Research Article

Kara E. Rudolph*, Nicholas T. Williams, Caleb H. Miles, Joseph Antonelli, and Ivan Diaz

All models are wrong, but which are useful? Comparing parametric and nonparametric estimation of causal effects in finite samples

https://doi.org/10.1515/jci-2023-0022
received April 20, 2023; accepted September 09, 2023

Abstract: There is a long-standing debate in the statistical, epidemiological, and econometric fields as to whether nonparametric estimation that uses machine learning in model fitting confers any meaningful advantage over simpler, parametric approaches in finite sample estimation of causal effects. We address the question: when estimating the effect of a treatment on an outcome, how much does the choice of nonparametric vs parametric estimation matter? Instead of answering this question with simulations that reflect a few chosen data scenarios, we propose a novel approach to compare estimators across a large number of data-generating mechanisms drawn from nonparametric models with semi-informative priors. We apply this proposed approach and compare the performance of two nonparametric estimators (Bayesian adaptive regression tree and a targeted minimum loss-based estimator) to two parametric estimators (a logistic regression-based plug-in estimator and a propensity score estimator) in terms of estimating the average treatment effect across thousands of data-generating mechanisms. We summarize performance in terms of bias, confidence interval coverage, and mean squared error. We find that the two nonparametric estimators can substantially reduce bias as compared to the two parametric estimators in large-sample settings characterized by interactions and nonlinearities while compromising very little in terms of performance even in simple, small-sample settings.

Keywords: parametric, nonparametric, causal inference

MSC 2020: 62-XX, 62D20, 62G05, 00A72

1 Introduction

In the past two decades, work emerging from the nonparametric statistics literature has allowed for the incorporation of machine learning algorithms in estimation while, in many cases, preserving theoretically valid statistical inference [e.g., 1–4]. A growing faction, though still a subset, of statisticians, epidemiologists, econometricians, and other applied scientists now take as a given that these nonparametric estimators, if appropriately applied, are typically superior to parametric estimators, particularly in terms of reducing or eliminating model misspecification bias, which can be substantial when using an overly simplistic parametric model in complex settings [e.g., 4–14]. (We note that we use the term "nonparametric estimator" to mean an estimator where each component of the causal model used in estimation [also called a "nuisance parameter,"

* **Corresponding author: Kara E. Rudolph,** Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA, e-mail: kr2854@cumc.columbia.edu, tel: +12123422926
Nicholas T. Williams: Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA
Caleb H. Miles: Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, USA
Joseph Antonelli: Department of Statistics, University of Florida, Gainesville, USA
Ivan Diaz: Division of Biostatistics, Department of Population Health, New York University Grossman School of Medicine, New York, USA

e.g., an outcome regression] is estimated nonparametrically.) Minimizing all sources of bias is essential, especially because when bias remains on the same order as standard error, the probability of a corresponding hypothesis test rejecting the null will tend to one even when no effect is present.

Even so, an attitude of caution and skepticism remains prevalent, with many favoring simpler, parametric model-based approaches over approaches that are deemed by some to be unnecessarily complex [e.g., 15]. For example, Little (2013) observes that "Much modern nonparametric and semiparametric statistical theory lives in the 'land of asymptotia' [15]." This is accurate, and few have contributed evidence of the extent to which the assumption of nonparametric estimator superiority is borne out in *real-world*, *finite-sample* analyses, leaving the debate as to whether nonparametric methods confer any meaningful practical advantage in typical real-world applications largely unresolved [16].

Both nonparametric and parametric estimation methods generally involve using theory and subject-matter knowledge to inform the underlying causal model/graph [17] and the variables input into the model [18]. However, whereas parametric estimation methods would proceed by specifying parametric models (e.g., linear regression) for each nuisance parameter, specifying its functional form, interactions, higher-order variable forms, etc., nonparametric methods would typically involve using data-adaptive methods, such as machine learning algorithms, in flexibly fitting nuisance parameters.

Statistical theory informs when we would expect parametric versus nonparametric estimators to perform better in terms of bias. The bias of parametric estimators is a function of how close the parametric regression model is to the truth. For a simple parametric plug-in estimator (discussed further in Section 3 and Section S1 in the Supplement), bias is a function of how close the parametric outcome model is to the truth. For parametric estimators based on the propensity score, bias is a function of how close the treatment model is to the truth. For example, if the response and/or treatment assignment surfaces are nonlinear, then a parametric linear regression-based estimator that does not incorporate such nonlinearities would be expected to have appreciable bias [19]. The bias of nonparametric estimators also depends on how close the nonparametric regression models are to the truth, but the reliance on correct specification of the regression models is relaxed in at least one or two respects, depending on the estimator. First, the bias of nonparametric estimators that are also semi-parametric efficient estimators (e.g., the augmented inverse probability of treatment weighted [A-IPTW] estimator, the one-step estimator, double/de-biased machine learning estimator, targeted minimum loss-based estimator [TMLE]) is an integral of a product of two regression errors – e.g., one for the outcome regression and one for the treatment regression, a property called "double robustness" [20–22]. If the two errors are small, this product will generally be smaller than the single error involved in the definition of the parametric estimator bias. We discuss this further in Section S1 in the Supplement and TMLE in more detail in Section 3. Second, nonparametric estimators use machine learning algorithms in fitting nuisance parameters, which reduces reliance on correct parametric model specification, and therefore, can improve model fit and reduce bias even further. Consequently, because correctly specifying parametric models *a-priori* seems unlikely in the absence of clear mechanistic understanding, based on this asymptotic theory, nonparametric estimators are expected to have less bias, on average, than parametric estimators in very large samples when an accurate model is important to obtain an unbiased estimate, e.g., in the presence of significant, complex confounding or censoring [19]. However, asymptotic approximations may be poor in small or even moderately sized real-world samples that are afflicted by the curse of dimensionality, resulting in bias and under-coverage of confidence intervals [23]. The sample size at which asymptotic approximations reflect reality is generally problem dependent and unknown.

Previous simulation studies comparing performance between parametric and nonparametric estimation have generally shown that nonparametric estimators incorporating data-adaptive, machine learning algorithms outperform parametric approaches, especially under the data structures enumerated earlier [19,24–26]. However, common critiques of simulation studies are that they (a) are typically designed to illustrate a difference in performance (or lack thereof), and thereby reflect hand-picked settings that are favorable to the method of interest; (b) are oversimplified "stylized models of reality"; and (c) may hold little value if the goal is to make a general statement about the degree to which the choice of nonparametric versus parametric estimation matters [27–31]. A similar criticism regarding the lack of generalizability can apply to the few

previous real data analysis examples comparing nonparametric versus parametric estimators, sometimes finding meaningful differences and other times, not [14,32].

Responding to the critique that simulation-based evaluations of estimator performance have been conducted on oversimplified data too far removed from the complexity of real-world datasets, several groups recently proposed more comprehensive, yet tailored, evaluations of estimator performance by simulating datasets that closely mimic an observed dataset in all its complexity (similar to the idea of "plasmode simulation" [33]) [e.g., 27,28,34,35]. This work is premised on the belief that the optimal estimator will differ for different data scenarios; hence, the focus is on generating simulated data to match a particular observed dataset. However, by using the same data to both select an estimator and evaluate this estimator, these approaches are vulnerable to problems with post-selection inference. Given sufficiently large sample sizes, this problem can be circumvented via sample splitting [36]; however, practitioners can seldom afford to discard large amounts of data to perform estimator selection.

The current debate about how much the choice of a nonparametric estimator versus a parametric estimator may matter in the real world asks a more general question than the current literature can answer. We propose an approach to help answer such a question: what we call a large-scale Monte-Carlo simulation that instead of considering just a few data-generating mechanisms, summarizes estimator performance across thousands of data-generating mechanisms within a user-defined data setting. Unlike the plasmode simulation approaches referenced earlier, ours does not depend on the observed data. We apply this approach, comparing several parametric and nonparametric estimators to demonstrate how one can generate broader evidence of the degree to which choosing among these estimators can impact bias, confidence interval coverage, and mean squared error (MSE) in finite sample causal effect estimates.

Specifically, in our application of the proposed approach, we consider data-generating distributions drawn from the following setting: a binary treatment, a binary outcome, five binary covariates, and one multi-valued covariate. We compare performance in estimating the average treatment effect (ATE, formally defined in Section 2). We consider: (1) a parametric "plug-in" estimator (also called a "g-computation estimator"), (2) a propensity score estimator that uses a parametric treatment model, (3) a nonparametric "plug-in" estimator, and (4) a nonparametric efficient estimator.

This article is organized as follows. In Section 2, we introduce notation and the causal estimand we consider. In Section 3, we describe our proposed large-scale Monte-Carlo simulation for comparing parametric and nonparametric estimators. Also in this section, we provide the specific parameters that define the space of data-generating mechanisms we consider in applying the proposed approach, the parametric and nonparametric estimators we compare, and how we quantify and summarize estimator performance in finite samples. In Section 4, we provide and discuss results. Section 5 concludes.

2 Notation and estimands

For simplicity, in this section, we focus on one component of the ATE contrast – the expected counterfactual outcome had treatment been set to some value, t, possibly contrary to fact, denoted $\mathsf{E}(Y^t)$, where T denotes a binary treatment variable, and Y denotes the outcome variable. The ATE would then be denoted $\mathsf{E}(Y^1) - \mathsf{E}(Y^0)$. We assume that the exchangeability assumption holds, $Y^t \perp\!\!\!\perp T|X$ for $t \in \{0, 1\}$, where X denotes the confounding variables. This assumption is necessary for the causal parameter, $\mathsf{E}(Y^t)$, to be identified from the observed data, $Z = (X, T, Y)$, by the statistical parameter $\theta(\mathsf{P}) = \mathsf{E}[\mathsf{E}(Y|T = t, X)]$, where θ denotes the statistical parameter, which is a function of the data distribution, P. We let P be an element of the nonparametric statistical model defined as the set of all possible observed data distributions, noting that nonparametric statistical models can be correct. We also assume positivity, $\mathsf{P}(T = t|x) > 0$ for all x in the support of X, and for each treatment value in the causal parameter of interest. This is necessary for the statistical parameter to be well-defined.

We note that $\theta(\mathsf{P})$ constitutes the building block for the estimation of common marginal causal effects: the ATE, identified by $\mathsf{E}\{\mathsf{m}(1, X) - \mathsf{m}(0, X)\}$; the relative risk (RR), identified by $\frac{\mathsf{E}\{\mathsf{m}(1,X)\}}{\mathsf{E}\{\mathsf{m}(0,X)\}}$; and the odds ratio (OR),

identified by $\frac{\mathsf{E}\{\mathsf{m}(1,X)\}\,/\,\mathsf{E}\{1-\mathsf{m}(1,X)\}}{\mathsf{E}\{\mathsf{m}(0,X)\}\,/\,\mathsf{E}\{1-\mathsf{m}(0,X)\}}$, where $\mathsf{m}(t,x)$ represents the true conditional density function (i.e., regression) $\mathsf{E}(Y|T=t,X=x)$, and where we use subscripts with $\mathsf{m}(t,x)$ to denote the estimates of that conditional density based on a particular statistical model. For example, we use $\mathsf{m}_{\hat{\beta}}(t,x)$ to denote an estimate of that distribution that is a function of fitted parametric regression coefficients, $\hat{\beta}$. Similarly, we use subscripts with θ to denote the estimates of the statistical parameter with the subscript denoting the particular estimator used. For example, we use θ_{TMLE} to denote a TMLE estimate of the parameter $\theta(\mathsf{P})$.

When the exposure takes values in a discrete set, like the binary exposure we consider here, the common marginal causal effects of the ATE, RR, and OR can be nonparametrically estimated, yet with the same asymptotic properties as if they were parametrically estimated, as long as the dimension of X is fixed [4]. Taking the ATE as an example, this means that it will be possible to find an estimator $\widehat{\mathsf{ATE}}$ such that as the sample size n grows, $\sqrt{n}(\widehat{\mathsf{ATE}}-\mathsf{ATE})$ converges to a random variable that is normally distributed with mean zero and variance equal to the nonparametric efficiency bound. The efficiency bound is the smallest possible variance attainable by a regular estimator [1]. This means that the normal distribution can be used to approximate the sampling distribution of the estimator $\widehat{\mathsf{ATE}}$ for finite sample sizes, which allows us to construct approximately correct confidence intervals and hypothesis tests.

In the next sections, we develop a simulation-based approach to evaluate if and when nonparametric estimators of $\theta(\mathsf{P})$ can be expected to outperform parametric estimators across a range of finite samples under minimal but reasonable assumptions on the nature of the true data-generating mechanism.

3 A large-scale Monte-Carlo simulation approach to systematically evaluate the finite sample performance of estimators

We propose the following general method to systematically evaluate and compare the performance of a set of candidate estimators in finite samples. Our method, which draws on the nonparametric Bayesian literature, compares the performance across a large number of data distributions within a setting defined based on the following: the number and type of variables, range of confounding bias, treatment effect heterogeneity, and degree of nonlinearity.

- For each $j \in \{1, \ldots, J\}$, we draw a probability distribution P_j using a minimally informative prior across the space of distributions (i.e., nonparametric model), $\mathcal{M}$. The minimally informative distribution on $\mathcal{M}$ reflects the fact that investigators often have little knowledge of the data-generating mechanism before seeing the data. We describe this part of the procedure in more detail in Section 3.1.
- Then, for each of the J data-generating mechanisms, we generate S simulated datasets. So, for each j and s we have a dataset $\mathsf{D}_{j,s}$ of some finite sample size n from P_j.
- Then, for each of the L estimators being considered, we compute the parameter estimate $\hat{\theta}_{l,j,s}$ and the standard error estimate $\hat{\sigma}_{l,j,s}$[1]. Let $\hat{\theta}_{1,j,s},\ldots,\hat{\theta}_{L,j,s}$ denote the parameter estimates and $\hat{\sigma}_{1,j,s},\ldots,\hat{\sigma}_{L,j,s}$ denote the standard error estimates.
- We next compute the performance metrics based on sample averages across the S simulated draws. For distribution P_j and sample size n, the Monte-Carlo bias, confidence interval coverage, and MSE of a given estimator $\hat{\theta}_l$ are

1 We note that BART returns the credible interval – not the standard error estimate.

$$\mathrm{B}_l^{(n)}(\mathrm{P}_j) = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}_{l,j,s} - \theta(\mathrm{P}_j),$$

$$\mathrm{CICov}_l^{(n)}(\mathrm{P}_j) = \frac{1}{S}\sum_{s=1}^{S}\mathbb{1}\left\{\hat{\theta}_{l,j,s} - \frac{c\hat{\sigma}_{l,j,s}}{\sqrt{n}} < \theta(\mathrm{P}_j) < \hat{\theta}_{l,j,s} + \frac{c\hat{\sigma}_{l,j,s}}{\sqrt{n}}\right\},$$

$$\mathrm{MSE}_l^{(n)}(\mathrm{P}_j) = \frac{1}{S}\sum_{s=1}^{S}\{\hat{\theta}_{l,j,s} - \theta(\mathrm{P}_j)\}^2,$$

where c denotes the 97.5th percentile of the standard normal distribution.

- Finally, we summarize the performance across the J data distributions with a probability distribution of each performance metric (e.g., the bias). Specifically, the bias for estimator l at sample size n is summarized by:

$$\mathbb{P}(|\mathrm{B}_l^{(n)}| > b) = \frac{1}{J}\sum_{j=1}^{J}\mathbb{1}\{|\mathrm{B}_l^{(n)}(\mathrm{P}_j)| > b\}.$$

This function can be seen as an approximation of the so-called survival or reliability function, where the former name is common in the statistics literature and the latter in the engineering literature. We use the name reliability function, because it is closer to the intended interpretation for each estimation procedure. Because we consider the sample space of probability distributions as all possible data-generating mechanisms (i.e., phenomena under study), the probability distribution can be heuristically interpreted as probabilities across all possible studies.

In contrast to standard simulation studies, which often focus on a few data-generating mechanisms, this large-scale Monte-Carlo simulation will provide evidence of the behavior of the estimators across many data-generating mechanisms and, consequently, will be informative for assessing how the estimators perform across a broader range of problems that may be encountered within the defined data setting, $\mathcal{M}$.

3.1 Specifying and sampling from the space of probability distributions

Our simulations will be restricted to binary treatments and binary outcomes. Although we do allow for non-binary confounding variables, we restrict to discrete, ordinal variables. These restrictions ensure computational tractability but generalizations can, in principle, be devised for any data structure. Our models for P will be characterized by the following user-given, minimally informative priors, representing an analyst's prior knowledge of the data generating distribution:

- An integer u representing the number of binary confounding variables.
- An integer h representing the number of nonbinary confounding variables, together with an integer c denoting the cardinality of the support of each. Without loss of generality, we assume that each is supported in the set $\mathcal{N} = \{0, 1/(c-1), 2/(c-1), \ldots, 1\}$.
- A parameter η that controls the non-linearity of the effect of the non-binary confounding variables in the data generating process (additional detail later).
- A parameter ρ that controls the smoothness of the effect of the non-binary confounding variables in the data generating process (additional detail later).
- An interaction order k (additional detail later).
- A boolean value hte $\in$ {TRUE, FALSE} indicating whether there is treatment effect heterogeneity, meaning that the effect of the treatment on the outcome varies by the level of one or more confounding variables.
- A positive number q bounding the treatment probabilities as

$$\max_{t\in\{0,1\},x\in\mathcal{X}} \frac{\mathsf{P}(T=t)}{\mathsf{P}(T=t|X=x)} \leq q,$$

where $\mathcal{X}$ is the support of X.

• A real number b representing the desired *confounding bias*: $\mathsf{E}(Y|T=1) - \mathsf{E}(Y|T=0) - \mathsf{E}\{\mathsf{m}(1,X) - \mathsf{m}(0,X)\}$.

Our sampling scheme proceeds by first sampling $\mathsf{P}(X=x)$, then $\mathsf{P}(T=1|X=x)$, and finally $\mathsf{P}(Y=1|T=t,X=x)$. Note that the vector X takes values on the set $\mathcal{X} = \{0,1\}^u \times \mathcal{N}^u$, which has cardinality $2^u \times c^h$. In the first step, we sample the vector $\{\mathsf{P}(X=x) : x \in \mathcal{X}\}$ using a Dirichlet uniform distribution in $[0,1]^{2^u \times c^h}$. The Dirichlet distribution is used as it is the most commonly used prior distribution for categorical data and will allow us to explore many possible covariate distributions. Other distributions could be used that would induce more or less correlation among the covariates; see Dunson and Xing (2009) for distributions over the space of multivariate, categorical data [37].

3.1.1 Sampling treatment probabilities

Let $X = (X_{\text{bin}}, X_{\text{num}})$ represent the partitioning of binary and non-binary confounding variables. Treatment probabilities are generated using a linear probability model where we consider:
• interactions up to k-th order for X_{bin} and
• interactions of each of the aforementioned terms with a non-linear transformation of X_{num}.

Let $L = \{l = (l_1, \ldots, l_u) : l_j \in \{0,1\}; \sum_j l_j \leq k\}$. For instance, when $u = 3$ and $k = 2$, $L = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1)\}$. Note that this set has cardinality $|L| = \sum_{j=0}^{k} \binom{u}{j}$. Mathematically, the model takes the form $\mathsf{P}(T=1|X=x) = \mathsf{g}(x;\alpha)$, where

$$\mathsf{g}(x;\alpha) = \sum_{l \in L} \{\alpha_{0,l} + \alpha_{1,l} f_l(x_{\text{num}})\} \prod_{j=1}^{u} x_{\text{bin},j}^{l_j}.$$

For example, when $u = 2$ and $k = 1$,

$$\mathsf{g}(x;\alpha) = \alpha_{0,1} + \alpha_{1,1} f_1(x_{\text{num}}) + \alpha_{0,2} x_{\text{bin},1} f_2(x_{\text{num}}) + \alpha_{0,3} x_{\text{bin},1} + \alpha_{1,3} x_{\text{bin},1} f_3(x_{\text{num}}).$$

The parameters $(\alpha_{0,l}, \alpha_{1,l}, f_l)$ are sampled as follows. First, each $f_l(x_{\text{num}})$ is sampled from a Gaussian process with linear mean $\gamma^\top x_{\text{num}}$ and covariance function equal to:

$$K(x_{\text{num},i}, x_{\text{num},j}) = \eta \exp\{-\rho ||x_{\text{num},i} - x_{\text{num},j}||^2\},$$

where each coefficient in γ is independently drawn from a standard normal distribution. Then, the vector of coefficients $\{(\alpha_{0,l}, \alpha_{1,l}) : l \in L\}$ is sampled uniformly from a convex polytope defined by the following linear constraints for all $x \in \mathcal{X}$:

$$\begin{aligned}
1 &\geq \mathsf{g}(x;\alpha), \\
0 &\leq \mathsf{g}(x;\alpha), \\
q &\geq \frac{\sum_x \mathsf{g}(x;\alpha)\mathsf{P}(X=x)}{\mathsf{g}(x;\alpha)}, \\
q &\geq \frac{\sum_x \{1 - \mathsf{g}(x;\alpha)\}\mathsf{P}(X=x)}{1 - \mathsf{g}(x;\alpha)},
\end{aligned}$$

where the first two constraints ensure that $\mathsf{g}(x;\alpha)$ is a well-defined probability. The third and fourth constraints prevent near-positivity violations governed by the parameter q, i.e., they ensure the conditional probability of each treatment level given each covariate value is not too small relative to the marginal probability of that treatment level. Note that the third and fourth constraints are linear; e.g., the third constraint can be rewritten as $\sum_x \mathsf{g}(x)\mathsf{P}(X=x) - q \times \mathsf{g}(x;\alpha) \leq 0$. The aforementioned sampling is performed using the `volesti` R package [38].

Once a vector $\{(\alpha_{0,l}, \alpha_{1,l}) : l \in L\}$ is sampled, it is checked for whether it can possibly yield a desired confounding bias b, the details of which are given in the following. If not, the current draw is rejected.

The confounding bias for a given distribution is equal to:

$$C = \sum_t (2t-1) \sum_x \frac{\mathsf{P}(X=x)}{\mathsf{P}(T=t)} \{\mathsf{P}(T=t|X=x) - \mathsf{P}(T=t)\} \quad \mathsf{P}(Y=1|T=t, X=x), \tag{1}$$

so we must ensure that $\mathsf{P}(X=x)$ and $\mathsf{P}(T=t|X=x)$ are such that it is possible to find values $0 \leq \mathsf{P}(Y=1|T=t, X=x) \leq 1$ such that $C=b$. This occurs if $C_{\text{low}} \leq b \leq C_{\text{high}}$, where

$$\begin{aligned} C_{\text{high}} &= \sum_x \max\left\{\frac{\mathsf{P}(X=x)}{\mathsf{P}(T=1)}\{\mathsf{P}(T=1|X=x) - \mathsf{P}(T=1)\}, 0\right\} \\ &\quad - \sum_x \min\left\{\frac{\mathsf{P}(X=x)}{\mathsf{P}(T=0)}\{\mathsf{P}(T=0|X=x) - \mathsf{P}(T=0)\}, 0\right\}, \\ C_{\text{low}} &= \sum_x \min\left\{\frac{\mathsf{P}(X=x)}{\mathsf{P}(T=1)}\{\mathsf{P}(T=1|X=x) - \mathsf{P}(T=1)\}, 0\right\} \\ &\quad - \sum_x \max\left\{\frac{\mathsf{P}(X=x)}{\mathsf{P}(T=0)}\{\mathsf{P}(T=0|X=x) - \mathsf{P}(T=0)\}, 0\right\}. \end{aligned}$$

If $C_{\text{low}} \leq b \leq C_{\text{high}}$ is false, we reject the current draw of $\{\mathsf{P}(X=x) : x \in \mathcal{X}\}$ and $\{(f_l, \alpha_{0,l}, \alpha_{1,l}) : l \in L\}$ and repeat the process until $C_{\text{low}} \leq b \leq C_{\text{high}}$. If this condition is not achievable after 1,000 iterations, this is an indicator that the initial conditions may be infeasible. In this case, the algorithm fails and does not return a sampled distribution.

3.1.2 Sampling the outcome mechanism

Outcomes are also generated using a linear probability model: $\mathsf{P}(Y=1|T=t, X=x) = \mathsf{m}_{\lambda,\beta}(t,x)$, where

$$\mathsf{m}_{\lambda,\beta}(t,x) = t \sum_{l \in L} \{\lambda_{0,l} + \lambda_{1,l} h_l(x_{\text{num}})\} \tilde{x}_l + \sum_{l \in L} \{\beta_{0,l} + \beta_{1,l} w_l(x_{\text{num}})\} \tilde{x}_l,$$

if there is treatment effect heterogeneity, and

$$\mathsf{m}_{\lambda,\beta}(t,x) = t\lambda + \{\beta_{0,l} + \beta_{1,l} w_l(x_{\text{num}})\} \tilde{x}_l,$$

if there is no treatment effect heterogeneity, where $\tilde{x}_l = \prod_{j=1}^{u} x_{\text{bin},j}^{l_j}$ to simplify notation. The functions $h_l(x_{\text{num}})$ and $w_l(x_{\text{num}})$ are drawn from Gaussian processes as mentioned earlier. The confounding bias for a given distribution is equal to:

$$C = \sum_t (2t-1) \sum_x \frac{\mathsf{P}(X=x)}{\mathsf{P}(T=t)} \{\mathsf{P}(T=t|X=x) - \mathsf{P}(T=t)\} \times \mathsf{P}(Y=1|T=t, X=x) \tag{2}$$

and is linear in the coefficients $\{\lambda_{t,l}, \beta_{t,l} : t \in \{0,1\}, l \in L\}$. So, for a tolerance tol, we can draw the coefficients from a uniform distribution in the polytope defined by the following linear constraints:

$$\begin{aligned} 1 &\geq \mathsf{m}(t, x; \lambda, \beta), \\ 0 &\leq \mathsf{m}(t, x; \lambda, \beta), \\ b + \text{tol} &\geq C(\lambda, \beta), \\ b - \text{tol} &\leq C(\lambda, \beta), \end{aligned}$$

where $C(\lambda, \beta)$ denotes equation (2) with $\mathsf{P}(T=1|T=t, X=x)$ replaced by $\mathsf{m}(t, x; \lambda, \beta)$.

3.2 Application of the large-scale Monte-Carlo simulation approach

We define the data setting, $\mathcal{M}$, from which we sample data-generating distributions as follows. In addition to one binary treatment and one binary outcome, we considered five binary covariates ($u = 5$) and one numerical

covariate ($h = 1$) with cardinality 100. We considered interactions between covariates of order $k = \{1, 2, 3\}$. We considered distributions both with and without treatment effect heterogeneity (hte $\in$ {TRUE, FALSE}), and limiting the treatment probabilities as being ≥0.001 ($q = 1{,}000$). The parameter, η, controlling the nonlinearity of the numerical confounder was sampled from a uniform distribution $U(0.1, 10)$; the parameter, ρ, that controls the smoothness of this numerical confounder was sampled from a uniform distribution $U(0.1, 10)$. These ranges for η and ρ were chosen to result in some complex nonlinearities in the nuisance functions, but not to a degree that we felt most would consider to be unreasonable. Figure S1 shows how nonlinearity/smoothness in the propensity score can be varied by changing these values of η and ρ; we include the values of up to 100 for illustration. Finally, the parameter, b, that controls the confounding bias was sampled from a uniform distribution $U(-0.3, 0.3)$.

Within this defined setting, and for each combination of k and indicator of treatment effect heterogeneity, we sampled $J = 500$ distinct data-generating distributions, resulting in $3 \times 2 \times 500 = 3,000$ distributions. For each of these 3,000 distributions, we then sampled $S = 250$ data sets for each of the sample sizes $N \in \{100, 500, 1{,}000\}$. Figure S2 shows the Monte-Carlo simulation error – the distribution of the maximum Monte-Carlo biases, variances, and MSEs by sample size. Table 1 reports the averages and standard deviations of these maximum errors, providing evidence that $S = 250$ was a sufficient number of draws. However, approximately 21% of the 2,250,000 data sets (3,000 unique distributions × 3 sample sizes × 250 data sets = 2,250,000 data sets) we attempted to sample did not satisfy our propensity score and confounding bias constraints after 1,000 iterations. The final number of data sets was 1,773,000.

For each data set, we then estimated the ATE using $L = 4$ types of parametric and nonparametric estimators.

(1) First, we considered a parametric plug-in estimator (also called a g-computation estimator) based on fitting a model for $\mathsf{m}(t, x)$. We considered two implementations of this estimator: (1) logistic regression with only main effect terms and (2) logistic regression including all treatment × covariate interactions. For each subject $i = 1, \ldots, n$, the predicted outcome $\mathsf{m}_{\hat{\beta}}(t, X_i)$ is computed in a hypothetical world where treatment is set to $T_i = t$ for everyone. Then, the parametric plug-in estimate of the statistical parameter, θ, is given by $\hat{\theta}_{\text{sub}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathsf{m}}_{\hat{\beta}}(t, X_i)$. We make several notes about our consideration of these parametric plug-in estimators. First, although many would consider the main-terms-only model overly simplistic, it represents the approach of much current epidemiologic, medical, and social science research [e.g., 39–45]. Second, much of this research actually reports a conditional treatment effect estimate based on a coefficient of this main-terms parametric outcome regression model [e.g., 39–42,44,45], but to target the same parameter across estimators, we choose to include the marginal parametric plug-in regression estimator instead of the conditional parametric regression estimator.

(2) We also considered a parametric inverse probability of treatment weighting (IPTW) estimator with weights determined using the covariate balancing propensity score (CBPS), which optimizes the balance of covariates across the treatment and control groups [46]. For each subject $i = 1, \ldots, n$, let the inverse weight be denoted $\mathsf{g}_{\hat{\alpha}}(T = t, X_i) = \hat{\mathsf{P}}(T = t|X_i)$. Then, the estimate $\theta_{\text{IPTW-CBPS}} = \frac{1}{n}\sum_{i=1}^{n}\frac{T=t}{\hat{\mathsf{g}}_{\hat{\alpha}}(t, X_i)}Y_i$. Parametric propensity score estimators are popular in applied research, particularly in medical research [e.g., 12,47,48], despite their inefficiencies [49]. The CBPS estimator is one of the best-performing propensity-score-based estimators in finite samples and is robust to mild misspecification of the parametric model [46].

(3) Third, we considered a nonparametric plug-in estimator: Bayesian adaptive regression tree (BART, [3,50]), where BART is used as a model for $m(t, x)$ within a plug-in estimator, and posterior samples are drawn

Table 1: Average (and standard deviation) of the distributions of the maximum Monte-Carlo error among the five considered estimators

n	Bias	Variance	MSE
100	0.0063 (0.0006)	0.0009 (0.0002)	0.0014 (0.0007)
500	0.0027 (0.0003)	0.0002 (0.0000)	0.0004 (0.0003)
1,000	0.0019 (0.0002)	0.0001 (0.0000)	0.0003 (0.0002)

(additional detail is given in Section S1). We chose BART due to its excellent performance in recent data analysis competitions that has shown this type of estimator to work well in a wide range of settings [19]. The general BART model is given by $m_{\text{BART}}(t,x) = \sum_{m=1}^{M} g(t,x;\, T_m, M_m)$, where T_m represents a decision tree structure for tree m, while M_m corresponds to the predictions in the terminal nodes for tree m. The prior distribution for T_m encourages shallow decision trees, while the prior distribution for M_m enforces shrinkage so that overfitting does not occur for a large number of trees and that each tree is a weak-learner in the sense that it does not explain the full outcome regression on its own. One can then proceed with inference in a standard way using the posterior mean of θ_{BART} as a point estimator, and the relevant quantiles of the posterior distribution to construct credible intervals. For more details on the prior specification, or recent improvements to the original BART prior, we point readers to Hill et al. [51].

(4) Finally, we considered a nonparametric efficient estimator: TMLE [2,52], using machine learning for fitting nuisance parameters. Frequentist nonparametric efficient estimators – TMLE [5,52], double/ debiased machine learning [4], augmented inverse probability-weighted estimators [22], and other one-step bias-correction estimators [53] – rely on the estimation of both the treatment regression and outcome regression. Under regularity conditions, it can be shown that if the estimators of both the treatment and outcome regressions converge to the true functions in $L_2(\mathsf{P})$-norm at $n^{1/4}$-rate, then these estimators are "doubly robust" [5]. When both models are consistent, these estimators are asymptotically normal and semiparametric efficient, and the asymptotic variance can be consistently estimated by the sample variance of the efficient influence function [2,4]. Additional details are given in Section S1 of the Supplement. To fit the treatment and outcome regressions, we use an ensemble learner called Super Learner [54] with a library of estimators consisting of main-effects generalized linear models, BART [50], light gradient-boosting machine [55], and multivariate adaptive regression splines [56]. As recommended, we used a cross-fitted TMLE (using 10-folds for the cross-fitting and, within those, 5-folds for the Super Learner cross-validation), which typically results in better finite sample performance due to avoidance of the Donkser class condition required for asymptotic normality [4,57,58].

4 Results

We stratify results by each estimator, the three sample sizes, and by complexity of the data-generating mechanism. To vary complexity, we examine (a) no interactions between variables and no treatment effect heterogeneity, (b) treatment effect heterogeneity but no interactions between variables, (c) up to third-order interactions between variables and no treatment effect heterogeneity, and (d) up to third-order interactions and treatment effect heterogeneity. We also stratify results by degree of practical violations of the positivity assumption.

We see in Figures 1 and 2 that at least one, and often, both of the nonparametric estimators, BART and CV-TMLE, perform better than the parametric estimators in terms of both bias and MSE. The one exception is for the smallest sample size of $N = 100$ and the simplest two data-generating mechanisms that do not have any interactions between variables, the parametric plug-in estimators perform similar to TMLE in terms of bias (Figure 1(a) and (b)), and BART performs relatively worse. In terms of MSE, however, at least one of the nonparametric estimators performs best across all settings (Figure 2), with BART dominating in all cases when $N = 100$ and performing at least close to the best in all other scenarios. We also include a figure summarizing variance in this simulation study in Figure S3, and note that in general, BART has the smallest variance and CV-TMLE has the largest.

The largest separation between the parametric versus nonparametric reliability curves in Figures 1 and 2 occurs for sample sizes $N \in \{500, 1{,}000\}$ and in the more complex settings of variable interactions and possibly treatment effect heterogeneity. For example, in the most complex setting with $N = 1{,}000$ (Figure 1(d)), the nonparametric estimators result in absolute bias >0.1 in 0% of data distributions, but the parametric estimators result in absolute bias >0.1 in 11.1 and 13.3% of data distributions for the parametric main-terms plug-in estimator and IPTW estimator, respectively. This result was anticipated, because in larger sample sizes, the asymptotic advantages of nonparametric estimators to model complexity and nonlinearities may take hold.

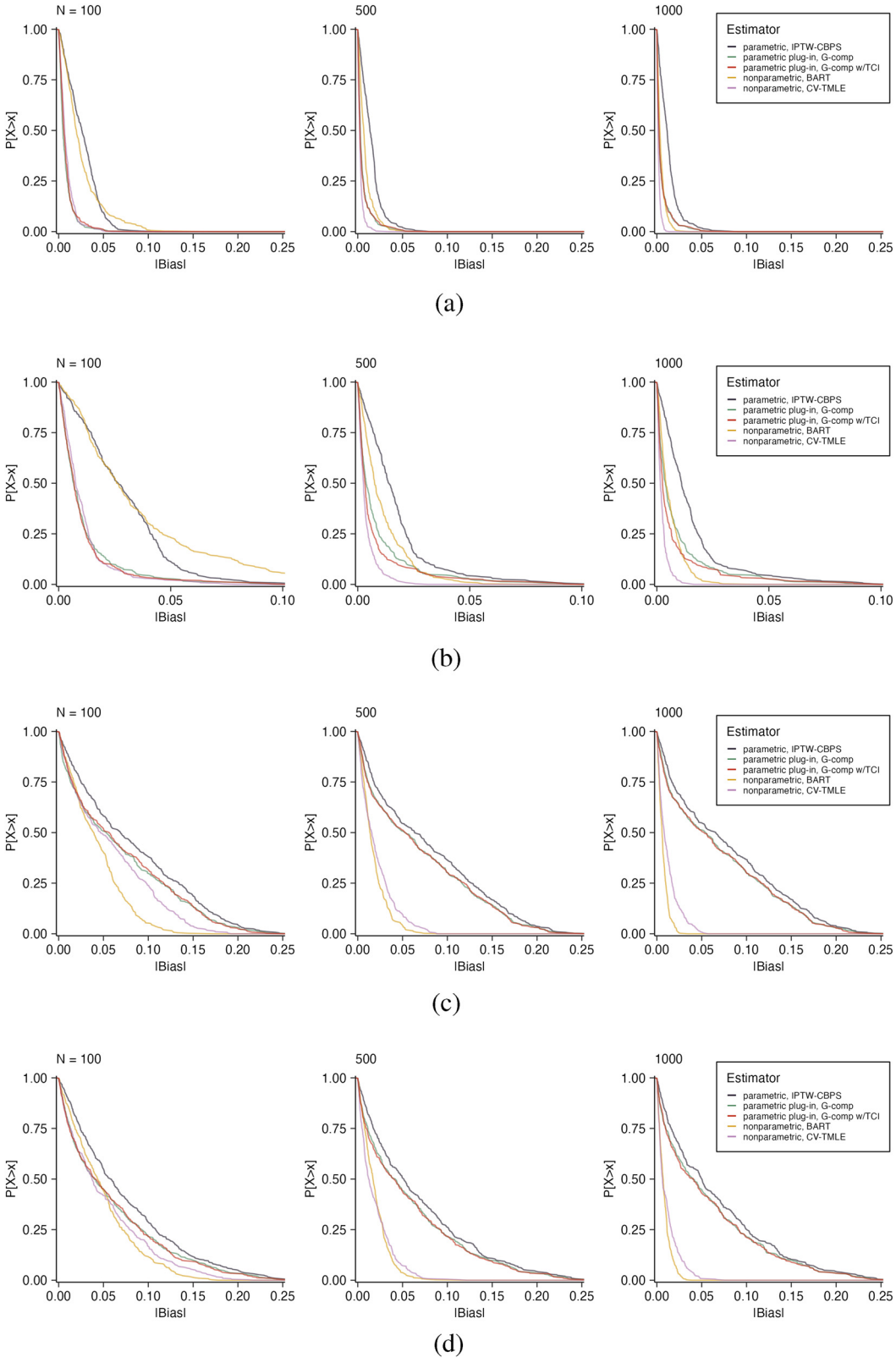

Figure 1: Reliability function for the absolute bias of each of the four estimators (parametric IPTW, parametric plug-in [g-computation: (i) main terms only and (ii) including treatment–covariate interactions (TCI)], nonparametric plug-in [BART], and nonparametric efficient estimator [CV-TMLE]) when considering data-generating mechanisms characterized by one binary treatment, one binary outcome, five binary covariates, one numeric covariate, and various levels of model complexity. (a) No treatment effect heterogeneity, no interactions, (b) treatment effect heterogeneity, no interactions, (c) no treatment effect heterogeneity, up to three-way interactions, and (d) treatment effect heterogeneity, up to three-way interactions.

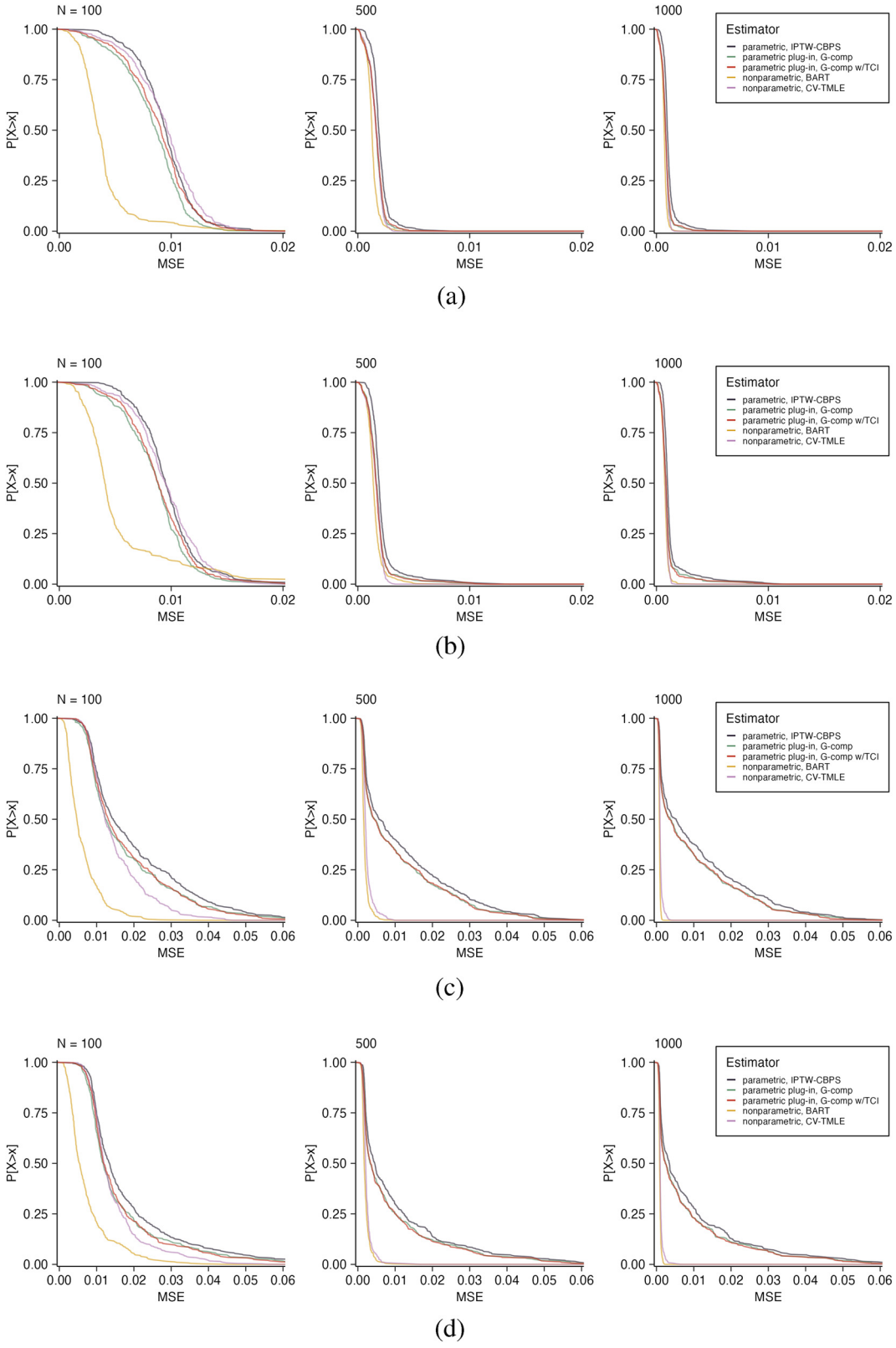

Figure 2: Reliability function for the absolute MSE of each of the four estimators (parametric IPTW, parametric plug-in [g-computation: (i) main terms only and (ii) including treatment–covariate interactions (TCI)], nonparametric plug-in [BART], and nonparametric efficient estimator [CV-TMLE]) when considering data-generating mechanisms characterized by one binary treatment, one binary outcome, five binary covariates, one numeric covariate, and various levels of model complexity: (a) no treatment effect heterogeneity, no interactions, (b) treatment effect heterogeneity, no interactions, (c) no treatment effect heterogeneity, up to three-way interactions, (d) treatment effect heterogeneity, up to three-way interactions.

We also stratify by degree of practical positivity violations in Figures S4 and S5. We categorize practical positivity violations as follows: "minimal violations" are those in which the minimum (across all iterations) predicted conditional probabilities of treatment ≥0.02; "moderate violations" are those in which the minimum predicted conditional probabilities of treatment <0.02 but ≥0.01; and "severe violations" are those in which the minimum predicted conditional probabilities of treatment <0.01. Although estimator performance degrades slightly in the presence of severe positivity violations, particularly for the IPTW estimator, the relative performance of the estimators remains the same.

In terms of 95% confidence interval (CI) coverage (Table 2), we see that all estimators perform well for sample size $N = 1{,}000$ in the simplest scenarios without interactions, with slight over-coverage by IPTW and BART. However, with more complexity in terms of interactions, coverage suffers for the parametric estimators, particularly with increasing sample size; in contrast, the nonparametric BART and CV-TMLE estimators continue to attain at least the nominal coverage rate. We show in Table S1 the importance of using cross-fitting in attaining nominal coverage for TMLE when using data-adaptive regression estimators for asymptotic normality [4,57,58]. Interestingly, BART tends to over-estimate uncertainty leading to coverage above the 95% level. This is not generally expected to hold for BART, though potentially this could be addressed using improvements to the BART prior, such as the SoftBART prior [59], which is better suited to many realistic data-generating processes and has better asymptotic properties.

5 Discussion

We proposed a large-scale Monte-Carlo simulation method to bring more comprehensive evidence to the debate [16] as to whether nonparametric estimators that use data-adaptive machine learning algorithms in

Table 2: Median (interquartile range) values of the 95% CI coverage distributions across all estimators and sample sizes. Int = interaction, HTE = treatment effect heterogeneity, TCI = treatment-covariate interactions

N	IPTW	G comp.	G comp. w/TCI	BART	TMLE	CV-TMLE
No int., no HTE						
100	0.96 (0.95, 0.98)	0.93 (0.92, 0.94)	0.90 (0.88, 0.92)	1.00 (0.99, 1.00)	0.82 (0.80, 0.85)	0.97 (0.96, 0.98)
500	0.96 (0.95, 0.98)	0.94 (0.94, 0.96)	0.94 (0.93, 0.95)	0.99 (0.98, 0.99)	0.92 (0.90, 0.93)	0.96 (0.95, 0.96)
1,000	0.96 (0.94, 0.98)	0.95 (0.93, 0.96)	0.94 (0.93, 0.96)	0.98 (0.98, 0.99)	0.93 (0.92, 0.94)	0.95 (0.94, 0.96)
No int., HTE						
100	0.96 (0.95, 0.98)	0.92 (0.91, 0.94)	0.90 (0.88, 0.92)	1.00 (0.99, 1.00)	0.83 (0.80, 0.85)	0.97 (0.96, 0.98)
500	0.96 (0.95, 0.98)	0.94 (0.93, 0.95)	0.94 (0.92, 0.95)	0.98 (0.97, 0.99)	0.91 (0.90, 0.93)	0.96 (0.95, 0.96)
1,000	0.96 (0.94, 0.98)	0.94 (0.92, 0.95)	0.94 (0.92, 0.95)	0.98 (0.97, 0.99)	0.93 (0.91, 0.94)	0.96 (0.94, 0.96)
Two-way int., no HTE						
100	0.93 (0.83, 0.96)	0.91 (0.80, 0.93)	0.88 (0.77, 0.91)	0.99 (0.98, 1.00)	0.78 (0.71, 0.81)	0.96 (0.93, 0.97)
500	0.86 (0.38, 0.95)	0.87 (0.37, 0.94)	0.88 (0.36, 0.94)	0.98 (0.98, 0.99)	0.89 (0.87, 0.91)	0.95 (0.94, 0.96)
1,000	0.74 (0.10, 0.94)	0.80 (0.11, 0.93)	0.80 (0.11, 0.93)	0.98 (0.98, 0.99)	0.91 (0.90, 0.92)	0.95 (0.94, 0.96)
Two-way int., HTE						
100	0.94 (0.87, 0.96)	0.91 (0.85, 0.94)	0.88 (0.83, 0.91)	0.99 (0.98, 1.00)	0.79 (0.73, 0.83)	0.96 (0.94, 0.97)
500	0.88 (0.48, 0.95)	0.89 (0.52, 0.94)	0.88 (0.52, 0.94)	0.98 (0.97, 0.99)	0.89 (0.87, 0.90)	0.95 (0.94, 0.96)
1,000	0.80 (0.17, 0.94)	0.83 (0.22, 0.93)	0.83 (0.22, 0.93)	0.98 (0.97, 0.98)	0.91 (0.90, 0.92)	0.95 (0.94, 0.96)
Three-way int., no HTE						
100	0.91 (0.77, 0.95)	0.89 (0.76, 0.93)	0.86 (0.72, 0.90)	0.99 (0.97, 1.00)	0.76 (0.65, 0.81)	0.96 (0.91, 0.97)
500	0.70 (0.20, 0.94)	0.74 (0.22, 0.94)	0.76 (0.19, 0.93)	0.98 (0.96, 0.99)	0.87 (0.80, 0.89)	0.94 (0.90, 0.96)
1,000	0.47 (0.02, 0.93)	0.55 (0.03, 0.93)	0.55 (0.02, 0.92)	0.98 (0.97, 0.99)	0.89 (0.86, 0.91)	0.94 (0.91, 0.96)
Three-way int., HTE						
100	0.92 (0.84, 0.95)	0.90 (0.82, 0.93)	0.88 (0.80, 0.91)	0.99 (0.96, 1.00)	0.78 (0.70, 0.82)	0.96 (0.93, 0.97)
500	0.82 (0.38, 0.94)	0.84 (0.42, 0.94)	0.82 (0.41, 0.93)	0.98 (0.96, 0.98)	0.87 (0.82, 0.90)	0.94 (0.90, 0.96)
1,000	0.66 (0.10, 0.93)	0.70 (0.13, 0.93)	0.72 (0.12, 0.93)	0.98 (0.97, 0.98)	0.90 (0.87, 0.92)	0.94 (0.92, 0.96)

fitting nuisance parameters confer any meaningful advantage over simpler parametric methods in real-world finite sample analyses. Previously, others contributed finite sample evidence in favor of one estimator or class of estimators over another by conducting simulation studies across a smaller, more bespoke set of data-generating processes [19,24,25,27,28], which may not be representative of performance in general. Consequently, our proposed large-scale approach greatly expands the number of data-generating mechanisms considered from a small few to thousands within a data setting that a user can define based on some minimal knowledge about observed variables in their data (but without the need to first look at their data), thereby likely resulting in more generalizable – and thus, informative – evidence of estimator performance in finite samples where asymptotic properties learned from theoretical results may not provide good approximations.

We applied our proposed approach to compare the performance of two nonparametric estimators, BART and CV-TMLE, to two parametric estimators, an outcome regression-based plug-in estimator and IPTW, in finite samples of sizes N = 100, 500, and 1,000 and across different degrees of model complexity. Corroborating findings from previous, smaller-scale simulation studies [19,24–26], in our application of the proposed approach, we found that even in small samples, nonparametric estimators nearly always outperform the parametric estimators in terms of bias and MSE. However, the advantage of nonparametric estimation attenuated with decreasing sample size and decreasing complexity, and in the simplest data-generating mechanisms and samples of N = 100, the parametric plug-in estimator performed similar to the nonparametric efficient estimator in terms of bias and between the nonparametric estimators in terms of MSE.

Even though our results were learned from 1,773,000 data sets across 3,000 data-generating mechanisms, the space of data-generating mechanisms we considered was nonetheless limited. In particular, we only considered settings with a binary treatment, a binary outcome, and six confounding variables, only one of which was multi-valued. It is certainly possible that our conclusions would differ for more complex settings. In future work, we will develop a software tool for running simulations with user-specified outcome and covariate types and covariate dimensions. This way, users would be able to compare estimator performances over a space of data-generating mechanisms that are likely to contain the probability distribution corresponding to their real-world data sampling setting or a close approximation thereof. Such future work could be further extended to accommodate longitudinal data as well as common data complexities such as missing data, measurement error, and selection bias.

Finally, although we have shown that the choice of nonparametric vs parametric estimator may matter – in some cases more than others – all estimators are limited by the data input. Unmeasured variables and variables measured with error are significant and near-ubiquitous limitations that can thwart accurate causal effect estimation and inference. A relatively recent high-profile and high-stakes example involved data errors leading to inaccurate algorithmic predictions in the criminal justice system that resulted in unintended parole denials [60,61]. We can work to improve the estimation step of answering research questions, but the accuracy of our answers will be limited by the weakest link, highlighting the importance of theory, subject matter knowledge, identification, and data quality, in addition to estimation.

In this article, we have focused on strengthening the estimation link. Our results show that in the large space of settings we have considered, this can be accomplished by employing nonparametric estimators grounded in asymptotic theory to substantially reduce bias in large-sample settings with interactions and nonlinearities while compromising very little in terms of performance even in simple, small-sample settings.

Funding information: ID, KER, and NTW were supported through a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2021C2-23636-IC).

Author contributions: KER and ID originally devised the study and wrote the main manuscript text. NTW conducted the simulations and prepared figures and tables. CHM contributed to study design and writing the main text. JA contributed to study design and writing the main text. All authors reviewed and critically revised the manuscript.

Conflict of interest: Prof. Iván Díaz is one of the Editors of the Journal of Causal Inference but was not involved in the review process of this article.

Data availability statement: The datasets generated and/or analyzed during this study are available in https://github.com/nt-williams/doesNPMatter.

References

[1] Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998;66(2):315–31.
[2] van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostatist. 2006;2(1):Article 11.
[3] Hill JL. Bayesian nonparametric modeling for causal inference. J Comput Graph Stat. 2011;20(1):217–40.
[4] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. Econometr J. 2018;21(1):C1–C68.
[5] Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. New York, New York: Springer; 2011.
[6] Balzer LB, Ayieko J, Kwarisiima D, Chamie G, Charlebois ED, Schwab J, et al. Far from MCAR: obtaining population-level estimates of HIV viral suppression. Epidemiology (Cambridge, Mass). 2020;31(5):620.
[7] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Amer Stat Assoc. 2018;113(523):1228–42.
[8] Athey S. The impact of machine learning on economics. In: The Economics of artificial intelligence: An Agenda. Chicago, Illinois: University of Chicago Press; 2018. p. 507–47.
[9] Ahern J, Balzer L, Galea S. The roles of outlet density and norms in alcohol use disorder. Drug and Alcohol Dependence. 2015;151:144–50.
[10] Grimmer J, Roberts ME, Stewart BM. Machine learning for social science: An agnostic approach. Ann Rev Politic Sci. 2021;24:395–419.
[11] Egami N, Fong CJ, Grimmer J, Roberts ME, Stewart BM. How to make causal inferences using texts. Sci Adv. 2022;8(42):eabg2652.
[12] Pirracchio R, Petersen ML, Van Der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. Amer J Epidemiol. 2015;181(2):108–19.
[13] Brand JE, Zhou X, Xie Y. Recent developments in causal inference and machine learning. Ann Rev Sociol. 2023;49:81–110.
[14] Kreif N, Diaz Ordaz K. Machine learning in policy evaluation: new tools for causal inference. In: Oxford research encyclopedia of economics and finance. Oxford, United Kingdom: Oxford University Press; 2019.
[15] Little RJ. In praise of simplicity not mathematistry! Ten simple powerful ideas for the statistical scientist. J Amer Statist Assoc. 2013;108(502):359–69.
[16] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat. 2004;86(1):4–29.
[17] Pearl J. Causality. Cambridge, United Kingdom: Cambridge University Press; 2009.
[18] Zhao Q, Hastie T. Causal interpretations of black-box models. J Business Econ Stat. 2021;39(1):272–81.
[19] Dorie V, Hill J, Shalit U, Scott M, Cervone D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. Statist Sci. 2019;34(1):43–68.
[20] Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In: Proceedings of the American Statistical Association. vol. 1999. Indianapolis, IN; 2000. p. 6–10.
[21] Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: AIDS epidemiology: methodological issues. New York, New York: Springer; 1992. p. 297–331.
[22] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Amer Stat Assoc. 1994;89(427):846–66.
[23] Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Stat Med. 1997;16(3):285–319.
[24] Porter KE, Gruber S, van der Laan MJ, Sekhon JS. The relative performance of targeted maximum likelihood estimators. Int J Biostat. 2011;7(1):Article 31.
[25] Ozery-Flato M, Thodoroff P, Ninio M, Rosen-Zvi M, El-Hay T. Adversarial balancing for causal inference. 2018. arXiv: http://arXiv.org/abs/arXiv:181007406.
[26] Balzer LB, van der Laan M, Ayieko J, Kamya M, Chamie G, Schwab J, et al. Two-Stage TMLE to reduce bias and improve efficiency in cluster randomized trials. Biostatistics. 2023;24(2):502–17.
[27] Parikh H, Varjao C, Xu L, Tchetgen ET. Validating causal inference methods. In: International Conference on Machine Learning. PMLR; 2022. p. 17346–58.
[28] Schuler A, Jung K, Tibshirani R, Hastie T, Shah N. Synth-validation: Selecting the best causal inference method for a given dataset. 2017. arXiv: http://arXiv.org/abs/arXiv:171100083.
[29] Advani A, Kitagawa T, Slłoczyński T. Mostly harmless simulations? Using Monte Carlo studies for estimator selection. J Appl Econom. 2019;34(6):893–910.

[30] Huber M, Lechner M, Wunsch C. The performance of estimators based on the propensity score. J Econom. 2013;175(1):1–21.
[31] Busso M, DiNardo J, McCrary J. New evidence on the finite sample properties of propensity score reweighting and matching estimators. Rev Econ Stat. 2014;96(5):885–97.
[32] Keele L, Small DS. Comparing covariate prioritization via matching to machine learning methods for causal inference using five empirical applications. Amer Statist. 2021;75(4):355–63.
[33] Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal. 2014;72:219–26.
[34] Athey S, Imbens GW, Metzger J, Munro E. Using Wasserstein generative adversarial networks for the design of Monte-Carlo simulations. J Econom. 2021;105076.
[35] Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. Tabddpm: Modelling tabular data with diffusion models. In: International Conference on Machine Learning. PMLR; 2023. p. 17564–79.
[36] van der Laan MJ, Dudoit S, van der Vaart AW. The cross-validated adaptive epsilon-net estimator. Statistics Decisions. 2006;24(3):373–95.
[37] Dunson DB, Xing C. Nonparametric Bayes modeling of multivariate categorical data. J Amer Stat Assoc. 2009;104(487):1042–51.
[38] Fisikopoulos V, Chalkis A. Contributors in file inst/AUTHORS. volesti: Volume Approximation and Sampling of Convex Polytopes; 2020. R package version 1.1.2. https://CRAN.R-project.org/package=volesti.
[39] Wise LA, Wang TR, Ncube CN, Lovett SM, Abrams J, Boynton-Jarrett R, et al. Use of chemical hair straighteners and fecundability in a North American preconception cohort. Amer J Epidemiol. 2023;192(7):1066–80.
[40] Belesova K, Gasparrini A, Wilkinson P, Sié A, Sauerborn R. Child survival and annual crop yield reductions in rural Burkina Faso: critical windows of vulnerability around early life development. Amer J Epidemiol. 2023;192(7):1116–27.
[41] Lu D, Yu Y, Ludvigsson JF, Oberg AS, Sooorensen HT, László KD, et al. Birth weight, gestational age, and risk of cardiovascular disease in early adulthood: influence of familial factors. Amer J Epidemiol. 2023;192(6):866–77.
[42] Khurshid S, Al-Alusi MA, Churchill TW, Guseh JS, Ellinor PT. Accelerometer-derived weekend warrior physical activity and incident cardiovascular disease. JAMA. 2023;330(3):247–52.
[43] Steenland MW, Fabi RE, Bellerose M, Desir A, White MS, Wherry LR. State public insurance coverage policies and postpartum care among immigrants. JAMA. 2023;330(3):238–46.
[44] Zhang L. Racial inequality in work environments. Amer Sociol Rev. 2023;88(2):252–83.
[45] Sharkey P, Torrats-Espinosa G, Takyar D. Community and the crime decline: The causal effect of local nonprofits on violent crime. Amer Sociol Rev. 2017;82(6):1214–40.
[46] Imai K, Ratkovic M. Covariate balancing propensity score. J R Stat Soc Ser B (Stat Meth). 2014;76(1):243–63.
[47] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006;59(5):437-e1.
[48] Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary JY, Porcher R. Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Medicine. 2010;36:1993–2003.
[49] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of double-robust estimators when inverse probability weights are highly variable. Stat Sci. 2007;22(4):544–59.
[50] Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat. 2010;4(1):266–98.
[51] Hill J, Linero A, Murray J. Bayesian additive regression trees: a review and look forward. Ann Rev Stat Appl. 2020;7:251–78.
[52] van der Laan MJ, Rose S. Targeted learning in data science. New York, New York: Springer; 2018.
[53] Pfanzagl J, Wefelmeyer W. Contributions to a general asymptotic statistical theory. Stat Risk Model. 1985;3(3–4):379–88.
[54] van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genetics Mol Biol. 2007;6(1):Article 25.
[55] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inform Process Syst. 2017;30:3146–54.
[56] Friedman JH. Multivariate adaptive regression splines. Ann Stat. 1991;19(1):1–67.
[57] Klaassen CA. Consistent estimation of the influence function of locally asymptotically linear estimators. Ann Stat. 1987:1548–62.
[58] Zheng W, van der Laan MJ. Cross-validated targeted minimum-loss-based estimation. In: Targeted learning. New York, New York: Springer; 2011. p. 459–74.
[59] Linero AR, Yang Y. Bayesian regression tree ensembles that adapt to smoothness and sparsity. J R Stat Soc Ser B Stat Methodol. 2018;80(5):1087–110.
[60] Rudin C, Carlson D. The secrets of machine learning: ten things you wish you had known earlier to be more effective at data analysis. In: Operations research & management science in the age of analytics. Seattle, Washington: INFORMS; 2019. p. 44–72.
[61] Wexler R. When a computer program keeps you in jail: How computers are harming criminal justice. New York Times. 2017: Available online: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html.