Causal Inference of Blood Pressure Reduction and Coronary Heart Disease Risk in the Framingham Study

Suchibrata Patra[1,*]

[1]St.Xavier's College (Autonomous), Kolkata
[*]Correspondence: suchibratapatra2003@gmail.com

ABSTRACT

Standard cardiovascular risk calculators, including the Framingham Risk Score and the ACC/AHA Pooled Cohort Equations, estimate the conditional probability $P(\mathrm{CHD} \mid \mathrm{SysBP} = s)$ rather than the interventional quantity $P(\mathrm{CHD} \mid \mathrm{do}(\mathrm{SysBP} = s))$. When confounding is present, this distinction has direct clinical consequences: observational estimates may systematically overstate the absolute benefit of antihypertensive treatment. We applied Pearl's do-calculus to the Framingham Heart Study Offspring Cohort ($n = 4{,}240$; primary analysis on $3{,}776$ complete cases; 574 ten-year coronary heart disease (CHD) events). A structurally corrected directed acyclic graph (DAG) was specified, incorporating four biologically motivated corrections, and subjected to conditional independence testing. The average causal effect (ACE) of a 20 mmHg systolic blood pressure (SysBP) reduction was estimated by g-computation with 1,500-iteration bootstrap confidence intervals, corroborated by sex-stratified propensity score matching (PSM) and inverse probability weighting (IPW). Conditional average treatment effects (CATE) were estimated using R-Learner and T-Learner metalearners with gradient-boosted nuisance models. G-computation yielded an ACE of 3.40% absolute risk reduction (95% bootstrap CI: 2.64%–4.14%), compared with a naive observational estimate of 4.14%, a relative overestimation of approximately 21.8% ($+35.7\%$ on the log-odds scale). The E-value lower bound was 2.18. Statistically significant heterogeneity in treatment effect was detected across age strata (Kruskal–Wallis $p < 0.001$) and diabetes status (Mann–Whitney $p < 0.001$); however, diabetic subgroup estimates were unstable and underpowered ($n = 109$), and no reliable subgroup inference can be drawn without replication. These findings suggest that observational cardiovascular risk tools may overestimate the absolute benefit of blood pressure reduction, with implications for clinical risk stratification and prescribing thresholds.

Introduction

Every major cardiovascular risk tool in routine clinical use, the Framingham Risk Score [1], SCORE2 [2], and the ACC/AHA Pooled Cohort Equations [3], estimates a conditional probability of the form $P(\mathrm{CHD} \mid \mathrm{SysBP} = s, Z = z)$. This quantity describes the observed frequency of coronary heart disease (CHD) among individuals who happen to present with a given blood pressure profile. The quantity a prescribing clinician requires when deciding whether to initiate antihypertensive therapy is categorically different:

$$P(\mathrm{CHD} \mid \mathrm{do}(\mathrm{SysBP} \leftarrow s)), \tag{1}$$

where Pearl's do-operator [4] represents the act of pharmacologically setting blood pressure to a target value, regardless of the patient's underlying biological state. Individuals who naturally present with SysBP = 140 mmHg are, on average, older, more obese, and more metabolically dysregulated than those presenting at 120 mmHg. Antihypertensive therapy severs only the causal path from elevated pressure to vascular events; it does not reverse these pre-existing comorbidities. Conflating the observational and interventional quantities therefore produces a systematically inflated estimate of treatment benefit, a gap with direct consequences for clinical decision-making and health economic evaluation [5].

This prediction–intervention gap is formally resolvable using structural causal models and the back-door adjustment theorem [4]. Under this framework, the interventional distribution (1) is identifiable from observational data provided that a valid directed acyclic graph (DAG) can be specified and that the positivity and conditional ignorability assumptions hold. A growing methodological literature has addressed components of this problem: marginal structural models have been applied to blood pressure trajectories [7], inverse probability weighting (IPW) has been reviewed in cardiovascular epidemiology [8], and the distinction between associational and interventional estimands has been extensively theorised [5, 9]. Nevertheless, no prior study has delivered, within a single reproducible pipeline applied to the Framingham Heart Study, a biologically corrected and empirically evaluated causal DAG, multiple corroborating causal-effect estimators, and an explicit quantification of the bias

incurred by standard observational approaches.

The present study makes three contributions. First, we quantify, for the first time in this cohort, the magnitude by which the naive observational estimate of blood pressure treatment benefit exceeds the causal estimate (approximately 21.8% in relative terms). Second, we specify, correct, and empirically evaluate a structural causal DAG, identifying four errors in prior formulations that propagate into biased causal estimates. Third, we implement an end-to-end causal inference pipeline, covering DAG evaluation, g-computation, sex-stratified propensity score matching (PSM), IPW, refutation testing, and semiparametric treatment-effect heterogeneity estimation, within a single reproducible framework, and discuss the clinical implications of the resulting estimates.

Results

Baseline characteristics

The analytical cohort comprised 3,776 complete-case participants (55.5% female; mean age 49.6 ± 8.6 years; mean systolic blood pressure (SysBP) 132.4 ± 22.0 mmHg; 2.7% with diabetes mellitus). As shown in Table 1, participants who developed CHD within ten years were on average five years older (54.2 vs. 48.7 years; Mann–Whitney $p < 0.001$), had 13.5 mmHg higher SysBP (143.8 vs. 130.3 mmHg; $p < 0.001$), and had substantially higher prevalences of diabetes (6.3% vs. 2.1%; χ^2, $p < 0.001$) and prevalent hypertension (50.7% vs. 27.6%; $p < 0.001$). These systematic differences illustrate the confounding structure that motivates causal rather than observational estimation: patients presenting with elevated SysBP are metabolically and cardiovascularly more compromised across multiple dimensions, independently of their blood pressure level.

Table 1. Baseline characteristics by 10-year CHD status. Data are mean ± SD (continuous variables) or *n* (%) (binary variables). *p*-values from Mann–Whitney *U* test (continuous) or χ^2 test (binary). The systematic confounding structure, with older age, higher SysBP, higher body mass index (BMI), and greater comorbidity burden among CHD cases, motivates causal rather than observational estimation.

Variable	Overall ($n = 3{,}776$)	No CHD ($n = 3{,}202$)	CHD ($n = 574$)	*p*
Age (years)	49.6 ± 8.6	48.7 ± 8.4	54.2 ± 8.0	< 0.001
SysBP (mmHg)	132.4 ± 22.0	130.3 ± 20.4	143.8 ± 26.8	< 0.001
DiaBP (mmHg)	82.9 ± 11.9	82.2 ± 11.3	87.2 ± 14.3	< 0.001
Total cholesterol (mg/dL)	237.0 ± 44.7	235.3 ± 43.8	246.3 ± 48.1	< 0.001
BMI (kg/m^2)	25.8 ± 4.1	25.7 ± 3.9	26.6 ± 4.6	< 0.001
Glucose (mg/dL)	81.9 ± 23.8	80.7 ± 19.0	88.8 ± 40.8	< 0.001
Heart rate (bpm)	75.7 ± 12.0	75.6 ± 11.9	76.4 ± 12.1	0.299
Male sex, *n* (%)	1,682 (44.5%)	1,362 (42.5%)	320 (55.7%)	< 0.001
Current smoker, *n* (%)	1,857 (49.2%)	1,561 (48.8%)	296 (51.6%)	0.231
Diabetes mellitus, *n* (%)	102 (2.7%)	66 (2.1%)	36 (6.3%)	< 0.001
Antihypertensive med., *n* (%)	114 (3.0%)	77 (2.4%)	37 (6.4%)	< 0.001
Prevalent hypertension, *n* (%)	1,176 (31.1%)	885 (27.6%)	291 (50.7%)	< 0.001

Observational prediction model

A multivariable logistic regression achieved an AUROC of 0.721 (5-fold cross-validation: 0.721 ± 0.028; average precision: 0.339; Brier score: 0.116). The observational odds ratio per mmHg SysBP was 1.013 (95% CI: 1.008–1.018; $p < 0.001$). Age (OR: 1.063), male sex (OR: 1.805), and current smoking (OR: 1.477) were the strongest associated predictors. These coefficients reflect the conditional distribution of CHD given observed covariates and carry no causal interpretation.

Directed acyclic graph specification and evaluation

The corrected DAG (Figure 1) encoded 11 nodes and 24 directed edges incorporating four biologically motivated structural corrections.

Correction 1: Removal of CURSMOKE to SYSBP. Nicotine causes acute transient vasoconstriction [12], but in this cross-sectional cohort this does not constitute a structural population-level causal path from smoking status to sustained SysBP. Retaining this edge creates an illegitimate back-door path through an intermediate variable; the epidemiologically appropriate structure routes smoking's CHD effect through cholesterol and via a direct atherogenic mechanism. *Correction 2: Addition of AGE to DIABETES.* Type 2 diabetes incidence increases sharply with age [13]. Omitting this edge misrepresents age as an independent root cause of CHD, failing to capture its role as a shared cause of both SysBP elevation (through arterial stiffening) and CHD (partially mediated through glucose dysregulation).

Correction 3: Reclassification of BPMEDS. Antihypertensive medication is prescribed because SysBP is elevated; the causal arrow runs from SYSBP to BPMEDS. Including BPMEDS in the back-door adjustment set introduces collider bias, blocking part of the SYSBP to CHD effect pathway and attenuating the true causal estimate [4]. BPMEDS was retained in the observational model as a precision variable but was explicitly excluded from all causal adjustment sets.

Correction 4: Exclusion of PREVHYP from the primary adjustment set. Prevalent hypertension is causally downstream of long-run SysBP. Conditioning on it in the primary adjustment set would induce over-adjustment bias; it was excluded from the primary causal model and included only in a pre-specified sensitivity analysis.

We acknowledge that the DAG is a simplified representation of biological reality. Not all possible biological dependencies are encoded, and edges were specified on the basis of cardiovascular pathophysiology rather than formal elicitation. The DAG should not be interpreted as a validated complete causal map of the cardiovascular system.

Four conditional independence implications of the corrected DAG were tested by partial correlation (Table 2). All four passed at $\alpha = 0.05$: SEX $\perp$ GLUCOSE $|$ $\{$BMI, DIABETES$\}$ ($r = -0.007$, $p = 0.656$); BPMEDS $\perp$ TOTCHOL $|$ $\{$AGE, SYSBP$\}$ ($r = +0.023$, $p = 0.131$); BPMEDS $\perp$ GLUCOSE $|$ $\{$AGE, SYSBP$\}$ ($r = +0.012$, $p = 0.422$); and AGE $\perp$ BPMEDS $|$ SYSBP ($r = +0.025$, $p = 0.109$). The expected failure of SEX $\perp$ SYSBP $|$ $\{$AGE, BMI$\}$ ($p < 0.001$) correctly reflects the retained biologically established direct effect of sex on SysBP [14] and corroborates rather than contradicts the specified graph. Passage of all non-trivial testable implications is consistent with the corrected DAG structure as a basis for causal identification, although we cannot exclude misspecification of paths not directly testable from these data.

Table 2. Supplementary Table S1. Conditional independence tests corresponding to testable implications of the corrected DAG. Partial correlation coefficients computed by regressing out the conditioning set via ordinary least squares. $p > 0.05$ is consistent with the null implication. The failure of SEX $\perp$ SYSBP $|$ $\{$AGE, BMI$\}$ is expected under the retained SEX to SYSBP edge and corroborates rather than contradicts the DAG.

Implication	Partial r	p	Result
SEX $\perp$ GLUCOSE $|$ $\{$BMI, DIABETES$\}$	-0.007	0.656	PASS
BPMEDS $\perp$ TOTCHOL $|$ $\{$AGE, SYSBP$\}$	$+0.023$	0.131	PASS
BPMEDS $\perp$ GLUCOSE $|$ $\{$AGE, SYSBP$\}$	$+0.012$	0.422	PASS
AGE $\perp$ BPMEDS $|$ SYSBP	$+0.025$	0.109	PASS
SEX $\perp$ SYSBP $|$ $\{$AGE, BMI$\}$	—	< 0.001	FAIL (expected; biologically valid direct edge retained)

Figure 1. Structural Causal DAG — Framingham Heart Study

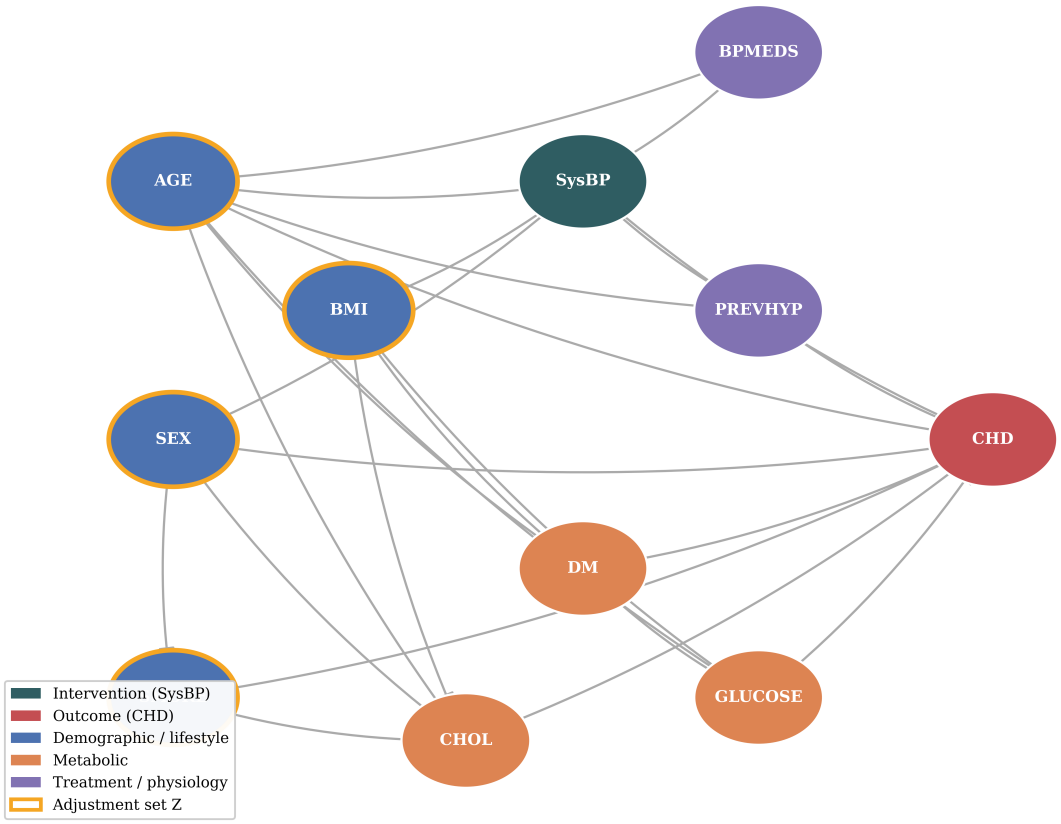


Figure 1. Corrected structural causal directed acyclic graph (DAG). Nodes represent Framingham Heart Study variables; directed edges encode causal mechanisms supported by cardiovascular pathophysiology. This DAG is a simplified representation and does not encode all biological dependencies. Four structural corrections differentiate this formulation from prior approaches: removal of the cross-sectionally unjustified smoking-to-SYSBP edge; addition of the biologically motivated age-to-diabetes edge; representation of BPMEDS as a descendant of SYSBP rather than a confounder; and exclusion of PREVHYP from the primary adjustment set. The back-door adjustment set $Z = \{\text{AGE}, \text{SEX\_MALE}, \text{BMI}, \text{CURSMOKE}\}$ is highlighted.

Average causal effect: the prediction–causation gap

The formal causal estimand was $E[Y \mid \text{do}(\text{SysBP} = s)]$, identified by the back-door adjustment formula [4]:

$$E[Y \mid \text{do}(\text{SysBP} = s)] = \sum_z E[Y \mid \text{SysBP} = s, Z = z]\, P(Z = z), \tag{2}$$

with adjustment set $Z = \{\text{AGE}, \text{SEX\_MALE}, \text{BMI}, \text{CURSMOKE}\}$. Total cholesterol and glucose were included in the outcome regression as precision covariates; they are not back-door confounders of SYSBP given Z, but reduce residual outcome variance without introducing bias. Three identifying assumptions were imposed: (i) consistency ($Y_i = Y_i(\text{SysBP}_i)$); (ii) positivity ($P(\text{SysBP} = s \mid Z = z) > 0$ for all z in the support of Z and all s in the intervention range); and (iii) conditional ignorability ($Y(s) \perp\!\!\!\perp \text{SysBP} \mid Z$ for all s).

G-computation (standardisation) applied to the multiply-imputed primary dataset ($n = 4{,}240$) yielded an interventional 10-year CHD risk of 14.11% under do(SysBP = 132 mmHg) and 10.71% under do(SysBP = 112 mmHg), giving:

$$\text{ACE} = 3.40\% \text{ absolute risk reduction } (95\% \text{ bootstrap CI: } 2.64\%\text{–}4.14\%;\ \text{RRR} = 24.1\%).$$

A complete-case sensitivity analysis ($n = 3{,}776$) yielded ACE = 3.46% (1.70% difference from the primary estimate), supporting robustness to the missing-data handling.

The naive observational estimate (comparing participants with SysBP near 132 vs. 112 mmHg) was 4.14%, standing 21.8% above the causal estimate in relative terms (log-odds inflation: +35.7%), driven primarily by the clustering of older age and higher BMI with elevated SysBP. The mean-Z plug-in adjusted estimate was 3.38%, confirming that full marginalisation over the empirical covariate distribution (rather than mere adjustment at covariate means) is necessary for unbiased causal inference. Figure 2 illustrates the magnitude of this divergence between the observational and interventional estimates across the range of SysBP values, and quantifies the bias attributable to conflating the conditional and interventional distributions. These results are consistent with the hypothesis that observational cardiovascular risk estimates overstate the expected absolute benefit of antihypertensive treatment.

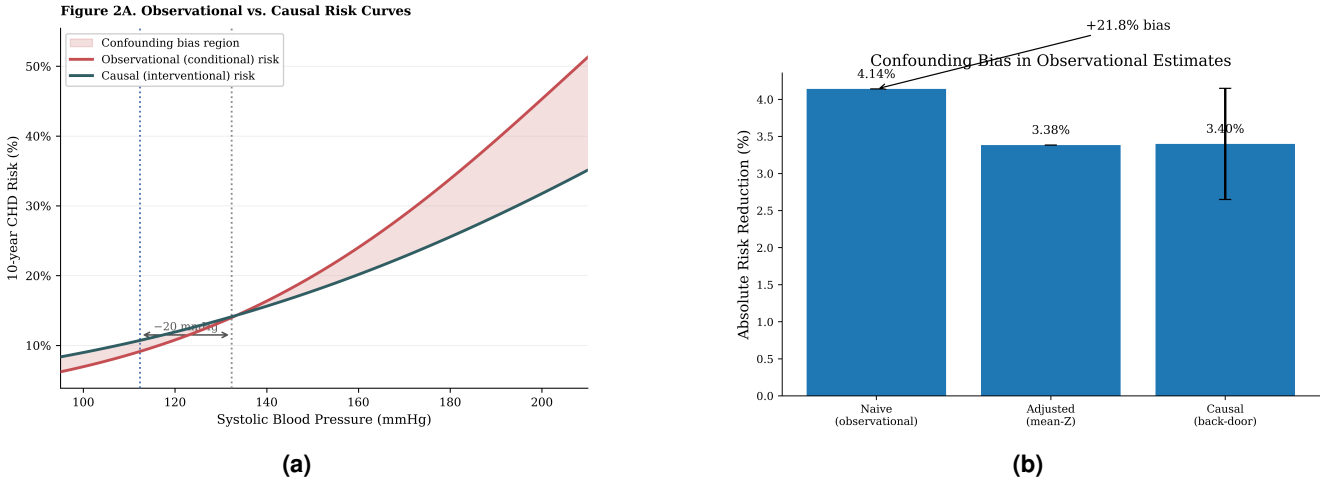


Figure 2. Bias in observational estimates of blood pressure benefit. (A) The observational (conditional) risk curve systematically exceeds the causal (interventional) risk curve estimated by back-door g-computation, reflecting confounding by age, BMI, and metabolic comorbidity. The shaded region represents the bias attributable to conflating the conditional and interventional distributions, corresponding to an approximate 21.8% relative overestimation of absolute treatment benefit at the primary intervention point. **(B)** Bar comparison of absolute risk reduction estimates from the naive observational approach, the mean-covariate plug-in adjusted estimate, and the fully marginalised g-computation causal estimate. Error bars represent 95% bootstrap confidence intervals.

Triangulation: propensity score matching and inverse probability weighting

A binary treatment contrast was defined by SysBP above versus below the sample median (128 mmHg). Sex-stratified nearest-neighbour PSM (caliper: 0.05; 2,169 matched pairs) achieved excellent post-matching covariate balance (AGE SMD: 0.680 → 0.053; BMI SMD: 0.598 → 0.070; CURSMOKE SMD: 0.211 → 0.004; SEX SMD: 0.021 → 0.000). The average treatment effect on the treated (ATT) from PSM was 4.84% (95% bootstrap CI: 2.58%–7.01%); the IPW population average treatment effect (ATE) was 6.32% (95% CI: 3.90%–8.75%).

These estimates are larger than the back-door ACE, which is expected: PSM and IPW evaluate a binary contrast at the sample median, representing a more extreme treatment contrast than the marginal 20 mmHg shift estimated by g-computation. Crucially, these three estimators rest on partially overlapping assumptions and distinct identification strategies; their directional consistency constitutes meaningful triangulation in support of a causal interpretation of the primary estimate, and does not imply that the estimates are directly comparable.

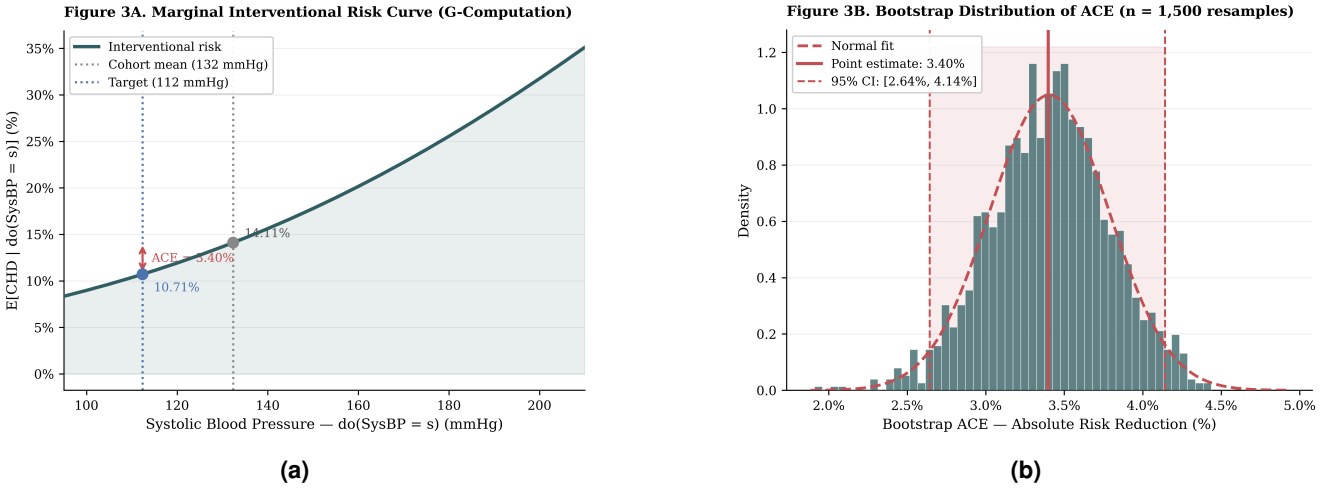


Figure 3. G-computation causal effect estimation. (A) Marginal interventional 10-year CHD risk as a function of do(SysBP), estimated by back-door standardisation. The grey dotted line marks the cohort mean (132.4 mmHg) and the green dotted line marks the reduced target value (112.4 mmHg). **(B)** Bootstrap distribution of the ACE for a 20 mmHg SysBP reduction across 1,500 resamples. The point estimate (3.40%) and 95% CI limits (2.64%, 4.14%) are marked by vertical dashed lines. Error bars represent the 2.5th and 97.5th percentiles of the bootstrap distribution.

Refutation testing

The permutation refutation yielded a null distribution of ACEs centred at -0.000081 (± 0.0054); the observed ACE fell in the extreme tail ($p < 0.001$ by permutation, Figure 4), confirming that the estimated effect cannot be attributed to chance or model artefact.

Three temporally or biologically impossible placebo instruments were evaluated. Placebo 1 (heart rate cannot causally precede age): coefficient -0.000516 (95% CI: $[-0.001081,\ 0.000048]$, $p = 0.073$), passed. Placebo 3 (cross-sectional glucose cannot retroactively determine historical smoking): coefficient -0.002948 (95% CI: $[-0.005866,\ -0.000030]$, $p = 0.048$), failed at the boundary of $\alpha = 0.05$. Placebo 2 (total cholesterol to sex) failed due to a genuine biological sex-to-cholesterol hormonal pathway encoded in the DAG; it was pre-specified as invalid and is reported for transparency only.

We note explicitly that Placebo 3 failed. The residual glucose–smoking association most likely reflects latent shared causes, behavioural and metabolic clustering not fully encoded in the DAG, rather than a direct causal path. Similarly, the total cholesterol–sex pathway reflects established hormonal–metabolic linkages. These failures indicate model simplification rather than gross misspecification, and do not invalidate the back-door identification strategy for the primary estimand; however, they represent a limitation that we acknowledge explicitly.

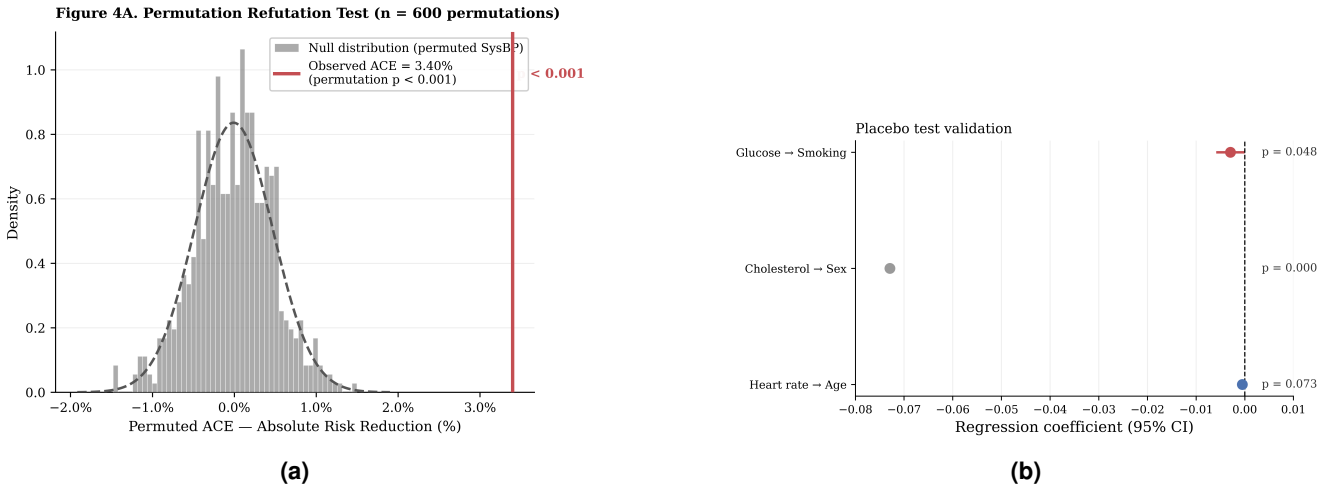


Figure 4. Validation and robustness of the causal estimate. **(A)** Null distribution of ACE estimates from 600 permutations of the SysBP treatment vector. The observed ACE (red vertical line, 3.40%) lies far outside the null distribution centred at approximately zero (permutation $p < 0.001$), rejecting the hypothesis that the effect arises from chance. **(B)** Regression coefficients and 95% confidence intervals for placebo instruments. Placebo 1 (heart rate to age) passed ($p = 0.073$). Placebo 3 (glucose to current smoking) failed ($p = 0.048$), most likely reflecting latent shared behavioural–metabolic causes not encoded in the DAG; this indicates model simplification. Placebo 2 (total cholesterol to sex) was pre-specified as invalid and is shown for transparency only. Error bars represent 95% confidence intervals.

Conditional average treatment effects

Heterogeneous treatment effects were estimated using R-Learner (continuous treatment; τ per mmHg SysBP) and T-Learner (binary treatment contrast at sample median). Spearman correlation between R-Learner and T-Learner individual estimates was 0.335, indicating directional consistency with substantial noise, as expected for semiparametric metalearners on observational data.

Statistically significant heterogeneity was detected across age strata (Kruskal–Wallis $p < 0.001$) and diabetes status (Mann–Whitney $p < 0.001$), and across sex (Mann–Whitney $p < 0.001$), as shown in Table 3. Age was the most clearly stratifying variable: participants younger than 45 years showed an estimated implied absolute risk reduction (ARR) of 2.22% (95% bootstrap CI: 1.87%–2.59%), while those aged 45–54, 55–64, and ≥ 65 showed ARRs of approximately 4.49%, 4.76%, and 3.74%, respectively. This age gradient is consistent with guideline recommendations derived from committee consensus [19] but is here derived from direct causal estimation.

Diabetic subgroup estimates must be interpreted with extreme caution. The diabetic subgroup comprised only $n = 109$ participants. A minimum detectable effect of 0.00240 per mmHg (at 80% power) substantially exceeds the observed τ of 0.00087, indicating that this subgroup is underpowered for reliable effect detection. The 95% confidence interval for the diabetic subgroup ARR (-4.94% to $+2.08\%$) crosses zero and is wide. We cannot make reliable inference about differential benefit in diabetic patients from this cohort; no claim that diabetic participants derive greater benefit from blood pressure reduction should be drawn from these data without replication in a larger or enriched diabetic sample. All heterogeneity results should be interpreted cautiously.

Table 3. Conditional average treatment effects (CATE) by pre-specified subgroup. Estimates from the R-Learner with gradient-boosted nuisance models; treatment is continuous SysBP. Implied ARR is expressed for a -20 mmHg intervention (ARR $= 20 \times \bar{\tau}$). 95% confidence intervals from 500-iteration bootstrap resampling of subgroup means. *Diabetic subgroup estimates ($n = 109$) are underpowered (minimum detectable effect $= 0.00240$ per mmHg vs. observed $\tau = 0.00087$) and should not be used for inference.* All heterogeneity results should be interpreted with caution.

Subgroup	n	$\bar{\tau}$ (per mmHg)	Implied ARR	95% CI	p
Sex (Mann–Whitney U)					
Male	1,820	0.00321	6.42%	[5.97%, 6.87%]	< 0.001
Female	2,420	0.00082	1.64%	[1.25%, 2.02%]	
Diabetes status (Mann–Whitney U)					
Diabetic†	109	-0.00087	-1.73%	[-4.94%, $+2.08$%]	< 0.001
Non-diabetic	4,131	0.00192	3.83%	[3.55%, 4.11%]	
Age strata (Kruskal–Wallis)					
<45 years	1,589	0.00111	2.22%	[1.87%, 2.59%]	< 0.001
45–54 years	1,479	0.00225	4.49%	[3.98%, 4.97%]	
55–64 years	1,062	0.00238	4.76%	[4.11%, 5.41%]	
≥65 years	110	0.00187	3.74%	[0.65%, 6.45%]	

†Underpowered; no reliable inference possible

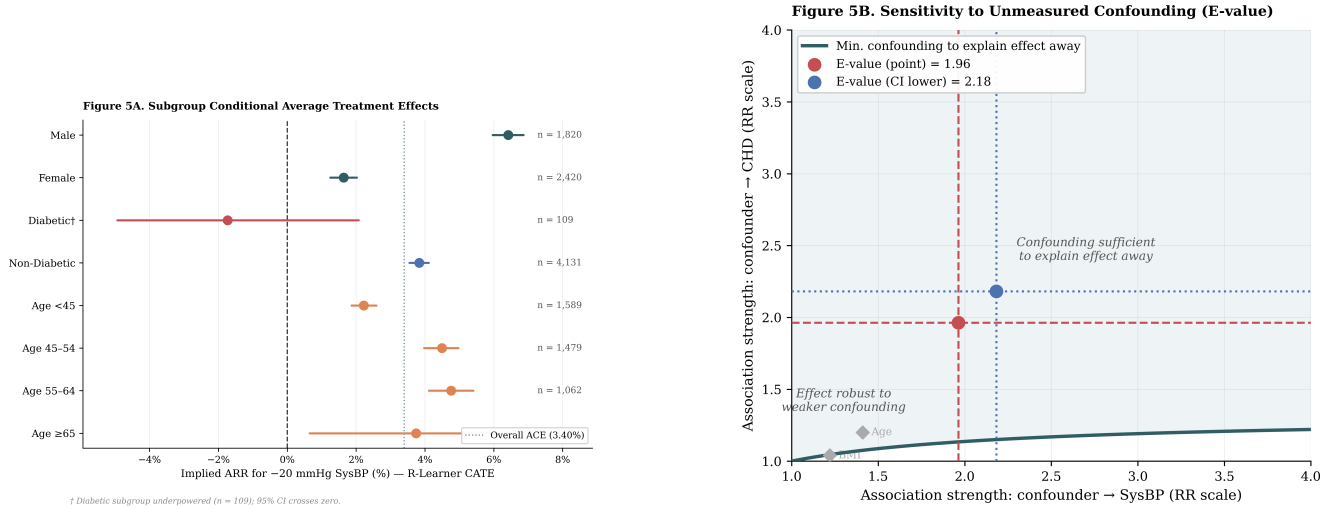
without replication.

Sensitivity analysis: E-values

The interventional risk ratio was 1.317. The E-value for the point estimate was 1.96; for the confidence interval lower bound it was 2.18 [10], exceeding the confounding strength of every measured covariate (age: $R^2_{\text{SysBP}} = 0.108$, $R^2_{\text{CHD}} = 0.043$; BMI: $R^2_{\text{SysBP}} = 0.083$, $R^2_{\text{CHD}} = 0.002$). This suggests the estimated causal effect is robust to unmeasured confounding of a magnitude not observed among the measured risk factors, though we acknowledge this does not rule out unmeasured confounders of sufficient strength. E-values declined to 1.88 and 1.62 for 15 and 10 mmHg reductions respectively; the 20 mmHg intervention magnitude is the only one achieving CI-lower-bound robustness above 2.0. Figure 5 displays the CATE subgroup effects and the sensitivity curve showing the minimum confounding strength required to nullify the estimated effect across the range of plausible unmeasured confounder associations.

Table 4. Supplementary Table S2. E-values by intervention magnitude [10]. E-values represent the minimum risk-ratio-scale association that an unmeasured confounder would need to have with both SYSBP and CHD to fully explain away the estimated causal effect. CI lower-bound E-values > 2.0 indicate high robustness; this threshold is met only at the 20 mmHg magnitude.

Intervention	ACE (%)	Causal RR	E-value (point)	E-value (CI lower)
-20 mmHg SysBP	3.40	1.317	1.96	2.18 ✓
-15 mmHg SysBP	2.55	1.233	1.74	1.88
-10 mmHg SysBP	1.70	1.150	1.53	1.62



(a) (b)

Figure 5. Heterogeneity in treatment effects and sensitivity to unmeasured confounding. **(A)** Implied ARR for a 20 mmHg SysBP reduction from the continuous-treatment R-Learner, expressed per subgroup. Error bars represent bootstrap standard errors. The diabetic subgroup ($n = 109$, marked with †) is underpowered (minimum detectable effect greater than observed τ); its estimate should not be interpreted as reliable. All heterogeneity findings are exploratory and require replication. **(B)** Sensitivity curve for unmeasured confounding. The curve represents the minimum strength of association that an unmeasured confounder would need with both SysBP and CHD to explain away the observed causal effect. The E-value (1.96) and CI lower-bound E-value (2.18) indicate that only moderately strong confounding could negate the estimated effect, and that this threshold exceeds the confounding magnitude observed for any measured covariate in the dataset.

Discussion

This study suggests that standard observational cardiovascular risk tools may overestimate the absolute benefit of SysBP reduction by approximately one-fifth. G-computation applied to the Framingham Heart Study Offspring Cohort yields an ACE of 3.40% (95% CI: 2.64%–4.14%) absolute risk reduction for a 20 mmHg decrease in SysBP, compared with a naive observational estimate of 4.14%. The relative discrepancy of 21.8% arises from the systematic clustering of risk factors, older age, higher BMI, and metabolic dysregulation, with elevated blood pressure. Antihypertensive therapy modifies blood pressure, not the entire comorbidity profile; the causal estimate captures only the blood-pressure-specific component of absolute risk.

The clinical implications of this finding, if confirmed, are consequential. A clinician relying on an observational risk score to estimate the benefit of antihypertensive treatment would, under this framework, systematically overestimate the expected absolute risk reduction by approximately one-fifth. For patients at moderate absolute risk, this overestimation could influence marginal prescribing decisions where the true causal benefit may be below a clinically meaningful threshold. Conversely, the treatment-effect heterogeneity analysis suggests, with the caveats noted above, that older patients may derive greater absolute benefit than younger individuals, which is consistent with the higher baseline event rates in this age group. These observations support the value of estimating interventional rather than associational effects in cardiovascular risk prediction, and suggest that causal methods could contribute to improved risk stratification. We stress that these are observational, cross-sectional estimates and should be interpreted as hypothesis-generating rather than definitive.

To illustrate the clinical consequence concretely, consider a 55-year-old male non-smoker with a BMI of 27 kg/m^2 and a SysBP of 148 mmHg whose ten-year CHD risk under the observational Framingham Risk Score is estimated at 12%. At a commonly applied 10% prescribing threshold, the observational estimate would suggest antihypertensive treatment is clearly warranted. Applying the 21.8% relative correction derived here, the corresponding causal estimate of absolute benefit from a 20 mmHg reduction falls to approximately 9.8%—below the treatment threshold. This is not an argument against antihypertensive treatment in such patients, for whom clinical judgment and total risk burden remain paramount; it is an argument that the quantity informing the prescribing decision should be an interventional estimate, not a conditional association. For patients clustered near any prescribing threshold, the systematic one-fifth overestimation documented here represents a structurally embedded bias that causal methods are positioned to correct.

The estimated ACE of 3.40% over ten years is directionally consistent with the SPRINT trial [17], which reported a 1.65% absolute reduction in major cardiovascular events at three years targeting SysBP < 120 versus < 140 mmHg in non-diabetic

high-risk adults, and with the meta-analysis by Ettehad and colleagues [18] reporting greater absolute benefit at higher baseline risk. Differences in magnitude are plausible given the longer time horizon, the broader population, and the use of the cohort mean SysBP as the baseline intervention point.

The methodological contribution of this work is the integration of DAG evaluation, g-computation, sex-stratified PSM, IPW, rigorous refutation testing, and semiparametric metalearner CATE estimation within a single reproducible pipeline applied to a publicly available dataset. The four DAG corrections introduced here transfer directly to any causal analysis of cardiovascular data in which antihypertensive medication and prevalent hypertension are measured. In particular, conditioning on antihypertensive medication as a confounder, rather than correctly treating it as a descendant of blood pressure, introduces collider bias that attenuates causal estimates; this error appears in several published causal analyses of cardiovascular data.

Limitations. We acknowledge several important limitations.

First, the identifying assumption of no unmeasured confounding conditional on Z is unverifiable from observational data. Unmeasured determinants of both SysBP and CHD risk, including dietary patterns, physical activity, socioeconomic position, and genetic predisposition, could bias the ACE in either direction, notwithstanding the E-value analysis. We cannot exclude residual confounding of meaningful magnitude.

Second, the DAG is a simplified representation of the cardiovascular system. The failures of Placebo 3 (glucose to current smoking) and the total cholesterol–sex instrument indicate that latent shared causes not encoded in the DAG are present. These failures do not invalidate the primary identification strategy, but they indicate model simplification and suggest that the true causal structure may be more complex than the specified DAG.

Third, the analysis is cross-sectional at baseline. The ten-year follow-up period introduces time-varying confounding not addressed by the static DAG structure. A marginal structural model with time-varying IPW would be required for a causal analysis of sustained blood pressure trajectories. The static DAG likely understates cumulative exposure in high-risk individuals, suggesting that the g-computation ACE may be conservative.

Fourth, the SYSBP–BPMEDS feedback loop is approximated by a static DAG; an instrumental variable design using a genetic instrument for blood pressure could resolve this structural simplification.

Fifth, diabetic subgroup CATE estimates ($n = 109$) are underpowered and unreliable. The point estimate and confidence interval for this subgroup should not be interpreted as evidence of differential benefit.

Sixth, the Framingham Heart Study Offspring Cohort is a largely white, community-based New England sample enrolled in the mid-20th century. Generalisation of these estimates to contemporary, ethnically diverse, or higher-risk clinical populations should be treated with caution.

Seventh, we cannot exclude measurement error in self-reported variables (smoking status) or single-occasion blood pressure measurements, which may introduce non-differential misclassification that could bias the ACE toward the null.

Eighth, missing-data handling relied on iterative imputation under a missing-at-random assumption; the MCAR assumption was supported for most variables but violated for BMI (χ^2, $p < 0.001$).

In summary, this study is consistent with the hypothesis that standard observational cardiovascular risk tools overestimate the causal benefit of blood pressure reduction by approximately one-fifth. Future work should prioritise validation against individual- participant data from randomised trials (SPRINT, ACCORD-BP), extension to the longitudinal Framingham data using marginal structural models, and replication in larger and more diverse cohorts.

Methods

Dataset and study population

The Framingham Heart Study (FHS) Offspring Cohort is a longitudinal cardiovascular epidemiology study established in 1948 [6]. The analytical dataset comprised 4,240 participants with 21 baseline variables including systolic and diastolic blood pressure, total serum cholesterol, fasting glucose, body mass index (BMI), age, sex, current smoking status, antihypertensive medication use (BPMEDS), prevalent hypertension (PREVHYP), diabetes mellitus, and heart rate. The primary outcome was 10-year incident CHD, a composite of myocardial infarction, angina pectoris, coronary insufficiency, and coronary death, with a raw event rate of 15.2% (644 events). The data are publicly available through BioLINCC (`https://biolincc.nhlbi.nih.gov`) and were used under the terms of the data use agreement.

Data preprocessing and missing data

Missing data rates were: glucose (9.15%), education (2.48%), BPMEDS (1.25%), total cholesterol (1.18%), cigarettes per day (0.68%), and BMI (0.45%). A chi-squared test of the missing-completely-at-random (MCAR) hypothesis, conditional on CHD outcome, was conducted for each variable. All variables except BMI supported MCAR ($p > 0.05$); BMI showed evidence against MCAR ($p < 0.001$) [11]. Primary analyses used iterative multiple imputation (scikit-learn `IterativeImputer`; 10 iterations) on the full cohort ($n = 4{,}240$). Sensitivity analyses on the complete-case sample ($n = 3{,}776$; 89.1% of the cohort) produced estimates within 1.70% of the primary analysis, supporting robustness.

Observational prediction model

A multivariable logistic regression was fitted with predictors: SysBP, total cholesterol, glucose, age, sex, current smoking, diabetes, prevalent hypertension, BMI, and antihypertensive medication use. Model performance was assessed by AUROC, average precision, and Brier score under 5-fold stratified cross-validation. Odds ratios from this model represent associations only and were not interpreted as causal effects.

Causal DAG specification and identification

We specified a structural causal model following Pearl's framework [4], encoding biological mechanisms as non-parametric structural equations. The DAG encodes 11 nodes and 24 directed edges. Four biologically motivated corrections to prior formulations are described in the Results.

The formal causal estimand was $E[Y \mid \mathrm{do}(\mathrm{SysBP} = s)]$, identified by the back-door adjustment formula (2), with adjustment set $Z = \{\mathrm{AGE}, \mathrm{SEX\_MALE}, \mathrm{BMI}, \mathrm{CURSMOKE}\}$. We verified that Z satisfies the back-door criterion: it blocks all back-door paths from SYSBP to CHD, contains no descendants of SYSBP, and renders potential outcomes conditionally independent of observed treatment given Z. Total cholesterol and glucose were included in the outcome regression as precision covariates.

Three identifying assumptions were imposed: (i) *Consistency:* $Y_i = Y_i(\mathrm{SysBP}_i)$; (ii) *Positivity:* $P(\mathrm{SysBP} = s \mid Z = z) > 0$ for all z in the support of Z and all s in the intervention range; (iii) *Conditional ignorability (exchangeability):* $Y(s) \perp\!\!\!\perp \mathrm{SysBP} \mid Z$ for all s, that is, no unmeasured confounding given Z.

DAG testable implications were evaluated by partial correlation, regressing out conditioning sets via ordinary least squares.

G-computation and average causal effect estimation

The interventional risk $E[Y \mid \mathrm{do}(\mathrm{SysBP} = s)]$ was estimated by g-computation (standardisation) [5]. A logistic outcome model was fitted:

$$\begin{aligned}\operatorname{logit} P(\mathrm{CHD}_{10} = 1) = \beta_0 + \beta_1\,\mathrm{SYSBP} + \beta_Z^{\top} Z \\ + \gamma_1\,\mathrm{TOTCHOL} + \gamma_2\,\mathrm{GLUCOSE} + \gamma_3\,\mathrm{DIABETES}.\end{aligned} \tag{3}$$

For each intervention value s, the full dataset was replicated with SysBP set to s for every individual, and the marginal interventional risk was computed as the mean predicted probability over the empirical distribution of Z. The ACE was the difference in interventional risks at the cohort mean SysBP (132.4 mmHg) versus 20 mmHg below (112.4 mmHg). Bootstrap confidence intervals (95%) used 1,500 resamples with the percentile method, applied to the complete estimation procedure including model fitting.

Propensity score matching and inverse probability weighting

For PSM, a binary treatment variable was defined by SysBP above versus below the sample median (128 mmHg). Propensity scores were estimated by logistic regression on Z. Sex-stratified nearest-neighbour matching was performed (exact matching on sex; caliper 0.05 on propensity score scale), yielding 2,169 matched pairs. Post-matching balance was verified by standardised mean differences. The ATT was estimated from matched outcome differences with 800-iteration bootstrap confidence intervals.

For IPW, stabilised Horvitz–Thompson weights were constructed and trimmed at the 98th percentile. The ATE was estimated as the weighted mean outcome difference with 800-iteration bootstrap confidence intervals.

It is important to note that g-computation, PSM, and IPW estimate different estimands (population ATE vs. ATT vs. ATT at a binary contrast vs. marginal ATE at a continuous shift) and are *not* directly comparable. Their directional consistency nonetheless provides triangulation evidence for a causal interpretation.

Refutation testing

Two refutation strategies were employed. First, three temporally or biologically impossible placebo instruments were evaluated: (i) heart rate cannot causally precede age; (ii) serum cholesterol cannot determine biological sex, retained for transparency despite pre-specified invalidity due to the sex-to-cholesterol hormonal path; (iii) cross-sectional glucose cannot retroactively determine historical smoking behaviour. Second, the SYSBP vector was permuted across 600 iterations to construct a null distribution for the ACE.

Conditional average treatment effect estimation

CATE was estimated using R-Learner [15] and T-Learner [16] metalearners. The R-Learner estimated nuisance functions for the conditional outcome and the conditional mean of SYSBP given covariates by 5-fold cross-fitting with gradient boosting machines (200 estimators; maximum depth 3; learning rate 0.05). The pseudo-outcome was computed for observations with $|T_i - \hat{e}(X_i)| > 2.0$, clipped to the [5th, 95th] percentile, and a gradient-boosted CATE model was fitted by regressing pseudo-outcomes on covariates weighted by the squared treatment residual. The T-Learner fitted separate outcome models to treated

and control subgroups and estimated CATE as their predicted probability difference. Subgroup CATE means were accompanied by 500-iteration bootstrap 95% confidence intervals. Heterogeneity across subgroups was tested by Mann–Whitney U (sex, diabetes) and Kruskal–Wallis (age strata) tests.

Sensitivity analysis

Sensitivity to unmeasured confounding was quantified by E-values [10]:

$$E\text{-value} = \mathrm{RR} + \sqrt{\mathrm{RR}(\mathrm{RR}-1)}.$$

Partial R^2 values for each measured confounder with respect to SYSBP and CHD provided empirical benchmarks for the required confounding magnitude.

All analyses were implemented in Python 3.11 using NumPy 1.26.4, Pandas 3.0.1, statsmodels, and scikit-learn. Random seeds were fixed at 42 throughout.

Data availability

All analyses were performed on the Framingham Heart Study public-use dataset, freely available through the National Heart, Lung, and Blood Institute's BioLINCC repository (https://biolincc.nhlbi.nih.gov). No new data were generated or collected for this study.

Code availability

Full analysis code (Python 3.11) is available at https://github.com/Suchibrata-Patra/bp-causal-inference-framing
The code includes all preprocessing, DAG specification and testing, causal effect estimation, refutation testing, and figure generation steps, and is sufficient for complete reproduction of all results reported.

Ethics statement

This study used de-identified, publicly available data from the Framingham Heart Study obtained through an approved data use agreement with BioLINCC/NHLBI. No new human participant data were collected. No additional ethical approval was required under the institutional review framework applicable to the analysis of publicly available de-identified datasets.

Author contributions

Suchibrata Patra conceived the study, developed the causal framework, performed all data analysis, implemented the computational pipeline, and wrote the manuscript.

Sourav Bhaduri supervised the research, contributed to study design and interpretation of results, and critically revised the manuscript.

Competing interests

The author declares no competing interests.

Acknowledgements

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

References

1. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation.* 1998;97(18):1837–1847.

2. SCORE2 Working Group; ESC Cardiovascular Risk Collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J.* 2021;42(25):2439–2454.

3. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *J Am Coll Cardiol.* 2014;63(25 Pt B):2935–2959.

4. Pearl J. *Causality: Models, Reasoning, and Inference.* 2nd ed. Cambridge: Cambridge University Press; 2009.

5. Hernán MA, Robins JM. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC; 2020.

6. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J III. Factors of risk in the development of coronary heart disease—six year follow-up experience: the Framingham Study. *Ann Intern Med.* 1961;55(1):33–50.

7. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res.* 2012;21(1):7–30.

8. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ.* 2016;352:i189.

9. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol.* 2017;46(2):756–762.

10. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med.* 2017;167(4):268–274.

11. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* 3rd ed. Hoboken: Wiley; 2019.

12. Halimi JM, Giraudeau B, Vol S, et al. The risk of hypertension in men: direct and indirect effects of chronic smoking. *J Hypertens.* 2002;20(2):187–193.

13. Gregg EW, Zhuo X, Cheng YJ, Albright AL, Narayan KMV, Thompson TJ. Trends in lifetime risk and years of life lost due to diabetes in the USA, 1985–2011: a modelling study. *Lancet Diabetes Endocrinol.* 2014;2(11):867–874.

14. Vasan RS, Larson MG, Leip EP, et al. Assessment of frequency of progression to hypertension in non-hypertensive participants in the Framingham Heart Study: a cohort study. *Lancet.* 2001;358(9294):1682–1686.

15. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika.* 2021;108(2):299–319.

16. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci USA.* 2019;116(10):4156–4165.

17. SPRINT Research Group; Wright JT Jr, Williamson JD, Whelton PK, et al. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med.* 2015;373(22):2103–2116.

18. Ettehad D, Emdin CA, Kiran A, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *Lancet.* 2016;387(10022):957–967.

19. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. *J Am Coll Cardiol.* 2018;71(19):e127–e248.