://doi.org/10.1016/j.ypmed.2019.105826> PMID: 31473218
33. Sims M, Glover LSM, Gebreab SY, Spruill TM. Cumulative psychosocial factors are associated with cardiovascular disease risk factors and management among African Americans in the Jackson Heart Study. *BMC Public Health*. 2020; 20(1):566. <https://doi.org/10.1186/s12889-020-08573-0> PMID: 32345300
34. Sims M, Liptford KJ, Patel N, Ford CD, Min YI, Wyatt SB. Psychosocial Factors and Behaviors in African Americans: The Jackson Heart Study. *American Journal of Preventive Medicine*. 2017; 52(1):S48–S55. <https://doi.org/10.1016/j.amepre.2016.09.020> PMID: 27989292
35. Ford CD, Sims M, Higginbotham JC, Crowther MR, Wyatt SB, Musani SK, et al. Psychosocial factors are associated with blood pressure progression among African Americans in the Jackson Heart Study. *American Journal of Hypertension*. 2016; 29(8):913–24. <https://doi.org/10.1093/ajh/hpw013> PMID: 26964661
36. Gebreab SY, Diez-Roux AV, Hickson DA, Boykin S, Sims M, Sarpong DF, et al. The contribution of stress to the social patterning of clinical and subclinical CVD risk factors in African Americans: the Jackson Heart Study. *Social science & medicine*. 2012; 75(9):1697–707.
37. National Institute on Minority Health and Health Disparities. NIMHD Research Framework. 2017.
38. Kino S, Hsu Y-T, Shiba K, Chien Y-S, Mita C, Kawachi I, et al. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM-population health*. 2021; 15:100836. <https://doi.org/10.1016/j.ssmph.2021.100836> PMID: 34169138
39. Berkowitz SA, Basu S, Venkataramani A, Reznor G, Fleegler EW, Atlas SJ. Association between access to social service resources and cardiometabolic risk factors: a machine learning and multilevel modeling analysis. *BMJ open*. 2019; 9(3):e025281. <https://doi.org/10.1136/bmjopen-2018-025281> PMID: 30862634
40. Galobardes B, Smith GD, Lynch JW. Systematic review of the influence of childhood socioeconomic circumstances on risk for cardiovascular disease in adulthood. *Annals of Epidemiology*. 2006; 16(2):91–104. <https://doi.org/10.1016/j.annepidem.2005.06.053> PMID: 16257232
41. Carter RT. Racism and Psychological and Emotional Injury: Recognizing and Assessing Race-Based Traumatic Stress. *The Counseling Psychologist*. 2007; 35(1):13–105.
42. Sorenson SB. Violence against women—Examining ethnic differences and commonalities. *Eval Rev*. 1996; 20(2):123–45.
43. Dyson JL. The effect of family violence on children's academic performance and behavior. *J Natl Med Assoc*. 1990; 82(1):17–22. PMID: 2304094
44. Williams DR, Lawrence JA, Davis BA. Racism and health: evidence and needed research. *Annual review of public health*. 2019; 40:105–25. <https://doi.org/10.1146/annurev-publhealth-040218-043750> PMID: 30601726
45. Taylor HA Jr, Wilson JG, Jones DW, Sarpong DF, Srinivasan A, Garrison RJ, et al. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn Dis*. 2005; 15(4 Suppl 6):S6–4. PMID: 16320381
46. Gebreab SY, Diez Roux AV, Brenner AB, Hickson DA, Sims M, Subramanyam M, et al. The impact of lifecourse socioeconomic position on cardiovascular disease events in African Americans: the Jackson Heart Study. *Journal of the American Heart Association*. 2015; 4(6):e001553. <https://doi.org/10.1161/JAHA.114.001553> PMID: 26019130
47. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr, et al. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: the JNC 7 report. *Jama*. 2003; 289(19):2560–71. <https://doi.org/10.1001/jama.289.19.2560> PMID: 12748199
48. Association AD. Diagnosis and classification of diabetes mellitus. *Diabetes care*. 2010; 33(Supplement 1):S62–S9.

49. Mannan H, Stevenson C, Peeters A, Walls H, McNeil J. Framingham risk prediction equations for incidence of cardiovascular disease using detailed measures for smoking. *Heart international*. 2010; 5(2). <https://doi.org/10.4081/hi.2010.e11> PMID: 21977296
50. Honjo K, Iso H, Tsugane S, Tamakoshi A, Satoh H, Tajima K, et al. The effects of smoking and smoking cessation on mortality from cardiovascular disease among Japanese: pooled analysis of three large-scale cohort studies in Japan. *Tobacco control*. 2010; 19(1):50–7. <https://doi.org/10.1136/tc.2009.029751> PMID: 20008160
51. Smitherman TA, Dubbert PM, Grothe KB, Sung JH, Kendzor DE, Reis JP, et al. Validation of the Jackson Heart Study physical activity survey in African Americans. *Journal of Physical Activity and Health*. 2009; 6(s1):S124–S32. <https://doi.org/10.1123/jpah.6.s1.s124> PMID: 19998858
52. Lloyd-Jones DM, Hong Y, Labarthe D, Mozaffarian D, Appel LJ, Van Horn L, et al. Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond. *Circulation*. 2010; 121(4):586–613. <https://doi.org/10.1161/CIRCULATIONAHA.109.192703> PMID: 20089546
53. Williams DR, Yu Y, Jackson JS, Anderson NB. Racial differences in physical and mental health: Socio-economic status, stress and discrimination. *Journal of health psychology*. 1997; 2(3):335–51. <https://doi.org/10.1177/135910539700200305> PMID: 22013026
54. Sims M, Wyatt SB, Gutierrez ML, Taylor HA, Williams DR. Development and psychometric testing of a multidimensional instrument of perceived discrimination among African Americans in the Jackson Heart Study. *Ethnicity & disease*. 2009; 19(1):56. PMID: 19341164
55. Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. *Applied psychological measurement*. 1977; 1(3):385–401.
56. Payne TJ, Wyatt SB, Mosley TH, Dubbert PM, Guterrez-Mohammed ML, Calvin RL, et al. Sociocultural methods in the Jackson Heart Study: conceptual and descriptive overview. *Ethn Dis*. 2005; 15(4 Suppl 6):S6–38. PMID: 16317984
57. Brantley PJ, Waggoner CD, Jones GN, Rappaport NB. A daily stress inventory: Development, reliability, and validity. *Journal of behavioral medicine*. 1987; 10(1):61–73. <https://doi.org/10.1007/BF00845128> PMID: 3586002
58. Barber S, Hickson DA, Kawachi I, Subramanian S, Earls F. Neighborhood disadvantage and cumulative biological risk among a socioeconomically diverse sample of African American adults: an examination in the Jackson Heart Study. *Journal of racial and ethnic health disparities*. 2016; 3(3):444–56. <https://doi.org/10.1007/s40615-015-0157-0> PMID: 27294737
59. Gebreab SY, Hickson DA, Sims M, Wyatt SB, Davis SK, Correa A, et al. Neighborhood social and physical environments and type 2 diabetes mellitus in African Americans: The Jackson Heart Study. *Health & place*. 2017; 43:128–37. <https://doi.org/10.1016/j.healthplace.2016.12.001> PMID: 28033588
60. Auchincloss AH, Moore KA, Moore LV, Roux AVD. Improving retrospective characterization of the food environment for a large region in the United States during a historic time period. *Health & place*. 2012; 18(6):1341–7. <https://doi.org/10.1016/j.healthplace.2012.06.016> PMID: 22883050
61. Josse J, Prost N, Scornet E, Varoquaux G. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:190206931*. 2019.
62. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. 2011; 12:2825–30.
63. Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. *Statistics in medicine*. 2005; 24(24):3927–44. <https://doi.org/10.1002/sim.2427> PMID: 16320281
64. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *arXiv preprint arXiv:190700825*. 2019.
65. Lee C, Zame W, Yoon J, Van Der Schaar M, editors. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI conference on artificial intelligence*; 2018.
66. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Jama*. 1982; 247(18):2543–6. PMID: 7069920
67. Bergstra J, Yamins D, Cox D, editors. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *International conference on machine learning*; 2013: PMLR.
68. Xiang A, Lapuerta P, Ryutov A, Buckley J, Azen S. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Computational statistics & data analysis*. 2000; 34(2):243–57.
69. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *The annals of applied statistics*. 2008; 2(3):841–60.

70. Pölsterl S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *J Mach Learn Res.* 2020; 21(212):1–6.
71. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access.* 2018; 6:52138–60.
72. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems.* 2017; 30.
73. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, Van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PloS one.* 2019; 14(5):e0213653. <https://doi.org/10.1371/journal.pone.0213653> PMID: 31091238
74. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one.* 2017; 12(4):e0174944. <https://doi.org/10.1371/journal.pone.0174944> PMID: 28376093
75. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, et al. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *Journal of the American Heart Association.* 2018; 7(8): e008081. <https://doi.org/10.1161/JAHA.117.008081> PMID: 29650709
76. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circulation research.* 2017; 121(9):1092–101. <https://doi.org/10.1161/CIRCRESAHA.117.311312> PMID: 28794054
77. Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC medical research methodology.* 2018; 18(1):1–11.
78. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. *Journal of the American Heart Association.* 2018; 7(22): e009476. <https://doi.org/10.1161/JAHA.118.009476> PMID: 30571498
79. Li Y, Liu SH, Niu L, Liu B. Unhealthy behaviors, prevention measures, and neighborhood cardiovascular health: a machine learning approach. *Journal of Public Health Management and Practice.* 2019; 25(1): E25–E8. <https://doi.org/10.1097/PHH.0000000000000817> PMID: 29889182
80. Ramírez-Vélez R, Saavedra JM, Lobelo F, Celis-Morales CA, del Pozo-Cruz B, García-Hermoso A, editors. Ideal cardiovascular health and incident cardiovascular disease among adults: a systematic review and meta-analysis. *Mayo Clinic Proceedings*; 2018: Elsevier.
81. Jilani MH, Javed Z, Yahya T, Valero-Elizondo J, Khan SU, Kash B, et al. Social determinants of health and cardiovascular disease: current state and future directions towards healthcare equity. *Current atherosclerosis reports.* 2021; 23(9):1–11. <https://doi.org/10.1007/s11883-021-00949-w> PMID: 34308497
82. Powell-Wiley TM, Baumer Y, Baah FO, Baez AS, Farmer N, Mahlobo CT, et al. Social Determinants of Cardiovascular Disease. *Circulation Research.* 2022; 130(5):782–99. <https://doi.org/10.1161/CIRCRESAHA.121.319811> PMID: 35239404
83. Fox ER, Samdarshi TE, Musani SK, Pencina MJ, Sung JH, Bertoni AG, et al. Development and Validation of Risk Prediction Models for Cardiovascular Events in Black Adults: The Jackson Heart Study Cohort. *JAMA Cardiol.* 2016; 1(1):15–25. <https://doi.org/10.1001/jamacardio.2015.0300> PMID: 27437649