Algorithm Selection for Estimating Causal Effects

Nulliparous Pregnancy Outcomes Study: Monitoring Mothers to Be

Zhaohua Zeng,[a] *Lisa M. Bodnar,*[b] *and Ashley I. Naimi*[a]

Background: The Super Learner is an ensemble learning method that has been widely used with doubly robust causal effect estimators. It is recommended to deploy the Super Learner with a diverse library of algorithms. To our knowledge, however, the magnitude of the improvements gained by including many algorithms has not yet been systematically evaluated in common epidemiologic research settings.
Methods: We applied Super Learning with two doubly robust estimators, augmented inverse probability weighting (AIPW) and targeted minimum loss-based estimation (TMLE), to estimate the average treatment effect (ATE) of high periconceptional dietary fruit and vegetable density on the risk of preeclampsia among 7,923 women from the nuMoM2b study. Using a reference ensemble with a diverse library of algorithms, we compared estimates under different sets of algorithms included in the Super Learner to evaluate whether ATE estimates were sensitive to library choices.
Results: The doubly robust estimators fitted with the reference Super Learner ensemble suggested ≥2.5 cups/1,000 kcal of total fruit and vegetable density was associated with a lower risk of preeclampsia. ATE estimated on the risk difference scale by AIPW was −0.019 (95% confidence interval = −0.036, −0.003) and by TMLE was −0.023 (95% confidence interval = −0.039, −0.007). Excluding any individual algorithm from the reference ensemble had little impact on estimates from either AIPW or TMLE. However, relying on a single algorithm (e.g., extreme gradient boosting) yielded results that were much more variable.
Conclusion: Our empirical findings support recommendations to build ensemble learners for doubly robust estimators using a diverse array of flexible machine learning algorithms.

Keywords: Average treatment effect; Dietary intake; Doubly robust estimator; Machine learning; Nutrition; Preeclampsia; Super Learning

(Epidemiology 2025;36: 760–768)

Doubly robust estimators, including augmented inverse probability weighting (AIPW) and targeted minimum loss-based estimation (TMLE), are increasingly being used to estimate average treatment effects (ATE) in a range of epidemiologic contexts.[1,2] In contrast to singly robust causal estimators (e.g., marginal standardization or inverse probability weighting), doubly robust estimators can be deployed using machine learning algorithms and still retain optimal statistical properties, such as low bias, honest confidence interval (CI) coverage, and root-n convergence.[3] Machine learning methods rely less on correct model specification assumptions, which often require knowledge of the true underlying functional form between the variables included in a parametric regression model. However, this knowledge does not usually exist in observational data.

The Super Learner has been widely used in fitting doubly robust estimators.[4,5] Super Learner allows for a wide degree of flexibility in capturing the underlying but unknown functional forms via inclusion of a variety of algorithms and regression models into a single meta-learner.[6] The included algorithms are combined into the Super Learner via a cross-validated loss function.[7] Theoretical results (i.e., the oracle inequality) show that Super Learner can perform asymptotically as well as the best algorithm in the ensemble.[8] This result suggests that researchers should deploy the Super Learner with a large and diverse library of algorithms, including parametric regression, penalized regression, spline-based algorithms, tree-based algorithms, and variations of these algorithms under different tuning parameter specifications.[9] In many software environments,[9–11] the number of algorithms that

Submitted October 02, 2024; accepted July 25, 2025
From the [a]Department of Epidemiology, Emory University, Atlanta, GA; and [b]Department of Epidemiology, University of Pittsburgh, Pittsburgh, PA.
This study is supported by grant funding from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD): R01HD102313 and R01HL174652 to L.M.B. and A.I.N. and U10 HD063036, RTI International; U10 HD063072, Case Western Reserve University; U10 HD063047, Columbia University; U10 HD063037, Indiana University; U10 HD063041, University of Pittsburgh; U10 HD063020, Northwestern University; U10 HD063046, University of California Irvine; U10 HD063048, University of Pennsylvania; and U10 HD063053, University of Utah. Support was also provided by respective Clinical and Translational Science Institutes to Indiana University (UL1TR001108) and University of California, Irvine (UL1TR000153).
Disclosure: The authors report no conflicts of interest.
The data used for this analysis contain protected health information and cannot be made publicly available. Code used to carry out the analysis is available on GitHub.
SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).
Correspondence: Ashley I. Naimi, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Rd., NE, Atlanta, GA 30322. E-mail: ashley.naimi@emory.edu.

ISSN: 1044-3983/25/366-760768
DOI: 10.1097/EDE.0000000000001906

can be included in any given Super Learner algorithm can be considerable.

However, including a large and diverse library of algorithms in the Super Learner can be associated with important tradeoffs. Notably, in big datasets, including many algorithms in the Super Learner library, can lead to considerably long computing times. This often leads to questions about which algorithms should be prioritized in any given setting. Few empirical studies have evaluated the impact of algorithm inclusion on the variability of causal effect estimates obtained from doubly robust estimators. Additionally, algorithm selection for the Super Learner ensemble is mostly arbitrary. While a diverse library of algorithms is often recommended,[9,12,13] the magnitude of the improvements that accrue from a given set of algorithms over another set has not yet been systematically evaluated in specific datasets commonly used in epidemiologic research settings.

Here, we use 7,923 observations from the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers to Be (nuMoM2b) cohort to estimate the ATE of high dietary fruit and vegetable density on the risk of preeclampsia using AIPW and TMLE with the Super Learner. Specifically, we compare ATE estimates under different sets of algorithms included in the Super Learner and evaluate whether they are sensitive to including different library sets in the ensemble.

METHODS

Study Population

We used data from the nuMoM2b, a large prospective pregnancy cohort study, which has been described previously.[14] In brief, 10,037 women from eight medical centers in the United States were enrolled from 2010 to 2013 during their first trimester of pregnancy. Participants were eligible for inclusion if they were 6–13 weeks pregnant with a singleton gestation, had ≤3 prior miscarriages, and had no pregnancy history that lasted ≥20 weeks of gestation. At enrollment, trained research personnel collected and ascertained baseline characteristics via physical examination and detailed interviews. Pregnancy outcomes were retrieved from medical records at least 30 days after delivery. The study protocol was approved by local institutional review boards at each study site, and all participants provided written informed consent.

Usual dietary intake in the 3 months around conception was assessed using a self-administered modified Block 2005 food frequency questionnaire (FFQ). The methodology for dietary assessment has been described in detail previously.[15,16] The FFQ food list included approximately 120 food and beverage items. We modified the FFQ to assess usual intake in the 3 months around conception and to add foods eaten commonly in Spanish-speaking populations. The FFQ has acceptable validity (most correlations range from 0.5 to 0.6 for nutrients in comparison to 4-day food records), including in racially diverse samples.[17–20]

Block Dietary Data Systems (Berkeley, CA) performed scanning, nutrient and food group mapping, and summary analysis of the FFQ data using software developed at the National Cancer Institute.[21] The food and beverage items were linked to the nutrient database, based on the USDA Food and Nutrient Database for Dietary Studies and the Food Patterns Equivalents Database,[22,23] to generate nutrient and food group variables.

Our exposure of interest was total periconceptional fruit and vegetable density. We dichotomized the density of fruit and vegetable intake at 2.5 cups/1,000 kcal per day, reflecting the 80th percentile of the distribution, and approximated the recommended intake as defined by the US Department of Agriculture Healthy US-Style Eating Pattern.[24] We used the food and nutrient estimates to calculate the Healthy Eating Index-2015 (HEI-2015) components. Our outcome of interest was preeclampsia as defined by the 2013 American College of Obstetricians and Gynecologists diagnostic criteria and adapted for the nuMoM2b study.[4,25]

We identified confounders using causal diagrams and adjusted for them in all analytic models.[26] Specifically, we accounted for demographic and reproductive health-related characteristics including maternal age, race–ethnicity, marital status, insurance status, education, acculturation, neighborhood walkability,[27] neighborhood area deprivation index,[28] percent of neighborhood with income below the federal poverty line, gravidity, prepregnancy body mass index, preconceptional smoking and alcohol consumption, preexisting hypertension and diabetes, sleep quality, health literacy level,[29] planned pregnancy, usage of assisted reproductive technologies, and symptoms of nausea/vomiting,[30] depressive,[31] stress,[32] and anxiety during the first trimester.[33] In addition, we included the total HEI-2015 score that excluded the fruit and vegetable components to account for the residual influence of periconceptional diet quality.[34] In total, 26 confounders (13 continuous, 13 categorical) were adjusted for in our analyses.

The original nuMoM2b cohort was n = 10,037. For the present analysis, we excluded participants without information on dietary intake (n = 1,786) and preeclampsia diagnosis (n = 571). Participants with missingness in potential confounders were retained using median and mode imputation for continuous and categorical variables, respectively, and adjusted for missingness indicators as covariates in all models.[35] Our final analytic dataset included 7,923 participants.

Statistical Analysis

As an empirical exploration on the impact of Super Learner ensemble algorithm selection on causal effect estimation, we estimated the ATE, which can be identified using our nuMoM2b data under exchangeability, counterfactual

consistency, positivity, and no interference.[36] All estimation procedures were implemented in two steps.

First, we fit both outcome and propensity score models using the Super Learner with squared (L2) loss function, which is minimized via the non-negative least squares method. Our library included up to 10 algorithm sets: (1) random forests with 500 trees, at least 50 observations in each node, and 5, 6, or 7 predictor variables randomly selected at each split; (2) extreme gradient boosting (XGBoost) with 500 trees, maximum tree depth of 4, 5, or 6, and shrinkage parameters of 0.01, 0.1, or 0.3; (3) generalized linear models (GLM); (4) elastic-net regularized generalized linear models (GLMNET) with mixing parameter $\alpha = 0$ (ridge regression), 0.25, 0.5, 0.75, or 1.0 (LASSO regression); (5) Bayesian generalized linear model with a t-distribution prior; (6) GLM with forward, backward, and stepwise variable selection by Akaike information criterion (AIC); (7) multivariate adaptive regression spline (MARS) with a maximum allowable degree of interaction of 3, 4, or 5; (8) single-hidden-layer neural networks with logistic activation function, 5, 7, or 9 units in the hidden layer, weight decay parameter of 0.01 or 0.1, and with or without skip-layer connections; (9) the k-nearest neighbor algorithm with 5, 10, or 50 nearest neighbors; and (10) the simple mean. Overall, our library included up to 41 subalgorithms with distinctive tuning parameter specifications, representing tree-based, regression-based, and penalized modeling approaches covering a wide range of algorithm use cases.

We defined our reference Super Learner ensemble as the meta learner that included the entire set of candidate algorithms under all distinct tuning parameter specifications, which was expected to minimize misspecification for both nuisance models and provide us with "pseudo-unbiased" ATE estimates. We then explored the impact of two versions of subensembles built from the reference ensemble library. The first version consisted of a Super Learner ensemble created by excluding a single candidate algorithm from the reference library at a time. This led to a total of 10 different Super Learner ensembles, which we generically denote $\mathrm{SL}_{(-x)}$. For example, $\mathrm{SL}_{(-xgboost)}$ denotes the Super Learner ensemble fit with all algorithms listed above, except the extreme gradient boosting algorithms.

The second version consisted of a set of algorithm-specific ensembles that included only one candidate algorithm with associated hyperparameter specifications per ensemble. This led to 10 different Super Learner algorithms, which we generically denote $\mathrm{SL}_{(x)}$. For example, $\mathrm{SL}_{(xgboost)}$ denotes the Super Learner ensemble fit with only the extreme gradient boosting algorithms including all variations constructed via different tuning parameter specifications.

All Super Learner ensembles were fit using 10-fold internal cross-validation and were embedded in an outer 10-fold sample-splitting scheme to avoid the empirical process conditions for the doubly robust estimators.[3,37,38] For both propensity score and outcome models fitted with the reference ensemble, we computed the 10-fold averages of algorithmic coefficients (weights) and corresponding mean squared errors (MSEs) to evaluate what proportion of each model could be explained by each candidate algorithm, respectively. To leverage the full sample size in modeling, we fitted one outcome model for all subjects in the training sample of each sample-splitting fold.

In this work, our estimand was the ATE of high total fruit and vegetable density on preeclampsia risk:

$$\psi = E\left(Y^{a=1} - Y^{a=0}\right)$$

where $Y^{a=1}$ denotes the preeclampsia outcome that would be observed if a pregnant woman consumed at least 2.5 cups/1,000 kcal of fruits and vegetables before conception ($Y^{a=0}$ otherwise). With the propensity score and outcome models fit using the referent and sub-Super Learner ensembles, we constructed AIPW and TMLE estimators of the ATE of high fruit and vegetable density on preeclampsia risk. For any participant i, we let Y_i denote the observed outcome, and A_i and W_i denote the treatment and confounder set received. Then, we can define $\hat{g}_a(W_i)$ and $\hat{Q}(A_i = a, W_i)$ as the estimated propensity score and outcome model prediction for participant i, respectively. The AIPW estimator can thus be denoted as:[39]

$$\hat{\psi}_{\mathrm{AIPW}} = \hat{E}\left[Y^{a=1}\right] - \hat{E}\left[Y^{a=0}\right]$$

where the counterfactual mean $E[Y^a]$ is identified and computed via:

$$\hat{E}[Y^a] = \frac{1}{n}\sum_{i=1}^{n}\{Q(A_i = a, W_i) + \frac{I(A_i = a)}{g_a(W_i)}[Y_i - Q(A_i = a, W_i)]\}$$

under the causal identifiability assumptions. Alternatively, the TMLE estimator can be implemented as:[40]

$$\hat{\psi}_{\mathrm{TMLE}} = \frac{1}{n}\sum_{i=1}^{n}\{Q^*(A_i = 1, W_i) - Q^*(A_i = 0, W_i)\}$$

where the updated Q^* can be obtained from a least favorable submodel, in our case defined as:

$$\mathrm{logit}\, Q^*(A_i = a, W_i) = \mathrm{logit}\, Q(A_i = a, W_i) + \epsilon H(A_i, W_i)$$

and the clever covariate $H(A_i, W_i)$ is defined as:

$$H(A_i, W_i) = \frac{I(A_i = 1)}{g_1(W_i)} - \frac{I(A_i = 0)}{g_0(W_i)}$$

The 95% CIs for the AIPW and TMLE estimates were calculated using the standard error of the corresponding estimated efficient influence function.[40,41]

To assess the impact of each sub-Super Learner ensemble, we first computed the ATE estimates of AIPW and TMLE with the reference ensemble as our base comparators, denoted

as $\psi_{\text{AIPW,SL}}$ and $\psi_{\text{TMLE,SL}}$, respectively. Similarly, we let $\psi_{\text{AIPW,SL}_{(\cdot)}}$ and $\psi_{\text{TMLE,SL}_{(\cdot)}}$ denote the estimates of subensembles (i.e., $\text{SL}_{(x)}$ and $\text{SL}_{(-x)}$), then we define metrics as:

$$\Delta_{\text{AIPW, SL}_{(\cdot)}} = \left|\psi_{\text{AIPW,SL}_{(\cdot)}} - \psi_{\text{AIPW,SL}}\right|$$

$$\Delta_{\text{TMLE,SL}_{(\cdot)}} = \left|\psi_{\text{TMLE,SL}_{(\cdot)}} - \psi_{\text{TMLE,SL}}\right|$$

where $\Delta_{AIPW,SL_{(\cdot)}}$ and $\Delta_{TMLE,SL_{(\cdot)}}$ denote the deviation in point estimates from the estimates of the reference for each subensemble under AIPW and TMLE applied. For example,

$$\Delta_{\text{TMLE,SL}_{(-\text{xgboost})}} = \left|\psi_{\text{TMLE,SL}_{(\text{-}xgboost)}} - \psi_{\text{TMLE,SL}}\right|$$

denotes the difference in ATE estimates from TMLE when the propensity score and outcome models were fit with all algorithms in the reference super learner ensemble except the extreme gradient boosting algorithms ($\psi_{\text{TMLE,SL}_{(-\text{xgboost})}}$), compared with the ATE estimates from TMLE when the propensity score and outcome models were fit with all ten algorithms in the referent ensemble ($\psi_{\text{TMLE,SL}}$). To explore the potential setting of data with a smaller sample size and similar data-generating mechanism, we repeated our analysis under a 10% subset randomly drawn from the analytic data (n = 792), details are provided in eAppendix 2; https://links.lww.com/EDE/C266.

All analyses were conducted using R with version 4.3.1,[42] Super Learner ensembles were fit using R package *SuperLearner* with version 2.0-29.[43]

RESULTS

Most participants were aged 25–34 years, non-Hispanic White, married, normal weight, planned their pregnancies, and had private insurance, some college or higher education, no prepregnancy smoking, and no history of preexisting diabetes and chronic hypertension. About 17% of participants had a dietary fruit and vegetable density of ≥2.5 cups per 1,000 kcal. Participants with higher density were more likely than their counterparts to be over the age of 25, non-Hispanic White, college or higher educated, married, privately insured, nonsmokers before pregnancy, and planned to be pregnant (Table 1).

Preeclampsia had occurred in 8.6% of the cohort. Women who consumed ≥2.5 cups of fruits and vegetables per 1,000 kcal were less likely to develop preeclampsia compared with those whose intake was <2.5 cups per l000 kcal (6.7% vs. 9.0%). Using doubly robust estimators with the reference Super Learner ensemble to adjust for covariates, we obtained the ATE estimates $\psi_{\text{AIPW,SL}} = -0.019$ (95% CI = −0.036, −0.003) and $\psi_{\text{TMLE,SL}} = -0.023$ (95% CI = −0.039, −0.007) (Table 2).

The distribution of algorithmic coefficients for the reference Super Learner ensemble differed in fitting the outcome and propensity score models. As shown in Tables 3 and 4, each coefficient represented the sum of the coefficients over the whole hyperparameter grid for a given algorithm. The predominant algorithms of the outcome model were GLM with stepwise selection by AIC, random forests, and GLMNET, while the propensity score model was weighted towards random forests, MARS, GLM, and neural networks. Tables 3 and 4 also summarized the subalgorithm performance (measured by MSE) for each algorithm included in our reference Super Learner ensemble, while the MSE for the entire reference ensemble was 0.077 for the outcome model and 0.120 for the propensity score model. The best subalgorithm performance (lowest MSE) was approximate for different algorithms included. Different tuning parameter specifications for a given algorithm in the ensemble led to a wide range of estimated MSEs within the same algorithm, especially for flexible machine learning algorithms that impose minimal functional form assumptions such as neural networks and k-nearest neighbors.

Figures 1 and 2 and eTables 1-1 and 1-2; https://links.lww.com/EDE/C266 showed the doubly robust estimates via each sub Super Learner ensemble (i.e., $\psi_{\text{AIPW},(\cdot)}$ and $\psi_{\text{TMLE},(\cdot)}$), accompanied by their deviation compared with the reference estimates (i.e., $\Delta_{\text{AIPW},(\cdot)}$ and $\Delta_{\text{TMLE},(\cdot)}$). Seven of the 10 ensembles that included only one candidate algorithm (SL_{mean}, SL_{GLMNET}, SL_{GLM}, SL_{BayesGLM}, $SL_{\text{random forest}}$, and $SL_{\text{neural network}}$) estimated a strong protective ATE when applying either of the estimators, with estimates ranging from −0.035 to −0.016 via AIPW, and −0.025 to −0.017 via TMLE. Both AIPW and TMLE estimates from $SL_{\text{random forest}}$ and the GLM-based SL_x ensembles were close to the corresponding reference estimates, whereas both doubly robust estimates of SL_{xgboost} and SL_{KNN} deviated away from the reference with relatively wide 95% CIs. The estimates of SL_{MARS} approximated the reference only when AIPW was used, while the estimates of the $SL_{\text{neural network}}$ approximated the reference only when TMLE was used (Figure 1, eTable 1-1; https://links.lww.com/EDE/C266).

All AIPW and TMLE estimates of the ensembles created by excluding a single candidate algorithm from the reference library (i.e., SL_{-x}) were nearly identical to the reference in both point estimates and 95% CIs. Excluding any individual algorithm from the reference ensemble had little impact on the ATE estimates, regardless of which doubly robust estimator was used (Figure 2, eTable 1-2; https://links.lww.com/EDE/C266). Our repeated analysis using the 10% randomly drawn subset yielded similar point effect estimates compared with the full-sample analysis (eAppendix 2; https://links.lww.com/EDE/C266).

DISCUSSION

Doubly robust estimators and the Super Learner ensemble method provide epidemiologists with a generalized framework to implement flexible, data-adaptive methods into causal effect estimation. Here, we compared the performance of doubly robust estimators fitted with different Super Learner

TABLE 1. Selected Characteristics of 7,923 Pregnant Women in the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be (nuMoM2b)

		Total Fruit and Vegetable Density	
	Total (N = 7,923)	**<2.5 cups/1,000 kcal (N = 6,612)**	**≥2.5 cups/1,000 kcal (N = 1,311)**
Maternal age (years)			
<25	2,598 (33%)	2,408 (36%)	190 (14%)
25–34	4,557 (58%)	3,633 (55%)	924 (70%)
≥35	768 (10%)	571 (9%)	197 (15%)
Maternal race/ethnicity			
Hispanic	1,353 (17%)	1,172 (18%)	181 (14%)
Non-Hispanic Black	892 (11%)	833 (13%)	59 (5%)
Non-Hispanic White	5,023 (63%)	4,059 (61%)	964 (74%)
Others	655 (8%)	548 (8%)	107 (8%)
Maternal education			
High school or less	1,397 (18%)	1,314 (20%)	83 (6%)
Some college/associate	2,246 (28%)	2,017 (31%)	229 (17%)
College graduate	2,337 (29%)	1,863 (28%)	474 (36%)
Graduate degree	1,943 (25%)	1,418 (21%)	525 (40%)
Marital status			
Married	5,147 (65%)	4,052 (61%)	1,095 (84%)
Not married	2,776 (35%)	2,560 (39%)	216 (16%)
Insurance			
Private	5,643 (71%)	4,510 (68%)	1,133 (86%)
Public	2,146 (27%)	1,987 (30%)	159 (12%)
Self-pay	134 (2%)	115 (2%)	19 (1%)
Prepregnancy body mass index			
Underweight	314 (4%)	272 (4%)	42 (3%)
Normal weight	4,453 (56%)	3,633 (55%)	820 (63%)
Overweight	1,702 (21%)	1,434 (22%)	268 (20%)
Obese	1,448 (18%)	1,267 (19%)	181 (14%)
Prepregnancy smoking			
No	6,591 (83%)	5,381 (81%)	1,210 (92%)
Yes	1,326 (17%)	1,225 (19%)	101 (8%)
Preexisting chronic hypertension			
No	7,692 (97%)	6,410 (97%)	1,282 (98%)
Yes	231 (3%)	202 (3%)	29 (2%)
Preexisting diabetes			
No	7,813 (99%)	6,524 (99%)	1,289 (98%)
Yes	110 (1%)	88 (1%)	22 (2%)
Planned pregnancy			
No	3,039 (38%)	2,744 (42%)	295 (23%)
Yes	4,880 (62%)	3,864 (58%)	1,016 (77%)

Counts may not sum to the total due to missing data.

ensembles in estimating the ATE of high dietary fruit and vegetable density on the risk of preeclampsia.

Doubly robust estimators allow for consistent and efficient causal effect estimation when machine learning algorithms are used to estimate nuisance functions, such as the propensity score and outcome model. Here, "double robustness" is often used to refer to the property that the estimators would be asymptotically consistent if at least one of the nuisance models (i.e., outcome and propensity models) is correctly specified. In addition, doubly robust estimators help to mitigate the influence of the curse of dimensionality for nonparametric machine learning algorithms, which results in slow estimator convergence rates in high-dimensional settings and leads to imprecise and biased estimates in finite samples.[44] By applying with sample-splitting or cross-fitting to avoid empirical process assumptions (e.g., the Donsker condition), doubly robust estimators can achieve a better convergence than singly robust methods, especially when the estimators of each nuisance function converge at slower (e.g., $n^{-1/4}$) rates. This facilitates obtaining valid inferences even when flexible machine learning methods are used to estimate effects.[41]

TABLE 2. Estimates of the Average Treatment Effect of ≥2.5 cups/1,000 kcal Total Fruit and Vegetable Density on the Risk of Preeclampsia Using the Reference Super Learner Ensemble

			ATE Estimates (95% CI)	
Fruit and Vegetable Intake	**Population at Risk**	**Preeclampsia (%)**	**AIPW**	**TMLE**
<2.5 cups/1,000 kcal	6,612	594 (9.0%)	Ref.	Ref.
≥2.5 cups/1,000 kcal	1,311	88 (6.7%)	−0.019 (−0.036, −0.003)	−0.023 (−0.039, −0.007)

AIPW indicates augmented inverse probability weighting; ATE, average treatment effect; TMLE, targeted minimum loss-based estimation.

TABLE 3. Coefficient and Subalgorithm MSE of Outcome Model Fitted With the Reference Super Learner Ensemble Using nuMoM2b Data (n = 7,923)

	Outcome Model				
			Subalgorithm MSE		
Algorithm	**Coefficient**	**N[a]**	**Min**	**Max**	**Geometric Mean**
Stepwise by AIC	0.589	3	0.077	0.079	0.078
Random forests	0.163	3	0.078	0.078	0.078
GLMNET	0.123	5	0.077	0.077	0.077
Neural network	0.041	12	0.082	0.109	0.094
MARS	0.034	3	0.081	0.083	0.082
XGBoost	0.027	9	0.084	0.085	0.085
K Nearest	0.013	3	0.080	0.093	0.086
Mean	0.010	1	0.079	0.079	0.079
BayesGLM	0.000	1	0.078	0.078	0.078
GLM	0.000	1	0.078	0.078	0.078

[a]N, number of subalgorithms in each candidate subensemble.

AIC indicates Akaike information criterion; BayesGLM, Bayesian generalized linear model; GLM, generalized linear model; GLMNET, elastic-net regularized generalized linear models; MARS, multivariate adaptive regression spline; MSE, mean squared error; nuMoM2b, Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be; XGBoost, extreme gradient boosting.

Using both AIPW and TMLE estimators with our "pseudo-unbiased" reference Super Learner ensemble, we found ≥ 2.5 cups/1,000 kcal of total fruit and vegetable density was associated with approximately 2 reduced cases of preeclampsia per 100 pregnancies. This aligns with our previous findings,[26] as we obtained roughly the same estimates of ATE but using an augmented ensemble with a more diverse library of candidate algorithms and an optimized coverage of tuning parameter specifications.

In our reference ensemble, we noted different algorithmic coefficients (weights) between the propensity score and outcome models. This suggests the importance of the data-adaptive property of the Super Learning method: different algorithms are driving the overall fit of the propensity score versus outcome model. Super Learning provides a generalized approach that can combine parametric, semiparametric, and nonparametric algorithms together by leveraging multifold cross-validation to assign weights for each algorithm, yielding a convex combination to minimize the overall cross-validated risk.[6,7,9] In this work, the MSE of the reference Super Learner ensemble for both outcome regression and the propensity score model was no greater than any candidate subalgorithms, indicating that the ensemble performed at least as well as the best-performing subalgorithm in minimizing the L2 loss.

The Super Learner can be applied with a variety of loss functions, such as L2, cross-entropy, and AUC losses.[7,13] We restricted our Super Learners to the L2 loss minimized via non-negative least squares for two reasons. First, in our experience, this is the most commonly deployed optimization approach when researchers use the Super Learner with AIPW or TMLE to estimate ATEs, and is supported by theoretical work in statistics.[41,45] Additionally, a recent paper suggests that, when nuisance functions are optimized via L2 loss, first-order ATE estimators such as AIPW and TMLE are not improvable (in a minimax sense) without additional structural assumptions.[46]

One fundamental motivation for applying flexible machine learning algorithms in causal inference is to overcome the reliance on parametric assumptions and thus avoid model misspecification. However, when we applied each candidate algorithm with AIPW and TMLE estimators, the individual performance of all flexible machine learning algorithms (except random forests) was inferior, with point estimates deviating away from the reference and relatively wide 95% CIs. In contrast, parametric GLM-based algorithms,

TABLE 4. Coefficient and Subalgorithm MSE of Propensity Score Model Fitted With the Reference Super Learner Ensemble Using nuMoM2b Data (n = 7,923)

	Propensity Score Model				
			Subalgorithm MSE		
Algorithm	Coefficient	N[a]	Min	Max	Geometric Mean
Random forests	0.374	3	0.121	0.121	0.121
MARS	0.154	3	0.125	0.128	0.127
GLM	0.136	1	0.121	0.121	0.121
Neural network	0.112	12	0.125	0.153	0.137
BayesGLM	0.090	1	0.121	0.121	0.121
GLMNET	0.062	5	0.121	0.121	0.121
XGBoost	0.046	9	0.135	0.140	0.138
Stepwise by AIC	0.018	3	0.122	0.138	0.127
K Nearest	0.009	3	0.129	0.150	0.139
Mean	0.000	1	0.138	0.138	0.138

[a]N, number of subalgorithms in each candidate subensemble.

AIC, Akaike information criterion; BayesGLM, Bayesian generalized linear model; GLM, generalized linear model; GLMNET, elastic-net regularized generalized linear models; MARS, multivariate adaptive regression spline; MSE, mean squared error; nuMoM2b, Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be; XGBoost, extreme gradient boosting.

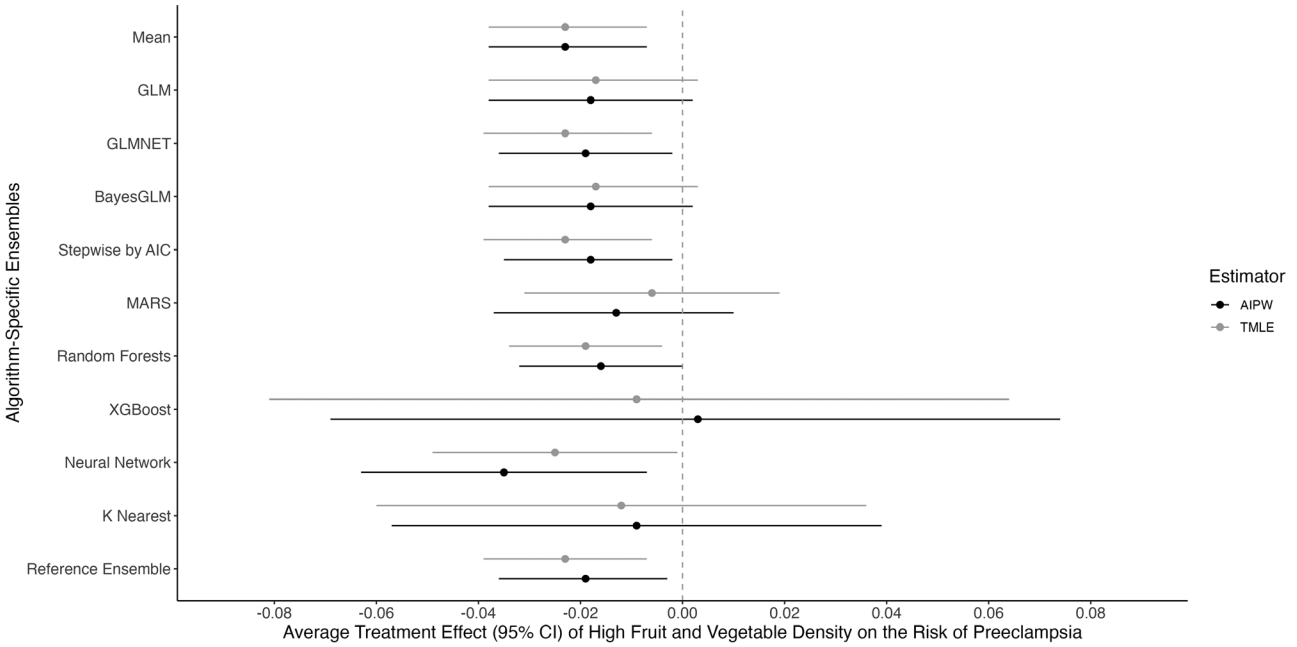

FIGURE 1. Estimates of the average treatment effect of ≥2.5 cups/1,000 kcal total fruit and vegetable density on the risk of preeclampsia using algorithm-specific Super Learner ensembles in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b).

especially GLM with elastic net regularization, obtained estimates that approximated the reference. In this case, the improvement brought by using a Super Learner ensemble with a diverse library that includes flexible machine learning algorithms was limited compared with the conventional parametric modeling approach.

Our work revealed one advantage of incorporating a wide variety of algorithms into the Super Learner in epidemiologic practice, that is, to maximize the likelihood of capturing the underlying functional format of both nuisance models in the doubly robust estimator. In our analysis, the estimates remained unchanged for all $\mathrm{SL}_{(-x)}$ subensembles, even when we removed the most dominant candidate algorithm in the outcome and propensity score models from the reference ensemble (Figure 2). However, this result should be interpreted conditional on the arbitrarily selected reference

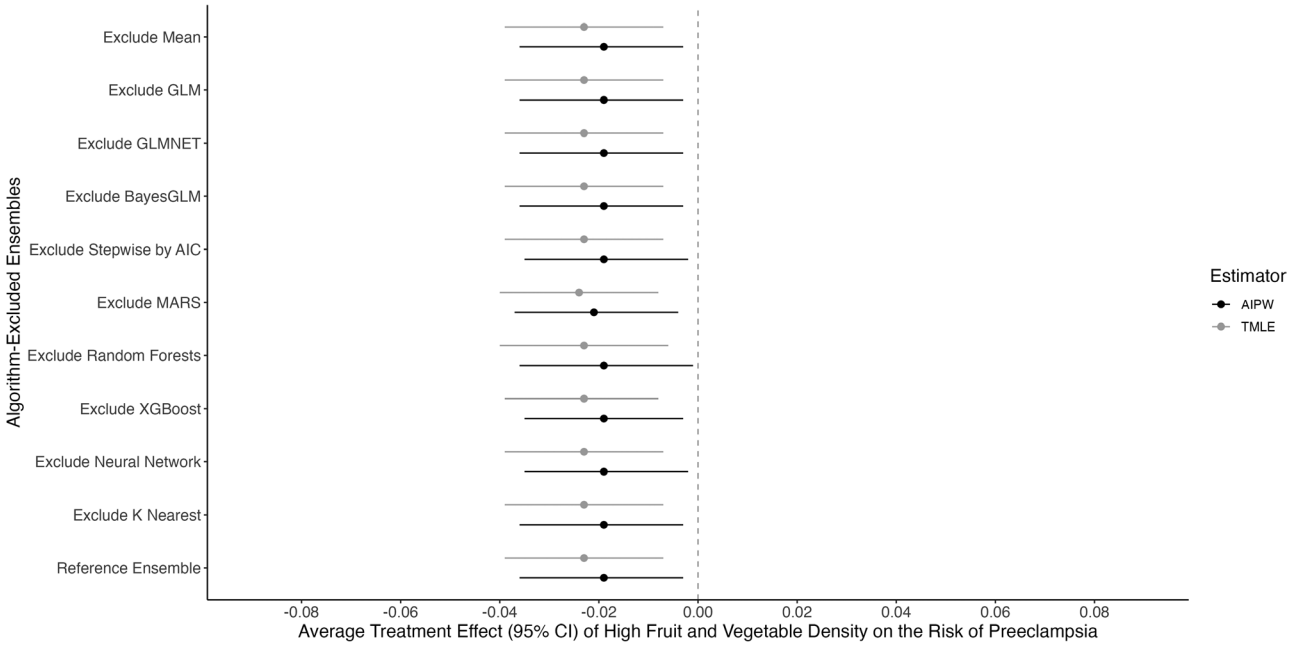

FIGURE 2. Estimates of the average treatment effect of ≥2.5 cups/1,000 kcal total fruit and vegetable density on the risk of preeclampsia when each algorithm was excluded from the reference Super Learner ensemble in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b).

Super Learner, as we included "mutually substitutable" candidate algorithms with significant overlap in strengths (e.g., GLM and GLMNET). The impact of excluding certain algorithm from a Super Learner ensemble with such overlapped algorithms on the effect estimates would therefore be much smaller than the impact of excluding the same algorithm from an ensemble without overlapping. Overall, our result is suggestive of the importance of using a wide array of diverse candidate algorithms in the Super Learner, despite the cost-effectiveness of this strategy may become another concern under limited computational resources.

Our findings should be considered with several key limitations in mind. First, we estimated effects in empirical data, and thus do not know the true ATE. In our setting, we used the effects obtained under a full reference library as a comparator for all other candidate libraries. However, it is possible that our referent estimates were biased due to an inappropriately specified Super Learner algorithm. Second, our findings are specific to our data and may not be generalizable to other settings of different topics, especially those with smaller sample sizes and different underlying effect sizes. We partially addressed the concern on sample size by repeating the analysis using 10% subset (eAppendix 2; https://links.lww.com/EDE/C266). Additionally, our overall approach only consisted of including or excluding one candidate algorithm each time from the reference ensemble. For example, all GLMNET subalgorithms with different tuning parameter specifications were included or excluded at once. However, we did not evaluate the impact of including or excluding a set of different algorithms at a time.

Our findings suggest that when applying doubly robust estimators with the Super Learning ensemble method in large epidemiologic data, a good performance in estimation can be achieved by including finite algorithms. Building ensembles with an extremely large array of flexible machine learning algorithms may only yield minimal improvement in precision and accuracy of doubly robust estimation and is therefore not cost-effective under limited computational resources. Despite this, using a diverse Super Learner ensemble does benefit the doubly robust estimation in practice as consistent effect estimates can be obtained with different algorithm choices.

REFERENCES

1. Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Polit Anal*. 2010;18:36–56.
2. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2:Article 11.
3. Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology*. 2021;32:393–401.
4. Bodnar LM, Cartus AR, Kirkpatrick SI, et al. Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes. *Am J Clin Nutr*. 2020;111:1235–1243.
5. Bosch NA, Teja B, Law AC, Pang B, Jafarzadeh SR, Walkey AJ. Comparative effectiveness of fludrocortisone and hydrocortisone vs hydrocortisone alone among patients with septic shock. *JAMA Intern Med*. 2023;183:451–459.
6. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33:459–464.
7. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:Article25.
8. van der Vaart AW, Dudoit S, van der Laan MJ. Oracle inequalities for multi-fold cross validation. *Stat Decis*. 2006;24:51–71.
9. Polley EC, van der Laan MJ. Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*. 2010;266.

10. Coyle JR, Hejazi NS, Malencia I, et al. sl3: Modern pipelines for machine learning and super learning. 2021. Available at: https://github.com/tlverse/sl3. Accessed 21 March 2025.
11. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
12. Balzer LB, Westling T. Demystifying statistical inference when using machine learning in causal research. *Am J Epidemiol*. 2023;192:1545–1549.
13. Phillips RV, van der Laan MJ, Lee H, Gruber S. Practical considerations for specifying a super learner. *Int J Epidemiol*. 2023;52:1276–1285.
14. Haas DM, Parker CB, Wing DA, et al; NuMoM2b study. A description of the methods of the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b). *Am J Obstet Gynecol*. 2015;212:539.e1–539.e24.
15. Bodnar LM, Kirkpatrick SI, Parisi SM, Jin Q, Naimi AI. Periconceptional dietary patterns and adverse pregnancy and birth outcomes. *J Nutr*. 2024;154:680–690.
16. Petersen JM, Naimi AI, Bodnar LM. Does heterogeneity underlie differences in treatment effects estimated from SuperLearner versus logistic regression? an application in nutritional epidemiology. *Ann Epidemiol*. 2023;83:30–34.
17. Block G, Woods M, Potosky A, Clifford C. Validation of a self-administered diet history questionnaire using multiple diet records. *J Clin Epidemiol*. 1990;43:1327–1335.
18. Siega-Riz AM, Savitz DA, Zeisel SH, Thorp JM, Herring A. Second trimester folate status and preterm birth. *Am J Obstet Gynecol*. 2004;191:1851–1857.
19. Block G, Thompson FE, Hartman AM, Larkin FA, Guire KE. Comparison of two dietary questionnaires validated against multiple dietary records collected during a 1-year period. *J Am Diet Assoc*. 1992;92:686–693.
20. Kristal AR, Feng Z, Coates RJ, Oberman A, George V. Associations of race/ethnicity, education, and dietary intervention with the validity and reliability of a food frequency questionnaire: the Women's Health Trial Feasibility Study in minority populations. *Am J Epidemiol*. 1997;146:856–869.
21. Epidemiology and Genomics Research Program, National Cancer Institute. *Diet*Calc Analysis Program, Version 1.5.0*. National Cancer Institute; 2012.
22. Agricultural Research Service Food Surveys Research Group, US Department of Agriculture. *USDA Food and Nutrient Database for Dietary Studies, Version 1.0*. US Department of Agriculture; 2004.
23. Agricultural Research Service Food Surveys Research Group, US Department of Agriculture. *Food Patterns Equivalents Database 2011-12*. US Department of Agriculture; 2014.
24. U.S. Department of Agriculture and U.S. Department of Health and Human Services. *Dietary Guidelines for Americans, 2020-2025*. 9th ed. US Government Publishing Office; 2020. Available at: DietaryGuidelines.gov. Accessed 23 May 2025.
25. Hypertension in pregnancy. Report of the American College of Obstetricians and gynecologists' task force on hypertension in pregnancy. *Obstet Gynecol*. 2013;122:1122–1131.
26. Bodnar LM, Kirkpatrick SI, Roberts JM, Kennedy EH, Naimi AI. Is the association between fruits and vegetables and preeclampsia due to higher dietary vitamin C and carotenoid intakes? *Am J Clin Nutr*. 2023;118:459–467.
27. Giles-Corti B, Macaulay G, Middleton N, et al. Developing a research and practice tool to measure walkability: a demonstration project. *Health Promot J Austr*. 2014;25:160–166.
28. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible—the neighborhood atlas. *N Engl J Med*. 2018;378:2456–2458.
29. Davis TC, Crouch MA, Long SW, et al. Rapid assessment of literacy levels of adult primary care patients. *Fam Med*. 1991;23:433–435.
30. Koren G, Boskovic R, Hard M, Maltepe C, Navioz Y, Einarson A. Motherisk-PUQE scoring system for nausea and vomiting of pregnancy. *Am J Obstet Gynecol*. 2002;186:S228–S231.
31. Cox JL, Chapman G, Murray D, Jones P. Validation of the Edinburgh Postnatal Depression Scale (EPDS) in non-postnatal women. *J Affect Disord*. 1996;39:185–189.
32. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. *J Health Soc Behav*. 1983;24:385–396.
33. Spielberger CD, Gorsuch RL, Lushene RE. *Manual for State-Trait Anxiety Inventory (Self-Evaluation Questionnaire)*. Consulting Psychologists Press; 1970.
34. Krebs-Smith SM, Pannucci TE, Subar AF, et al. Update of the healthy eating index: HEI-2015. *J Acad Nutr Diet*. 2018;118:1591–1602.
35. Coyle J. The TMLE framework. Available at: https://tlverse.org/tlverse-handbook/tmle3.html. In: van der Laan MJ, Coyle J, Hejazi NS, Malenica I, Phillips R, Hubbard A, eds. *Targeted Learning in R: Causal Data Science with the tlverse Software Ecosystem*; 2022. Available at: https://tlverse.org/tlverse-handbook/index.html. Accessed 21 March 2025.
36. Naimi AI, Whitcomb BW. Defining and identifying average treatment effects. *Am J Epidemiol*. 2023;192:685–687.
37. Zheng W, van der Laan MJ. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*. 2010;273.
38. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J*. 2018;21:C1–C68.
39. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846–866.
40. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer; 2011.
41. Kennedy EH. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint*. 2022. arXiv:2203.06469 [stat.ME].
42. R Core Team. *R: A Language and Environment for Statistical Computing*. Version 4.3.1. R Foundation for Statistical Computing; 2023. Available at: https://www.R-project.org/. Accessed 21 March 2025.
43. Polley EC, LeDell E, Kennedy C, Lendle S, van der Laan MJ. *SuperLearner: Super Learner Prediction*. R package version 2.0-29. Published 20 February 2024. Available at: https://cran.r-project.org/package=SuperLearner. Accessed 21 March 2025.
44. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat Med*. 1997;16:285–319.
45. Hines O, Dukes O, Diaz-Ordaz K, Vansteelandt S. Demystifying statistical learning based on efficient influence functions. *Am Stat*. 2022;76:292–304.
46. Balakrishnan S, Kennedy EH, Wasserman L. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint*. 2023. arXiv:2305.04116 [math.ST].