Hu *et al.*
BMC Medical Informatics and Decision Making (2025) 25:53
https://doi.org/10.1186/s12911-025-02873-4

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Accounting for racial bias and social determinants of health in a model of hypertension control

Yang Hu[1], Nicholas Cordella[2], Rebecca G. Mishuris[3] and Ioannis Ch. Paschalidis[1*]

Abstract

Background Hypertension control remains a critical problem and most of the existing literature views it from a clinical perspective, overlooking the role of sociodemographic factors. This study aims to identify patients with not well-controlled hypertension using readily available demographic and socioeconomic features and elucidate important predictive variables.

Methods In this retrospective cohort study, records from 1/1/2012 to 1/1/2020 at the Boston Medical Center were used. Patients with either a hypertension diagnosis or related records (≥ 130 mmHg systolic or ≥ 90 mmHg diastolic, $n = 164{,}041$) were selected. Models were developed to predict which patients had uncontrolled hypertension defined as systolic blood pressure (SBP) records exceeding 160 mmHg.

Results The predictive model of high SBP reached an Area Under the Receiver Operating Characteristic Curve of 74.49% ± 0.23%. Age, race, Social Determinants of Health (SDoH), mental health, and cigarette use were predictive of high SBP. Being Black or having critical social needs led to higher probability of uncontrolled SBP. To mitigate model bias and elucidate differences in predictive variables, two separate models were trained for Black and White patients. Black patients face a 4.7 × higher False Positive Rate (FPR) and a 0.58 × lower False Negative Rate (FNR) compared to White patients. Decision threshold differentiation was implemented to equalize FNR. Race-specific models revealed different sets of social variables predicting high SBP, with Black patients being affected by structural barriers (e.g., food and transportation) and White patients by personal and demographic factors (e.g., marital status).

Conclusions Models using non-clinical factors can predict which patients exhibit poorly controlled hypertension. Racial and SDoH variables are significant predictors but lead to biased predictive models. Race-specific models are not sufficient to resolve such biases and require further decision threshold tuning. A host of structural socioeconomic factors are identified to be targeted to reduce disparities in hypertension control.

Keywords Hypertension, Social determinants of health, Racial bias, Machine learning

*Correspondence:
Ioannis Ch. Paschalidis
yannisp@bu.edu
Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Backgroound

Hypertension is one of the most prevalent chronic cardiovascular disorders. While nearly half of the adults with hypertension in the U.S. have Systolic Blood Pressure (SBP) over 140 mmHg, only 24% of them have their hypertension under control [1]. Lack of control can lead to adverse outcomes, such as ischemic heart disease [2], stroke, and heart failure [3]. Severely hypertensive patients typically need long-term medical care with significant cost implications.

It is also hard to ignore longstanding health inequity issues in hypertension. There is evidence linking high blood pressure and Social Determinants of Health (SDoH), including income measures [4–7]. Multiple studies have found that Black people have a higher incidence of hypertension than any other racial group in the U.S [8–11]. The confluence of all these factors creates persistent health disparities, impacting hypertension control. It becomes important to identify severe hypertension and target patients with a more urgent need of better BP management.

Most of the current hypertension predictive models focus on predicting the emergence of hypertension [12–15]. Less attention has been given to predicting poorly managed hypertension. Several of the existing prognostic models use clinical variables (e.g., past BP measurements) and laboratory results to improve prediction accuracy [14, 16]. A drawback to using such variables is that model applicability is restricted to (resource-rich) settings, where such information is available for most patients. However, in these settings lack of good hypertension control is less prevalent. Also, models with more features and higher complexity may become difficult and more costly to implement in clinical practice [17]. The goal of the present study is to develop models that rely on readily available and self-reported sociodemographic information and assess whether such information can predict poor hypertension management and enable targeted interventions.

An additional goal is to elucidate the role of racial and SDoH variables [18]. Whereas these factors are important to study, to the best of our knowledge there has been less attention on how existing disparities may affect predictive models and introduce biases in the way predictions are made. The existing studies are either related to other diseases (e.g., heart failure [19]), or focus on causal inference models in different applications and without the level of detailed SDoH information we have access to for this study [20]. In particular, there is not enough work that considered an array of SDoH variables and sought to understand their relative importance in making predictions to inform subsequent targeted support programs that promote health equity.

The definition of uncontrolled hypertension has changed over time and differs from earlier hypertension guidelines. We defined very high blood pressure as SBP exceeding 160 mmHg. Historically, that was the threshold for Stage-2 hypertension, but it has been lowered to 140 mmHg with the most recent guidelines [21]. While other thresholds such as 150 or 170 mm Hg could have been used, the specific value is less critical than the overarching goal of identifying those with uncontrolled hypertension.

The major contributions and novelties of this study are:

- We developed an easily implementable and interpretable predictive model of very high blood pressure among hypertensive patients, based only on demographic features and socioeconomic factors. We opted for not using clinical or laboratory variables, not even blood pressure measurements, so that the model could be used for most individuals, even those seen infrequently and with a limited clinical record in a health care system. The model could enable preventive actions that could range from contact with a care management professional, inviting the patient in the clinic for a hypertension consultation, to broader efforts in partnership with community support groups to address patients' individual social needs.
- We elucidated the importance of sociodemographic features in assessing hypertension control at a population level. The clinical predictive models have a significant performance gap across different racial groups and needed to be vetted and corrected to mitigate this bias.
- We identified a host of non-clinical structural SDoH factors that could targeted to help reduce disparities in hypertension control (which build on an existing body of literature linking disparities to social and environment factors).

Materials and methods

Cohort design

We extracted data from Boston Medical Center (BMC) Electronic Health Records (EHRs) from January 1, 2012 to January 1, 2020. The dataset included all patients who satisfied one or more of the following conditions: (1) patients who had a hypertension diagnosis; (2) patients with high blood pressure in their problem list; and (3) patients with at least two recorded SBP measurement exceeding 130 mmHg or a Diastolic Blood Pressure (DBP) measurement exceeding 90 mmHg.

In total, the dataset included 164,041 patients, with 85,924 female (52.38%) and 78,110 male (47.62%). Features extracted included: age, sex, race, language, marital status, SDoH factors, depression scale, cigarette use, and

ZIP code. Notably, 86.5% of the patients in the dataset do not have frequent SBP measurements, which substantiates our decision to not use clinical variables as predictive variables in the model. SDoH factors were extracted from the THRIVE survey, a custom screening program created by BMC which surveys patients on their unmet social needs in eight different domains: transportation, ability to secure caregiving for family members, ability to pay for utilities, education, food, housing, employment issues, and ability to pay for medications [18]. The details of feature collection and corresponding pre-processing steps are provided in Appendix A. All records were de-identified before analysis. The study was approved by the Boston University Medical Campus (IRB #H-32061) and Boston University Institutional Review Boards.

Sociodemographic variables

Demographics were readily available in the patients' EHRs; while basic information like age and sex were determined by birth record documentation; race, language and other sociodemographic features were self-reported. Among all patients in this dataset, 10 categories of race were recorded: American Indian, Asian, Asian Pacific Islander, Black, Hispanic (yes/no for all races), Indian, Middle Eastern, Multiracial, White, and 'other' for the rest non-mentioned races. A total of 8 languages were included: African, American, Asian, English, European, Spanish, Middle Eastern language, and other. Information on social needs was extracted from the THRIVE SDoH assessment and resource connection program developed and implemented in all ambulatory care settings at BMC [22]. We combined standardized Patient Health Questionnaires (PHQ2 and PHQ9) scores to indicate patients' mental health state as it pertains to depression symptoms [23]. The PHQ2 and PHQ9 questionnaires are part of standard screenings in primary care at BMC. The marital status variables take the following six values: 'single', 'married', 'separated', 'divorced', 'widow', and 'other'. Two cigarette smoking related variables were included in the data: whether a patient ever was a cigarette user and whether they were subjected to passive cigarette smoke exposure.

Pre-processing

We extracted all patient answers for the THRIVE survey and created a binary indicator variable for each of the 8 domains; specifically, a value of '1' implies that the patient reports the corresponding social need and a value of '0' otherwise. We combined and encoded the patients' depression test scores into a mental health indicator variable as follows. If a patient had a record of a PHQ2 score larger than 2 or a PHQ9 score larger than 4, their 'Depression' feature was recorded as '1', otherwise it was set to '0'. These thresholds on the PHQ scores are consistent with the scoring system originally described by Spitzer et al. [24]. All categorical features were then encoded into indicator variables for each category (in what is often referred to as 'one-hot' encoding). Specifically, a value of '1' is assigned to the indicator variable corresponding to the category of a patient and a value of '0' for all other categories. Features were standardized by subtracting their mean and dividing with their standard deviation. We also estimated the median household income and the distance to BMC for each patient based on the provided ZIP code. This was based on the USA ZIP code database 'uszipcode' [25], which utilizes up-to-date census information. Median values were used to impute the missing values for each numerical feature. For categorical variables, and given the type of variables we consider in this study, we considered missing values as not-true (assigning a value of '0' to the corresponding indicator variable). These pre-processing procedures yielded 40 features for each patient.

Models and metrics

We developed machine-learning models to predict whether a patient with hypertension has at least one SBP record exceeding 160 mmHg. Predictor variables included: age, sex, race, language, marital status, estimated median household income, distance to BMC, depression scale, cigarette use, and SDoH variables.

We trained both nonlinear models and linear models. Nonlinear ensemble methods included Random Forests (RF) [26], Oblique Random Forests (ORF) [27], and gradient boosted decision trees (XGBoost) [28]. All typically provide excellent performance but do not yield interpretable models as the classifier may combine hundreds of decision trees. ORF is an advanced ensemble learning technique that enhances the traditional Random Forest algorithm by allowing decision trees to split data using oblique hyperplanes rather than the standard axis-aligned splits. This method enables ORF to capture complex interactions and relationships within the data more effectively, leading to improved predictive accuracy, especially in high-dimensional spaces. The accelerated ORF algorithm package 'aorsf' was used for our results [27]. Linear models included Support Vector Machine (SVM) [29] and Logistic Regression (LR) [30]; both yield interpretable models and produce feature weights that can be used to elucidate the relative predictive power of different features. Features with higher absolute coefficient values can be viewed as more significant in predicting high blood pressure. The sign of the coefficient indicates whether the corresponding feature is positively or negatively correlated with the outcome. Regularization (with $\ell 1$

or ℓ2-norm) was added to the linear models to boost the robustness of the model against outliers/noise [31, 32] and avoid models that involve arbitrary linear combinations of highly correlated features that may limit interpretability. In addition to the coefficients, we also reported the p-value and the odds ratio for the top features with the highest absolute LR coefficients. The p-value was computed using a chi-squared test for categorical variables and a Kolmogorov–Smirnov (KS) test for continuous variables. The null hypothesis was that the distribution is the same in the two cohorts. A low p-value supports rejection of the null hypothesis, implying that the corresponding variable is statistically different in the cohort of those with high blood pressure compared to its complement. Odds ratios were reported together with their 95% confidence intervals (CI).

To evaluate the models' ability to predict high blood pressure, the Area Under the Receiver Operating Characteristic Curve (AUC) and the weighted F1 score was calculated out-of-sample (i.e., in a test set not used for training the models). AUC measures the area underneath the ROC curve and represents the probability of the model to assign to a random positive example a higher score than that of random negative example. AUC is classification threshold-invariant since it provides an aggregate measure of performance over all possible thresholds. The F1 score is the harmonic mean of precision and recall. Precision (or positive predictive value) is defined as the ratio of true positives among all true and false positives. Recall (a.k.a. sensitivity) is the ratio of true positives over true positives and false negatives. The weighted-F1 score is computed by weighting the F1-score of each class by the number of patients in that class. The prediction thresholds were optimized to achieve the best weighted-F1 score.

The data were randomly split into training (80%) and test sets (20%). Algorithm parameters were optimized on the training set using fivefold cross-validation and grid search. The regularization strength parameter was defined on a logarithmic scale [0.001, 0.01, 0.1, 1, 10, 100]. For the RF and ORF, the maximum depth of each decision tree and the number of features per split have been optimized, and the number of decision trees was set to 100. The corresponding candidate ranges were defined as [0.05, 0.1, 0.2, 0.3] and with an integer range from [2, 10]. For XGBoost, the maximum depth of the individual trees and the subsampling of the features (colsample_bytree) were optimized on the candidate range [0.05, 0.1, 0.2, 0.3] and with an integer range [0.05, 0.075, 0.1, 0.15, 0.2, 0.3], respectively. The performance metrics were computed on the test set. We report the mean and standard deviation of the test performance metrics over 5 random splits.

Model robustness

In this study, linear models with regularization were used to identify patients with uncontrolled hypertension. Chen and Paschalidis(31) have established a connection between regularization and robustness, through a Distributionally Robust Optimization (DRO) formulation of the classification problem. Consider, for instance, an LR model with response $y \in \{-1, +1\}$, and a predictor vector $\boldsymbol{x} \in \mathbb{R}^p$. The Maximum Likelihood Estimator of (MLE) coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is found by minimizing the negative log-likelihood (logloss):

$$h_\beta(\boldsymbol{x}, y) = log\big(1 + exp\big(-y\boldsymbol{\beta}'\boldsymbol{x}\big)\big).$$

Define the Wasserstein distance metric on the data space as follows:

$$s\big((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2)\big) = ||\boldsymbol{x}_1 - \boldsymbol{x}_2|| + M(y_1 - y_2), \forall (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \in \mathcal{Z},$$

where $\mathcal{Z} = \mathbb{R}^p \times \{-1, +1\}$. Then the robust LR problem can be formulated as:

$$\inf_\beta \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}}\big[log\big(1 + exp\big(-y\boldsymbol{\beta}'\boldsymbol{x}\big)\big)\big],$$

where $\mathbb{Q}$ can be any probability distribution of $(\boldsymbol{x}, y)$ within certain Wasserstein distance of the uniform empirical distribution over the training samples. By defining M to be an infinitely large positive number, which emphasizes the distance cost along the y direction, it can be shown that the robust LR problem can be reformulated as a regularized LR problem with a regularizer equal to the dual norm of the coefficient vector $||\boldsymbol{\beta}||_*$. Therefore, given N samples $(\boldsymbol{x_i}, y_i), i = 1, \ldots, N$, in the training set, the robust LR problem can be reformulated as:

$$\inf_\beta \frac{1}{N} \sum_{i=1}^{N} log\big(1 + exp\big(-y_i\boldsymbol{\beta}'\boldsymbol{x_i}\big)\big) + \epsilon\, ||\boldsymbol{\beta}||_*$$

where ϵ is a hyper-parameter trading off empirical risk (first term) and the regularization term.

Results

Among 164,041 patients in our dataset, 59,306 (36.15%) have a record of SBP over 160 mmHg ("very high blood pressure"). Table 1 provides basic statistics on the entire dataset and on specific groups identified. The summary statistics for the distribution of features across all 40 features have been included in the Supplement. Supplement Table 1 presents the statistics for the entire cohort, while Supplement Table 2 and Supplement Table 3 provide statistics specifically for Black and White patients, respectively.

Table 1 Dataset summary statistics. The 'high blood pressure group' refers to patients with systolic blood pressure (SBP) over 160 mmHg. Patients who report SDoH refers to patients who indicate a social need in any of the 8 domains surveyed by Thrive. Lower income patients are those whose household income is below the median among all patients in the dataset

Demographics	Value		*P*-value
	All patients *N* = 164041	**Patients with SBP record > 160 mmHg M = 59306 (36.15% of N)**	
Mean age	57.7	64.2	0
Median age	59.0	64.0	
Female	85924 (52.38% of N)	31677 (53.41% of M)	< 0.001
Male	78110 (47.62% of N)	27628 (46.59% of M)	< .001
White Patients	56218 (34.27% of N)	16443 (27.73% of M)	0
Black Patients	59353 (36.18% of N)	27339 (46.10% of M)	0
Patients who report SDoH	9317 (5.68% of N)	5296 (8.93% of M)	0
Patients with lower income (< $53,798 per year)	65967 (40.21% of N)	26937 (45.42% of M)	< 0.001

There were 10 different races considered by the predictive model, with the majority being Black or White patients. Black patients comprised 36.18% of all patients but their percentage in the very high SBP cohort was 46.10%. White patients comprise 34.27% of all patients but only 27.73% of the very high SBP group. The computed p-value (chi-square test) compares the distribution of Black/White patients in the two cohorts (patients with no SBP record > 160 mmHg vs. patients with SBP record > 160 mmHg). The p-values were 0 for both variables, which does refute the null hypothesis (the same distribution in the two cohorts).

Similar disparities appear in patients who reported SDoH in the THRIVE survey. Specifically, 5.68% of the patients reported SDoH among all patients but the percentage rose to 8.93% among the very high SBP group. Specifically, 56.84% of those with reported SDoH exhibit very high SBP. The p-value associated with SDoH among the two cohorts was 0, indicating statistical significance.

The estimated median household income of all patients in the dataset was $53,798 U.S. dollars per year. 40.21% of all patients had an estimated household income strictly lower than the median, with that percentage rising to 45.42% among the very high SBP group. The p-value also indicates statistical significance of the influence of household income.

Our algorithms predicting who has very high SBP (> 160 mmHg) reached 74.49% ± 0.23% AUC for nonlinear models (XGBoost) and 73.42% ± 0.23% AUC for linear models. Additional performance metrics are listed in Table 2.

Table 2 also reports the top 20 variables, ranked according to their corresponding absolute LR model coefficient. In accordance with the mean age difference observed between all patients and the ones in the high SBP cohort (57.7 vs. 64.2), the most important variable is 'Age'.

The second top feature is cigarette smoking history with positive coefficient 0.41. Smokers had 5.06 times the odds (CI: 4.89—5.22) of exhibiting very high SBP relative to nonsmokers, indicating strong association between smoking status and very high blood pressure.

Race is also significant, with 'Race_Black' being the third most important feature and 'Race_White' the fourth; each with LR coefficients 0.21 (p-value 0) and -0.15 (p-value 0). Being Black has an associated Odds Ratio (OR) of 1.94 (CI: 1.90—1.98), implying the ratio of the odds for very high SBP versus not is almost twice as high in Black patients compared to other races.

Mental health, and specifically 'Depression', plays a significant role in the model with a coefficient equal to 0.09. The corresponding OR is 1.81 (CI: 1.74—1.88). Among the eight domains of SDoH variables, 'transportation' and 'food' were the most significant with the same coefficient 0.05 (p-value < 0.001 respectively). Additional SDoH variables included housing and employment needs, both inducing a higher likelihood of very high SBP. Further, 'Household_income' also appeared in the top contributing variables with coefficient -0.06 (p-value < 0.001).

The analysis revealed disparities in very high blood pressure between Black and White patients. In the very high SBP cohort, 46.10% (27,339 out of 59,306) are Black and only 27.73% (16,443 out of 59,306) are White. As shown in Table 2, 'Race_Black' and 'Race_White' are in the top 5 predictive features. A more detailed comparison between Black and White patients is given in Table 3. The incidence of very high SBP among Black patients is significantly higher than among White patients. Similarly, Black patients have a larger fraction of lower income

Table 2 Predictive model of High SBP record (over 160 mmHg): validation performance metrics and top variables. SD refers to the standard deviation of the corresponding metric. LR-L2 and SVM-L2 refer to the ℓ_2-norm regularized LR and SVM models. The top 20 features of the predictive model are listed and ranked by the absolute LR coefficients (Coef). We also listed the p-value, correlation of the variable with the outcome (Y-corr), the mean of the variable (Y1-mean) in the patient cohort with high SBP records over 160 mmHg, and the mean of the variable (Y0-mean) in the patients with no SBP record above 160 mmHg. We report the corresponding odds ratios (OR) and their 95% confidence intervals

Algorithm	AUC		Weighted F1-score	
	Mean	SD	Mean	SD
LR-L2	73.42%	0.23%	69.17%	0.24%
SVM-L2	73.40%	0.23%	69.14%	0.19%
XGBoost	74.49%	0.23%	70.07%	0.20%
RF	73.53%	0.24%	69.27%	0.24%
ORF	73.86%	0.25%	78.67%	0.16%

Top 20 Variables in the LR model

	Variable	Coef	*p*-value	Y1-mean	Y0-mean	Y-corr	OR	95% OR CI	5% OR CI
1	Age	0.67	0	64.20	54.02	0.28	1.04	1.04	1.04
2	Ever Cigarette User-YES	0.41	0	0.23	0.05	0.26	5.06	5.22	4.89
3	Race_Black	0.21	0	0.46	0.31	0.16	1.94	1.98	1.90
4	Race_White	-0.15	0	0.28	0.38	-0.10	0.63	0.64	0.61
5	Depression	0.09	< 0.001	0.09	0.05	0.07	1.81	1.88	1.74
6	Language_English	-0.09	< 0.001	0.67	0.71	-0.04	0.85	0.87	0.84
7	Race_Hispanic	-0.09	< 0.001	0.01	0.02	-0.05	0.42	0.46	0.38
8	Marital_status:other	-0.07	< 0.001	0.03	0.05	-0.04	0.66	0.69	0.62
9	Race_Other	-0.07	< 0.001	0.01	0.02	-0.03	0.48	0.53	0.43
10	Household_income	-0.06	< 0.001	55586.96	58078.01	-0.06	0.99	0.99	0.99
11	SDoH_transportation	0.05	< 0.001	0.03	0.01	0.08	3.46	3.76	3.19
12	Race_Asian	-0.05	< 0.001	0.03	0.03	-0.03	0.72	0.77	0.68
13	SDoH_food	0.05	< 0.001	0.05	0.02	0.08	2.47	2.62	2.34
14	Race_Middle Eastern	-0.05	< 0.001	0.00	0.00	-0.02	0.37	0.48	0.29
15	Language_Other	-0.04	< 0.001	0.18	0.17	0.02	1.12	1.15	1.09
16	Language_African	0.03	< 0.001	0.06	0.03	0.06	1.76	1.85	1.68
17	SDoH_housing	0.03	< 0.001	0.01	0.00	0.05	3.05	3.44	2.70
18	Marital_status:divorced	0.03	< 0.001	0.07	0.05	0.04	1.48	1.54	1.42
19	Marital_status:separated	0.02	< 0.001	0.03	0.02	0.04	1.68	1.79	1.58
20	SDoH_job	0.02	< 0.001	0.02	0.01	0.05	2.09	2.27	1.93

patients and a higher percentage of patients with SDoH needs. Furthermore, Black patients live a shorter distance to BMC, on average, while 'Distance_to_BMC' has a negative coefficient in our models. The feature distribution comparison for each of the SDoH and marital status variables have been visualized in Figs. 1 and Fig. 2. Black patients in general have a much higher chance presenting with mental health issues and SDoH needs in all 8 domains. Furthermore, the percentage of single patients are higher among Black patients compared with White patients.

To further elucidate racial differences and the specific features associated with high SBP in each group, we trained models on Black and White patients separately. Since the cohort of Black patients has more people in the high SBP cohort, we downsampled the Black patients to induce the same positive sample rate compared to White patients (29.25%). The interesting finding is that for both AUC and weighted F1 score, the performance of the model for Black patients is higher. Specifically, the model for Black patients achieves AUC of 74.81% whereas the corresponding model for White patients achieves AUC of 70.11%. More detailed results are in Table 4. The performance gap between two racial groups has also been visualized in a heatmap shown in Fig. 3.

Table 4 and Fig. 3 indicate that when we use the standard decision threshold on the predicted probability of very high SBP (i.e., patients with a predicted probability above 0.5 are classified in the high SBP group), the corresponding false positive rates (FPR)

Table 3 Comparison between Black and White patients. Percentages in the entire dataset are calculated as a fraction of the total number of patients ($N = 164{,}041$). The population in the high SBP cohort consists of Black or White patients with SBP records exceeding 160 mmHg and the percentages are computed as fractions of M = 59,306, which is the size of the high SBP cohort. The population below the median household income level consists of Black or White patients who have household income below the overall annual median level in the entire dataset ($53,798); the corresponding percentages are computed as fractions of Black (respectively, White) patients in the entire dataset. The population with SDoH consists of patients who answered 'Yes' in any of the 8 domains of the Thrive survey, and percentages are calculated as fractions of Black (respectively, White) patients in the entire dataset

	Black patients	White patients
Population in the entire dataset	59,353 (36.18%)	56,218 (34.27%)
Population in the high SBP cohort	27,339 (46.10%)	16,443 (27.73%)
Population below the median household income level	30,537 (51.45%)	12,628 (22.46%)
Population with SDoH	5,941 (10.01%)	1,215 (2.16%)
Average distance to BMC (miles)	16.99	47.25

and false negative rates (FNR) are substantially different between the two models. In particular, the FPR is 27.71% for Black patients and 5.92% for White patients (a factor of 4.68). Correspondingly, the FNR for Black patients is 43.11% compared to 73.82% for White patients (a factor of 0.58). We also computed a so called 'Treatment Equality' metric, defined as the ratio FPR/FNR [33], to evaluate the model's racial disparity. Having this metric equalized among the two cohorts is considered desirable. Under the standard 0.5 prediction threshold, FPR/FNR for Black patients is 8 times higher than the ratio for White patients. These statistics demonstrate that the models are biased and are more likely to classify Black patients as belonging to the very high SBP group. In essence, the models "pick up" and exploit racial disparities present in the data.

To resolve the algorithmic racial disparity, we modified the decision thresholds used in the two models. Clearly there is a trade-off between FPR and FNR; higher FPR implies lower FNR and vice versa. One way to resolve the bias is to equate the FNR rates. Arguably, false negatives are more critical to hypertension control because they lead to worse longer-term outcomes and, presumably, greater burden associated with long-term poorly controlled hypertension. On the other hand, false positives only imply providing more support to patients who may not need it. The associated resource consumption favors the lower costs of primary prevention and care versus the consequences of longer-term poorly controlled disease.

Figure 4 plots the FNR for both models as a function of the corresponding decision threshold. We elected to set the FNR for either model to 25%, which is attained by a threshold for Black patients equal to 0.296 and a corresponding threshold for White patients equal to 0.245. Table 5 shows the performance metrics for both models after modifying the decision thresholds. The resulting FPR of the model for Black patients becomes 41.11% and the FPR for White patient model becomes 47.20%. Similarly, the treatment equality metric of FPR/FNR becomes 1.64 for the Black patients' model and 1.89 for the White patients' model (a factor of 0.87).

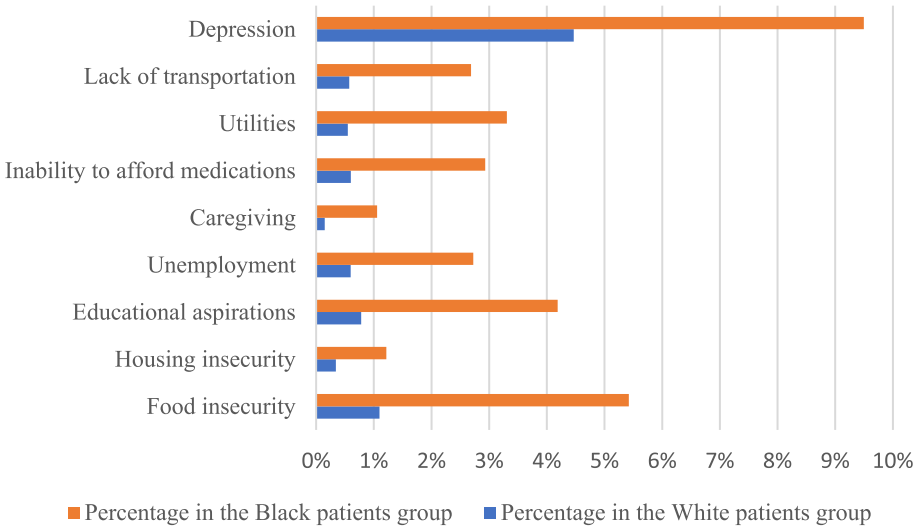

Fig. 1 SDoH features distribution percentages in the cohort of Black patients and in the cohort of White patients

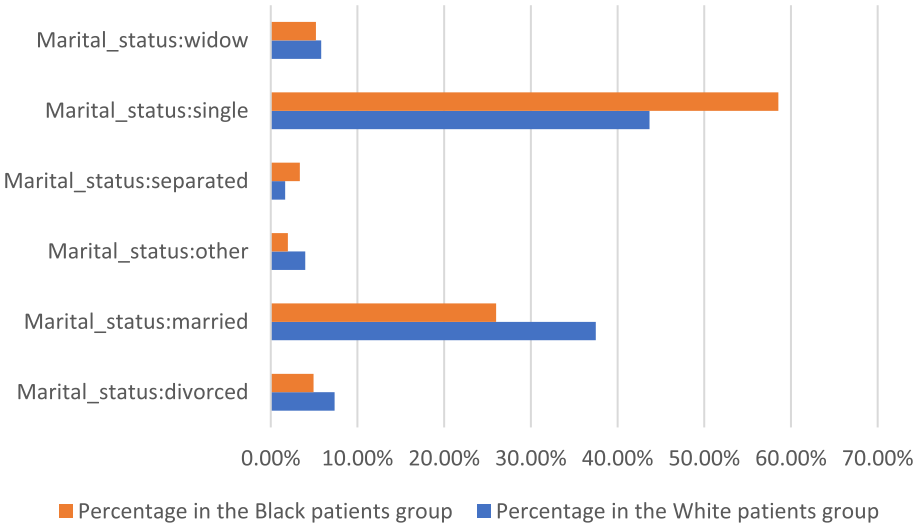

Fig. 2 Marital status features distribution percentages in the cohort of Black patients and in the cohort of White patients

Table 4 High SBP prediction models for Black and White patients trained separately: a comparison of validation performance metrics. LR-L2 refers to the ℓ_2-norm regularized LR model. FNR and FPR refer to the model's false positive and false negative rate when the decision threshold (on the predicted probability of high SBP) is set to 0.5

LR-L2	Black patients	White patients
AUC	74.07%	70.11%
Weighted F1-score	67.51%	68.11%
FPR (Threshold = 0.5)	27.71%	5.92%
FNR (Threshold = 0.5)	43.11%	73.82%
FPR/FNR (Threshold = 0.5)	0.64	0.08

Discussion

The discriminatory power of our models is consistent with the performance of existing hypertension prediction models which use additional information, such as familial history of hypertension and blood pressure (BP) measurements [12]. Our results, using a model that relies only on self-reported demographics and socioeconomic variables, outperform models that use BP measurements, laboratory results, and genetic variables (achieving an AUC of 66.4%) [34]. Other models achieve AUCs higher than 0.75 but incorporate multiple BP measurements over time [35] and risk scores [12]. Considering that we only used demographics and socioeconomic variables,

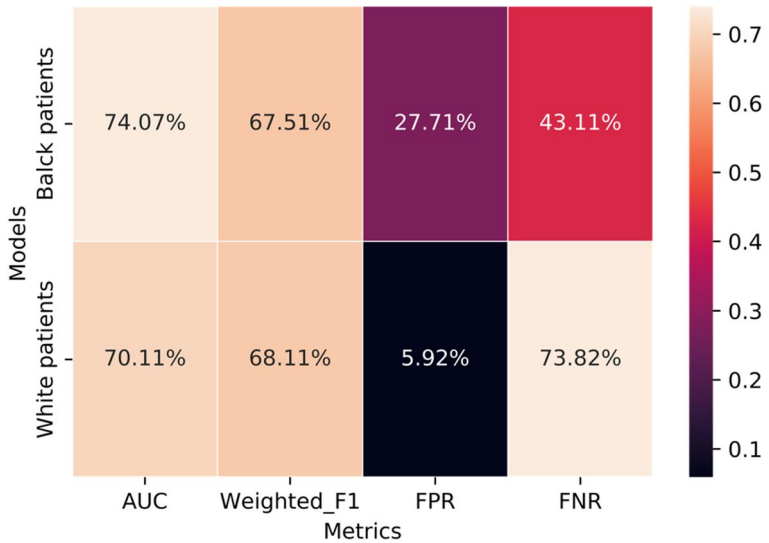

Fig. 3 Heatmap showing the performance gap between the Black patients' model and the White patients' model

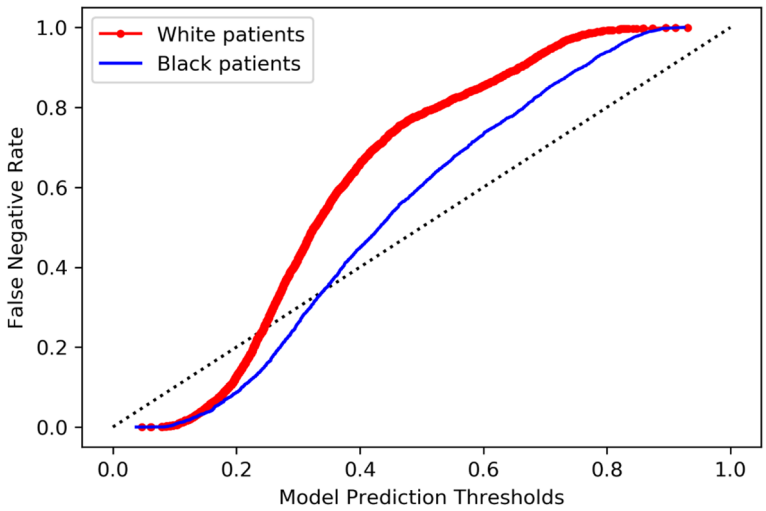

Fig. 4 Relationship between false negative rate (FNR) and decision threshold in the models for Black and White patients

Table 5 LR prediction thresholds used to tune the validation FNR of the separate models (set to 25%). Validation model performance metrics after tuning the thresholds to equate FNR

	Black patients	White patients
Weighted F1	65.77%	60.91%
FPR	41.11%	47.20%
FPR/FNR	1.64	1.89
Setting FNR to 25%		
	Prediction Thresholds	
White Patients	0.245	
Black Patients	0.296	

the discriminatory power of our proposed model is noteworthy.

The importance of age is not surprising since numerous studies have shown such association, providing plausible physiological mechanisms [36].

The harmful effects of cigarette smoking in hypertension control have been previously demonstrated. Moreover, hypertensive smokers have an increased risk of developing severe forms of hypertension, which may result from accelerated atherosclerosis due to smoking [37].

It is known that Black patients have significantly higher risk of hypertension than other racial groups [8]. Currently, there is no clear physiological explanation. The disparities observed may be rooted in systemic issues, including barriers to healthcare access and social determinants of health (SDoH), which could affect key aspects of hypertension management such as diet, stress management, and medication practices. Furthermore, evidence suggests that the increased levels of uncontrolled blood pressure among Black and Hispanic individuals could be associated with factors like inconsistent insurance coverage and levels of education [11, 38]. Our research, along with other studies, suggests that these disparities in hypertension continue to exist even when socioeconomic differences are taken into account [6, 39].

We sought to understand whether there are discernible differences in the predictive variables that lead to differences between Black and White patients and shed some light into the etiology of these differences. Table 6 lists the top 20 predictive variables for each model (ranked by the absolute value of their corresponding LR coefficient). Other than age, cigarette use and depression, which are common, the model for Black patients emphasizes SDoH needs in transportation, food, and housing resources. Transportation needs may explain poorer control of hypertension as patients face barriers in physical access to care. Further, needing help to pay for food may correlate with consumption of low-cost, high-calorie, high-sodium fast food. Similarly, housing needs are a cause for stress and overall lower quality of life which may contribute to hypertension. Other neighborhood-level socioeconomic factors like annual household income and distance to BMC also play significant roles in the Black patient predictive model, and they are negatively correlated with the outcome. In particular, 51.45% of Black patients have median household income less than the overall median value in the dataset. This percentage is significantly higher than the average low-income percentage of 40.21% computed over all patients in the dataset. This finding confirms previous research on the association between neighborhood environment and hypertension [40]. Namely,

Table 6 Top 20 features with largest absolute LR coefficients (Coef) in the predictive models for Black and White patients, respectively

	TOP 20 Predictive Variables for Black patients		TOP 20 Predictive Variables for White patients	
	Variable	**Coef**	**Variable**	**Coef**
1	Age	0.69	Age	0.56
2	Ever Cigarette User-YES	0.46	Ever Cigarette User-YES	0.41
3	Depression	0.10	Depression	0.11
4	Marital_status:other	-0.07	Marital_status:other	-0.08
5	Household_income	-0.07	Language_English	-0.06
6	SDoH_food	0.06	Language_Spanish	0.06
7	SDoH_transportation	0.06	Household_income	-0.05
8	Language_English	-0.05	Marital_status:divorced	0.05
9	SDoH_housing	0.03	Sex_M	0.04
10	SDoH_caring	-0.03	Marital_status:married	-0.04
11	SDoH_utilities	0.03	SDoH_job	0.03
12	Distance_to_BMC	-0.03	Language_the Middle East	0.03
13	Language_Spanish	-0.03	SDoH_food	0.03
14	Marital_status:married	-0.02	Marital_status:separated	0.03
15	Passive_Cigarette_Exposure_YES	0.02	SDoH_housing	0.03
16	Sex_F	-0.01	Language_European	0.02
17	SDoH_job	0.01	Sex_F	-0.02
18	Marital_status:widow	0.01	SDoH_transportation	0.02
19	Marital_status:single	-0.01	SDoH_utilities	0.02
20	Sex_M	0.01	Marital_status:widow	0.01

living in an area of high resource deprivation is often linked to various deleterious factors such as limited access to medical institutions, decreased neighborhood safety, and lower education level. All of these factors can in turn exacerbate health disparities.

Considering the model for White patients, the top features contain marital status (divorced and separated increase the likelihood of high SBP) and language. Arguably, the top predictive variables for Black patients correspond to structural social needs (food, transportation), whereas top predictive variables for White patients correspond to other types of demographic or personal issues.

Limitations

Our study has several limitations. First, it is based on patients who have clinical care records and present with elevated blood pressure, not all patients. Some patients may not have access to hospital-based medical care or simply may be unaware of their hypertension to seek help. Second, only records from BMC are included in the study. BMC is a tertiary care, academic, safety-net medical center in an urban area. Patients in rural areas with generally fewer medical resources may experience larger sociodemographic disparities. Third, median household incomes were estimated by patient ZIP code and do not fully account for individual factors.

Conclusion

This study developed an easy to implement, interpretable prognostic model for identifying patients with high SBP exceeding 160 mmHg among hypertensive or pre-hypertensive patients. The derived models exhibit strong predictive power based solely on non-clinical variables. The top predictive features identified are age, cigarette use, race, mental health, and SDoH factors. We found implicit racial disparity in the non-stratified predictive model. Race-specific models were trained for Black and White patients to investigate what underlies this bias. While Black patients may have a greater prevalence of high blood pressure, their conditions were more significantly associated with structural SDoH variables corresponding to food, transportation, housing, and lower household income. This finding may facilitate the targeted intervention on specific SDoH factors for Black patients in need. Addressing the most important social needs for Black and White patients alike may improve the overall quality of care for patients with hypertension. Moreover, to resolve the algorithmic bias and equate false positive

and false negative predictions across racial groups, one needs to select different decision thresholds for Black and White patients.

Predictive analytics of the type we developed can be used in various ways: (1) understanding the risk profiles can lead to better clinical decision making and more effective risk communication with patients; (2) tailoring prescriptions based not only on a BP measurement during a visit but also on the risk profile suggested by the model may lead to better hypertension control; (3) targeted interventions in partnership with national, state, and neighborhood support programs may address SDoH which are modifiable, leading to better hypertension control and lower medical care costs.

Abbreviations

SBP	Systolic Blood Pressure
SDoH	Social Determinants of Health
BMC	Boston Medical Center
EHRs	Electronic Health Records
DBP	Diastolic Blood Pressure
RF	Random Forests
ORF	Oblique Random Forest
XGBoost	Gradient Boosted Decision Trees
SVM	Support Vector Machine
LR	Logistic Regression
KS	Kolmogorov-Smirnov
CI	Confidence Intervals
AUC	Area Under the Receiver Operating Characteristic Curve
DRO	Distributionally Robust Optimization
FPR	False Positive Rates
FNR	False Negative Rates
BP	Blood Pressure

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02873-4.

Supplementary Material 1.

Acknowledgements
We thank Boston Medical Center for providing the EHR dataset.

Authors' contributions
Y.H. contributed to methods, performed the analysis, compiled the results, and co-wrote the manuscript. N.C. and R.M. provided input and medical intuition throughout the analysis and edited the manuscript. I.C.P. led the study, contributed to methods, provided input to the analysis, and co-wrote the manuscript.

Funding
This work was supported in part by the National Science Foundation grants IIS-1914792, DMS-1664644, DEB-2433726, ECCS-2317079, and CCF-2200052, by the Office of Naval Research grant N00014-19–1-2571, by the National Institutes of Health grant R01 GM135930, by the Boston University Clinical and Translational Science Award (CTSA) under NIH/NCATS grant UL54 TR004130, and by the Boston University Hariri Institute for Computing and Computational Science & Engineering.

Data availability
The data that support the findings of this study are available from Boston Medical Center (BMC) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the corresponding authors upon reasonable request and with permission of BMC.

Declarations

Ethics approval and consent to participate
The study adhered to the Helsinki Declaration and was approved by the Boston University Medical Campus (IRB #H-32061) and Boston University Institutional Review Boards. Since it analyzed de-identified records, it was classified as non-human subject research. Informed consent has been waived. The Boston University team has signed a Data Use Agreement (DUA) with the Boston Medical Center to access the data.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Electrical and Computer Engineering, Department of Biomedical Engineering, Division of Systems Engineering, and Faculty of Computing & Data Sciences, Boston University, 8 Saint Mary's St., Boston, MA 02215, USA. [2]Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, MA, USA. [3]Mass General Brigham and Harvard Medical School, Boston, MA, USA.

Received: 10 April 2024 Accepted: 16 January 2025
Published online: 03 February 2025

References

1. Control C for D, Prevention. Hypertension cascade: hypertension prevalence, treatment and control estimates among US adults aged 18 years and older applying the criteria from the American College of Cardiology and American Heart Association's 2017 Hypertension Guideline—NHANES 2013–2016. Atlanta GA US Dep Health Hum Serv. 2019.
2. Grassi G, Seravalle G, Mancia G. Cardiovascular consequences of poor compliance to antihypertensive therapy. Blood Press. 2011;20(4):196–203.
3. Chobanian AV. 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.
4. Leng B, Jin Y, Li G, Chen L, Jin N. Socioeconomic status and hypertension: a meta-analysis. J Hypertens. 2015;33(2):221–9.
5. Mills KT, Bundy JD, Kelly TN, Reed JE, Kearney PM, Reynolds K, et al. Global disparities of hypertension prevalence and control. Circulation. 2016;134(6):441–50.
6. 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.
7. Hu Y, Huerta J, Cordella N, Mishuris RG, Paschalidis ICh. Personalized hypertension treatment recommendations by a data-driven model. BMC Med Inform Decis Mak. 2023;23(1):44.
8. Hajjar I, Kotchen TA. Trends in prevalence, awareness, treatment, and control of hypertension in the United States, 1988–2000. JAMA. 2003;290(2):199–206.
9. Kramer H, Han C, Post W, Goff D, Diez-Roux A, Cooper R, et al. Racial/ethnic differences in hypertension and hypertension treatment and control in the multi-ethnic study of atherosclerosis (MESA). Am J Hypertens. 2004;17(10):963–70.
10. Carnethon MR, Pu J, Howard G, Albert MA, Anderson CAM, Bertoni AG, et al. Cardiovascular health in African Americans: a scientific statement from the American Heart Association. Circulation. 2017;136(21):e393-423.
11. Kressin NR, Terrin N, Hanchate AD, Price LL, Moreno-Koehler A, LeClair A, et al. Is insurance instability associated with hypertension outcomes and does this vary by race/ethnicity? BMC Health Serv Res. 2020;20(1):216.
12. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, et al. A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. Ann Intern Med. 2008;148(2):102–10.

13. Paynter NP, Cook NR, Everett BM, Sesso HD, Buring JE, Ridker PM. Prediction of incident hypertension risk in women with currently normal blood pressure. Am J Med. 2009;122(5):464–71.
14. Teixeira PL, Wei WQ, Cronin RM, Mo H, VanHouten JP, Carroll RJ, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. J Am Med Inform Assoc. 2017;24(1):162–71.
15. Muntner P, Woodward M, Mann DM, Shimbo D, Michos ED, Blumenthal RS, et al. Comparison of the Framingham Heart Study hypertension model with blood pressure alone in the prediction of risk of hypertension: the Multi-Ethnic Study of Atherosclerosis. Hypertension. 2010;55(6):1339–45.
16. Ye X, Zeng QT, Facelli JC, Brixner DI, Conway M, Bray BE. Predicting optimal hypertension treatment pathways using recurrent neural networks. Int J Med Inf. 2020;1(139):104122.
17. John LH, Kors JA, Reps JM, Ryan PB, Rijnbeek PR. Logistic regression models for patient-level prediction based on massive observational data: do we need all data? Int J Med Inf. 2022;1(163):104762.
18. de la Vega PB, Losi S, Martinez LS, Bovell-Ammon A, Garg A, James T, et al. Implementing an EHR-based screening and referral system to address social determinants of health in primary care. Med Care. 2019;57:S133–9.
19. Li Y, Wang H, Luo Y. Improving fairness in the prediction of heart failure length of stay and mortality by integrating social determinants of health. Circ Heart Fail. 2022;15(11):e009473.
20. Mhasawade V, Chunara R. Causal multi-level fairness. arXiv. 2021. Available from: http://arxiv.org/abs/2010.07343. Cited 2023 Mar 10.
21. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison HC, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. J Am Coll Cardiol. 2018;71(19):e127-248.
22. Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. Ann Fam Med. 2010;8(4):348–53.
23. Spitzer RL, Kroenke K, Williams JB, Group PHQPCS, Group PHQPCS. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Jama. 1999;282(18):1737–44.
24. Welcome to uszipcode Documentation — uszipcode 0.2.6 documentation. Available from: https://uszipcode.readthedocs.io/. Cited 2021 Nov 3.
25. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
26. Jaeger BC, Welden S, Lenoir K, Speiser JL, Segar MW, Pandey A, et al. Accelerated and interpretable oblique random survival forests. J Comput Graph Stat. 2024;33(1):192–207.
27. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California: Association for Computing Machinery; 2016. p. 785–94. (KDD '16). https://doi.org/10.1145/2939672.2939785. Cited 2020 Jun 19.
28. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
29. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction, vol. 1. Berlin: Springer series in statistics Springer; 2001.
30. Chen R, Paschalidis IC. Distributionally robust learning. Found Trends® Optim. 2020;4(1–2):1–243.
31. Chen R, Paschalidis IC. A robust learning approach for regression models based on distributionally robust optimization. J Mach Learn Res JMLR. 2018;19(1):517–64.
32. Caton S, Haas C. Fairness in machine learning: A survey. ACM Comput Surv. 2024;56(7):166:1–38. https://doi.org/10.1145/3616865.
33. Fava C, Sjögren M, Montagnana M, Danese E, Almgren P, Engström G, et al. Prediction of blood pressure changes over time and incidence of hypertension by a genetic risk score in Swedes. Hypertension. 2013;61(2):319–26.
34. Kivimäki M, Tabak AG, Batty GD, Ferrie JE, Nabi H, Marmot MG, et al. Incremental predictive value of adding past blood pressure measurements to the Framingham hypertension risk equation: the Whitehall II Study. Hypertension. 2010;55(4):1058–62.
35. Dohi Y, Thiel MA, Bühler FR, Lüscher TF. Activation of endothelial L-arginine pathway in resistance arteries Effect of age and hypertension. Hypertension. 1990;16(2):170–9.
36. Virdis A, Giannarelli C, Fritsch Neves M, Taddei S, Ghiadoni L. Cigarette smoking and hypertension. Curr Pharm Des. 2010;16(23):2518–25.
37. Alves RFS, Faerstein E. Educational inequalities in hypertension: complex patterns in intersections with gender and race in Brazil. Int J Equity Health. 2016;15(1):146.
38. Hertz RP, Unger AN, Cornell JA, Saunders E. Racial disparities in hypertension prevalence, awareness, and management. Arch Intern Med. 2005;165(18):2098–104.
39. Morenoff JD, House JS, Hansen BB, Williams DR, Kaplan GA, Hunte HE. Understanding social disparities in hypertension prevalence, awareness, treatment, and control: the role of neighborhood context. Soc Sci Med. 2007;65(9):1853–66.
40. Mujahid MS, Roux AVD, Morenoff JD, Raghunathan TE, Cooper RS, Ni H, et al. Neighborhood characteristics and hypertension. Epidemiology. 2008;19:590–8.

Publisher's Note

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