Pseudo-random Number Generator Influences on Average Treatment Effect Estimates Obtained with Machine Learning

Ashley I. Naimi,[a] Ya-Hui Yu,[a] and Lisa M. Bodnar[b]

Background: The use of machine learning to estimate exposure effects introduces a dependence between the results of an empirical study and the value of the seed used to fix the pseudo-random number generator.

Methods: We used data from 10,038 pregnant women and a 10% subsample (N = 1004) to examine the extent to which the risk difference for the relation between fruit and vegetable consumption and preeclampsia risk changes under different seed values. We fit an augmented inverse probability weighted estimator with two Super Learner algorithms: a simple algorithm including random forests and single-layer neural networks and a more complex algorithm with a mix of tree-based, regression-based, penalized, and simple algorithms. We evaluated the distributions of risk differences, standard errors, and P values that result from 5000 different seed value selections.

Results: Our findings suggest important variability in the risk difference estimates, as well as an important effect of the stacking algorithm used. The interquartile range width of the risk differences in the full sample with the simple algorithm was 13 per 1000. However, all other interquartile ranges were roughly an order of magnitude lower. The medians of the distributions of risk differences differed according to the sample size and the algorithm used.

Conclusions: Our findings add another dimension of concern regarding the potential for "p-hacking," and further warrant the need to move away from simplistic evidentiary thresholds in empirical research. When empirical results depend on pseudo-random number generator seed values, caution is warranted in interpreting these results.

Keywords: Causal inference; Machine learning; Random number generation; Random seed; Statistics

(Epidemiology 2024;35: 779–786)

Submitted October 23, 2023; accepted July 31, 2024

Editors' note: Related articles appear on pages 764 and 787.

From the [a]Department of Epidemiology, Emory University; and [b]Department of Epidemiology, University of Pittsburgh School of Public Health, Atlanta, GA.

A.I.N. and L.M.B. were supported by NIH grant number R01HD098130.

The authors report no conflicts of interest.

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: 1044-3983/24/356-779786

DOI: 10.1097/EDE.0000000000001785

Machine learning methods are increasingly being used to estimate treatment or exposure effects in epidemiologic data. These methods make fewer assumptions about how data are generated and are thus presumed to be less susceptible to biases due to model misspecification. Several techniques are being deployed in a range of areas, including single robust learners such as the so-called "S," "T," and "X" learners,[1] as well as double robust methods such as augmented inverse probability weighting (augmented IPW)[2] and targeted maximum likelihood estimation (TMLE).[3] To improve the chances of correctly specifying the needed regression models, combining these estimators with a stacked regression algorithm such as the Super Learner is recommended.[4,5] Commonly used algorithms include random forests, the least absolute selection and shrinkage operator (LASSO), neural networks, and a range of other tree-based, regression-based, penalized, or smoothing methods.

When machine learning algorithms are used to estimate causal effects, pseudo-random number generators are required at a number of stages: First, many algorithms often included in a stacked regression rely on pseudo-random number generation for deployment. For example, random forests and single-layer neural networks both rely on pseudo-random number generators to be fit to a given dataset (level-0 seed dependence). Second, stacking combines several algorithms into a single learner using a cross-validated criterion (e.g., mean squared error), with cross-validation folds selected based on a pseudo-random number generator[5] (level-1 seed dependence). Finally, use of some form of out-of-sample estimation (e.g., sample-splitting, cross-fitting, or cross-validation) with double robust methods is often advised,[6–8] which again requires splitting data into folds, with folds once again selected using pseudo-random number generation (level-2 seed dependence).

Consequently, estimating causal effects with machine learning algorithms creates a potentially important dependence between the seed value used to set the pseudo-random number generator, and the overall estimate obtained from the data. While it is known that different seed values will result in different treatment effect estimates and potentially different inferences about the exposure-outcome relation of interest, little research has been done on the impact this variability might have on inferences in empirical data.

In this work, we quantify the variability in average treatment effect estimates of the relation between fruit and vegetable intake around conception and the risk of preeclampsia in a cohort of women recruited to a major national pregnancy study.[9] Specifically, we evaluate this risk difference under different seed values when using a cross-validated augmented inverse probability weighted estimator, with the nuisance functions (here, propensity score and outcome models) fit using a complex stacking algorithm with several machine learning algorithms included. We assess the impact of different seeds on the magnitude and variability of results in the full data, as well as a 10% subset of the data. Finally, we explore the "heavy-tail" properties of the distribution of point estimates and standard errors obtained from the augmented IPW estimator under different stacking algorithms and scenarios.

METHODS

We analyzed data from the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b), a pregnancy cohort study conducted in eight US medical centers from 2010 to 2013 that recruited 10,038 women.[9] Each site's local institutional review board approved the study and all women gave written informed consent. To explore the role of sample size in seed-induced variability, we replicated all analyses in a 10% subsample of these data (N = 1004). Individuals with singleton pregnancies at 6–13 weeks gestation with no previous pregnancy lasting ≥20 weeks' gestation were eligible. Participants completed study visits at 6–13 weeks (enrollment, visit 1), 16–21 weeks (visit 2), and 22–29 weeks (visit 3), when research personnel ascertained data on demographics, medical history, behaviors, social factors, psychosocial assessments, and events and complications of pregnancy. Pregnancy and birth outcomes and delivery diagnoses were recorded by study personnel from medical records at ≥30 days after delivery.

At enrollment, usual dietary intake in the 3 months around conception was assessed using a self-administered modified Block 2005 food frequency questionnaire (FFQ, available in English and Spanish). Details of the dietary assessment have been published previously.[10] Analysis of the National Health and Nutrition Examination Survey 1999–2002 24-hour dietary recall data formed the basis of the FFQ's list of approximately 120 food and beverage items. NutritionQuest (Berkeley, CA) performed scanning, nutrient and food group mapping, and summary analysis of the FFQ data.[11] The food and beverage items were linked to the Food Patterns Equivalents Database,[12] to generate food group variables.

Our exposure was total fruit and vegetable intake. Grams of fruits and vegetables were summed and then defined as a density (cups per 1000 kcal). We dichotomized the density of fruit and vegetable intake at 2.5 cups/1000 kcal of fruit per day, which reflects the 80th percentile of the distribution and approximates the recommended intake as defined by the US Department of Agriculture Healthy US-Style Eating Pattern.[13] To account for aspects of dietary patterns independent of fruit and vegetable intake in modeling, we calculated the total Healthy Eating Index—2015 score excluding the fruit and vegetable components[14]

Our outcome was preeclampsia, based on the 2013 American College of Obstetricians and Gynecologists' diagnostic criteria,[15] which was adapted by the nuMoM2b investigators to the data the study collected.[16] We adjusted for a host of potential confounders, including a range of demographic, nutritional, anthropometric, and behavioral variables (eAppendix 1; http://links.lww.com/EDE/C176).

Statistical Analysis: Estimator

We estimated the average treatment effect of fruit and vegetable intake on preeclampsia risk using a cross-validated augmented inverse probability weighted estimator, defined as:

$$\hat{\psi}_{cv} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} I\left(G_i = k\right) \left\{ \frac{\left(2X_i - 1\right)\left[Y_i - \hat{g}^{-k}\left(X_i, C_i\right)\right]}{\left(2X_i - 1\right)\hat{f}^{-k}\left(C_i\right) + \left(1 - X_i\right)} + \hat{g}^{-k}\left(X_i = 1, C_i\right) - \hat{g}^{-k}\left(X_i = 0, C_i\right) \right\}$$

where i denotes women in the sample, and k denotes the cross-validation fold, X_i denotes the fruit and vegetable density variable, Y_i denotes the preeclampsia outcome, and C denotes the set of confounders adjusted for in our analysis. Additionally, $\hat{f}^{-k}\left(C_i\right)$ denotes predictions for woman i from a propensity score model fit to the cross-validation folds that exclude woman i regressing the exposure X against all confounders C. Similarly, $\hat{g}^{-k}\left(X_i, C_i\right)$ denotes predictions for woman i from an outcome model fit to the cross-validation folds that exclude woman i, regressing preeclampsia Y against the exposure X and all confounders C. Under relevant identification assumptions,[17] $\hat{\psi}_{cv}$ can be interpreted as the average treatment effect (on the difference scale) of fruit and vegetable consumption on preeclampsia risk.

We used a stacked generalization (i.e., super learner[18]) to estimate the propensity score and outcome models $\hat{f}_i\left(C\right)$ and $\hat{g}_i\left(X = x,\ C\right)$. We explored the impact of seed variation under two versions of a super learner. The first super learner included a standard mean estimator, and a standard generalized linear model (GLM) estimator, the LASSO algorithm (with the penalty parameter estimated via cross-validation[19]), random forests with 500 trees, and a random subspace selection number of six,[20] extreme gradient boosting with 1000 trees, a shrinkage parameter of 0.1, and a max tree depth of four, and a single-layer neural network, with a layer size of 5.[21] In this first super learner, some algorithms were included that do not rely themselves on pseudo-random number generation for deployment (standard mean, standard GLM, the LASSO, and xgboost). Thus, while this first scenario was characterized by seed dependence at levels 0, 1, and 2, the level 0 dependence was potentially muted or removed by including algorithms that do not depend on the seed.

The second super learner included only two algorithms that internally rely on pseudo-random number generation for deployment: random forests with 500 trees and a random subspace selection number of six and a single-layer neural network, with a layer size of 5. Thus, this second super learner relied at all levels fully on the specified seed value.

For these algorithms, we used the Super Learner fit using 10-fold cross-validation and a non-negative least squares loss function.[5] We also implemented 10-fold cross-validation for the augmented IPW estimator by fixing the folds for the Super Learner algorithm used for the propensity score and the outcome model, and constructing augmented IPW estimates within each fold. Specifically, the same individuals are used to estimate the propensity score and outcome model, and the same out-of-fold set is used to generate propensity score and outcome model predictions to compute the augmented IPW estimate $\hat{\psi}_{cv}$. This reduces the computation time needed to obtain an estimate, reduces seed dependence levels 1 and 2 to a single level, and yields valid out-of-sample (cross-validated) estimates of the effect of interest. Finally, for all analyses, we bound the propensity score to lie within values of 0.001 and 0.999.

In total, we computed 5000 different estimates $\hat{\psi}_{cv,\,j}$ for each sample size (N = [10,038, 1,004]), where each estimate was obtained under a different seed value. Integer seed values were obtained by sampling (without replacement) 5000 seeds from a set of sequential integers ranging from 1 to 10×10^6 (sample min = 585, sample max = 9,999,469, sample median = 5,078,502). Throughout, we index these seed values as $j = 1 \text{ to } J = 5,000$.

Implementing the cross-validated augmented IPW estimator with the super learner under these seeds yielded a distribution of risk differences, standard errors, and *P* values for the effect of fruit and vegetable density on preeclampsia risk. First, we evaluated summary statistics and plotted the distributions to capture how much the risk differences, standard errors, and *P* values between fruit and vegetable intake and preeclampsia changes under different seeds. We also computed summary statistics (median and interquartile range) of these distributions to numerically capture the variability across seeds. Finally, we assessed if the distributions of risk differences and standard errors showed evidence of "heavy-tailed" behavior using maximum-to-sum plots.[22] If the distributions of the risk differences or standard errors are heavy-tailed, this can lead to volatility in the distribution of estimates that might threaten the credibility of treatment effects estimated using machine learning methods.[23]

RESULTS

Table 1 presents summary statistics for key variables in both samples used to estimate the association of interest. Overall, summary measures of the variables in the original and reduced subsample were similar. Our analyses of the properties of the risk differences, standard errors, and *P* values suggested that the results were highly sensitive to the version of the Super Learner algorithm, as well as to the choice of seed values used. These sensitivities were borne out in a number of ways.

First, by definition, the risk difference must lie between values of −1 and 1. However, in our simulations, there were seed values that resulted in risk differences beyond this range (a known limitation of the augmented IPW estimator[24]). Table 2 shows how many seed values yielded risk differences beyond (−1,1) stratified by the sample size and super learner explored. The worst-case scenario arose with the limited super learner version that included only random forests and single-layer neural networks (both of which depend on seed values for implementation beyond their inclusion in the super learner), with $\frac{293}{5,000} \approx 6\%$ risk differences falling beyond the (−1,1) range. We removed all seed value scenarios in which the risk difference was outside of the (−1,1) range to present all subsequent results (total N removed = 309).

Figure 1A–F demonstrates the distributions of risk differences, standard errors, and *P* values for the remaining seed values used to specify the random number generator for the full super learner approach. Notably, the sample size had an

TABLE 1. Distribution of Key Variables by Preeclampsia Status in the Full (N = 10,038) and Reduced (1004) Sample of Women From Pregnant Women Recruited From 8 US Medical Centers From 2010 to 2013 for the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be (nuMoM2b)

	Full Sample			Reduced Sample (10%)		
	Non Preeclamptic	Preeclamptic	Overall	Non Preeclamptic	Preeclamptic	Overall
	(N = 9204)	(N = 834)	(N = 10038)	(N = 928)	(N = 76)	(N = 1004)
Fruit and vegetable consumption, n (%)						
≥1.5 cups/ day/1000 kcals	1291 (14%)	88 (11%)	1379 (14%)	139 (15%)	12 (16%)	151 (15%)
Maternal age						
Median (Min, Max)	26.0 (12.0, 44.0)	26.0 (14.0, 41.0)	26.0 (12.0, 44.0)	26.0 (13.0, 43.0)	25.5 (16.0, 39.0)	26.0 (13.0, 43.0)
Prepregnancy BMI (kg/m^2)						
Median (Min, Max)	22.9 (11.5, 72.5)	25.7 (13.4, 65.2)	23.0 (11.5, 72.5)	22.7 (13.7, 65.4)	27.4 (16.2, 61.9)	22.8 (13.7, 65.4)
Prepregnancy smoking status, n (%)						
Ever smoker	1473 (16%)	157 (19%)	1630 (16%)	154 (17%)	13 (17%)	167 (17%)

TABLE 2. Number of Seed Value Runs that Returned Risk Differences Greater Than 1 or Less Than −1, Stratified by Sample Size and Super Learner Version

Sample Size	Super Learner	Count
10,038	Version 1	0
1,004	Version 1	5
10,038	Version 2	293
1,004	Version 2	11

Version 1: mean, GLM, LASSO, random forests, xgboost, single-layer nnet.
Version 2: random forests, single-layer nnet.

important impact on the distribution of results. Reducing the sample size led to an increase in variability and a shift in the overall magnitude of the point estimate. However, the shape of the distribution of risk differences from both sample size scenarios was roughly symmetric and bell-shaped. On the other hand, Figure 2A–F shows the corresponding distributions when the reduced super algorithm is used. These Figures show a much wider degree of volatility in the distribution of risk differences, particularly in the full sample (Figure 2A), where a cluster of results was obtained at the lower bound of the risk difference scale.

Table 3 shows medians and interquartile range values of the distributions in Figures 1 and 2, demonstrating important summary-level differences across scenarios. For example, the risk difference in the original sample with the full super learner version suggests that fruit and vegetable consumption at ≥2.5 cups per day led to a reduction of two preeclampsia cases per 100 (interquartile range width [IQR_w] = 0.001).

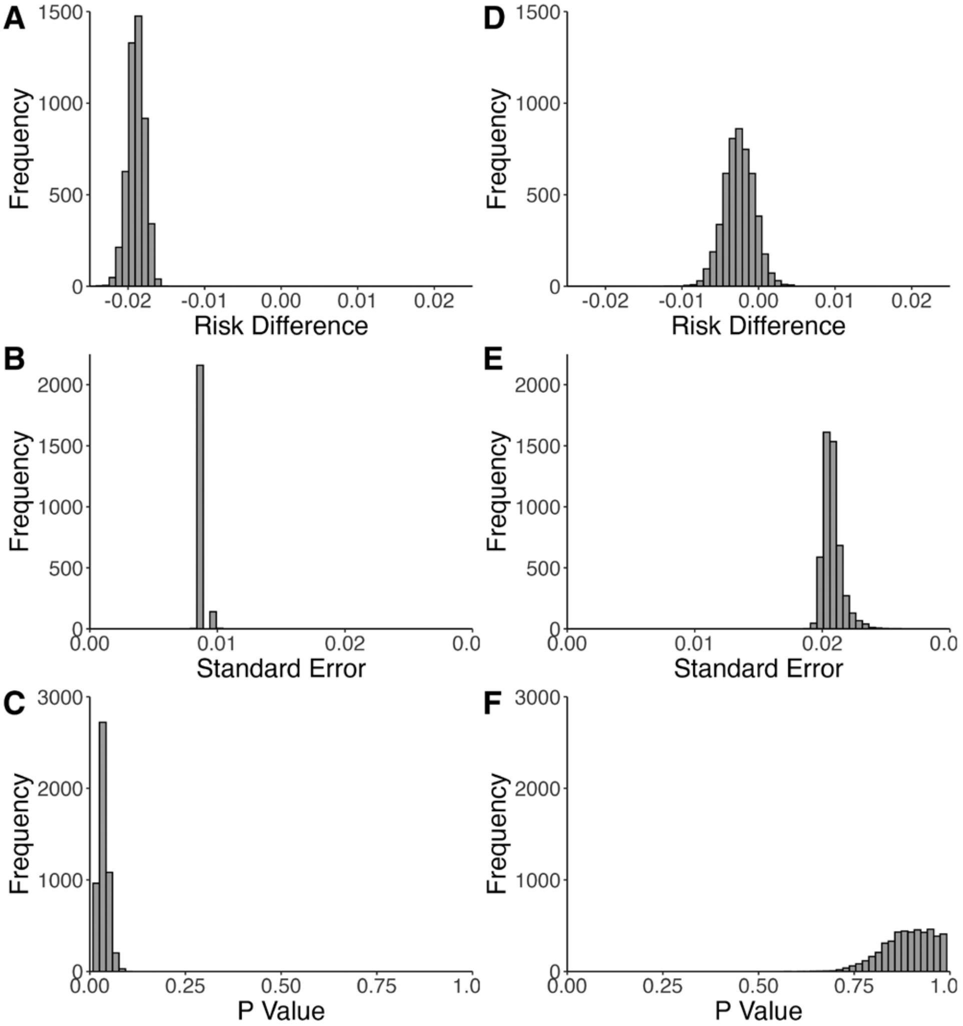

Figure 1. Distributions of risk differences, standard errors, and *P* values obtained from an augmented inverse probability weighted estimator with a super learner algorithm that included the standard mean, standard GLM, LASSO, random forests, extreme gradient boosting, and single layer neural networks for the relation between fruit and vegetable intake and preeclampsia risk implemented under 5000 different seed values. A–C, Full sample, N = 10,038; D–F, reduced sample, N = 1004. Results presented are based on subset of 5000 seed values after removing realizations with risk differences greater than 1 or less than −1.

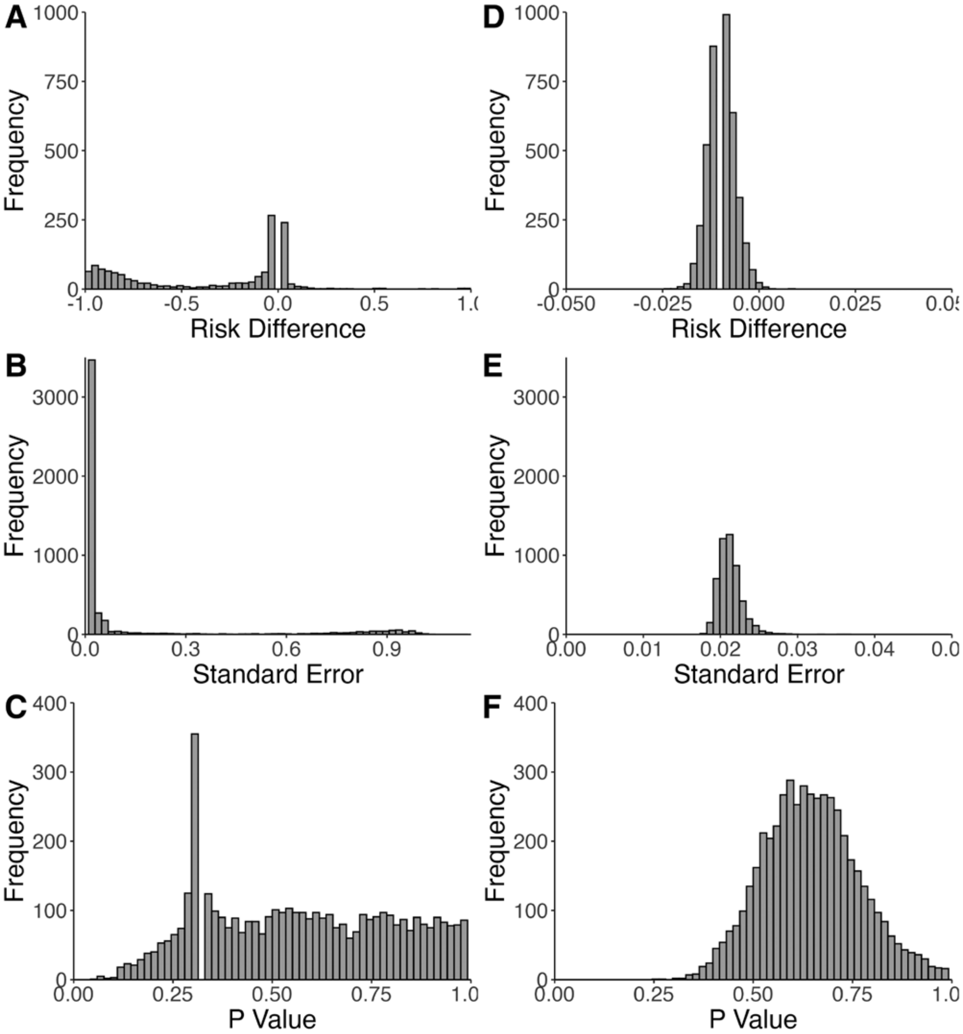

Figure 2. Distributions of risk differences, standard errors, and *P* values obtained from an augmented inverse probability weighted estimator with a super learner algorithm that included only random forests and single layer neural networks for the relation between fruit and vegetable intake and preeclampsia risk implemented under 5000 different seed values. A–C, Full sample, N = 10,038; D–F, reduced sample, N = 1004. Results presented are based on subset of 5000 seed values after removing realizations with risk differences greater than 1 or less than –1.

TABLE 3. Median (Interquartile Range Width) of the Distributions of Risk Differences, Standard Errors, and *P* Values Obtained From an Augmented Inverse Probability Weighted Estimator for the Relation Between Fruit and Vegetable Intake and Preeclampsia Risk Implemented Under 5000 Different Seed Values by Sample Size and Super Learner

Sample Size	Super Learner	Risk Difference	Standard Error	*P* Value
10,038	Version 1	−0.019 (0.001)	0.009 (0.0003)	0.034 (0.0158)
1004	Version 1	−0.003 (0.002)	0.021 (0.0008)	0.900 (0.0977)
10,038	Version 2	−0.009 (0.013)	0.019 (0.0194)	0.529 (0.441)
1004	Version 2	−0.010 (0.004)	0.021 (0.0018)	0.640 (0.162)

Version Results presented are based on subset of 5,000 seed values after removing realizations with risk differences greater than 1 or less than −1.
Version 1: mean, GLM, LASSO, random forests, xgboost, single-layer nnet.
Version 2: random forests, single-layer nnet.

However, in the original sample with the reduced super learner version, this risk difference changed to a reduction of nine preeclampsia cases per 1000 (IQR_w = 0.013). In contrast to the risk differences, the patterns across standard errors differed. The median standard error in the original sample with the full super learner was 0.009 (IQR = 0.0003).

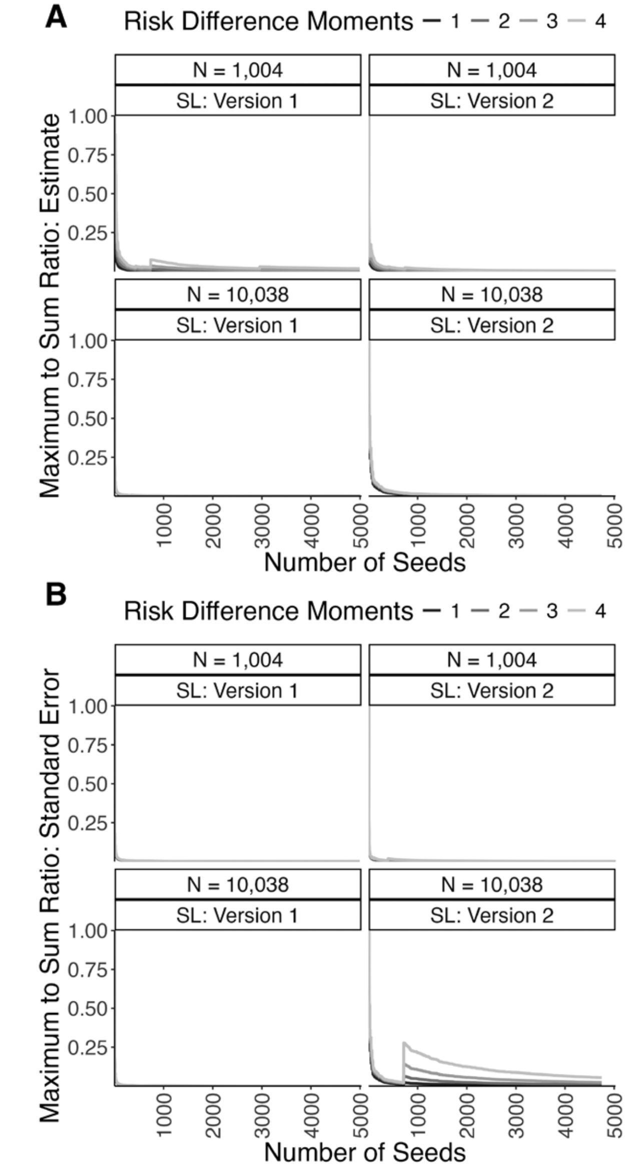

Figure 3. Maximum-to-Sum Plots for the risk differences (A) and standard errors (B) after removing all seed values that yielded risk differences greater than 1 or less than –1, stratified by sample size and super learner version. SL: Version 1 included the mean, GLM, LASSO, random forests, xgboost, single layer nnet; SL: Version 1 included random forests and single layer neural networks.

However, across all other scenarios, the median standard error increased to roughly 0.02 (with IQR_w ranging from 0.0008 to 0.0194).

At an individual seed level, there were important differences in our findings. For example, in the original sample and the full super learner, a seed value of 204,189 yielded a risk difference of two fewer preeclampsia cases per 100 women in the sample, with 95% confidence intervals of −3.7 and −0.3 (*P* value = 0.022). Using a seed value of 129,202 yielded a risk difference of 2 fewer preeclampsia cases per 100 women in the sample, with 95% confidence intervals of −3.6 and 0 (*P* value = 0.052). Finally, for the same contrast, using a seed value of 203,256 yielded a risk difference of 0.2 more preeclampsia cases per 100 women in the sample, with 95% confidence intervals of −4.3 and 4.7 (*P* value = 0.922).

As a final analysis, we used maximum-to-sum plots to explore whether there was evidence suggesting that the distributions of risk differences or standard errors were "heavy-tailed." For example, if the true distribution of risk differences or standard errors across all seed values was a member of (e.g.) certain power-law families, maximum-to-sum plots would demonstrate that, as the sample size increases, key moments of these distributions would fail to converge to zero. This would, in turn, suggest that the volatility across seed values is so high that standard summary measures across seeds might fail to convey useful information. Figure 3 shows that, in the data in which risk difference values outside the (−1,1) range are removed, maximum-to-sum plots for the first four moments of the risk difference and the standard errors from both Super Learner versions demonstrate convergence to zero. However, in eAppendix 2, eFigure 2; http://links.lww.com/EDE/C177, we reproduce these maximum-to-sum plots in the data that include risk differences outside the (−1,1) range, which shows general nonconvergence for both versions of the super learner in the reduced sample.

DISCUSSION

As researchers continue using machine learning algorithms, it is becoming more important to consider how inferences from this work are affected by random number generators that pervade machine learning methods. This issue has not gone completely without consideration. Chernozhukov et al[25] incorporate a degree of seed variability in their implementation of the double-debiased machine learning algorithm.[26] Furthermore, Zivich and Breskin[7] show how repeating the cross-fitting process can help stabilize point estimates across seed repetitions. In both, the median is used as a summary measure. However, despite this work, to our knowledge, no research has systematically explored the sensitivity of causal effect estimates to pseudo-random number generator seeds.

Related research on the variability in machine learning algorithm results as a result of changes in random seed values has been conducted outside of causal inference settings. For example, Dodge et al[27] demonstrated how varying seed values affect the fine-tuning of contextual embedding models commonly used in natural language processing settings. In this work, they treat the seed values as a hypertuning parameter and thus optimize the performance of their language models as a function of the seed value. Bethard[28] extends the various ways of handling seed values in the context of general machine learning by classifying practices into "safe" and "unsafe" uses of seed values. Specifically, he argues that safe seed uses include their inclusion as model hyperparameters, creating an ensemble algorithm by including identical algorithms under different seed values or treating the seed as a sensitivity parameter. Unsafe uses include specifying a single seed value to replicate study results. This latter practice is, to our knowledge, commonly used in causal inference analyses.

Using data from a large pregnancy cohort in the United States, we demonstrated the sensitivity of the estimate of the average treatment effect to setting the seed value for the random number generator used throughout the deployment of the algorithms in use. Overall, we found that seed values played an important role in generating variability in the risk differences, standard errors, and P values estimated in our data. Furthermore, this variability that resulted from seeds depended heavily on which machine learning algorithm was used to model the nuisance functions needed to fit the augmented inverse probability weighted estimator, as well as on the sample size available to fit these models. Notably, when using a version of the Super Learner that included only random forests and single-layer neural networks, the variability of our results increased dramatically. Additionally, in this case, we noted more variability in the larger sample of 10,038 observations than in the reduced subsample of 1004 observations (Figure 2). These results add to the growing evidence and recommendations on why it is important to use a wide variety of machine learning algorithms to model nuisance functions when estimating exposure effects, instead of relying on a narrow set of select algorithms.

We also found it relatively easy to identify scenarios in the same dataset with the same Super Learner algorithm where seed values could be used to generate completely different conclusions about the exposure effect. Simply changing the seed from 204,189 to 129,202 yielded results with similar point estimates and P values but with the former P value below 0.05 and the latter above 0.05. Alternatively, changing the seed to 203,256 yielded a point estimate that was nearly null and a P value near one. More generally, our finding of wide variation in the proportion of P values ≤ 0.05 from the reliance on random number generators for causal effect estimation introduces a new potential form of P-hacking (seed hacking), further warranting the need to move away from simplistic evidentiary thresholds.[29]

We also conducted preliminary explorations about whether there was evidence to suggest that any of the distributions generated under the range of seeds we explored were “heavy-tailed.” Such distributions can pose problems when summarizing the distribution of estimates using standard summary measures (e.g., mean). In extreme scenarios, fundamental theorems central to most statistical analyses, such as the standard central limit theorem or the weak law of large numbers, can fail to hold in the presence of heavy-tailed distributions.[23] This can have important implications on the validity of the interpretation of a summary of point estimates. We are unaware of any theoretical or applied analyses that evaluate whether the distribution of point estimates obtained from doubly robust estimators when different machine learning algorithms are used to model the nuisance functions might fall in a class of heavy-tailed distributions (e.g., super cubic, Lévy-stable, or certain Power-Law distributions).

In our analyses, we only observed evidence of heavy tails if anomalous risk differences of less than −1 or greater than 1 were retained. After removing these clearly problematic results, we found no evidence of heavy tails. However, interpreting this finding and reasoning about its implications is challenging because anomalies in our analyses could be easily identified and removed. Less clear is whether point estimates on scales not bounded by hard limits (e.g., the mean difference for a continuous outcome) will be affected differently.

Our analysis is subject to important limitations that should be noted. First, there are a wide array of random number generators available in most statistical software packages. We used the Mersenne-Twister,[30] which is the default in the R programming language[31] and thus arguably the most commonly used approach. However, several others exist and these may lead to different properties than we observed here. Second, we explored augmented IPW as our treatment effect estimator. We did not explore targeted minimum loss-based estimation,[32] double machine learning,[25] or other estimators that can be deployed with machine learning methods.

Relatedly, we limited our exploration of machine learning algorithms to two versions of a stacking algorithm. The full version included the simple mean, GLM, LASSO, random forests, extreme gradient boosting, and single-layer neural networks under standard tuning parameters. The reduced version included only random forests and single-layer neural networks. This latter version was included to explore the impact of seed variability at all levels of deployment, including the fitting of the level-0 algorithm to our data, the cross-validation of the super learner loss function for both nuisance functions (the outcome model and the propensity score), and the cross-fitting of the augmented IPW estimator. In principle, our complex super learner could have avoided seed dependence in the fitting of the level-0 algorithms entirely if they were weighted towards algorithms that do not depend on seeds (e.g., GLM).

Despite these limitations, our analysis is characterized by strengths as well, including the fact that we evaluated the impact of random number generation on a complex combination of algorithms stacked into a meta-learner and deployed via a cross-validated augmented IPW estimator. Additionally, we explored these effects in real data that have been used extensively with machine learning methods.[10,33,34]

The need to rely on random number generators to estimate causal effects with machine learning algorithms introduces variability in the observed results. This variability can result in new forms of P hacking, data dredging, or other untoward practices that researchers should be aware of as they review and interpret scientific findings. More generally, when causal effect estimates from machine learning analyses are interpreted, researchers should be wary of their dependence on selected seed values. Our findings suggest that the variability due to seed values can lead to striking differences in the estimated effects in a given dataset.

ACKNOWLEDGMENTS

We thank an anonymous peer reviewer for helpful comments on a previous version of this manuscript.

REFERENCES

1. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci USA*. 2019;116:4156–4165.
2. Jonsson-Funk M, Westreich D, Wiesen C, et al. Doubly robust estimation of causal effects. *Am J Epidemiol*. 2011;173:761–767.
3. Gruber S, van der Laan MJ. *Targeted Maximum Likelihood Estimation: A Gentle Introduction*. UC Berkeley; 2009.
4. Wolpert D. Stacked generalization. *Neural Netw*. 1992;5:241–259.
5. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33:459–464.
6. Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *Am J Epidemiol*. 2023;192:1536–1544.
7. Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology*. 2021;32:393–401.
8. Zivich PN, Breskin A, Kennedy EH. *Machine Learning and Causal Inference*. *Wiley StatsRef: Statistics Reference Online*:1–8.
9. 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.
10. 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.
11. Epidemiology and Genomics Research Program, National Cancer Institute. *Diet*Calc Analysis Program, Version 1.5.0*. Bethesda, MD: National Cancer Institute; 2012.
12. U.S. Department of Agriculture. *Food Patterns Equivalents Database 2011-12*. Beltsville, MD: Agricultural Research Service, Food Surveys Research Group; 2014.
13. U.S. Department of Agriculture, U.S. Department of Health and Human Service. Dietary Guidelines for Americans, 2020-2025. 2020. Available at: DietaryGuidelines.gov. Accessed 25 August 2024.
14. 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.
15. Hypertension in pregnancy. Report of the American College of Obstetricians and Gynecologists' task force on hypertension in pregnancy. *Obstet Gynecol*. 2013;122:1122–1131.
16. Facco FL, Parker CB, Reddy UM, et al. Association between sleep-disordered breathing and hypertensive disorders of pregnancy and gestational diabetes mellitus. *Obstet Gynecol*. 2017;129:31–41.
17. Naimi AI, Whitcomb BW. Defining and identifying average treatment effects. *Am J Epidemiol*. 2023;192:685–687.
18. Polley E, LeDell E, Kennedy C, et al. SuperLearner: Super Learner Prediction. R package version 20-29. 2024. Available at: https://CRAN.R-project.org/package/SuperLearner. Accessed 25 August 2024.
19. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
20. Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*. 2017;77:1–17.
21. Chen T, He T, Benesty M, et al. xgboost: Extreme Gradient Boosting. R package version 1771 2024; Available at: https://CRAN.R-project.org/package/xgboost. Accessed 25 August 2024.
22. Cirillo P, Taleb NN. Tail risk of contagious diseases. *Nat Phys*. 2020;16:606–613.
23. Taleb NN. *Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications (Technical Incerto)*. New York, NY: STEM Academic Press; 2020.
24. Glynn A, Quinn K. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*. 2010;18:36–56.
25. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J*. 2018;21:C1–C68.
26. Bach P, Chernozhukov V, Kurz M, et al. DoubleML - an object-oriented implementation of double machine learning in R. *J Stat Softw*. 2024;108:1–56.
27. Dodge J, Ilharco G, Schwartz R, et al. Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. *arXiv*. 2020:2002.06305.
28. Bethard S. We need to talk about random seeds. *arXiv*. 2022: 2210.13393.
29. Greenland S. Analysis goals, error-cost sensitivity, and analysis hacking: essential considerations in hypothesis testing and multiple comparisons. *Paediatr Perinat Epidemiol*. 2021;35:8–23.
30. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans Model Comput Simul*. 1998;8:3–30.
31. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014.
32. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2:Article 11.
33. Bodnar LM, Cartus AR, Kennedy EH, et al. Use of a doubly robust machine-learning-based approach to evaluate body mass index as a modifier of the association between fruit and vegetable intake and preeclampsia. *Am J Epidemiol*. 2022;191:1396–1406.
34. 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.