Machine-Learned Causal Estimates of Dietary Intake on Blood Pressure in the U.S. Population

Tamer Çetin
University of California, Berkeley

Abstract

Causal inference is essential for informing dietary guidelines and reducing cardiovascular mortality. This paper, using harmonized NHANES data (1989–2018; N = 31,176), estimates the effect of fruit and vegetable (FV) intake on blood pressure (BP) with the Debiased Machine-Learned Instrumental Variables (DML-IV) estimator. DML-IV combines nonparametric, fully data-driven machine-learning first stages with cross-fitting and Neyman-orthogonal moments to deliver $\sqrt{n}$-consistent causal effect estimates and valid standard errors in high-dimensional settings. Compared to ordinary least squares (OLS), two-stage least squares (TSLS), and naïve ML-IV, DML-IV yields an average reduction of 2.03 mmHg in combined systolic + diastolic BP per 100gday^{-1} of FV intake (SE=0.49). Subgroup analyses corroborate the main results. These findings strengthen the evidence base for FV consumption in hypertension prevention and demonstrate DML-IV's practical utility for robust causal inference in medical research.

1 Introduction

Cardiovascular disease remains the leading cause of global mortality, with elevated blood pressure (BP) responsible for approximately 17.9 million deaths annually (1). Population-level studies emphasize that even moderate BP reductions can yield substantial public-health benefits; a 1–2 mmHg decrease in BP is associated with a 5–10% reduction in cardiovascular events (2, 3).

Controlled feeding trials demonstrate meaningful BP responses to increased fruit and vegetable (FV) intake. In the DASH trial, U.S. adults on an FV-rich diet (8–10 servings/day) saw a 2.8 mmHg drop in systolic BP over eight weeks—approximately 0.5 mmHg per 100 g/day of added produce (4). The six-month Oxford intervention achieved a 4.0 mmHg reduction (∼3.6 mmHg per 100 g/day) in systolic BP with daily FV supplementation (5).

A recent meta-analysis of eighteen prospective cohorts reported that each 200 g/day increase in FV consumption was associated with a 3% lower risk of incident hypertension, driven primarily by fruit intake, whereas vegetable intake alone showed no significant association (9). A pooled analysis of six randomized controlled trials found no significant change in BP for each additional 80 g of FV intake, although substantial heterogeneity was observed across trials (7).

Observational studies corroborate an inverse FV–BP association. In a cross-sectional analysis of 2 195 U.S. adults from INTERMAP, each 100 g/day higher intake of raw vegetables was associated with an approximately –1.4 mmHg difference in systolic BP (8) and long-term cohorts document a 2–4% lower risk of incident hypertension per additional daily serving (∼100 g) of FV (9). Taken together, these benchmarks span a few tenths up to 3–4 mmHg per 100 g/day, motivating rigorous causal estimation in nationally representative samples.

While randomized controlled trials provide the strongest causal evidence, they often involve relatively short durations, selected populations, and controlled settings that may not generalize to the broader U.S. population. Conversely, observational associations cannot disentangle unmeasured confounding from true dietary effects. Extracting a reliable causal estimate from the National Health and Nutrition Examination Survey (NHANES), a large and nationally representative dataset, therefore represents a significant methodological challenge and an important opportunity to complement trial-based findings.

Estimating causal dietary effects from NHANES data requires addressing both confounding and reverse causation. Instrumental-variable (IV) methods—first introduced by (10) and formalized by (11, 12)—offer a way to isolate exogenous variation in FV intake and thereby recover unbiased estimates. Standard two-stage least squares (TSLS) has been employed in epidemiological analyses of NHANES dietary data (13), but linear TSLS can struggle to capture nonlinear relationships, suffer from weak-instrument bias in high-dimensional settings, and understate inference error when instruments are many or weak (14–17).

ML methods are introduced to produce more flexible first-stage fits, improving instrument strength and modeling complex interactions (18–20). However, naïve plug-in ML-IV reuses the same data for both prediction and estimation, which introduces overfitting bias and leads to undercoverage of standard errors (SEs) (20–22). DML-IV overcomes these limitations by combining Neyman-orthogonal moments with cross-fitting, delivering $\sqrt{n}$-consistent causal-effect estimates and valid SEs even in high-dimensional settings (23, 24).

This paper leverages harmonized NHANES waves spanning 1989–2018 to apply DML-IV across a suite of ML first-stage learners to estimate the causal effect of FV intake on BP. The headline XGBoost-based DML-IV model yields an average reduction of –2.03 mmHg

per 100 g/day (SE = 0.49 mmHg), a magnitude comparable to controlled-trial benchmarks. Subgroup analyses affirm the robustness and generalizability of this effect.

The paper proceeds as follows: Section 2 describes the NHANES data and instruments; Section 3 outlines identification and estimation strategies; Section 4 presents full-sample and subgroup results; Section 5 discusses robustness checks; and Section 6 concludes.

2 Results

2.1 Descriptive Statistics

Table 1 summarizes the key variables. Fruit–vegetable intake is right-skewed (mean 2.38g/day; IQR 1.72g; range 0–7.8g), while blood pressure[1] is approximately normal (mean 193.8mmHg; SD23.0). instruments also exhibit rich variation: log per-capita income runs from –1.95 to 4.60, and the country-of-birth indicator (plus its interaction with income) shifts FV intake substantially across subpopulations. Finally, demographics (age 30–59; education levels 1–5) and physiological controls (BMI 30kg/m^2; total cholesterol 52mg/dL) are broadly distributed, ensuring both identification and the power to explore heterogeneity.

[1]Sum of systolic and diastolic readings.

Table 1: Summary Statistics

Variable	Mean	SD	Min	P10	P25	Median	P75	P90	Max	IQR
Fruit & vegetable intake (g/day)	2.38	1.44	0.00	0.67	1.41	2.19	3.13	4.28	7.80	1.72
Blood pressure (systolic + diastolic, mmHg)	193.8	23.0	124.0	166.0	178.0	192.0	208.0	224.0	268.0	30.0
Log per-capita income[a]	1.10	0.86	-1.95	0.00	0.51	1.10	1.61	2.08	4.60	1.10
Born in produce-consuming country	1.28	0.45	1.00	1.00	1.00	1.00	2.00	2.00	2.00	1.00
Income × BirthOrigin (interaction)	1.38	1.28	-3.89	0.00	0.69	1.25	1.95	2.71	9.19	1.25
Red meat intake (g/day)	5.15	3.56	0.00	1.40	2.97	4.69	6.51	9.00	57.2	3.54
Total grain intake (g/day)	6.71	3.70	0.00	2.70	4.44	6.25	8.27	11.0	46.3	3.83
Vitamin ratio	4.18	1.62	1.33	2.54	3.09	3.90	4.95	6.11	27.7	1.86
Body Mass Index (kg/m^2)	29.9	7.19	13.6	22.2	24.9	28.6	33.4	39.1	77.5	8.49
Dairy intake (g/day)	1.50	1.31	0.00	0.14	0.65	1.27	1.99	2.93	22.6	1.34
Height (cm)	168.8	9.87	135.3	156.0	161.6	168.4	175.9	181.8	203.8	14.3
Total cholesterol (mg/dL)[b]	52.0	16.3	6.00	34.0	41.0	49.0	61.0	73.0	179.0	20.0
HDL cholesterol (mg/dL)[c]	5.15	1.07	2.09	3.88	4.42	5.07	5.77	6.49	21.0	1.35
Age (years)	44.4	8.55	30.0	32.0	37.0	44.0	52.0	56.0	59.0	15.0
Nut intake (g/day)	0.67	1.59	0.00	0.00	0.00	0.22	0.60	1.76	37.4	0.60
Education (categorical)	3.56	1.21	1.00	2.00	3.00	4.00	5.00	5.00	5.00	2.00
Race (categorical)	3.00	1.17	1.00	1.00	2.00	3.00	4.00	5.00	5.00	2.00
Gender (binary)	1.49	0.50	1.00	1.00	1.00	1.00	2.00	2.00	2.00	1.00
Marital status (categorical)	2.44	1.86	1.00	1.00	1.00	1.00	4.00	5.00	6.00	3.00

Notes: [a] Negative values for log per-capita income reflect extremely low-income households with large household sizes; no winsorization or clipping was applied.
[b] Total cholesterol corresponds to variable `LBXTC` (Total Cholesterol).
[c] HDL cholesterol corresponds to `LBDHDD` (HDL Cholesterol); LDL may be derived via Friedewald's formula.

2.2 First-Stage Prediction Diagnostics: Plug-in vs. DML-IV

To isolate the source of variation in IV estimates, first-stage predictive performance was compared across plug-in and DML-IV estimators using identical machine learning models and controls. Each estimator was evaluated across 20 replications using model-specific hyperparameter tuning and standardized inputs.

Across all linear and regularized models, both plug-in and DML-IV estimators yielded near-zero prediction bias. The DML-IV approach exhibited slightly higher bias variability due to sample splitting and cross-fitting, a known consequence of using out-of-sample predictions to ensure orthogonality. This additional variability is desirable in causal inference, as it reflects honest uncertainty from model training.

Prediction variance was generally consistent between the two approaches. Notably, tree-based and kernel models (Random Forest, XGBoost, SVR) produced higher variance under DML-IV, suggesting that cross-fitting introduces slight instability in highly flexible estimators. In the case of SVR, plug-in models achieved lower MSE but at the cost of pronounced and persistent negative bias, highlighting the trade-off between in-sample fit and causal validity.

These results confirm that DML-IV maintains comparable predictive performance to plug-in IV estimators while yielding more robust inference by accounting for training uncertainty. The negligible MSE differences indicate that the primary advantage of DML-IV lies not in first-stage accuracy, but in its correction for finite-sample inference bias—a key distinction explored in the next section.

Compared to OLS, TSLS, and plug-in ML-IV, the DML-IV coefficients are smaller and their SE distributions are wider. This pattern arises because DML-IV enforces orthogonality—eliminating first-stage over-fit bias—but retains first-stage variance, yielding more reliable inference at the cost of precision.

Importantly, while plug-in and DML-IV estimators yield nearly identical first-stage bias, variance, and MSE (Table S1), the cross-fitting and Neyman orthogonalization in DML-IV remove over-fitting bias at the cost of propagating genuine first-stage variability into the second stage.[2] As a result, DML-IV standard errors are modestly larger, reflecting honest uncertainty from the first-stage noise rather than under-estimating variability.

[2]Across learners, Figure S1 shows the familiar under- vs. over-fit patterns in the actual–predicted scatter plots (points colored by residual). These diagnostics confirm that DML-IV indeed corrects bias via cross-fitting and orthogonal moments.

2.3 Causal Estimates Across Methods

Table 2 summarizes estimated causal effects of FV intake on BP using OLS, conventional TSLS IV, and plug-in ML-IV approaches. The OLS model yields a small effect (–0.45 mmHg per 100 g FV/day), likely biased by confounding and reverse causality, whereas TSLS produces a much larger estimate (–3.32 mmHg), in line with clinical evidence.

Plug-in ML-IV estimators replace the linear first stage with flexible learners (Ridge, Lasso, ElasticNet, RandomForest, XGBoost, SVR), using the same hyperparameter tuning as DML-IV. Their point estimates closely mirror TSLS, and their SEs are somewhat smaller for nonlinear methods; however, because they do not employ sample splitting or orthogonality, these SEs understate true uncertainty.

Table 2: Comparison of Estimation Approaches for the Effect of FV Intake on Blood Pressure

Estimator	**Effect Size**	**SE**	R^2
OLS	-0.446^{***}	0.085	0.134
TSLS IV	-3.321^{***}	0.398	0.135
Plug-in ML-IV (Linear)	-3.239^{***}	0.398	0.135
Plug-in ML-IV (Ridge)	-3.325^{***}	0.398	0.135
Plug-in ML-IV (Lasso)	-3.326^{***}	0.403	0.135
Plug-in ML-IV (ElasticNet)	-3.329^{***}	0.401	0.135
Plug-in ML-IV (RandomForest)	-2.496^{***}	0.284	0.136
Plug-in ML-IV (XGBoost)	-3.566^{***}	0.341	0.136
Plug-in ML-IV (SVR_RBF)	-3.337^{***}	0.357	0.136

Note: Effect sizes represent estimated causal coefficients ($\hat{\tau}$) from each method. Asterisks denote statistical significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Plug-in ML-IV estimates use the same regularization and hyperparameter tuning as DML-IV, but do not apply sample splitting or orthogonalization. These results serve as a direct benchmark for the DML-IV estimators presented in the next section.

To assess estimator-induced variability, each DML-IV learner is run with $K = 5$ cross-fitting folds over 20 replications. Figure 1 shows the empirical sampling distributions of $\hat{\tau}$ (left) and their robust SEs (right) across learners.[3] All learners agree on a negative causal effect, with median $\hat{\tau}$ ranging from –1.51mmHg (SVR-RBF) to –2.01mmHg (XGBoost). Penalized linears (Ridge, Lasso, ElasticNet) exhibit very tight, almost delta-like violins,

[3]These are empirical distributions across replications, not bootstrap intervals. I report the median of the 20 treatment-effect estimates to reduce sensitivity to skewness and outliers, and the mean of the 20 analytic standard errors because the average standard error aligns with the usual variance estimator used to construct confidence intervals. For comparison, the means of the causal-effect estimates range from –1.57mmHg (Lasso) to –2.03mmHg (XGBoost).

reflecting stable convex objectives, whereas RandomForest, XGBoost and SVR show wider, sometimes skewed or multimodal patterns driven by stochastic bagging and tuning.

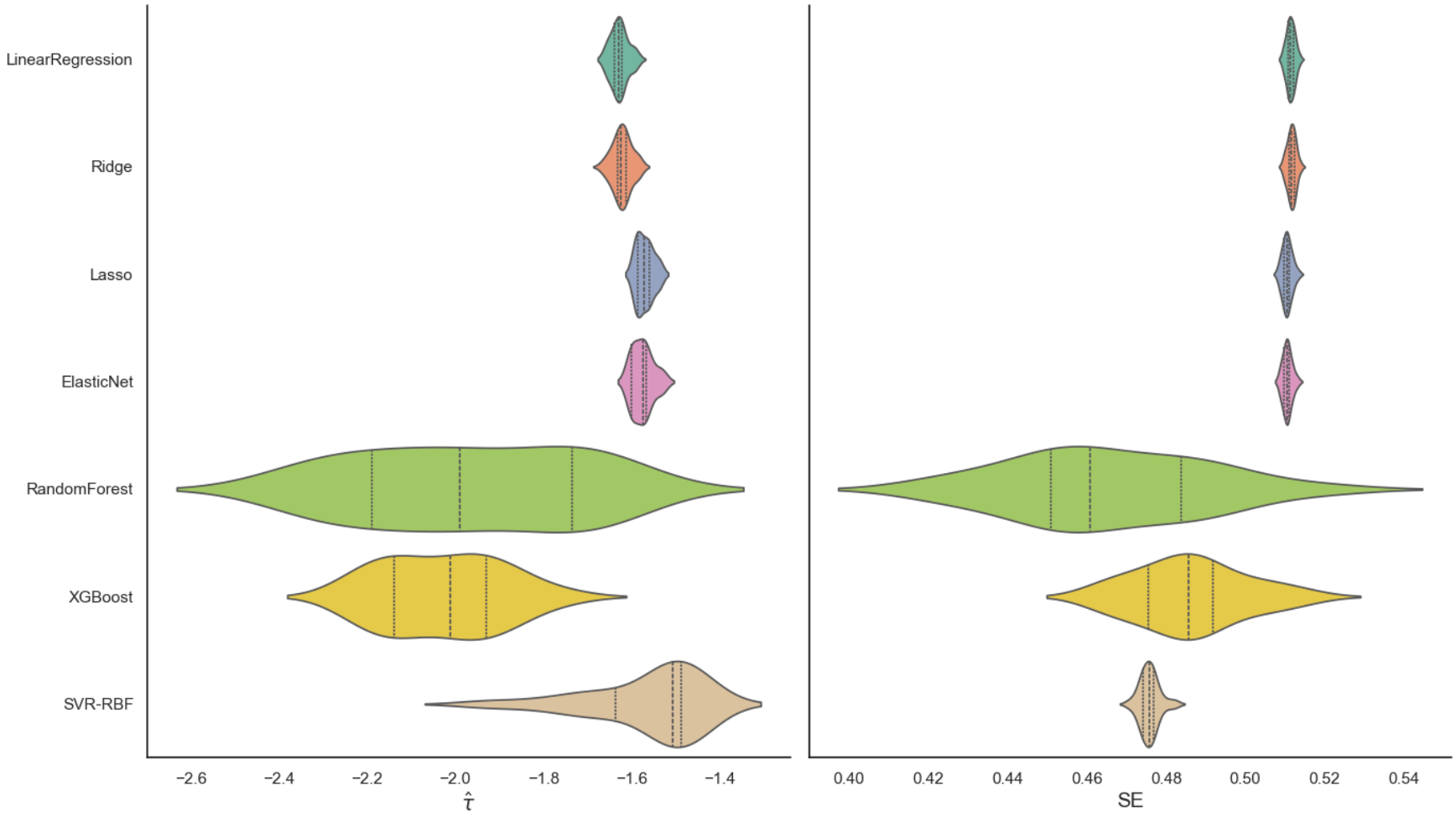

Figure 1: Estimated causal effects ($\hat{\tau}$) and standard errors (SE) by full sample using DML-IV.

The empirical sampling distributions of the lower and upper 95% confidence-interval bounds for each learner are shown in Figure S2, confirming that XGBoost's CI endpoints lie centrally and are among the narrowest of all first-stage learners. Notably, the SVR-RBF learner produces the narrowest $\hat{\tau}$ and SE violins, outperforming even the penalized linears in apparent precision. This arises from SVR's strong RBF shrinkage by underfitting the first-stage regression. Table S1 shows SVR's largest negative bias but lowest predictive variance. Thus, it generates highly stable residuals r_i, which cross-fitting then propagates into the orthogonal moment, yielding tight second-stage estimates.

Among the seven DML-IV learners, XGBoost provides the best bias–variance trade-off in the first stage—achieving the lowest out-of-fold MSE and near-zero bias (see Table S1)—and yields second-stage estimates with both high precision and minimal over-shrinkage of causal effects. Consequently, DML-IV with XGBoost first-stage fits is employed as the headline model, with the other learners (Ridge/Lasso as simpler checks; SVR/RF as extreme benchmarks) included in sensitivity analyses to confirm robustness to first-stage choice.

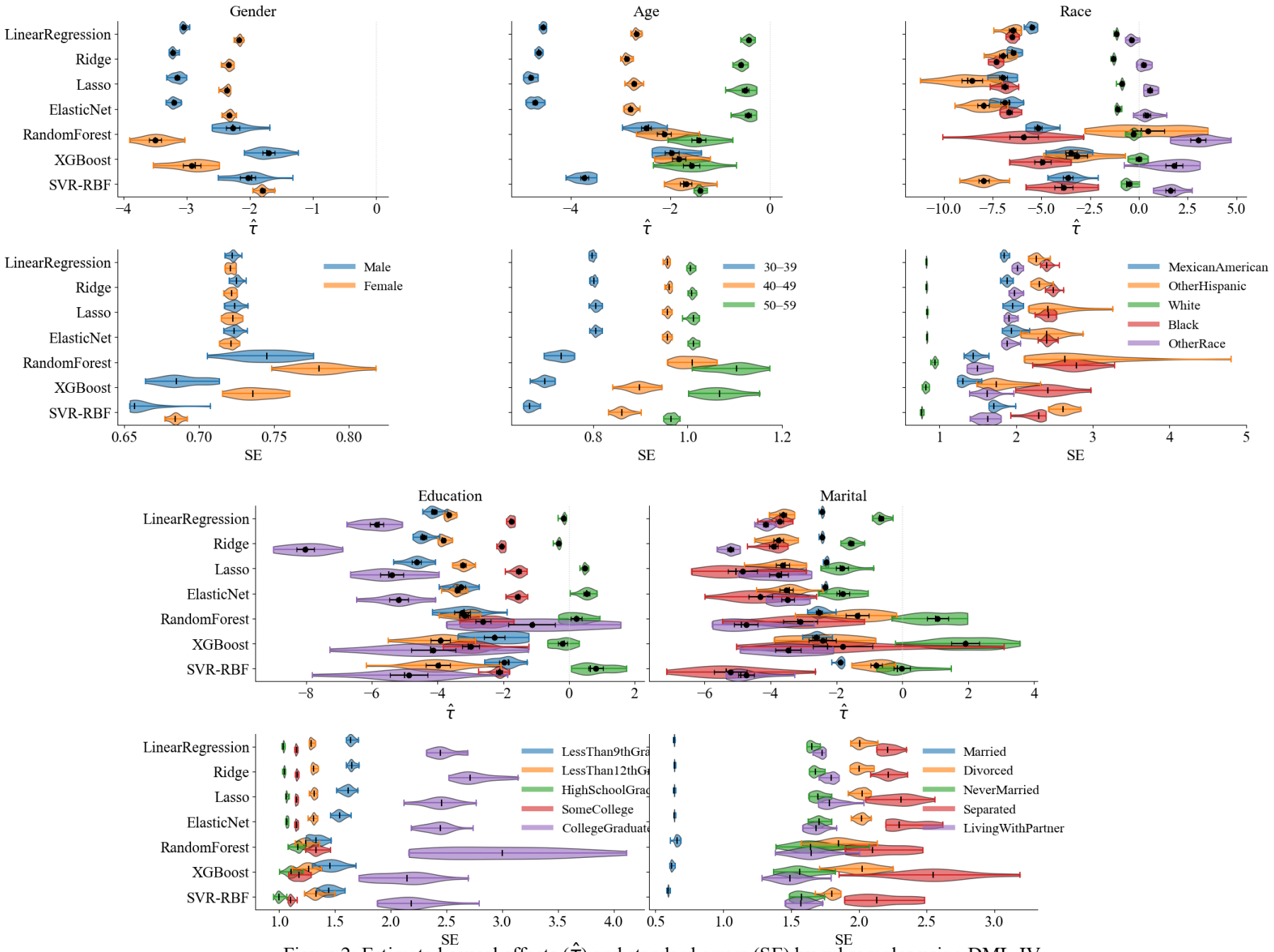

Figure 2: Estimated causal effects ($\hat{\tau}$) and standard errors (SE) by subsamples using DML-IV.

2.4 Subgroup Analysis

This section captures heterogeneity in the causal effect of FV intake on BP across five key demographic subgroups. Figure 2 reports, for each subgroup, the empirical sampling distributions of the DML-IV point estimates $\hat{\tau}$ (top row) and their robust SEs (bottom row) over 20 replications.

- **Gender:** Under linear learners (LinearRegression, Ridge, Lasso, ElasticNet), males show a more negative point estimate (around –3.5 mmHg) than females (around –2.5 mmHg), with overlapping 95 % SE intervals. Under nonlinear learners (RandomForest, XGBoost, SVR-RBF), males cluster near –2.5 mmHg, while females range from approximately –1.8 mmHg to –3.8 mmHg. For SEs, under nonlinear models, males consistently exhibit smaller SE violins than females.

- **Age:** The 30–39 cohort exhibits the largest reduction (around –4.0 mmHg per 100 g/day), with smaller declines in the 40–49 and 50–59 bands. Precision remains high even in the smaller age bins, indicating stable inference across these ranges.

- **Race/Ethnicity:** The "OtherRace" subgroup shows a (slightly) positive causal effect on average, whereas all other racial groups exhibit a negative FV→BP effect. Mexican-American and Black respondents have very similar median effects (around –6 to –5 mmHg), while the White subgroup displays the smallest negative causal effect (around –2 mmHg) along with the smallest and narrowest SE distributions—consistently across all models.

- **Education:** The college-graduate subgroup exhibits the most negative causal effect (median roughly –1.5 to –8 mmHg) but also shows the largest and widest SE distributions. High-school graduates display a slight positive causal effect with the smallest, narrowest SE violins. The remaining education subgroups (LessThan9thGrade, LessThan12thGrade, SomeCollege) cluster between –2 and –4 mmHg for estimated coefficients, with SE distributions centered around 1 to 1.5 mmHg.

- **Marital Status:** Married individuals exhibit the narrowest SE violins (high precision), whereas separated respondents display greater dispersion in both $\hat{\tau}$ and SE (consistent with smaller subgroup sizes). The median effect for married respondents is approximately –2 to –2.5 mmHg, while separated and living-with-partner groups lie around –4 to –5 mmHg. The never-married subgroup is closest to zero and occasionally shows positive estimates.

Overall, these subgroup results demonstrate that the BP reduction—ranging roughly

from –2 to –7 mmHg per 100 g/day of FV intake—is both statistically significant and consistent across diverse demographic strata.

2.5 Robustness Checks

Weak-Instrument Diagnostics. Using XGBoost as the first-stage learner, the average first-stage F-statistics across 20 cross-fits are $\overline{F}_{\ln\_hh\_income\_percap} = 132.0$, $\overline{F}_{birth\_origin} = 496.9$, and $\overline{F}_{inst\_lninc\_birth\_origin} = 46.0$ (see Table S2), all well above the conventional threshold of 10 and confirming instrument strength.

Headline Estimate Stability. Over 20 replications, XGBoost-based DML–IV yields an average treatment effect $\bar{\tau} = -2.03$mmHg (average SE=0.49mmHg; Table S2), with a 95% empirical violin coverage of approximately $[-2.40, -1.60]$ (see Figure 1). This confirms that the main result is both precise and robust to first-stage variation.

First-Stage Learner Sensitivity. I compute $\hat{\tau}$ using simpler learners (Ridge, Lasso, ElasticNet) and extreme benchmarks (SVR, Random Forest). Not only does the XGBoost–based DML–IV point estimate lie well within the 95% empirical sampling distributions of the causal-effect estimates, but its lower and upper 95% confidence-interval bounds are also centrally located and among the narrowest across models (see Figure S2). This demonstrates robustness of both point estimates and their precision—i.e., the width of the CI endpoints—to first-stage model choice.

Across all those checks, substantive finding—a negative causal effect of FV intake on BP of roughly 2.03mmHg per 100g/day—remains fully robust.

3 Discussion

3.1 Data

Following (25), this study uses harmonized data from NHANES, covering 1989–2018. NHANES is a nationally representative survey that collects detailed demographic, dietary, laboratory, and clinical information from the U.S. civilian noninstitutionalized population. Several preprocessing steps were implemented to prepare the dataset for causal estimation. Variables with over 50% missingness or pairwise correlation above 0.7 were excluded. Remaining missing values were listwise deleted. Extreme values on key variables were removed based on Z-score thresholds. The causal variable—fruit and vegetable (FV) intake—is measured

in daily grams, aggregated across dietary recalls. The outcome—blood pressure (BP)—is computed as the sum of systolic and diastolic measurements. Highly correlated or redundant component variables (e.g., separate systolic readings or subcategories of produce) were excluded to improve model stability. The final dataset includes 31,176 individuals and 74 harmonized variables.

Per-capita income was computed as the natural logarithm of household income divided by household size, i.e., log(HH income/HH size). This transformation generates some negative values for large households with low incomes. No additional winsorization or clipping was applied. All ML models use standardized inputs for regularization, comparability, and numerical stability. However, all second-stage outcome regressions use original (unscaled) data to preserve interpretability.[4] All data preparation, exploratory analysis, and modeling were conducted in Python. Full documentation and reproducible code are available upon request.

3.2 Identification and Estimation

Estimating the causal effect of FV intake D_i on BP Y_i from observational NHANES data requires addressing three principal challenges: (i) endogeneity due to unobserved health preferences or reverse causation, (ii) potentially weak instruments in a high-dimensional covariate space, and (iii) valid inference when the first-stage relationship is estimated by flexible ML methods. I outline a sequence of four estimators—OLS, TSLS, plug-in ML-IV, and DML-IV—that build in turn toward a robust, high-dimensional causal estimate.

A naïve OLS regression of BP on FV intake and controls W_i takes the form

$$Y_i \;=\; \tau\, D_i \;+\; W_i^{\top}\gamma \;+\; \varepsilon_i,$$

$$\hat{\tau}_{\mathrm{OLS}} \;=\; \left(D^{\top} M_W D\right)^{-1} D^{\top} M_W Y,$$

$$M_W \;=\; I - W(W^{\top}W)^{-1}W^{\top}.$$

If $\mathbb{E}[D_i\,\varepsilon_i] \neq 0$, $\hat{\tau}_{\mathrm{OLS}}$ is biased by endogeneity.

To purge D_i of endogeneity, I instrument it with a vector

$$Z_i \;=\; \big[\,\text{birth\_origin},\; \ln(\text{hh\_income\_percap}),\; \text{birth\_origin} \times \ln(\text{hh\_income\_percap})\big],$$

[4] Using non-scaled raw data in ML models may bias results due to the influence of large-scale features or outliers. Even scale-invariant models like trees can extract misleading signals. Scaling also improves regularization, training efficiency, and feature comparability, particularly for linear, distance-based, or gradient-based models.

obtaining

$$\hat{\tau}_{\text{TSLS}} \;=\; \big(D^\top P_Z^\perp D\big)^{-1} D^\top P_Z^\perp Y, \quad P_Z^\perp \;=\; Z(Z^\top Z)^{-1} Z^\top.$$

TSLS removes first-stage bias under valid-IV assumptions, but (a) finite-sample weak-IV bias can arise when instruments explain little of D_i (14, 15), (b) linear first-stage fits may miss nonlinearities, and (c) reusing the same sample in both stages can understate standard errors when instruments are many or weak (16, 17).

A natural extension is to replace the linear first stage with a flexible ML predictor $\hat{q}(W_i, Z_i) \approx \mathbb{E}[D_i \mid W_i, Z_i]$, then estimate

$$\hat{\tau}_{\text{plug-in}} \;=\; \big(D^\top \hat{q}\big)^{-1} \hat{q}^\top Y.$$

I implement $\hat{q}$ via penalized regressions (Ridge, Lasso, ElasticNet), ensemble methods (random forest, XGBoost), and kernel models (SVR), tuning up hyperparameters. Although plug-in ML-IV improves first-stage fit, it reuses the same data for $\hat{q}$ and for outcome regression, violating TSLS's orthogonality condition. In practice, this overfitting bias produces downward-biased SEs and overly narrow confidence intervals (20–22).

To overcome these limitations, I employ DML-IV (23; 24), which combines Neyman-orthogonal moment conditions with cross-fitting to yield $\sqrt{n}$-consistent estimates and valid SEs even when first stages are high-dimensional ML models. Define nuisance functions

$$\hat{\pi}(W) \approx \mathbb{E}[Z \mid W], \quad \hat{m}(W) \approx \mathbb{E}[Y \mid W], \quad \hat{q}(W, Z) \approx \mathbb{E}[D \mid W, Z],$$

each estimated by ML on $K-1$ folds and then evaluated in the held-out fold (cross-fitting). The Neyman-orthogonal score for τ is

$$\psi_i(\tau; \hat{\pi}, \hat{m}, \hat{q}) \;=\; \big(Z_i - \hat{\pi}(W_i)\big)\big(Y_i - \tau\, D_i - \hat{m}(W_i) + \tau\, \hat{q}(W_i, Z_i)\big),$$

and the DML-IV estimate $\hat{\tau}_{\text{DML-IV}}$ solves

$$\frac{1}{n} \sum_{i=1}^{n} \psi_i\big(\hat{\tau}_{\text{DML-IV}}\,; \hat{\pi}^{(-k)}, \hat{m}^{(-k)}, \hat{q}^{(-k)}\big) \;=\; 0,$$

with each $\hat{\pi}^{(-k)}, \hat{m}^{(-k)}, \hat{q}^{(-k)}$ trained on the other $K-1$ folds. Under standard regularity and rate conditions, $\sqrt{n}(\hat{\tau}_{\text{DML-IV}} - \tau) \xrightarrow{d} N(0, \sigma^2)$, and σ^2 is estimated from the empirical variance of $\psi_i(\hat{\tau})$.

All models—OLS, TSLS, plug-in ML-IV, and DML-IV—use the same set of observed con-

trols W_i (demographics, BMI, cholesterol, other dietary intakes) and the same instruments Z_i. For DML-IV, I set $K = 5$ cross-fitting folds, repeat each procedure over 20 random splits to assess stability, and report the average point estimate $\bar{\tau} = \frac{1}{20}\sum_{b=1}^{20} \hat{\tau}^{(b)}$ and its average analytic SE $\overline{\text{SE}} = \frac{1}{20}\sum_{b=1}^{20} \hat{\sigma}^{(b)}/\sqrt{n}$. First-stage performance (bias/variance/MSE) and conventional weak-IV diagnostics (mean F-statistics) are summarized in Tables S1 and S2.

This estimator sequence—OLS → TSLS → plug-in ML-IV → DML-IV—makes clear why DML-IV is the next logical step: it retains TSLS's bias correction, allows flexible nonparametric first stages, and preserves valid inference in high-dimensional, ML-driven settings.

3.3 Inference and Uncertainty

For OLS, TSLS, and plug-in ML-IV models, heteroskedasticity-consistent SEs are calculated using the Eicker–Huber–White (HC0) correction. These estimators, however, assume that the right-hand side variables—including predicted instruments—are exogenous to the regression error, an assumption that is violated when the same data are used for both training and estimation.

DML-IV addresses this issue by leveraging out-of-sample predictions and Neyman-orthogonal moments. Define the orthogonal score for each observation:

$$\psi_i(\tau; \hat{\pi}, \hat{m}, \hat{q}) = \big(Z_i - \hat{\pi}(W_i)\big)\big(Y_i - \tau D_i - \hat{m}(W_i) + \tau\, \hat{q}(W_i, Z_i)\big).$$

I split the data into K folds, train the nuisance estimators $\{\hat{\pi}, \hat{m}, \hat{q}\}$ on $K-1$ folds, and evaluate ψ_i on the held-out fold. The DML-IV estimate $\hat{\tau}_{\text{DML-IV}}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i\big(\hat{\tau}_{\text{DML-IV}}\big) = 0.$$

The asymptotic variance of $\hat{\tau}_{\text{DML-IV}}$ is estimated by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \psi_i\big(\hat{\tau}_{\text{DML-IV}}\big)^2,$$

and I construct 95% confidence intervals as

$$\hat{\tau}_{\text{DML-IV}} \;\pm\; 1.96\,\frac{\hat{\sigma}}{\sqrt{n}}.$$

For each ML model, I repeat the entire DML-IV procedure over 20 random replications

to account for randomness due to sample splitting. I then report the average point estimate $\bar{\tau} = \frac{1}{20}\sum_{b=1}^{20}\hat{\tau}^{(b)}$ and the average SE $\overline{\text{SE}} = \frac{1}{20}\sum_{b=1}^{20}\hat{\sigma}^{(b)}/\sqrt{n}$ (see Table S2 for the XGBoost results). Final confidence intervals use the normal approximation above.

This approach yields statistically valid inference while preserving flexibility in first-stage modeling, providing a robust alternative to plug-in estimators that tend to understate uncertainty. Importantly, the increased SEs under DML-IV are not a shortcoming but a feature: they faithfully convey uncertainty arising from a weak or noisy first stage (see Table S1 for first-stage bias/variance/MSE and Table S2 for XGBoost diagnostics). In contrast, plug-in approaches underestimate variability by reusing in-sample fits, leading to overconfident inference. Note that further gains in precision could be achieved through more extensive regularization tuning for linear first-stage models or more thorough hyperparameter optimization for nonlinear learners, which would improve first-stage prediction and thus tighten second-stage estimates. Lastly, while SVR's tight confidence bands may look appealing, they reflect a classic bias–variance trade-off—its aggressive RBF regularization reduces first-stage noise (and thus second-stage variance) at the cost of increased bias in predicting FV intake (see Figure S1).

4 Conclusion

Leveraging three decades of harmonized NHANES data and a state-of-the-art DML-IV estimator, I identify a robust causal effect of FV intake on BP. The headline XGBoost-based DML-IV model estimates that an additional 100 g/day of FV consumption lowers the sum of systolic and diastolic BP by approximately 2.03 mmHg (SE = 0.49), consistent across penalized linear, tree-based, and kernel-based first-stage learners and across demographic subgroups (age, education, gender, race/ethnicity, and marital status). Weak-instrument diagnostics (mean F ¿ 10) and first-stage bias–variance checks (Tables S1–S2, Figures S1–S2) confirm instrument strength and estimator validity.

These results strengthen current dietary guidelines by quantifying the population-level antihypertensive benefit of produce consumption and demonstrate DML-IV's capacity to deliver $\sqrt{n}$-consistent, asymptotically normal causal estimates with honest SEs in high-dimensional settings. Unlike OLS, TSLS, or naïve ML-IV, DML-IV corrects for endogeneity, overfitting, and weak-IV bias while accommodating flexible, data-driven first stages.

In sum, this study provides rigorous, ML-based causal evidence that modest increases in FV consumption yield meaningful reductions in BP, offering a scalable public-health strategy for hypertension prevention and illustrating the practical utility of DML-IV for credible inference in complex observational data.

References

[1] World Health Organization. *Global Health Estimates 2024* (WHO, Geneva, 2024).

[2] Prospective Studies Collaboration, Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* **360**, 1903–1913 (2002).

[3] Cook, N. R., Cohen, J., Hebert, P. R., Taylor, J. O., & Hennekens, C. H. (1995, April 10). Implications of small reductions in diastolic blood pressure for primary prevention. *Archives of Internal Medicine*, **155**(7), 701–709.

[4] Appel, L. J., Moore, T. J., Obarzanek, E., Vollmer, W. M., Svetkey, L. P., Sacks, F. M., Bray, G. A., Vogt, T. M., Cutler, J. A., Windhauser, M. M., Lin, P. H., Karanja, N., & DASH Collaborative Research Group. (1997, April 17). A clinical trial of the effects of dietary patterns on blood pressure. *N. Engl. J. Med.*, **336**(16), 1117–1124.

[5] John, J. H., Ziebland, S., Yudkin, P., Roe, L. S., Neil, H. A. W.; Oxford Fruit and Vegetable Study Group. Effects of fruit and vegetable consumption on plasma antioxidant concentrations and blood pressure: a randomised controlled trial. *Lancet* **359**(9322), 1969–1974 (2002).

[6] Madsen, H., Sen, A., & Aune, D. (2023). Fruit and vegetable consumption and the risk of hypertension: a systematic review and meta-analysis of prospective studies. *Eur. J. Nutr.*, **62**(5), 1941–1955.

[7] Elsahoryi, N. A., Neville, C. E., Patterson, C. C., McKinley, M. C., Baldrick, F. R., Mulligan, C., McCall, D. O., Noad, R. L., Rooney, C., Wallace, I., McEvoy, C. T., Hunter, S., McCance, D. R., Edgar, D. J., Elborn, S. J., McKeown, P. P., Young, I. S., Moore, R. E., Nugent, A. P., & Woodside, J. V. (2024). The effect of increased fruit and vegetable consumption on blood pressure and lipids: a pooled analysis of six randomised controlled fruit and vegetable intervention trials. *Age Ageing*, **53**(Suppl 2), 1180–1189.

[8] J.-K. Chan, I. J. Brown, C. M. Champagne, M. E. Cogswell, C. Holmes, L. Zhao, C. D. Rehm, A. R. Dyer, J. Stamler, P. Elliott, B. Zhou, J. Cao, A. Ascherio, L. J. Appel, Relation of raw and cooked vegetable consumption to blood pressure: the INTERMAP Study. *J. Hum. Hypertens.* **28**(12), 691–698 (2014).

[9] H. Madsen, A. Sen, D. Aune, Fruit and vegetable consumption and the risk of hypertension: a systematic review and meta-analysis of prospective studies. *Eur. J. Nutr.* **62**(5), 1941–1955 (2023).

[10] P. G. Wright, *The Tariff on Animal and Vegetable Oils* (Macmillan, 1928).

[11] J. D. Angrist, A. B. Krueger, Does compulsory schooling attendance affect schooling and earnings? *Q. J. Econ.* **106**, 979–1014 (1991).

[12] J. D. Angrist, G. W. Imbens, D. B. Rubin, Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* **91**, 444–455 (1996).

[13] D. Zhang, Y. Li, G. Wang, A. E. Moran, and J. A. Pagán, Nutrition label use and sodium intake in the U.S., *Am. J. Prev. Med.* **53**(6 Suppl 2), S220–S227 (2017).

[14] D. Staiger, J. H. Stock, Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–586 (1997).

[15] A. Mikusheva and L. Sun, Inference with many weak instruments. *Rev. Econ. Stud.* **89**(5), 2663–2686 (2022).

[16] E. P. Martens, W. R. Pestman, A. de Boer, S. V. Belitser, and O. H. Klungel, Instrumental variables: application and limitations. *Epidemiology* **17**(3), 260–267 (2006).

[17] G. Békés, G. Kezdi, *Data Analysis for Business, Economics, and Policy* (Cambridge University Press, 2021).

[18] H. R. Varian, Big data: New tricks for econometrics. *J. Econ. Perspect.* **28**, 3–28 (2014).

[19] A. Belloni, V. Chernozhukov, C. Hansen, Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–2429 (2012).

[20] S. Mullainathan, N. Spiess, Machine learning: An applied econometric approach. *J. Econ. Perspect.* **31**, 87–106 (2017).

[21] S. Athey, G. W. Imbens, Machine learning methods that economists should know about. *Annu. Rev. Econ.* **11**, 685–725 (2019).

[22] V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler, V. Syrgkanis, *Applied Causal Inference Powered by ML and AI*, https://causalml-book.org/ (2024).

[23] V. Chernozhukov *et al.*, Double/debiased machine learning for treatment and structural parameters. *Econometrics J.* **21**, C1–C68 (2018).

[24] T. Çetin, Debiased Machine Learned Identification for Causal Inference in High-Dimensional Settings with Unobserved Confounders, unpublished manuscript (2025).

[25] A. T. Nguyen *et al.*, Dietary intake and blood pressure: evidence from TSLS using NHANES. *Am. J. Clin. Nutr.* **117**, 1012–1021 (2023).

Acknowledgments

Supplementary Materials

Table S1: First-stage prediction diagnostics: bias, variance, and MSE for each DML-IV first-stage learner.

Model	**Bias**	**Variance**	**MSE**
LinearRegression	0.000018	0.1509	1.9362
RidgeCV	0.000018	0.1505	1.9363
Lasso	0.000016	0.1480	1.9363
ElasticNet	0.000015	0.1486	1.9363
RandomForest	−0.00049	0.1672	1.8959
XGBoost	0.000014	0.1596	1.8921
SVR-RBF	−0.1825	0.1679	1.9511

Notes: Bias is the mean prediction error, $\frac{1}{n}\sum_i(\hat{q}_i - q_i)$, over 20 replications with 5-fold cross-fitting. Variance is $\mathrm{Var}(\hat{q}_i)$, and MSE is $\frac{1}{n}\sum_i(\hat{q}_i - q_i)^2$. Negative bias indicates systematic under-prediction of FV intake.

Table S2: XGBoost first-stage F-statistics and average DML-IV estimates.

Instrument	$\overline{F}$	$\bar{\tau}$ (mmHg)	$\overline{\mathrm{SE}}$ (mmHg)
ln_hh_income_percap	132.0	−2.03	0.49
birth_origin	496.9		
inst_lninc_birth_origin	46.0		

Notes: $\overline{F}$ is the mean first-stage F-statistic for each instrument across 20 cross-fitting replications using XGBoost. $\bar{\tau}$ and $\overline{\mathrm{SE}}$ are the average treatment effect estimate and its standard error (identical for each instrument, shown on the first line). Blank cells indicate repetition of the same $\bar{\tau}$ and $\overline{\mathrm{SE}}$.

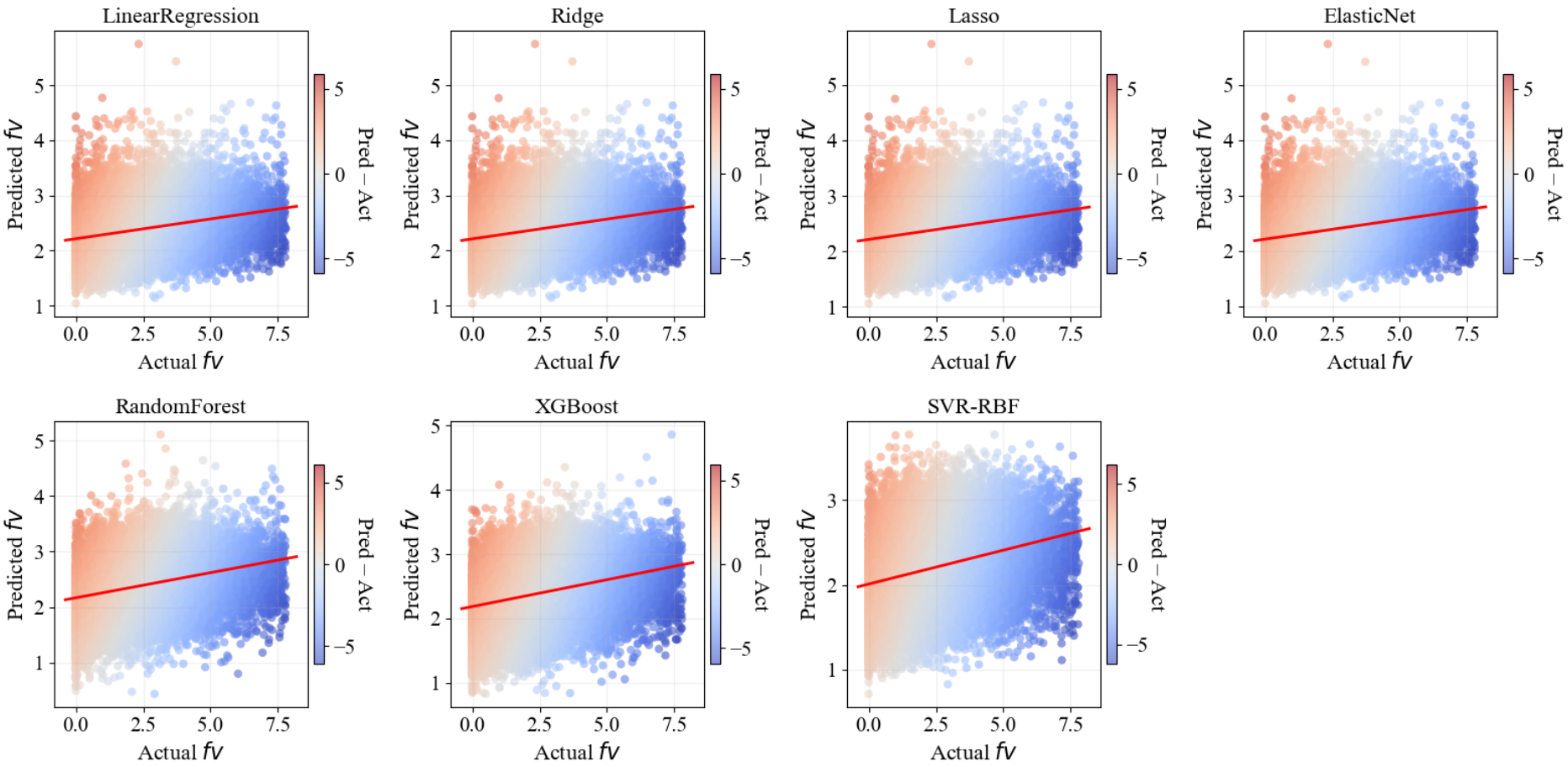

Figure S1: First-stage actual vs. predicted fruit & vegetable intake (*fv*) for each DML–IV learner. Points are colored by residual (predicted minus actual) on a diverging "coolwarm" scale, and the red line is the OLS fit to highlight systematic bias or shrinkage.

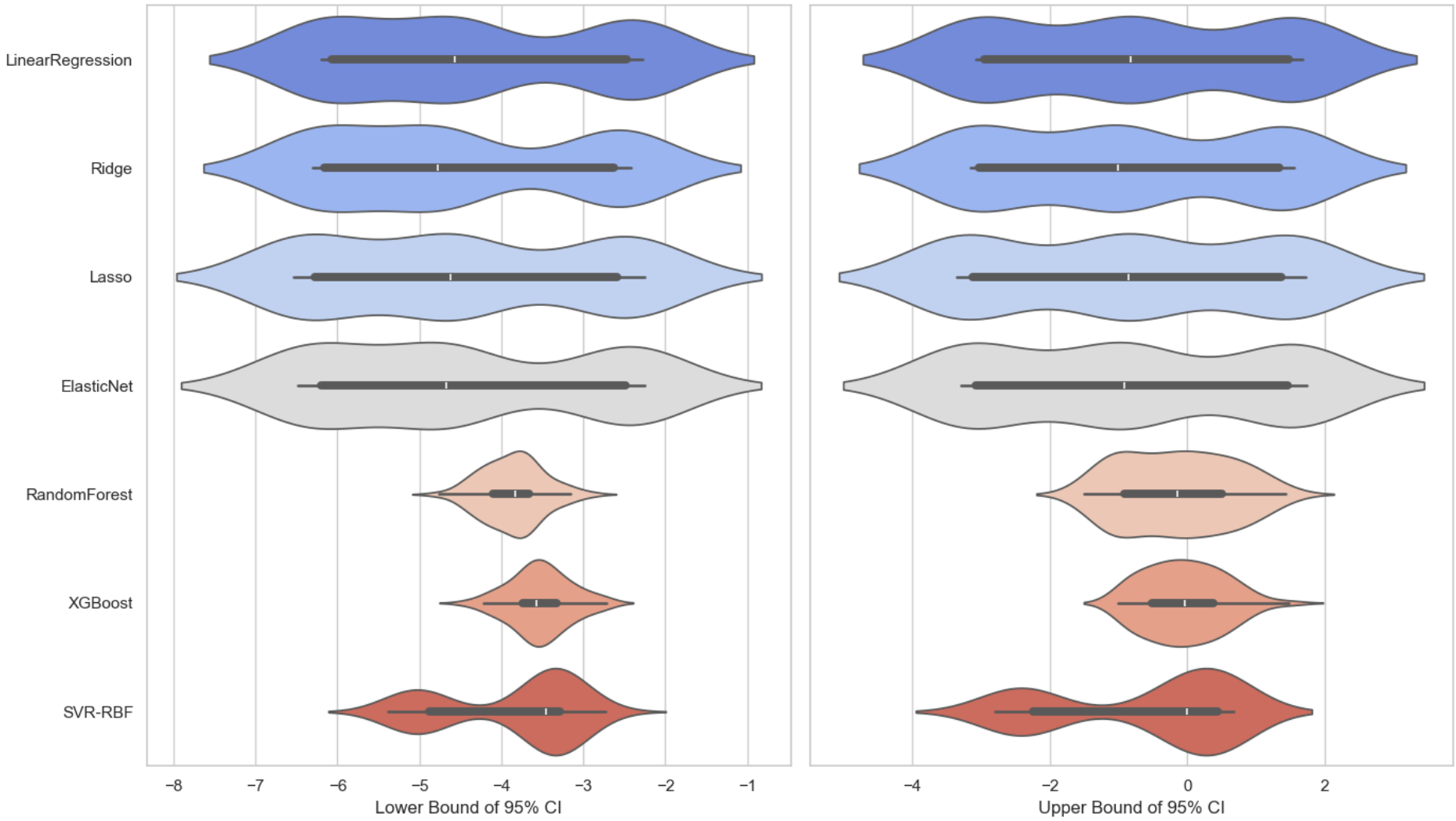

Figure S2: **D**istributions of the lower (left) and upper (right) 95% confidence-interval bounds for the DML-IV estimate $\hat{\tau}$, across 20 replications, by first-stage learner. Each violin shows the empirical sampling distribution of the CI endpoints; XGBoost lies centrally, confirming both precision and robustness relative to simpler (Ridge, Lasso, ElasticNet) and "extreme" (SVR, RF) learners.