Received: 19 November 2021 | Revised: 16 January 2023 | Accepted: 30 January 2023

DOI: 10.1111/biom.13837

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Improved inference for doubly robust estimators of heterogeneous treatment effects

Heejun Shin | **Joseph Antonelli**

Department of Statistics, University of Florida, Gainesville, Florida, USA

Correspondence
Heejun Shin, Department of Statistics, University of Florida, Gainesville, FL, USA.
Email: hshin1@ufl.edu

Abstract

We propose a doubly robust approach to characterizing treatment effect heterogeneity in observational studies. We develop a frequentist inferential procedure that utilizes posterior distributions for both the propensity score and outcome regression models to provide valid inference on the conditional average treatment effect even when high-dimensional or nonparametric models are used. We show that our approach leads to conservative inference in finite samples or under model misspecification and provides a consistent variance estimator when both models are correctly specified. In simulations, we illustrate the utility of these results in difficult settings such as high-dimensional covariate spaces or highly flexible models for the propensity score and outcome regression. Lastly, we analyze environmental exposure data from NHANES to identify how the effects of these exposures vary by subject-level characteristics.

KEYWORDS

Bayesian nonparametrics, causal inference, doubly robust estimation, high-dimensional statistics, treatment effect heterogeneity

1 | INTRODUCTION

Understanding how the effect of a treatment varies across subgroups of the population is a common scientific goal in a wide array of fields. This variation in the treatment effect, typically referred to as treatment effect heterogeneity, is important for learning which individuals are most likely to benefit from treatment. Analyzing treatment effect heterogeneity is, however, a challenging problem, especially when the dimension of the covariates is large or nonparametric models are employed. One difficulty is the ability to perform inference in such settings as the bootstrap may not be valid, and approaches relying on asymptotic approximations may not perform well, particularly when the sample size is small. To address these problems, we combine posterior distributions for nuisance parameters with doubly robust estimators to provide an estimate of the conditional average treatment effect (CATE) that is consistent if either the outcome or treatment models are correctly specified. We also introduce a novel variance estimation procedure which (i) is consistent when both nuisance models are correctly specified with sufficiently fast contraction rates, and (ii) is conservative in finite samples, under model misspecification, or when nuisance parameters are estimated with slow contraction rates.

There has been a significant increase in attention to treatment effect heterogeneity in recent years. A number of approaches fit into a category referred to as meta-learners, which permit the usage of existing estimators in the machine learning or high-dimensional literature when estimating the CATE. Examples of such approaches can be found in Künzel et al. (2019) and Nie and Wager (2021). Other approaches have specifically tailored algorithms toward treatment effect heterogeneity. The causal forest algorithm developed in Wager and Athey (2018) uses regression trees and the sample splitting approach of Athey

and Imbens (2016) to provide nonparametric estimation and honest inference of heterogeneous treatment effects. The Bayesian causal forest approach of Hahn et al. (2020) extends the initial work of Hill (2011) that uses Bayesian additive regression trees (BART; Chipman et al. 2012) to estimate heterogeneous treatment effects. They separate the outcome regression function into two components, one of which corresponds to the CATE, which allows them to focus one BART prior distribution on treatment effect heterogeneity and shrink effects toward an overall homogeneous effect. This approach, or variants of it, have also been shown to work well in recent causal inference data analysis competitions (Dorie et al., 2019).

Recent work has looked to extend these approaches to high-dimensional situations (Powers et al., 2018). Estimation of average treatment effects in high-dimensional scenarios has garnered substantial interest (Antonelli et al., 2018; Antonelli, Parmigiani & Dominici, 2019; Antonelli & Cefalu, 2020; Belloni, Chernozhukov & Hansen, 2014; Chernozhukov et al., 2018; Farrell, 2015; Ning, Sida & Imai, 2020; Tan, 2020b). When estimating average causal effects, adjusting for a high-dimensional set of covariates is a nuisance parameter, and the target of interest is a low-dimensional quantity. The problem is more difficult when estimating the CATE, because the parameter of interest can itself be high-dimensional. One approach is to allow the CATE to depend on a pre-specified subset of the covariates. Abrevaya et al. (2015) suggested using an inverse probability-weighted method and integrating out the remaining covariates. However, this estimator is unstable, especially when the propensity score is misspecified, which led to the development of doubly robust estimators in recent papers (Fan et al., 2020; Kennedy, 2020; Knaus, 2020; Lee, Okui & Whang, 2017; Semenova & Chernozhukov, 2021). There also has been work on finding effect modifiers using machine learning methods (Athey & Imbens, 2016), especially in the causal rule framework (Lee, Bargagli-Stoffi & Dominici, 2020; Lee, Small & Dominici, 2021; Wang & Rudin, 2022).

At the core of much of the work in nonparametric or high-dimensional causal inference are doubly robust estimators (Bang & Robins, 2005; Scharfstein, Rotnitzky & Robins, 1999). These estimators have been popular for many years due to the namesake property that only one of the propensity score or outcome regression models needs to be correctly specified to obtain consistent inference. In nonparametric or high-dimensional settings, they have an added advantage that parametric convergence rates can be obtained even when each of the propensity score or outcome regression models converge at slower rates. While this is advantageous for point estimation, inference is more challenging, as confidence intervals commonly rely on both models being correctly specified. Although recent work has aimed to alleviate this (Avagyan & Vansteelandt, 2021; Benkeser et al., 2017; Dukes, Avagyan & Vansteelandt, 2020; Dukes, Vansteelandt & Whitney, 2021; Tan, 2020a; Van der Laan, 2014), nearly all of this work has focused on estimating average treatment effects, while the focus of this paper is on CATEs.

In this paper, we develop a novel frequentist approach that combines Bayesian modeling of the propensity score and outcome regression models with meta learners for estimating heterogeneous treatment effects. This approach allows us to handle high-dimensional confounder spaces or utilize highly flexible nonparametric Bayesian models. This extends the literature on doubly robust estimation and doubly robust inference as we are able to provide improved inference on CATEs even in finite samples or model misspecification. We see empirically in Section 4 that this leads to substantial gains in confidence interval performance. In conjunction with flexible nonparametric Bayesian models that ensure small biases of resulting treatment effect estimates, this leads to a procedure with strong estimation and inferential properties.

2 | ESTIMANDS AND IDENTIFYING ASSUMPTIONS

Our goal will be to estimate the effect of a treatment T on an outcome Y, and to understand whether the treatment effect varies by observed characteristics. We observe a p-dimensional vector of covariates denoted by $\boldsymbol{X}$ that are used to adjust for confounding bias. Throughout, we will allow p to be large and potentially growing with the sample size, denoted by n. We are interested in learning the extent to which the treatment effect varies by covariates $\boldsymbol{V}$, where $\boldsymbol{V}$ is a q-dimensional set of covariates, and q is assumed to be relatively small and not growing with n. Typically, the covariates in $\boldsymbol{V}$ will be a subset of the covariates in $\boldsymbol{X}$, though the remaining ideas hold even if they are not a subset of the covariates in $\boldsymbol{X}$. Throughout, for the sake of simplicity, we introduce identifying assumptions based on the former setting, but these can be easily adapted to suit the latter using $\boldsymbol{X} \cup \boldsymbol{V}$. We assume that we observe n independent and identically distributed random variables denoted by $\boldsymbol{D}_i = [Y_i, T_i, \boldsymbol{X}_i, \boldsymbol{V}_i]$ for $i = 1, \dots, n$. We are interested in the CATE denoted by

$$E(Y(1) - Y(0) | \boldsymbol{V} = \boldsymbol{v}) = \tau(\boldsymbol{v}),$$

where $Y(t)$ is the potential outcome we would observe under treatment level t. In order to identify these treatment effects from the observed data we rely on certain assumptions. First, we assume the stable unit treatment value assumption (SUTVA; Little and Rubin 2000). This

assumption states that the treatment of one unit does not affect the outcomes of other units and that the treatment is well defined in the sense that $Y_i = Y_i(T_i)$. We also assume positivity and unconfoundedness, which are defined as

- *Positivity:* $0 < P(T = 1|\boldsymbol{X} = \boldsymbol{x}) < 1$ for all $\boldsymbol{x}$.
- *Unconfoundedness:* $Y(t) \perp\!\!\!\perp T|\boldsymbol{X}$ for $t = 0, 1$.

Here, $P(T = 1|\boldsymbol{X} = \boldsymbol{x})$ denotes the propensity score, and positivity states that all subjects have a positive probability of receiving either treatment level. Unconfoundedness effectively states that there are no unmeasured common causes of the treatment and outcome. While there are multiple ways to identify the CATE from the observed data under these assumptions, we will use an approach that relies on constructing a pseudo-outcome defined by

$$Z_i \equiv Z(\boldsymbol{D}_i, p_{ti}, m_{ti}) = \frac{1(T_i = 1)}{p_{1i}}(Y_i - m_{1i}) + m_{1i} - \frac{1(T_i = 0)}{p_{0i}}(Y_i - m_{0i}) - m_{0i},$$

where $p_{ti} = P(T = t|\boldsymbol{X} = \boldsymbol{X}_i)$ and $m_{ti} = E(Y|T = t, \boldsymbol{X} = \boldsymbol{X}_i)$. This pseudo-outcome was originally used for doubly robust estimation of average treatment effects (Robins & Rotnitzky, 1995) and is referred to as the augmented inverse propensity score weighting (AIPW) estimator. Note that for brevity we will typically refer to this quantity simply as Z_i, though it is implied throughout that it is dependent on the data $\boldsymbol{D}$ and unknown model parameters $\boldsymbol{\Psi}$. This pseudo-outcome is doubly robust in the sense that when either $P(T = t|\boldsymbol{X} = \boldsymbol{x})$ or $E(Y|T = t, \boldsymbol{X} = \boldsymbol{x})$ are correctly specified, the following result holds:

$$E(Z|\boldsymbol{V} = \boldsymbol{v}) = \tau(\boldsymbol{v}).$$

Identification through this pseudo-outcome points to a two-step strategy toward estimating heterogeneous causal effects, which is commonly referred to as the doubly robust (DR)-learner, and has been studied in Kennedy (2020). At the first step, the propensity score and outcome regression models are estimated and the pseudo-outcome is constructed. At the second stage, the pseudo-outcome is regressed against the covariates $\boldsymbol{V}$ to estimate $\tau(\cdot)$. We detail our strategy for these two steps, and how we account for all sources of uncertainty in the following section.

3 | METHODOLOGY

Throughout, we will be working under the assumption that both the treatment and outcome models are fit using Bayesian approaches and therefore we have posterior distributions of both p_{ti} and m_{ti} for $i = 1, \ldots, n$ and $t \in \{0, 1\}$. We denote all unknown parameters of these two models by $\boldsymbol{\Psi}$. Our framework is intended to work in the more difficult scenarios when $\boldsymbol{\Psi}$ is high-dimensional, either because the number of covariates is large relative to the sample size (e.g., $p > n$) or because very flexible, nonparametric approaches have been used to estimate the treatment model, outcome model, or both. These could include nonparametric approaches such as the BART prior, Gaussian process regression, and dirichlet process mixtures, or high-dimensional models such as those based on spike-and-slab prior distributions. Inference in this situation is complicated by three key factors: (1) the estimated treatment effect is not solely a function of the unknown parameters and therefore we cannot simply use the posterior distribution for inference, (2) the bootstrap frequently does not apply when using high-dimensional models (El Karoui & Purdom, 2018; see Web Appendix G for an empirical illustration), and (3) as we see in Section 4, estimators relying on asymptotic approximations to inference may perform poorly in difficult situations such as finite sample sizes, model misspecification, or high-dimensionality. Our goal is to construct an approach that can provide valid inference despite the presence of these three complicating factors.

3.1 | Estimation and inferential strategy

Once the posterior distribution of the treatment and outcome model parameters are obtained, which we denote by $\boldsymbol{\Psi}|\boldsymbol{D}$, we must construct an estimator of $\tau(\boldsymbol{v})$. Throughout we assume that $\tau(\boldsymbol{v}) = \boldsymbol{v}\boldsymbol{\alpha}^*$ for some $\boldsymbol{\alpha}^*$. This implies that the true CATE is a linear function of $\boldsymbol{v}$, though this can include nonlinear functions of the covariates. We can define an estimator of this quantity as follows:

$$\Delta(\boldsymbol{\Psi}, \boldsymbol{D}) = \boldsymbol{v}(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{V}^T\boldsymbol{Z}, \quad (1)$$

where $\boldsymbol{Z} = [Z_1, \ldots, Z_n]^T$ and $\boldsymbol{V}$ is the $n \times q$ matrix of treatment effect modifiers. Note that this is a function of both $\boldsymbol{\Psi}$ and $\boldsymbol{D}$ since Z_i is defined to be a function of unknown parameters $\boldsymbol{\Psi}$ and $\boldsymbol{D}_i$.

There are two natural estimators of $\tau(\boldsymbol{v})$ once the posterior distribution is obtained. The more common approach in the causal inference literature is to construct an estimate of both p_{ti} and m_{ti}. In our setting, we could use the posterior mean by setting $\hat{p}_{ti} = E_{\boldsymbol{\Psi}|\boldsymbol{D}}[p_{ti}]$ and $\hat{m}_{ti} = E_{\boldsymbol{\Psi}|\boldsymbol{D}}[m_{ti}]$. Then, we can plug-in these values to define the pseudo-outcomes, $\hat{Z}_i = Z(\boldsymbol{D}_i, \hat{p}_{ti}, \hat{m}_{ti})$ for $i = 1, \ldots, n$, which we can regress against the effect modifying covariates $\boldsymbol{V}$. Using the notation above, this can be defined as

$\hat{\tau}(\boldsymbol{v}) = \Delta(\hat{\boldsymbol{\Psi}}, \boldsymbol{D})$ with $\hat{\boldsymbol{\Psi}} = E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\boldsymbol{\Psi}]$. While this strategy is reasonable and potentially leads to good performance of the resulting estimates of $\tau(\boldsymbol{v})$, inference is more challenging. Typically, either the bootstrap would be used to account for uncertainty in both stages of the estimator, or the estimator's asymptotic distribution is obtained from which inference can proceed. As discussed above, however, these approaches either may not be valid or may not perform well in finite samples with complex models for the propensity score and outcome regression models. One may think that the posterior distribution can be used to help with regard to inference, but it is not clear how the posterior distribution can be used to properly account for uncertainty in this estimator. For these reasons, we take a second approach to estimating the CATE, which is to construct an estimator as

$$\hat{\tau}(\boldsymbol{v}) = E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{\Psi}, \boldsymbol{D})],$$

which is the posterior mean of the $\Delta(\boldsymbol{\Psi}, \boldsymbol{D})$ function. Intuitively, for every posterior draw of the propensity score and outcome regression models, a new pseudo-outcome is constructed, and this outcome is regressed against $\boldsymbol{V}$. This is done for every posterior draw, and the mean of these values is our estimator. We approximate this posterior mean using B posterior draws as follows:

$$\frac{1}{B}\sum_{b=1}^{B}\Delta(\boldsymbol{\Psi}^{(b)}, \boldsymbol{D}) = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{v}(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{V}^T\boldsymbol{Z}^{(b)}, \quad (2)$$

where $Z^{(b)}$ is the b^{th} posterior draw of the pseudo-outcome and is defined as

$$\frac{1(T_i = 1)}{p_{1i}^{(b)}}(Y_i - m_{1i}^{(b)}) + m_{1i}^{(b)} - \frac{1(T_i = 0)}{p_{0i}^{(b)}}(Y_i - m_{0i}^{(b)}) - m_{0i}^{(b)}.$$

Now that we have defined our estimator, we can describe our strategy to estimating the variance of this estimator in a way that accounts for all sources of uncertainty. The true variance of interest is the variance of the sampling distribution of this estimator and is defined by $\text{Var}_{\boldsymbol{D}}E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{\Psi}, \boldsymbol{D})]$. There are two main sources of variability in this estimator: (1) the uncertainty in parameter estimation for the propensity score and outcome regression models, and (2) sampling variability in $\boldsymbol{D}_i$ that is present even if we knew the true outcome and propensity score models. We extend the results seen in Antonelli et al. (2022) in order to construct an estimator of the variance that separately targets these two sources of variability. We will provide intuition for this variance estimator, and in the following section will show that it is a consistent estimator of the variance that will be conservative in finite samples or under model misspecification.

Before defining our variance estimator, we must introduce additional notation. We will let $\boldsymbol{D}^{(m)}$ be a resampled version of our original data $\boldsymbol{D}$, where resampling is done with replacement as in the nonparametric bootstrap (Efron & Tibshirani, 1994). Our variance estimator can then be defined as

$$\text{Var}_{\boldsymbol{D}^{(m)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{\Psi}, \boldsymbol{D}^{(m)})]\} + \text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{\Psi}, \boldsymbol{D})]. \quad (3)$$

The first of these two terms resembles the true variance, except the outer variance is no longer with respect to $\boldsymbol{D}$, but is now with respect to $\boldsymbol{D}^{(m)}$. This is a crucial difference, however, as the first term does not account for variability due to parameter estimation. The inner expectation of the first term is with respect to the posterior distribution of $\boldsymbol{\Psi}$ given the observed data $\boldsymbol{D}$, and not the resampled data $\boldsymbol{D}^{(m)}$. This means that this variance term does not account for variability that is caused by the fact that different datasets would lead to different posterior distributions. Ignoring this source of variability will likely lead to anticonservative inference as our estimated variance will be smaller than the true variance of our estimator. To counter this issue, the second term is added, which is the variability of the estimator due to parameter uncertainty. It makes sense to add posterior variability to the first term, which was ignoring uncertainty from parameter estimation, however, it is not clear that the summation of these two terms leads to a valid variance estimator. In Section 3.3, we detail how this variance estimator (1) leads to conservative estimates of the variance in general, and (2) is consistent when both the treatment and outcome models are correctly specified and the posterior distributions of the propensity score and the conditional mean outcome contract at sufficiently fast rates.

3.2 | Properties of the conditional average treatment effect estimator

First, we detail the properties of our estimator when either the propensity score or outcome regression models is correctly specified. Note that at times, we utilize sample splitting for theoretical proofs where half of the sample is used for estimating the unknown parameters $\boldsymbol{\Psi}$, and the remaining half is used in the second stage of our estimator of the CATE. This simplifies certain calculations substantially, though we expect the same results to hold in the absence of sample splitting. For this reason, we utilize both a cross-fitted estimator in the spirit of Chernozhukov et al. (2018), as well as one that does not use sample splitting. Empirically, we find that the theoretical results hold for both estimators, but the finite sample performance of the estimator without sample splitting is better.

Results comparing these two estimators can be found in Web Appendix E.

Let $\boldsymbol{p}_t = (p_{t1}, \ldots, p_{tn})$, $\boldsymbol{m}_t = (m_{t1}, \ldots, m_{tn})$, and let $\tilde{\boldsymbol{p}}_t$ and $\tilde{\boldsymbol{m}}_t$ denote their limiting values. We also let $\boldsymbol{p}_t^*$ and $\boldsymbol{m}_t^*$ denote their unknown true values. Further let P_0 represent the true data-generating distribution for $\boldsymbol{D}_i$, and let $\mathbb{P}_n$ denote the posterior distribution from a sample of size n. We first introduce key assumptions that are unique to our approach, while other common regulatory conditions can be found in Web Appendix A.1.

Assumption 1 (Bounds on the posterior distribution).

$$\text{(i)} \ \sup_{P_0} E_{P_0} \text{Var}_n\left(\frac{p_{ti} - p_{ti}^*}{p_{ti}} | \boldsymbol{D}_i\right) \leq K_p < \infty$$
$$\text{(ii)} \ \sup_{P_0} E_{P_0} \text{Var}_n\left(m_{ti} - m_{ti}^* | \boldsymbol{D}_i\right) \leq K_m < \infty.$$

This assumption effectively states that the posterior distribution of p_{ti} does not assign mass to neighborhoods of 0 and that the difference between the true conditional mean of the outcome and the corresponding posterior is bounded, which is reasonable in practice.

Assumption 2 (Posterior contraction rates). There exist two sequences of numbers $\epsilon_{nt} \to 0$ and $\epsilon_{ny} \to 0$, and constants $M_t > 0$ and $M_y > 0$ such that

$$\text{(i)} \ \sup_{P_0} E_{P_0} \mathbb{P}_n\left(\frac{1}{\sqrt{n}} ||\boldsymbol{p}_t - \tilde{\boldsymbol{p}}_t||_2 > M_t \epsilon_{nt} \mid \boldsymbol{D}\right) \to 0,$$
$$\text{(ii)} \ \sup_{P_0} E_{P_0} \mathbb{P}_n\left(\frac{1}{\sqrt{n}} ||\boldsymbol{m}_t - \tilde{\boldsymbol{m}}_t||_2 > M_y \epsilon_{ny} \mid \boldsymbol{D}\right) \to 0.$$

Note that we use $||\boldsymbol{a}||_2 = \sqrt{a_1^2 + \cdots + a_n^2}$. This assumption states that the posterior distributions of the propensity score and outcome regression models contract at rates ϵ_{nt} and ϵ_{ny}, respectively. The standard parametric rate of posterior contraction is $n^{-1/2}$, while slower rates are typically seen for nonparametric Bayesian models or high-dimensional scenarios, where the rates depend on the complexity of the model and true data-generating process.

Theorem 1. *Assume positivity, unconfoundedness, SUTVA, Assumptions 1 and 2, and additional regulatory conditions found in Web Appendix A.1. Additionally assume that data splitting is used such that half of the data are used for estimating the nuisance parameters, while the other half are used to estimate the CATE. If either* $\tilde{\boldsymbol{p}}_t = \boldsymbol{p}_t^*$ *or* $\tilde{\boldsymbol{m}}_t = \boldsymbol{m}_t^*$, *then*

$$\sup_{P_0} E_{P_0} \mathbb{P}_n(|\Delta(\boldsymbol{D}, \boldsymbol{\Psi}) - \tau(\boldsymbol{v})| > M \epsilon_n | \boldsymbol{D}) \to 0.$$

If both $\tilde{\boldsymbol{p}}_t = \boldsymbol{p}_t^*$ *and* $\tilde{\boldsymbol{m}}_t = \boldsymbol{m}_t^*$, *then* $\epsilon_n = \max(n^{-1/2}, \epsilon_{nt}\epsilon_{ny})$. *If only one model is correctly specified, then* ϵ_n *is equal to the contraction rate of the correctly specified model.*

This result importantly implies convergence rates of point estimators such as posterior medians and means (assuming bounded posterior variance). For instance, this result implies that our estimator converges at the $\sqrt{n}$ rate if the product of the contraction rates for the two posterior distributions is $n^{-1/2}$ or smaller.

3.3 | Theoretical justification for variance estimation

Now, we provide theoretical justification for the variance estimator in Equation (3). First, we describe that for general $\Delta(\boldsymbol{\Psi}, \boldsymbol{D})$ that are functions of both the observed data and unknown parameters, this will provide conservative inference on average. Next, we show that for the specific choice of $\Delta(\boldsymbol{\Psi}, \boldsymbol{D})$ defined in Equation (1), this variance estimator is consistent when both the propensity score and outcome regression posterior distributions contract sufficiently fast. Throughout this section, let $\widehat{V}$ be the variance estimator defined in Equation (3), and let V be the true variance defined by $\text{Var}_{\boldsymbol{D}} E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{\Psi}, \boldsymbol{D})]$. Our goal is to understand whether our variance estimator is conservative in the sense that

$$E_{\boldsymbol{D}}\{\widehat{V} - V\} \gtrapprox 0. \tag{4}$$

In Web Appendix B, we prove under sample splitting that this result holds for a particular class of functions $\Delta(\boldsymbol{\Psi}, \boldsymbol{D})$, and argue why it is expected to hold in general. Additionally, we provide insight into the specific form of the doubly robust estimator and why it is expected that this conservative result will hold for the estimator used in Equation (1). This result shows that the variance estimator is biased upward and may lead to conservative inference as our estimates of the variance will be too large. Importantly, this result holds in finite samples or under model misspecification of either the propensity score or outcome regression models. While it is useful to know that our variance estimator tends to be conservative, this could be problematic if the estimated variance is far larger than the true variance, which could lead to meaningless confidence intervals and low power to detect significant treatment effects. In Web Appendix B.2, we prove that the variance estimator has the following upper bound:

$$E_{\boldsymbol{D}}\{\widehat{V}\} \lessapprox 2V - \max\left(\mathrm{Var}_{\boldsymbol{D}_{(1)}}\left[E_{\boldsymbol{D}_{(2)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}_{(1)}}[\Delta(\boldsymbol{\Psi},\boldsymbol{D}_{(2)})]\}\right],\right.$$
$$\left.\mathrm{Var}_{\boldsymbol{D}_{(2)}}\left[E_{\boldsymbol{D}_{(1)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}_{(1)}}[\Delta(\boldsymbol{\Psi},\boldsymbol{D}_{(2)})]\}\right]\right)$$

where $\boldsymbol{D}_{(1)}$ and $\boldsymbol{D}_{(2)}$ are separate splits of the data, which are used to find the posterior distribution of $\boldsymbol{\Psi}$ and estimate the CATE, respectively. This shows that while the variance estimator is conservative, it is generally much smaller than two times the true variance and therefore should not lead to overly wide confidence intervals. Additionally, we show that for the doubly robust CATE estimator defined in Equation (1), the variance estimator is consistent. This requires an additional assumption on certain moments of the posterior distributions for the propensity score and outcome regression converging sufficiently fast to zero. For brevity we leave details to Web Appendix A.1, though this assumption is similar to those typically made in the high-dimensional and semiparametric causal inference literature (Farrell, 2015) and is expected to hold when both nuisance models contract at the $n^{-1/4}$ or faster rates.

Theorem 2. *Let $\Delta(\boldsymbol{\Psi},\boldsymbol{D})$ be defined as in Equation (1). Assume positivity, unconfoundedness, SUTVA, additional regulatory conditions found in Web Appendix A.1, and that data splitting is used. If Assumptions 1 and 2 hold with rates $\epsilon_{nt} \leq n^{-1/4}$ and $\epsilon_{ny} \leq n^{-1/4}$, and both $\tilde{\boldsymbol{p}}_t = \boldsymbol{p}_t^*$ and $\tilde{\boldsymbol{m}}_t = \boldsymbol{m}_t^*$, then $\widehat{V} - V = o_p(n^{-1})$.*

A proof of this can be found in Web Appendix A. This states that if both the propensity score and outcome regression models are correctly specified and their posterior distributions contract at rates faster than $n^{-1/4}$, then our variance estimator is consistent for the true variance. These rates can be obtained in high-dimensional settings under sparsity conditions (Castillo, Schmidt-Hieber & Van der Vaart, 2015) or with nonparametric prior distributions under smoothness constraints (van der Vaart & van Zanten, 2008). This result, combined with the previous result about the conservative nature of our variance estimator, provides strong justification for using our variance estimation procedure. Additionally, in Web Appendix D we show our estimator is asymptotically normal under the same conditions as Theorem 2, which justifies the use of a normal approximation for inference. Although this relies on correct specification of both nuisance models, we see empirically in Section 4 that we obtain valid inference even when one model is misspecified.

4 | SIMULATION STUDIES

Here, we present the results of simulation studies aimed at evaluating the proposed doubly robust estimator and corresponding variance estimator in nonlinear and high-dimensional modeling scenarios. Throughout this section, we consider the following data-generating process:

$$\boldsymbol{X}_i \overset{\text{i.i.d}}{\sim} N(\boldsymbol{0}_p, \mathbf{I}_p), \qquad T_i|\boldsymbol{X}_i \sim \text{Bernoulli}(p_i), \qquad Y_i|T_i,\boldsymbol{X}_i \sim N(\mu_i, 1)$$

for $i = 1, 2, \ldots, n$ where $\mu_i = \tau(\boldsymbol{V}_i)T_i + E(Y(0)|\boldsymbol{X}_i)$ and $\boldsymbol{V}_i$ consists of the first 10 elements of $\boldsymbol{X}_i$. We assume that the true treatment effect, the parameter of interest, is a linear function of covariates:

$$\tau(\boldsymbol{V}_i) = 0.3 + 0.4V_{1i} - 0.2V_{2i} + 0.7V_{8i},$$

where V_{ki} denotes the k^{th} element of $\boldsymbol{V}_i$. We explore estimating nonlinear treatment effect functions in Section 4.3. Further, all results for our approaches in this section do not utilize sample splitting or cross-fitting, but a comparison of our approaches with and without sample splitting can be found in Web Appendix E.

4.1 | Nonlinear nuisance functions

We first set $p = 10$ so that $\boldsymbol{V}_i = \boldsymbol{X}_i$ for each i and generate data under the following two scenarios for the nuisance functions:

$$\textbf{Linear}: \quad p_i = \Phi(0.3X_{1i} - 0.3X_{2i} + 0.3X_{3i} - 0.3X_{4i})$$
$$E(Y(0)|\boldsymbol{X}_i) = 0.9X_{1i} - 0.6X_{3i} + 0.6X_{4i} + 0.7X_{6i}$$
$$\textbf{Nonlinear}: \quad p_i = \Phi(1(X_{1i} > 0) - \cos(X_{2i}) + 0.3|X_{3i}| - \sin(X_{4i}))$$
$$E(Y(0)|\boldsymbol{X}_i) = \cos(X_{1i}) + 1(X_{2i} > 1) - 0.05X_{3i}^3 + 0.1e^{X_{4i}} + \frac{1}{X_{6i}^2 + 1}$$

We fit the outcome model using (a) a Bayesian generalized linear model (GLM) as in Gelman et al. (2008) (`DR-Linear`); and (b) Bayesian additive regression trees (`DR-BART`). For both methods, the propensity score is fit with a Bayesian GLM, and inference is performed as in Section 3.1. In addition to our methods, we use causal forest (`CF`), BART applied simply to the outcome model (`BART`), debiased machine learning using both linear regression (`DML-Linear`) and random forest (`DML-RF`) for nuisance parameter estimation (Semenova & Chernozhukov, 2021), and Bayesian causal forest of Hahn et al. (2020) (`BCF`) as our competitors. However, since there is no method in the `BCF` package for estimating treatment effects at specific locations, `BCF` is only evaluated in Web Appendix F, where we examine $\tau(\boldsymbol{V}_i)$ at the observed data locations $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$.

Our goal is to estimate $\tau(\boldsymbol{V})$ at 100 randomly drawn points from the distribution of $\boldsymbol{V}$, which are shared by all simulated datasets to facilitate comparison across

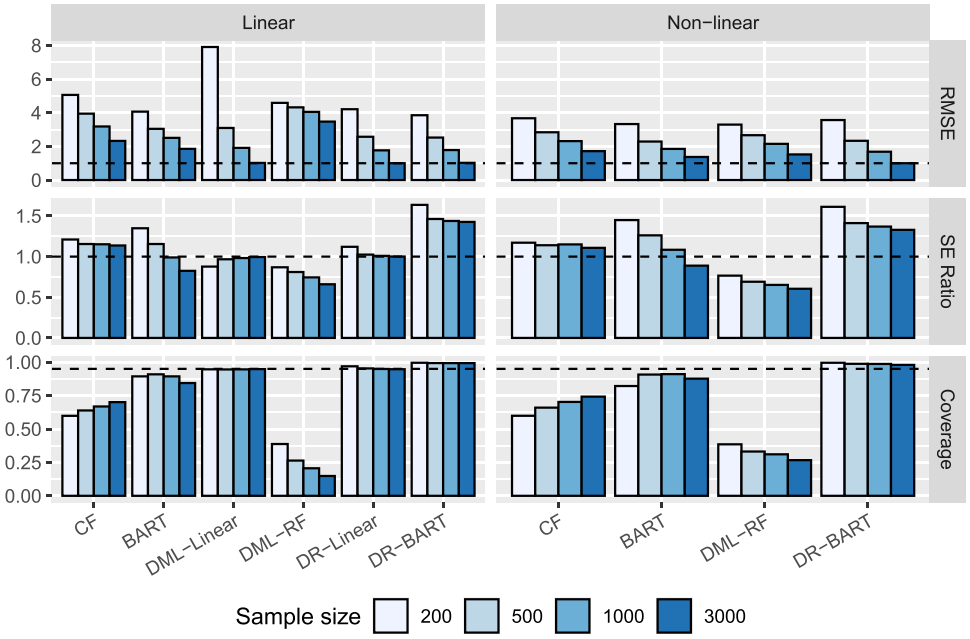

FIGURE 1 Results for the linear scenario (the first column) and the nonlinear scenario (the second column) for estimating the CATE at randomly chosen locations. The first row shows the scaled root mean-squared error, where values are divided by the smallest value within each scenario so that the best RMSE is 1. The second row shows the ratio of average estimated standard errors and Monte Carlo standard errors, and the last row shows the empirical coverage, where the dotted line represents the nominal 95% coverage. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

datasets. We compute empirical 95% interval coverage, root mean-squared error (RMSE) and the ratio of average estimated standard errors and Monte Carlo standard errors. Figure 1 shows the results of the simulations for both scenarios, where the values are averaged over the 100 locations for $\boldsymbol{V}$ and 500 simulations. We do not assess the linear methods (`DML-Linear` and `DR-Linear`) for the nonlinear scenario since those methods are not expected to perform well in this scenario.

The proposed methods (`DR-Linear` and `DR-BART`) generally have the lowest RMSE among all approaches considered. Regarding inference, the second row of Figure 1 suggests that our variance estimator is conservative as mentioned in Section 3.3. The `DR-Linear` approach has a standard error ratio approaching 1 as the sample size grows, which highlights the consistency of our variance estimator. `DR-BART` appears conservative regardless of sample size, though this is expected, because the contraction rate of BART is far slower than the $n^{-1/4}$ rates required for consistency (Ročková & van der Pas, 2020). Note that this stands in stark contrast to the debiased machine learning approaches, which have anti-conservative variance estimates and do not achieve nominal interval coverages. The `DML-Linear` approach has a consistent estimate of the standard error, though tends to underestimate it in small samples, while the `DML-RF` approach always underestimates the standard error. The last row shows that our confidence intervals achieve the nominal 95% coverage for both scenarios. It is notable that `DR-BART` is the only method that yields valid inference in all settings, and improves on standard `BART` in terms of RMSE and interval coverage. In short, our method succeeds in constructing valid confidence intervals with small RMSE regardless of the true nuisance functions. We see conservative inference when using BART for nuisance function estimation, but this can easily be reduced by using nonlinear models with faster contraction rates. While we utilized parametric models for the propensity score throughout, conservative inference would also be obtained if flexible propensity score models with slow contraction rates were incorporated.

We only considered performance at 100 randomly chosen locations, but we can also look at $\tau(\boldsymbol{V}_i)$ at the n observed data locations. We found similar results at the observed data locations, and therefore leave these results to Web Appendix F. One thing to note is that `BCF` outperforms `CF` at the observed data locations, but has a slightly larger MSE than the proposed estimators and is anti-conservative for certain sample sizes, though much closer to the nominal level than `CF`.

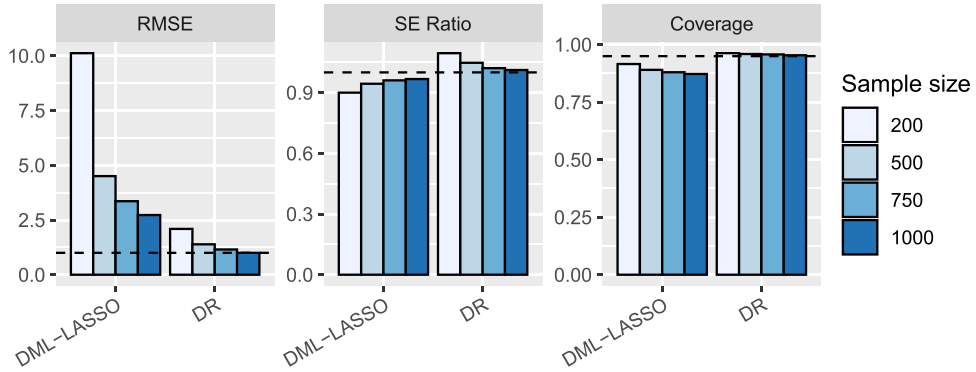

FIGURE 2 Results for the high-dimensional scenario. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

4.2 | High-dimensional nuisance functions

Now, we consider the case with the same linear nuisance functions in Section 4.1, but using a high-dimensional $\boldsymbol{X}$ and a low-dimensional $\boldsymbol{V}$. Specifically, set $n = 200, 500, 750$, and $1,000$, while $p = 2n$ is growing with the sample size. However, we are still interested in inferring $\tau(\boldsymbol{V})$ when $\boldsymbol{V}$ consists of the first 10 elements of $\boldsymbol{X}$. With high-dimensional $\boldsymbol{X}$, we compare our method using spike-and-slab priors to fit the nuisance functions (`DR`) with debiased machine learning using LASSO (`DML-LASSO`). The results for the randomly chosen locations are shown in Figure 2. Our method (`DR`) has a smaller RMSE with any sample size and achieves the nominal 95% coverage in all scenarios. Our variance estimator is slightly conservative for small-sample sizes, though converges to the truth as the sample size increases, while the `DML-LASSO` approach is anti-conservative at each sample size.

4.3 | Nonlinear $\tau(\boldsymbol{V}_i)$ case

Here, we explore empirically the performance of our approach for nonlinear $\tau(\boldsymbol{V})$ functions. In this case, we use a nonlinear parametric method such as natural cubic splines with a fixed degrees of freedom to regress the pseudo outcome on $\boldsymbol{V}$. We use the same linear nuisance functions simulation in Section 4.1 but replace the linear CATE function with a nonlinear function given by

$$\tau(\boldsymbol{V}_i) = 0.3 + 0.4\cos(V_{1i}) - 0.2V_{2i}^2 + 0.7|V_{8i}|,$$

and the results are shown in Figure 3. Both of our approaches, `DR-Linear` and `DR-BART`, are the only ones to achieve interval coverages that are at the nominal 95% level. `DR-Linear` achieves exactly nominal coverage, while the `DR-BART` approach is somewhat conservative as expected by our theoretical results, and seen previously in the earlier simulation results. In terms of RMSE, `CF` and `BART` have the lowest error when the sample size is small, while our approaches have the lowest error along with `BART` when the sample size is the largest. One might expect that `CF` and `BART` should have better coverage rates as they have standard error ratios above 1 and low RMSE, but as we see in the second row of Figure 3, these estimators tend to be biased and overly shrink heterogeneous treatment effect estimates toward an overall homogeneous treatment effect. The right panel of the second row shows the average estimated variability of the treatment effects, which for one particular dataset is defined to be $\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\tau}(\boldsymbol{V}_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}(\boldsymbol{V}_i)\right)^2$. We see that `CF` is substantially underestimating the degree of heterogeneity of the treatment effect, even at large-sample sizes, while our proposed estimators are able to capture this heterogeneity. Lastly, we also see that double machine learning estimators are either substantially biased (`DML-RF`) or have too much estimated heterogeneity (`DML-Linear`), which leads to the worsened performance.

5 | ANALYSIS OF ENVIRONMENTAL CHEMICALS

It is believed that a large portion of disease risk is attributable to differences in one's environment and the types of chemicals or pollutants they are exposed to on a daily basis (Patel & Ioannidis, 2014). Therefore, it is critical to understand the effects of environmental exposures on health, and whether these effects vary by subgroups of the population. To help address these questions, we will utilize data from The National Health and Nutrition Examination Survey (NHANES), which is a publicly

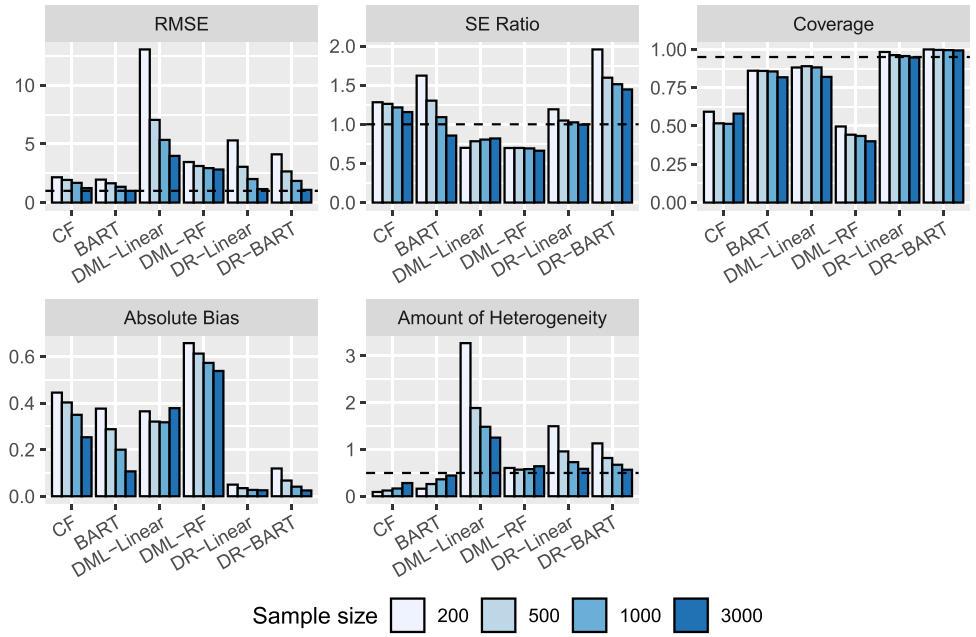

FIGURE 3 Results for the simulation with nonlinear $\tau(\boldsymbol{V})$. The left panel on the second row shows the average absolute bias of each estimator and the right panel shows the average estimated variability of the treatment effects, while the dotted line represents the variability of the true treatment effects. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

available data source provided by the United States Centers for Disease Control and Prevention (CDC). We will utilize data that was compiled and provided in Patel et al. (2016), which contains information from the 1999–2000, 2001–2002, 2003–2004, and 2005–2006 surveys. These data consist of demographic information, physical exam results, laboratory results, and answers from a questionnaire. We will focus on two distinct datasets provided by this database that examine the impacts of two separate environmental exposures: vitamin B intake, and dialkyl metabolite levels. Vitamin B is a nutrient that contributes to the overall well-being of an individual. Dialkyl metabolite levels can be used to estimate an individual's exposure to organophosphate pesticides, which are known to be toxic to humans. This leads to two distinct analyses on two separate populations from the NHANES data, and we provide details of each, along with results of our analyses below.

5.1 | The effect of vitamin B intake on triglycerides

We estimate the treatment effect of vitamin B intake on triglyceride levels, the main constituents of body fat, for 1,370 observations using the methods in Section 4.1. Since we do not address continuous treatments in this paper, we first transform the three continuous treatment variables (vitamin B12, serum folate, and red blood cell (RBC) folate levels) to a single binary treatment in the following manner: each observation is considered treated if at least two out of three treatment levels are greater than the average level of each treatment.

First, we use all variables in the second stage by setting $\boldsymbol{V} = \boldsymbol{X}$. Figure 4 shows the estimated $\tau(\boldsymbol{V}_i)$ for $i = 1, \ldots, n$, where the indices of the observations are sorted by estimates of each method. Note these are different from individual treatment effect estimates, and are simply estimates of the CATE evaluated at the observed covariate values $\boldsymbol{V}_i$. We see a big difference between the tree-based methods given by `CF` and `BCF`, and the remaining approaches. The tree-based approaches show very little variability in these treatment effect estimates, which suggests a homogeneous treatment effect that does not vary by $\boldsymbol{V}$. As discussed above in our simulations, we frequently saw the tree-based approaches overly shrinking treatment effect estimates toward homogeneity, and that appears to be occurring here as well. The remaining approaches, including our approaches, lead to far more variability in the estimated CATE indicating some level of variation in the treatment effect. The average treatment effect, which can be estimated by taking the sample average of the CATE, is negative for all approaches considered, which highlights an overall beneficial effect of vitamin B intake that reduces triglyceride levels.

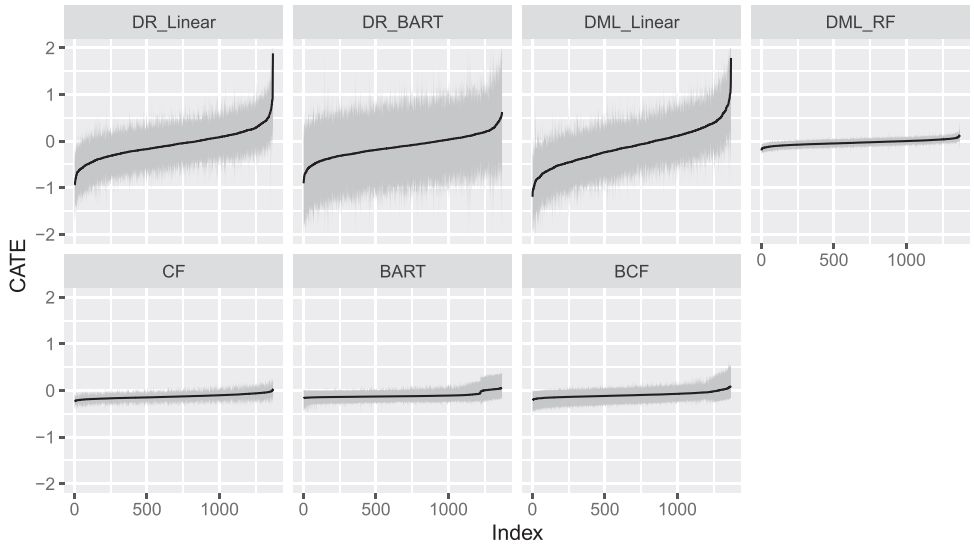

FIGURE 4 Comparing estimated conditional average treatment effects of vitamin B intake on triglycerides. The solid lines and gray areas represent the ordered CATE estimates and the 95% confidence intervals, respectively. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

5.2 | The effect of dialkyl metabolite levels on High-Density Lipoprotein (HDL) cholesterol

We now utilize our approach with high-dimensional models to estimate the treatment effect of dialkyl metabolite levels on HDL cholesterol, also known as "good" cholesterol, for 225 observations with 73 variables. Similar to the vitamin B intake analysis, we convert six continuous treatments to a single binary treatment. Due to the large number of variables in $\boldsymbol{X}$, we need to pre-specify a subset of variables $\boldsymbol{V}$ to investigate heterogeneity by. We select 20 variables that come from demographic information, a physical exam, or a questionnaire as variables to investigate heterogeneity by, since these are more interpretable than the remaining variables which come from laboratory tests.

We apply our method (`DR`) and debiased machine learning with LASSO models (`DML-LASSO`) to estimate the CATE, and Figure 5 shows the results. Both methods lead to a negative estimated average treatment effect indicating that higher levels of dialkyl metabolites decrease the level of HDL cholesterol. This is expected as dialkyl metabolite levels reflect exposure to toxic pesticides, and lower levels of HDL cholesterol can increase the risk for heart disease or stroke. In terms of heterogeneity, the proposed approach indicates less variability than the `DML-LASSO` approach as the estimated CATE values are roughly between −1.5 and 0.5 for our approach, while they are between −3 and 1 for `DML-LASSO`. Although `DML-LASSO` gives a larger degree of heterogeneity, this may not stem from true heterogeneity in the treatment effect, but rather from estimation uncertainty. In our high-dimensional

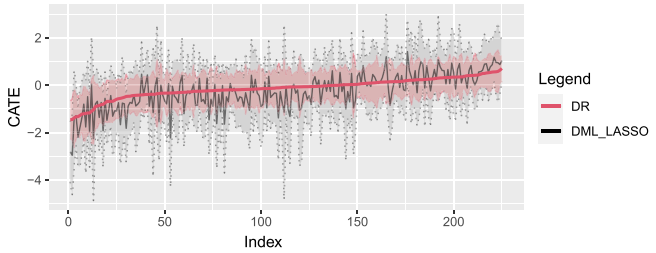

FIGURE 5 Comparing estimated conditional average treatment effect estimates of dialkyl metabolite levels on HDL cholesterol using 20 covariates. The solid lines and corresponding areas represent the ordered CATE estimates by `DR`, and the 95% confidence intervals, respectively. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

simulations of Section 4.2, we found that `DML-LASSO` yields a large RMSE compared with the proposed approach.

We need not investigate heterogeneity by all 20 variables in $\boldsymbol{V}$ jointly, but can rather investigate heterogeneity by one covariate at a time to see if any covariates are strongly associated with the treatment effect. This means we are now estimating $\tau_j(v) = E(Y(1) - Y(0)|V_j = v)$ for $j = 1, \ldots, p$. This estimand is interesting in its own right as it can provide intuition for any underlying mechanisms that are driving the treatment effect. Given that these are univariate functions, we use cubic splines with a fixed degrees of freedom for estimating the CATE. Figure 6 shows estimates of these univariate functions for the covariates that have the strongest association with the treatment effect. It appears that dialkyl metabolites have a particularly detrimental effect on the level of HDL

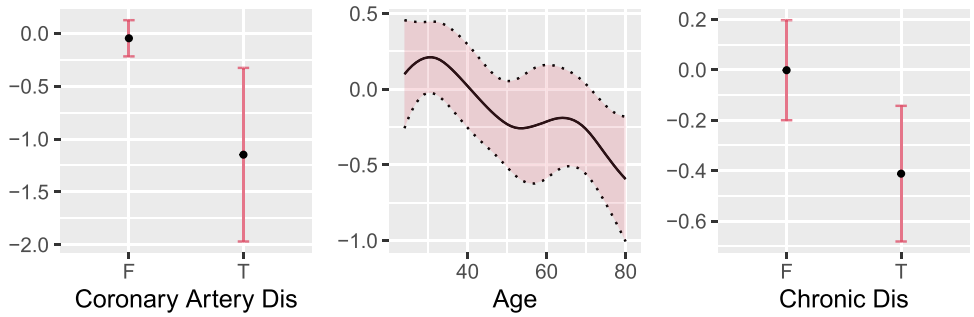

FIGURE 6 Estimated univariate CATE of dialkyl metabolite levels on the level of HDL cholesterol. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

cholesterol among subjects with coronary artery disease (CAD), chronic disease, or among older individuals. This leads to an overall understanding that the detrimental impact of dialkyl metabolites is more pronounced among those who are more vulnerable, and may not affect young, healthy individuals to the same degree.

6 | DISCUSSION

In this paper, we introduced a novel inferential procedure for doubly robust estimators of CATEs. We have shown that our variance estimator is consistent when both the propensity score and outcome regression models are correctly specified and contract at sufficiently fast rates, but is conservative in the more difficult settings of finite samples or model misspecification. We have seen empirically that this leads to improved performance in terms of interval coverage compared to existing state-of-the-art approaches, and provides valid inference in difficult high-dimensional or nonparametric modeling situations. The key difference in our approach is that we are able to account for parameter uncertainty by utilizing the posterior distribution of all unknown parameters. This source of uncertainty is difficult to account for in high-dimensional or nonparameteric settings, and is commonly ignored in asymptotic approximations for doubly robust estimators, which can lead to poor performance in finite samples. One might think that this can be solved by bootstrapping competing estimators, though in Web Appendix G, we illustrate that the bootstrap can lead to erratic behavior when trying to quantify uncertainty of existing estimators when either high-dimensional or nonparametric models are used.

There are a number of interesting extensions to be explored in the future. The first is to extend our methods to a continuous treatment, which will allow our method to be applied in a wider range of settings without having to dichotomize treatments as we did in the NHANES data analysis. One could potentially extend either the double machine learning framework (Knaus, 2020) or the two-stage procedure seen in Kennedy et al. (2017) for continuous treatments to allow for heterogeneity of the treatment effect while incorporating our proposed inferential procedure. It would also be interesting to develop theory combining our approach with nonparametric estimates of $\tau(\cdot)$ to reduce assumptions on the CATE. Lastly, we have implicitly assumed throughout that a researcher has a priori knowledge about which elements of $\boldsymbol{X}$ are of most interest and should be included in $\boldsymbol{V}$. While this is commonly assumed in models estimating treatment effect heterogeneity with a high-dimensional $\boldsymbol{X}$, it would be interesting to develop a data-driven approach to selecting the variables to include in $\boldsymbol{V}$ that still permits valid inference on the CATE.

ACKNOWLEDGMENTS

We would like to thank the associate editor and two anonymous reviewers who have substantially improved the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are openly available in Patel et al. (2016) and can be found at http://doi.org/10.1038/sdata.2016.96.

ORCID

Heejun Shin https://orcid.org/0000-0003-2733-6957
Joseph Antonelli https://orcid.org/0000-0001-7464-5766

REFERENCES

Abrevaya, J., Hsu, Y.-C. & Lieli, R.P. (2015) Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33, 485–505.

Antonelli, J. & Cefalu, M. (2020) Averaging causal estimators in high dimensions. *Journal of Causal Inference*, 8, 92–107.

Antonelli, J., Cefalu, M., Palmer, N. & Agniel, D. (2018) Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74, 1171–1179.

Antonelli, J., Papadogeorgou, G. & Dominici, F. (2022) Causal inference in high dimensions: a marriage between Bayesian modeling and good frequentist properties. *Biometrics*, 78, 100–114.

Antonelli, J., Parmigiani, G. & Dominici, F. (2019) High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Analysis*, 14, 805.

Athey, S. & Imbens, G. (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 7353–7360.

Avagyan, V. & Vansteelandt, S. (2021) High-dimensional inference for the average treatment effect under model misspecification using penalized bias-reduced double-robust estimation. *Biostatistics & Epidemiology*, 1–18. https://doi.org/10.1080/24709360.2021.1898730

Bang, H. & Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014) High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28, 29–50.

Benkeser, D., Carone, M., Laan, M. V.D. & Gilbert, P. (2017) Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104, 863–880.

Castillo, I., Schmidt-Hieber, J. & Van der Vaart, A. (2015) Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43, 1986–2018.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., et al. (2018) Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21, C1–C68.

Chipman, H.A., George, E.I. & McCulloch, R.E. (2012) BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4, 266–298.

Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. (2019) Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition 1. *Statistical Science*, 34, 43–68.

Dukes, O., Avagyan, V. & Vansteelandt, S. (2020) Doubly robust tests of exposure effects under high-dimensional confounding. *Biometrics*, 76, 1190–1200.

Dukes, O., Vansteelandt, S. & Whitney, D. (2021) On doubly robust inference for double machine learning. *arXiv preprint arXiv:2107.06124*.

Efron, B. & Tibshirani, R.J. (1994) *An introduction to the bootstrap*. CRC Press.

El Karoui, N. & Purdom, E. (2018) Can we trust the bootstrap in high-dimensions? The case of linear models. *The Journal of Machine Learning Research*, 19, 170–235.

Fan, Q., Hsu, Y.C., Lieli, R.P. & Zhang, Y. (2020) Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1), 313–327

Farrell, M.H. (2015) Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189, 1–23.

Gelman, A., Jakulin, A., Pittau, M.G. & Su, Y.S. (2008) A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2, 1360–1383.

Hahn, P.R., Murray, J.S. & Carvalho, C.M. (2020) Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15, 965–1056.

Hill, J.L. (2011) Bayesian Nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20, 217–240.

Kennedy, E.H. (2020) Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Kennedy, E.H., Ma, Z., McHugh, M.D. & Small, D.S. (2017) Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 1229–1245.

Knaus, M.C. (2020) Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25(3), 602–627.

Künzel, S.R., Sekhon, J.S., Bickel, P.J. & Yu, B. (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 4156–4165.

Lee, K., Bargagli-Stoffi, F.J. & Dominici, F. (2020) Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*.

Lee, K., Small, D.S. & Dominici, F. (2021) Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *Journal of the American Statistical Association*, *116*, 569–580.

Lee, S., Okui, R. & Whang, Y.-J. (2017) Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, 32, 1207–1225.

Little, R.J. & Rubin, D.B. (2000) Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21, 121–145.

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

Ning, Y., Sida, P. & Imai, K. (2020) Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107, 533–554.

Patel, C.J. & Ioannidis, J.P. (2014) Studying the elusive environment in large scale. *Jama*, 311, 2173–2174.

Patel, C.J., Pho, N., McDuffie, M., Easton-Marks, J., Kothari, C., Kohane, I.S. & Avillach, P. (2016) A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. *Scientific Data*, 3, 1–10.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N.H., Hastie, T. & Tibshirani, R. (2018) Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37, 1767–1787.

Robins, J.M. & Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90, 122–129.

Ročková, V. & van der Pas, S. (2020) Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48, 2108–2131.

Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096–1120.

Semenova, V. & Chernozhukov, V. (2021) Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24, 264–289.

Tan, Z. (2020a). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48, 811–837.

Tan, Z. (2020b). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107, 137–158.

Van der Laan, M.J. (2014) Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10, 29–57.

van der Vaart, A.W. & van Zanten, J.H. (2008) Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36, 1435–1463.

Wager, S. & Athey, S. (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113, 1228–1242.

Wang, T. & Rudin, C. (2022) Causal rule sets for identifying subgroups with enhanced treatment effects. *INFORMS Journal on Computing*, 34, 1626–1643.

SUPPORTING INFORMATION

Web Appendices A-G, referenced in Sections 3, 4, and 6, the NHANES data referenced in Section 5, and R codes to implement the proposed method are available with this paper at the Biometrics website on Wiley Online Library.

Supporting Information Data S1

How to cite this article: Shin, H. & Antonelli, J. (2023) Improved inference for doubly robust estimators of heterogeneous treatment effects. *Biometrics*, 79, 3140–3152. https://doi.org/10.1111/biom.13837