Author/s:
Ellul, S;Vansteelandt, S;Carlin, JB;Moreno-Betancur, M

Title:
Causal Machine Learning Methods and Use of Cross-Fitting in Settings With High-Dimensional Confounding

Date:
2025-09-01

Citation:
Ellul, S., Vansteelandt, S., Carlin, J. B. & Moreno-Betancur, M. (2025). Causal Machine Learning Methods and Use of Cross-Fitting in Settings With High-Dimensional Confounding. Statistics in Medicine, 44 (20-22), pp.e70272-. https://doi.org/10.1002/sim.70272.

Persistent Link:
https://hdl.handle.net/11343/362441

Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Causal Machine Learning Methods and Use of Cross-Fitting in Settings With High-Dimensional Confounding

Susan Ellul[1,2] | Stijn Vansteelandt[3] | John B. Carlin[1,2] | Margarita Moreno-Betancur[1,2]

[1]Murdoch Children's Research Institute, Parkville, Victoria, Australia | [2]Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia | [3]Department of Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

Correspondence: Susan Ellul (s.ellul@student.unimelb.edu.au)

Received: 17 October 2024 | **Revised:** 26 August 2025 | **Accepted:** 9 September 2025

Funding: This work was supported by the Australian National Health and Medical Research Council (NHMRC) Investigator Grant Emerging Leadership Level 2 (grant 2009572 awarded to M.M.-B.). S.E. is funded by an Australian Government Research Training Program Scholarship. Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

Keywords: augmented inverse probability weighting | causal inference | cross-fitting | doubly robust | high-dimensional confounding | targeted maximum likelihood estimation

ABSTRACT

Observational epidemiological studies commonly seek to estimate the causal effect of an exposure on an outcome. Adjustment for potential confounding bias in modern studies is challenging due to the presence of high-dimensional confounding, which occurs when there are many confounders relative to sample size or complex relationships between continuous confounders and exposure and outcome. Doubly robust methods such as Augmented Inverse Probability Weighting (AIPW) and Targeted Maximum Likelihood Estimation (TMLE) have the potential to address these challenges, using data-adaptive approaches and cross-fitting, but despite recent advances, limited evaluation and guidance are available on their implementation in realistic settings where high-dimensional confounding is present. Motivated by an early-life cohort study, we conducted an extensive simulation study to compare the relative performance of AIPW and TMLE using data-adaptive approaches for estimating the average causal effect (ACE). We evaluated the benefits of using cross-fitting with a varying number of folds, as well as the impact of using a reduced versus full (larger, more diverse) library in the Super Learner ensemble learning approach used for implementation. We found that AIPW and TMLE performed similarly in most cases for estimating the ACE, but TMLE was more stable. Cross-fitting improved the performance of both methods, but was more important for variance estimation and coverage than for point estimates, with the number of folds a less important consideration. Using a full Super Learner library was important to reduce bias and variance in complex scenarios typical of modern health research studies.

1 | Introduction

Estimating the causal effect of an intervention or exposure on an outcome in an observational study requires accounting for multiple sources of confounding. Modern-day studies are data-intensive, often collecting a large amount of data, including background, demographic, and biological factors, which in principle should allow for stronger inferences by enabling more extensive control of confounding. However, exploiting these opportunities requires addressing what we call the problem of high-dimensional confounding, which arises when there is a large number of confounders relative to sample size, or a few

Abbreviations: ACE, average causal effect; AIPW, augmented inverse probability weighting; CML, causal machine learning; TMLE, targeted maximum likelihood estimation.

continuous confounders that may have a complex relationship with the exposure and/or outcome of interest. Analytical adjustment for confounding in these settings can be challenging.

Commonly used methods for the estimation of causal effects, such as g-computation or inverse probability weighting (IPW) are singly robust (SR), meaning they rely upon a single model being correctly specified [1, 2]. G-computation involves the fitting of an outcome model (conditional expectation of the outcome given the exposure and confounders), and IPW an exposure model (probability of exposure given confounders). In the context of high-dimensional confounding, such methods are at risk of misspecification bias because the single model is commonly based on simplistic parametric functions [1, 3]. Specifically, in settings with a large number of confounders (or higher-order terms) and limited sample size, it can be challenging to apply the methods without some form of variable selection or model simplification, and attempting to do this naively can impact the validity of inferences [4, 5]. Indeed, many studies reduce to a restricted, manageable adjustment set that may not be sufficient to control confounding, resulting in bias. Additionally, the failure to capture complex nonlinear relationships between variables can potentially lead to further bias [2, 6–8].

To overcome these concerns, it is tempting to consider incorporating data-adaptive approaches to flexibly fit the single model required for SR methods. We define data-adaptive approaches as those where the modeling approach or algorithm used to fit the model is capable of learning from the data, in the sense that the full functional form is not fixed but can adapt. Data-adaptive approaches include machine learning (ML) approaches, and the terms are frequently used interchangeably [9] (p. 44), although data-adaptive can also refer to parametric variable selection approaches like stepwise regression, an approach that we do not consider here. Data-adaptive approaches could potentially provide protection against misspecification bias as well as a way of avoiding inappropriate a priori variable selection in high-dimensional confounding settings. However, data-adaptive estimators typically have non-standard asymptotic behavior, meaning that they are often non-normal, and their standard errors do not converge to zero at a rate of $1/\sqrt{n}$. As a result, the use of data-adaptive approaches with SR methods is problematic, with this behavior leading to bias in point estimates and with no valid approach to obtain standard errors [1, 10–12]. For example, for g-computation with data-adaptive approaches, the non-parametric bootstrap has been shown to be invalid for variance estimation [13]. In general, challenges in the construction of CIs [14], and underestimated variance and under-coverage have been reported, signaling that misleading inference is a concern [1, 12].

As promising alternatives to SR methods, Augmented Inverse Probability Weighting (AIPW) [15–17] and Targeted Maximum Likelihood Estimation (TMLE) [9, 18, 19] involve the fitting of both an outcome and an exposure model. AIPW and TMLE provide consistent estimation if at least one of the two models is consistently estimated, which is why they are referred to as doubly robust (DR). These methods also achieve optimal semi-parametric efficiency if both models are consistently estimated [20, 21]. Moreover, in contrast to SR methods, DR methods can validly incorporate the use of data-adaptive approaches to fit both models under some conditions. In particular, they are partially insulated against the slow convergence rates that affect data-adaptive estimators of the exposure and outcome models, provided that both are consistent. Consistency of both estimators is needed to ensure that AIPW and TMLE are root-n consistent when using data-adaptive approaches, meaning the double robustness property holds for consistency but not for the stronger requirement of root-n-consistency (and this in particular for the validity of standard errors). Cross-fitting (CF), whereby the outcome and exposure models are fit in subsets (folds) of the data and the causal effect estimate is obtained from the remaining data, has been proposed to improve the estimation of standard errors and for valid inference in the context of DR methods [22, 23].

However, there is limited evaluation, comparison, and guidance on the implementation of what we will henceforth refer to as "causal machine learning" (CML) (AIPW and TMLE with data-adaptive approaches) with CF in realistic settings encountered in modern observational studies with high-dimensional confounding [24]. Most methodological studies evaluate and compare the methods with and without CF in less realistic settings, with very few or binary confounders only [1, 3]. Empirical studies applying these methods often have large sample sizes, but many real-world studies have sample size limitations. In general, empirical studies applying these methods have been limited to settings where high-dimensional confounding has not been considered, when perhaps it should have been [18, 25–27]. To our knowledge, no studies have comprehensively evaluated and compared AIPW and TMLE, with and without CF, whilst considering varying sample sizes in the high-dimensional setting. In addition, there are limited studies that have evaluated how the number of folds used in CF may affect the performance of AIPW and/or TMLE [28].

Here, motivated by a real-world case study, we conduct an extensive simulation study to address these gaps in the context of estimating the average causal effect (ACE), to provide practical guidance on the application of CML methods in realistic settings. The manuscript is organized as follows. Section 2 introduces the motivating case study. Section 3 outlines relevant notation and assumptions, and Section 4 provides details on AIPW and TMLE for the estimation of causal effects. Section 5 outlines key considerations for the implementation and evaluation of DR methods. Sections 6 and 7 describe the simulation study design and results. In Section 8, to illustrate, we apply the methods to the case study, following which, in Section 9, we discuss our findings and their implications.

2 | Overview of the Motivating Case Study

The motivating example draws on data from the Barwon Infant Study (BIS), a birth cohort study with antenatal recruitment conducted in the south-east of Australia [29]. Infants included in the study were reviewed at multiple time points in early life (including at birth, 12 months, and 4 years of age). Further details regarding eligibility, recruitment criteria, and measures obtained for BIS are provided elsewhere [29].

One focus area for research in BIS is the developmental origins of cardiovascular disease (CVD). There is growing evidence for a role of inflammation in CVD [30, 31]. Chiesa et al. [32] found that inflammation (as measured by a biomarker called Glycoprotein acetyls (GlycA) obtained directly from blood samples) was associated with adverse cardiovascular profiles in adolescence and also predicted future risk. In adults, pulse wave velocity (PWV), a measure of arterial stiffness (higher PWV can indicate greater arterial stiffness), has been shown to predict CVD events [33, 34]. It is therefore of interest to examine the effect of early life inflammation on PWV at a later age in childhood (post infancy). The motivating example focuses specifically on the following research question: *What is the effect of early life inflammation, as measured by GlycA in 1-year-old infants, on PWV at 4 years of age?* We standardize the outcome (PWV) in the BIS data and consider GlycA as a dichotomous exposure, indicating either high or low inflammation. We dichotomize the exposure to make results relevant for the majority of applications that currently focus on binary exposures. The literature has not established a definitive cutoff for GlycA that would be indicative of high inflammation in infants, so we examined the impact of having inflammation in the top quartile of the distribution, dichotomizing GlycA at the 75th percentile.

To address the proposed causal question, one must consider the potential confounding role of background factors (e.g., demographic, environmental, familial, and perinatal) and other metabolomic factors, as depicted in the directed acyclic graph (DAG) of Figure 1 developed using subject-matter expertise. In BIS, metabolomic measures at 1 year of age and PWV at 4 years of age were only obtained for a subset of the participants by design. Details on missing data are provided in Table S1 of Section 1 of the Supporting Information. Only participants with complete data (no missing data on any of the relevant variables, $n = 252$) were included in the case study, as evaluation of challenges associated with the presence of missing data was not considered here (see Section 9). Characteristics of the infants included are outlined in Table 1. Here, we have clearly identified a problem of high-dimensional confounding because we have a sample size of 252 and up to 87 potential confounders, with 78 of these being continuous confounders.

3 | Notation and Assumptions

We consider a binary exposure indicator X, coded 1 for exposed (in the motivating example, 1 = high GlycA), and 0 for the unexposed (0 = low GlycA), continuous outcome Y (in the example, Y is the standardized PWV at 4 years of age) and a vector of confounders, W (based on subject-matter expertise). Let $Y^{X=x}$ be the potential outcome under exposure x.

3.1 | Causal Estimand

We focus on the average causal effect (ACE), defined as $E\left[Y^{X=1}\right] - E\left[Y^{X=0}\right]$, interpreted as the difference in average potential outcome in the population of interest when (a) everyone is exposed versus when (b) everyone is unexposed [35, 36]. In the motivating example, this is the difference in average PWV that would be seen at 4 years of age if everyone in the target population were set to have high versus low GlycA at 1 year of age.

3.2 | Identification

In the absence of missing data, under the assumptions of exchangeability ($Y^{X=x} \perp\!\!\!\perp X \mid W$ for $x = 0, 1$), consistency ($Y^{X=x} = Y$ when $X = x$ and $x = 0, 1$), and positivity ($P[P(X = x|W) > 0] = 1$ for $x = 0, 1$), we can identify the ACE from observable data as $E[E(Y|X = 1, W) - E(Y|X = 0, W)]$

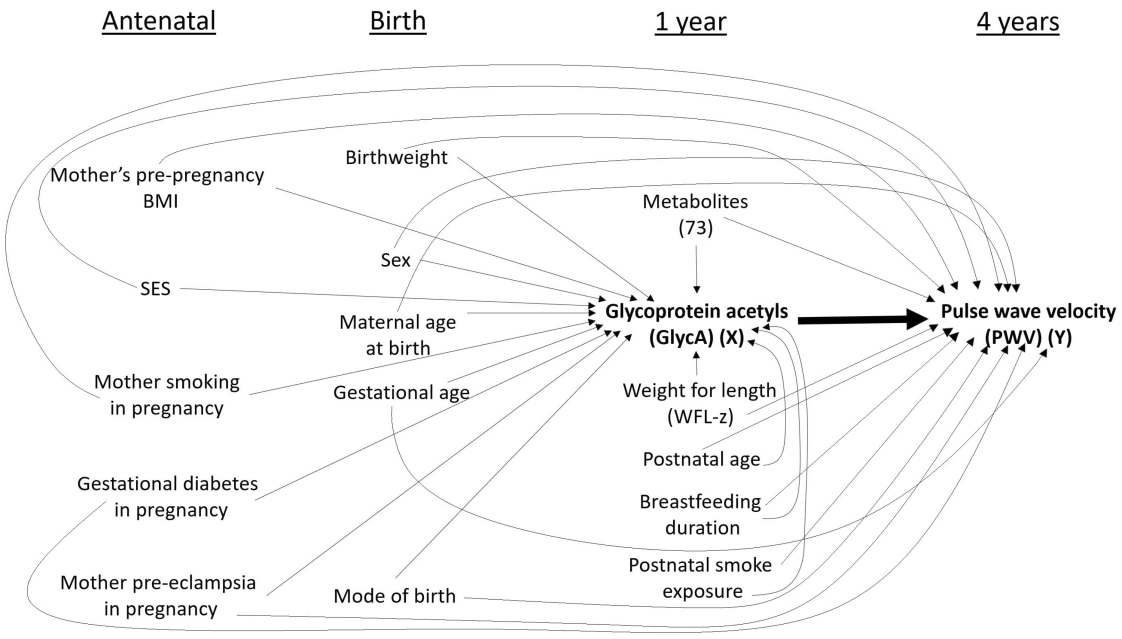

FIGURE 1 | Directed Acyclic Graph (BIS DAG) for the Barwon Infant Study motivating example, and as used for data-generation in the simulation study.

TABLE 1 | Characteristics of participants in the Barwon Infant Study (BIS) analytic sample used in the case study.

Characteristic	Inflammation[a]		
	Low, $N = 189$[b]	High, $N = 63$[b]	Overall, $N = 252$[b]
Pre-pregnancy BMI (kg/m^2)	24.5 [21.7, 27.8]	23.5 [21.3, 28.4]	24.3 [21.7, 27.9]
Socio-Economic Indexes for Areas (SEIFA)			
Low	65 (34%)	19 (30%)	84 (33%)
Med	63 (33%)	21 (33%)	84 (33%)
High	61 (32%)	23 (37%)	84 (33%)
Mother smoking in pregnancy	22 (12%)	10 (16%)	32 (13%)
Gestational diabetes in pregnancy	8 (4.2%)	3 (4.8%)	11 (4.4%)
Pre-eclampsia in pregnancy	7 (3.7%)	2 (3.2%)	9 (3.6%)
Birthweight (grams)	3531 (521)	3442 (581)	3508 (537)
Infant sex			
Female	82 (43%)	37 (59%)	119 (47%)
Male	107 (57%)	26 (41%)	133 (53%)
Maternal age at birth (years)	32.1 (4.3)	32.8 (3.9)	32.3 (4.2)
Gestational age at birth			
32–36 completed weeks	9 (4.8%)	4 (6.3%)	13 (5.2%)
37–42 completed weeks	180 (95%)	59 (94%)	239 (95%)
Mode of birth			
Caesarean	75 (40%)	20 (32%)	95 (38%)
Vaginal	114 (60%)	43 (68%)	157 (62%)
Weight-for-length z-score at 12 months	0.72 (1.09)	0.75 (1.04)	0.72 (1.08)
Age at 12-month measures (months)	12.93 (0.80)	12.97 (0.75)	12.94 (0.79)
Breastfeeding duration (exclusive weeks)[c]	8 [1, 22]	4 [0, 24]	7 [0, 22]
Postnatal smoke exposure	29 (15%)	10 (16%)	39 (15%)

[a]GlycA at 1 year of age, dichotomized using the 75th percentile of the observed distribution as the threshold.
[b]Mean (SD), Median [IQR] or Frequency (%) as appropriate.
[c]Number of weeks that infant was exclusively breastfed (i.e., no supplementary feeding).

which is the g-formula [2, 37]. It is worth noting that some of these assumptions are debatable in the motivating example, particularly the consistency assumption given the lack of a well-defined intervention. However, we assume they hold for the remainder of the manuscript given our focus is on examining the performance of estimators under those conditions (also see Section 9).

4 | Doubly Robust Methods

Under the assumptions outlined in Section 3, the ACE can be estimated using DR methods. We focus on two DR methods, AIPW and TMLE. For both of these methods, models are fitted for the conditional expectation of the outcome Y given the exposure X and confounders W, $E[Y|X,W]$ (the outcome model) and for the propensity score, $P(X=1|W)$ (the exposure model). The outcome and exposure models are often referred to as *nuisance* models because they are not of intrinsic interest but instead are used within AIPW and TMLE to estimate the target parameter [38].

DR estimators for the ACE are obtained by determining and then utilizing an efficient influence function (EIF), which has a unique form that is determined by the target parameter of interest [39]. DR estimators are constructed in a manner that allows them to attempt to correct for the bias (termed *plug-in bias*) that is induced when using g-computation with the data-adaptive approaches. Further details regarding the derivation and interpretation of the EIF, as well as plug-in bias, can be found elsewhere [40].

4.1 | Augmented Inverse Probability Weighting (AIPW)

AIPW is often referred to as a one-step correction or one-step estimation approach [41] and is based on directly subtracting an estimate of the bias term from the g-computation estimator. Based on the fitted outcome model, predicted values, $\widehat{E}_1(w) = \widehat{E}[Y|X=1, W=w]$ and $\widehat{E}_0(w) = \widehat{E}[Y|X=0, W=w]$ are obtained for each record, for $X=1$ and $X=0$ respectively. The estimate of the ACE is then calculated as

$$\widehat{\psi}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{E}_1(w_i) - \widehat{E}_0(w_i) + X_i\frac{\left(Y_i - \widehat{E}_1(w_i)\right)}{\widehat{P}(w_i)} - (1-X_i)\frac{\left(Y_i - \widehat{E}_0(w_i)\right)}{1-\widehat{P}(w_i)}\right] \quad (1)$$

where $\widehat{P}(w) = \widehat{P}(X=1|W=w)$ is an estimate of the propensity score based on the fitted exposure model. The estimated variance of the AIPW estimator is obtained as follows: [22, 42]

$$\widehat{var}(\widehat{\psi}_{AIPW}) = \frac{1}{n-1}\sum_{i=1}^{n}\left[\widehat{E}_1(w_i) - \widehat{E}_0(w_i) + X_i\frac{\left(Y_i - \widehat{E}_1(w_i)\right)}{\widehat{P}(w_i)} - (1-X_i)\frac{\left(Y_i - \widehat{E}_0(w_i)\right)}{1-\widehat{P}(w_i)} - \widehat{\psi}_{AIPW}\right]^2 \quad (2)$$

4.2 | Targeted Maximum Likelihood Estimation

In contrast to AIPW, after initial estimation of the outcome and exposure models, TMLE has a targeting or updating step, again with the purpose of correcting the bias of the g-computation estimator. Unlike AIPW, TMLE respects the bounds of the estimand's parameter space [23]. TMLE can be implemented using a procedure already well described in the literature (e.g., Schuler and Rose [2] and Luque-Fernandez et al. [20]), whereby the fitted exposure score model is used to generate a "clever" covariate. Using this, *targeted* or updated predicted outcomes $\widehat{E}_1^*(w)$ and $\widehat{E}_0^*(w)$ are constructed for $X=1$ and $X=0$, respectively, for each record, and these are used to estimate the ACE, with the estimator defined as

$$\hat{\psi}_{\text{TMLE}} = \frac{1}{n}\left[\sum_{i=1}^{n}\hat{E}_1^*(w_i) - \sum_{i=1}^{n}\hat{E}_0^*(w_i)\right] \quad (3)$$

The variance for TMLE is obtained similarly to AIPW by using (2), with $\hat{\psi}_{TMLE}$ in place of $\hat{\psi}_{AIPW}$, and with $\hat{E}_1^*(W)$ and $\hat{E}_0^*(W)$ in place of $\hat{E}_0(W)$ and $\hat{E}_0(W)$, respectively.

5 | Data-Adaptive Estimation of Nuisance Functions

When applying and evaluating the DR methods, there are key implementation considerations. In this section, we briefly outline key considerations that are explored in this paper.

5.1 | Data-Adaptive Approaches

Data-adaptive approaches can be used within the DR methods to estimate each of the nuisance functions, with the predictive performance of the approaches being the key criterion of interest. Therefore, nuisance function estimation can be viewed as a prediction modeling problem. One can consider parametric data-adaptive approaches, such as penalized regression methods (e.g., least absolute shrinkage and selection operator (LASSO) ([43, 44], p. 64–65)) wherein coefficients can be shrunk toward zero (or in some cases, set to zero), and are controlled or influenced by decision rules of the method. One can also consider non-parametric data-adaptive approaches, which include tree-based (e.g., random forest (RF) [45]) and non-tree-based (e.g., neural networks ([46], p. 141–145)), support vector machines ([44], p. 337–389) methods, and often allow greater flexibility than the parametric approaches. In this work, we consider both parametric and non-parametric data-adaptive approaches.

5.1.1 | Ensemble Learning

In seeking to improve predictive performance, it is common to use an ensemble learning approach ([47], p. 15), wherein multiple data-adaptive approaches, termed learners, are considered simultaneously. The group of learners considered for the prediction problem is called a candidate library, and the learners within are used to solve the same prediction problem, but predictions from the learners are combined to arrive at a solution with strengthened predictive ability.

In this work, we use the Super Learner (SL), a widely used ensembling approach in the context of DR estimation of causal effects [2, 3, 18, 24–27].

We consider two libraries of differing diversity containing non-adaptive parametric approaches (e.g., GLMs) as well as data-adaptive parametric and non-parametric approaches. Further details are provided in Section 6.3.

5.2 | Cross-Fitting

Cross-fitting (CF) has been proposed as an approach to help overcome issues that may arise from overfitting when using CML approaches [3, 39]. In this section, we outline the form of CF [11, 22, 25, 48, 49] that we consider in this work.

CF (or K-fold CF) is where the sample of size n is split randomly into $K \geq 2$ parts (folds) of roughly equal size (n/K). The nuisance models are fitted on all but one fold (the complement), and the remaining fold is used to get predictions from these models. The process is repeated K times, so that each fold $k = 1, 2, \ldots, K$ is used to obtain predictions once (by rotation). The predictions are then used within the given DR method to obtain a final estimate of the causal effect. In this study, when CF was applied, standard errors were calculated from the EIF (refer to Section 2 of the Supporting Information for implementation details) [11]. For TMLE with CF, targeting was performed in the remaining data that was not used to fit the nuisance models, an approach consistent with van der Laan and Rose [9], Motoya et al. [50] and Balzer and Westling [12], although we note that there are alternative implementations that retain the same theoretical properties [51].

6 | Simulation Study: Design and Methods

6.1 | Aim

We aimed to compare the performance of AIPW and TMLE using SL with different choices of library (reduced vs. full) and different implementations of CF (no CF and with 2, 5, and 10 folds) in realistic scenarios, largely informed by the BIS case study. We considered a range of scenarios, varying data generating mechanism complexity, confounder set size ($p = 14$ or 87), and sample size ($n = 200$, 500, 1000, and 2000). The *small* confounder set was modeled on demographic and background confounders in the BIS study, and the *large* set additionally considered metabolites [see Section 3 of the Supporting Information (Tables S2–S4) for more details].

6.2 | Data-Generating Mechanisms

The variables (confounders, exposure, and outcome) were generated in the order specified in Section 4 of the Supporting Information (Tables S5–S7). Parametric regression models were used to simulate variables for all five data-generating mechanisms, three of which involved the small confounder set (*simple-1*, *complex-1a*, and *complex-1b*), and two of which involved the large confounder set (*simple-2*, *complex-2*), detailed below. Parameters for the data-generation models were, unless otherwise stated, obtained by fitting analogous models to the BIS data.

Under the simple data-generating mechanisms (*simple-1* and *simple-2*), we considered exposure and outcome regression models with main effects only and a linear association for the continuous covariates. Complex data-generating mechanisms considered exposure and outcome models with main effects and two-way interactions (confounder-confounder and exposure-confounder as appropriate), with linear and non-linear associations for the continuous covariates. For the small confounder set, we used two complex scenarios, with coefficient values for interaction terms two times larger than observed in the BIS data (*complex-1a*) and another with coefficients for the interaction terms four

times larger (*complex-1b*) than observed in the BIS data. For the large confounder set, we considered one complex scenario with coefficient values as observed in the BIS data (*complex-2*). For full details regarding the data-generating mechanisms, refer to Section 4 of the Supporting Information (Tables S6–S11).

We simulated 2000 datasets for each data generation mechanism and sample size, with this number of simulation replications determined based on a requirement that the Monte Carlo SE for a coverage of 95% be no greater than 0.5%. Overall, 20 scenarios were considered (four sample sizes for five data-generating mechanisms). When generating the outcome in each scenario, we set the main effect to be of a size determined such that the null hypothesis of zero average causal effect was formally rejected (using the threshold of $p < 0.05$) in approximately 80% of the simulated datasets for the given sample size. We chose to keep the power constant to ensure that coverage probabilities, our main metric of interest, were comparable across sample size scenarios. The intercepts in all outcome models were modified so that the mean of Y remained at 0. Further details on the parameters used in the generation of confounders, exposure, and outcome are provided in Section 4 of the Supporting Information (Tables S12–S23).

6.3 | Estimand and Methods Compared

We estimated the average causal effect (ACE) of X on Y via AIPW and TMLE with no CF and with CF, using varying folds (2, 5 and 10). SL was used with 10-fold cross-validation for outcome and exposure prediction modeling. Reduced (less flexible) and full (more flexible) candidate libraries for SL were considered using default hyperparameters for each learner. The reduced library was chosen to include less flexible approaches. Specifically, it included the following non-adaptive and data-adaptive parametric approaches: GLM, GLM with pairwise interactions, Bayesian GLM, generalized additive modeling and GLM with Lasso/Elastic net regularization. The full library was comprised of the approaches in the reduced library and an additional set of more flexible, non-parametric data-adaptive approaches, including learners more likely to result in biased standard errors without the use of CF. Further details on the SL candidate libraries are provided in Section 5 of the Supporting Information (Table S24). To deal with extreme inverse probability weights, truncation of propensity score predictions at the 5th and 95th percentiles was undertaken.

6.4 | Performance Measures

For each scenario, to evaluate performance, we calculated the bias (the difference between the average of the ACE estimates across the 2000 simulations and the true value of the ACE), the relative bias (bias divided by the true ACE, as a percent) as well as the empirical standard error, average model-based standard error (model SE), and relative error of the model-based SE compared to the empirical SE (%). We also estimated the coverage probability of the 95% Wald CI. To calculate performance measures, the true values of the ACE used in each scenario were obtained empirically in a single very large simulated dataset [full details and true values used are provided in Section 6 of the Supporting Information (Table S25)]. We also estimated the Monte Carlo Standard Errors (MCSEs) for each performance measure. Morris et al. [52] provide more detail on the performance measures and how they are obtained.

7 | Simulation Study: Results

Our presentation and assessment of performance focus on the reporting of bias in point estimates (relative bias), the empirical SE, relative error in the model-based SE, and coverage probability. In general, patterns for the *complex-1b* scenario were similar to complex-1a. Hence, Figures 2–5 show performance measures for AIPW and TMLE when the reduced and full libraries were applied in SL across sample sizes, for the small confounder set (*simple-1* and *complex-1a*) and the large confounder set (*simple-2* and *complex-2*). Tables S26–S34 in Section 7 of the Supporting Information contain results for both the reduced and full libraries across all scenarios, for the comprehensive panel of performance measures, including associated Monte Carlo standard errors. This section mainly describes the results, and a fuller discussion of the findings is provided in Section 9.

7.1 | Method Comparison (AIPW vs. TMLE)

In general, both DR methods performed similarly across the key performance measures, except that AIPW failed to produce sensible results for a few datasets when CF was applied (2–12 datasets across scenarios). In these settings, the standard error was > 10 times the median standard error or the absolute value of the point estimate was > 5 times the absolute value of the median point estimate. These results were omitted from the figures, with details provided in Figures 2–5 and in the Supporting Information (Section 7, Table S26).

Relative biases in point estimates were very similar between methods (Figure 2). A small number of exceptions were noted for the variance (empirical SE) and coverage probability (Figures 3 and 5). For example, at the smallest sample size of 200 with *simple-1* and *simple-2*, the empirical SE appeared larger for AIPW than TMLE, a difference that appeared to dissipate when larger sample sizes were considered.

The relative error in model-SE was the indicator for which the most differences between methods were observed. As mentioned above, AIPW results obtained on a small number of datasets were not sensible when CF was applied, and the relative error in model-SE was large with the inclusion of these results. For example, for *simple-1* with a sample size of 1000, full library in SL, and 5-fold CF, the relative error in model-based SE was 87.5% and 4.5% prior to and after the exclusion of 2 datasets, respectively. The pattern of method differences for the relative error in model-SE was not clear, and those observed may potentially be due to the instability of AIPW.

7.2 | Impact of Using Cross-Fitting

In general, relative bias appeared to be of similar magnitude whether CF was applied or not (Figure 2). Empirical SE was

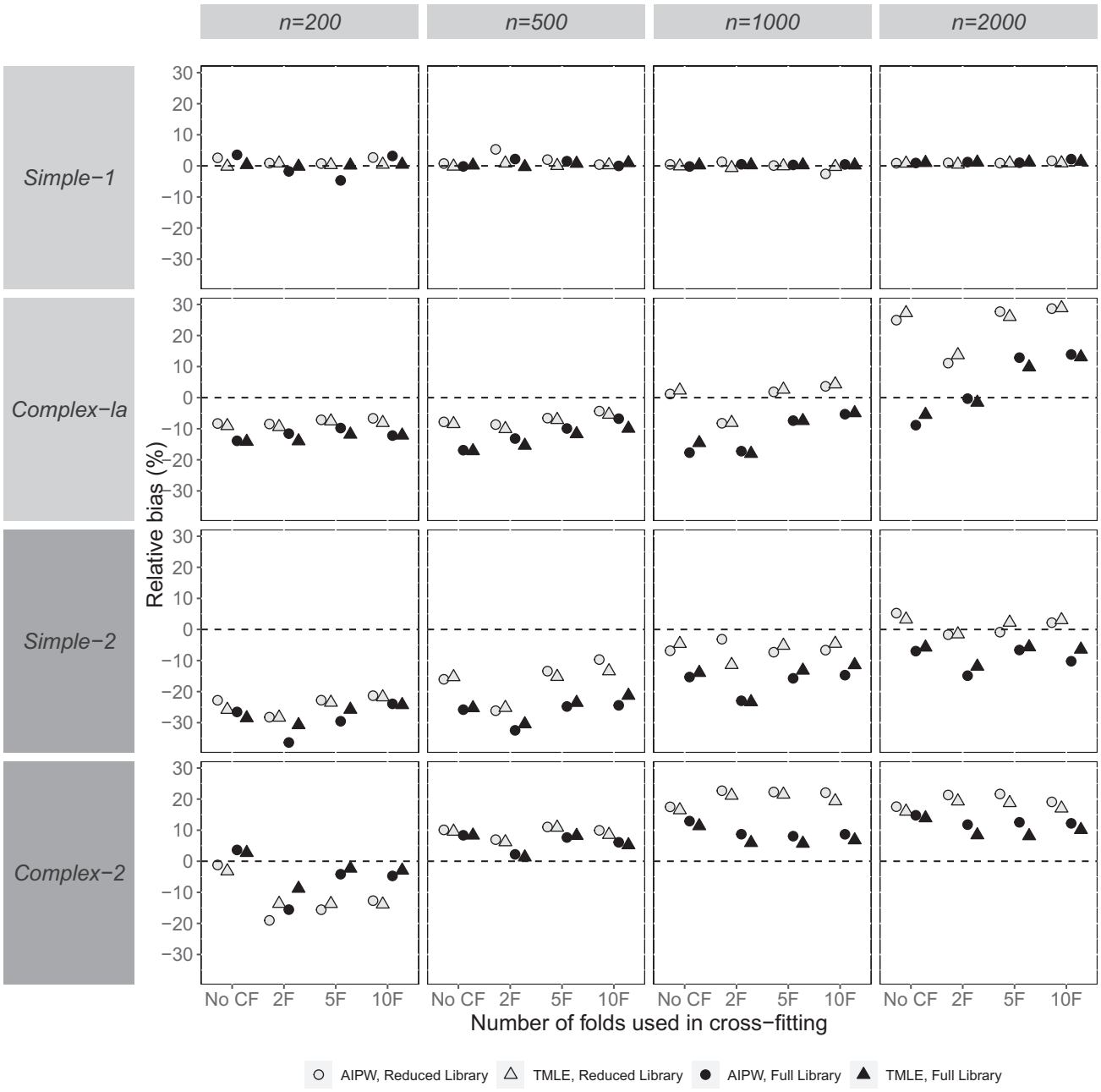

FIGURE 2 | Simulation study results for the relative bias of point estimates (%) for AIPW and TMLE with varying use of cross-fitting, by scenario (data generating mechanism and sample size). Section 5 of the Supporting Information (Table S25) provides the true value of the ACE used for each scenario. Results from a small number of datasets were excluded for AIPW (but not TMLE) (refer to Sections 7.1 and 5 of the Supporting Information for more details).

observed to be similar with or without CF, except for *complex-2* at smaller sample sizes ($n = 200$, 500), where it was larger with CF than without (Figure 3). In general, the use of CF resulted in lower relative error in model-SE than without CF (Figure 4). The use of CF appeared to improve coverage compared to not using CF, resulting in coverage probabilities that were closer to nominal (Figure 5).

Lastly, when CF was applied in scenarios with a large confounder set (*simple-2* and *complex-2*), the relative error in model-SE appeared to grow with increasing sample size considered, which resulted in lower coverage (further from nominal) (Figures 4 and 5). The same pattern was not observed for scenarios that considered the small confounder set (*simple-1* and *complex-1a*), or when CF was not applied. See Section 9 for discussion of these findings.

7.3 | Impact of Varying the Number of Folds Within Cross-Fitting

In general, method performance was similar regardless of the number of folds used within CF (Figures 2–5). An exception was noted at larger sample sizes (e.g., $n = 2000$) with *complex-1a*,

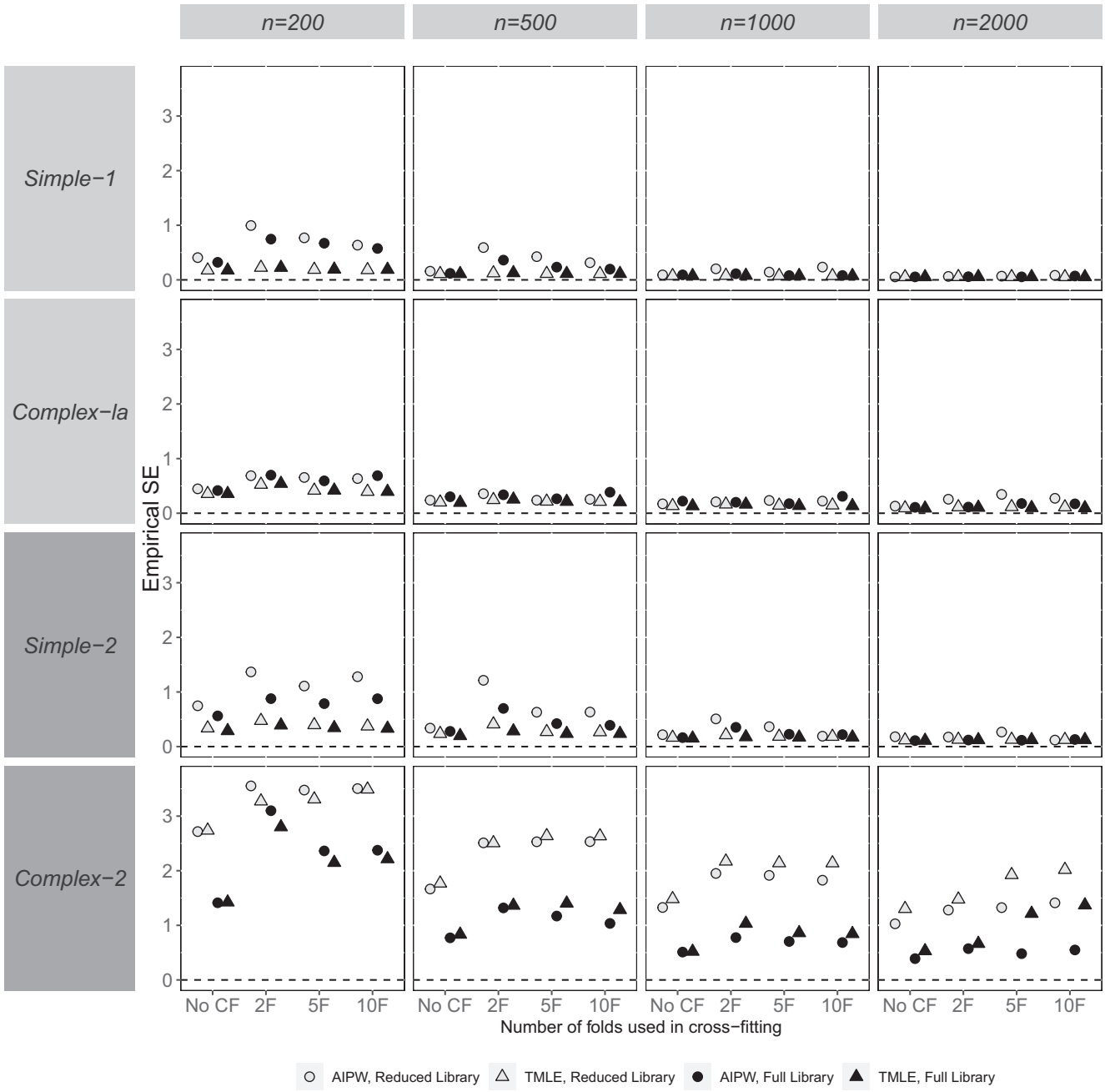

FIGURE 3 | Simulation study results for the Empirical SE for AIPW and TMLE with varying use of cross-fitting, by scenario (data generating mechanism and sample size). Results from a small number of datasets were excluded for AIPW (but not TMLE) due to the large effect that they had on some calculated performance measures (refer to Sections 7.1 and 5 of the Supporting Information for more details).

where using 2 folds appeared to result in lower bias compared to the use of ≥ 5 folds with CF (Figure 2). Across scenarios at smaller sample sizes, empirical SE was slightly smaller when using ≥ 5 folds than with 2 folds, but at larger sample sizes (e.g., $n = 2000$), empirical SE was similar regardless of the number of folds used (Figure 3). Patterns were harder to discern for the relative error in model-SE (Figure 4), with other factors appearing more important than the number of folds (e.g., the method). For coverage, with the larger confounder set (*simple-2* and *complex-2*), fewer folds appeared slightly better as sample size increased (Figure 5).

7.4 | Library Impact

Here we outline general observations regarding method performance according to the library used in SL. The full library produced substantially less bias in point estimates for *complex-1a* at the largest sample size ($n = 2000$) and for *complex-2* across all sample sizes compared to the reduced library. Otherwise, the reduced library performed as well as or better than the full library in terms of bias across all the sample sizes considered. Empirical SE did not appear to differ by the library used in SL except for *complex-2*, where empirical SE appeared smaller

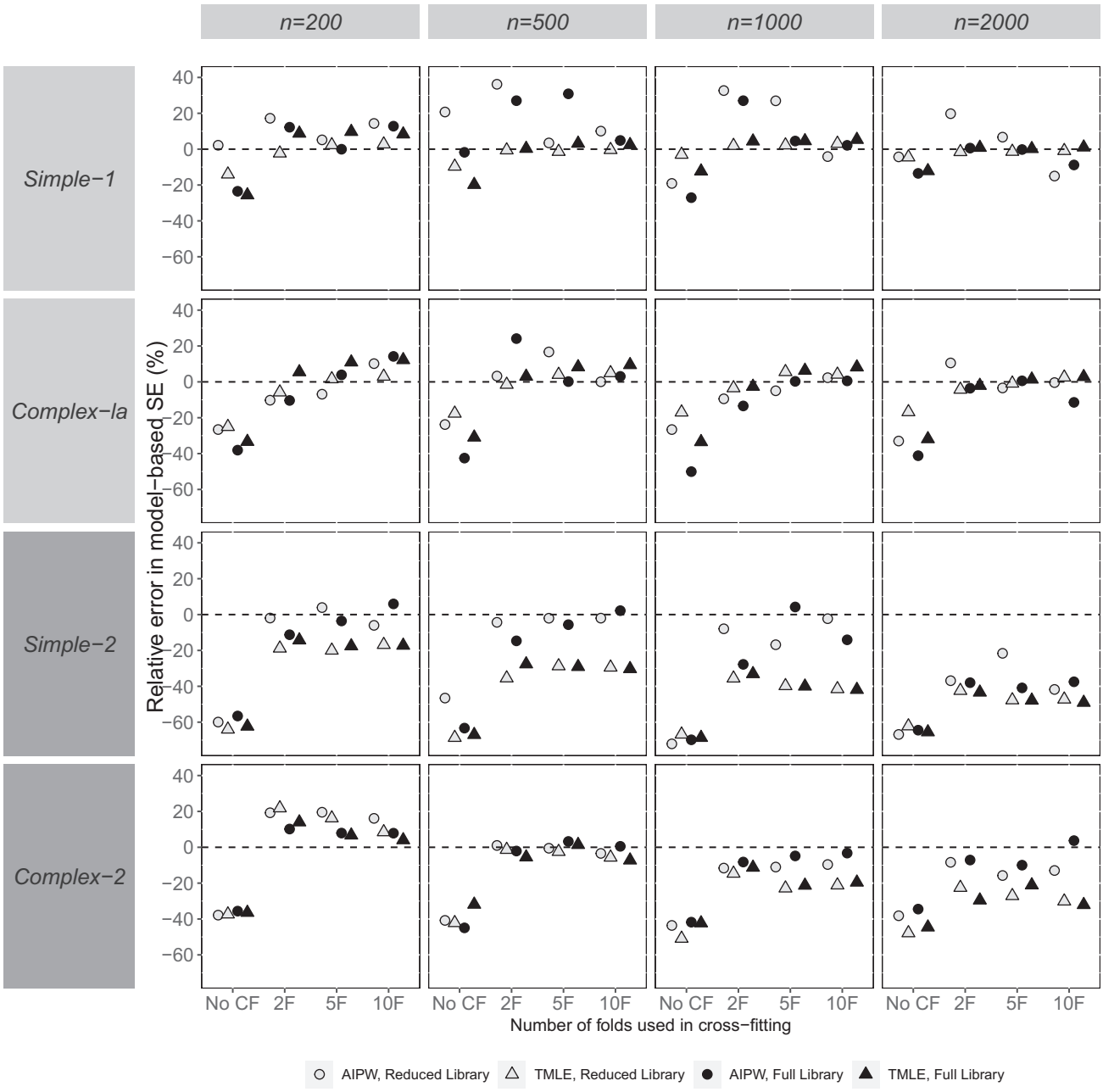

FIGURE 4 | Simulation study results for the bias in model-based SE (%) for AIPW and TMLE with varying use of cross-fitting, by scenario (data generating mechanism and sample size). Results from a small number of datasets were excluded for AIPW (but not TMLE) due to the large effect that they had on some calculated performance measures (refer to Sections 7.1 and 5 of the Supporting Information for more details).

with the full compared to the reduced library at all sample sizes.

For relative error in model-SE and coverage, the impact of library on method performance appeared to differ according to whether CF was applied or not (Figures 4 and 5). In general, but with some exceptions, smaller library differences were observed for relative error in model-SE and coverage when CF was applied. When CF was not applied, larger library differences were observed for *simple-1* and *complex-1a*, where higher relative error in model-SE and lower coverage (further from nominal) were observed when the full library was used compared to the reduced library.

8 | Application to the BIS Case Study

For the BIS case study example, we applied both AIPW and TMLE without CF and with CF, using 2, 5, and 10 folds. We used reduced and full libraries in SL [composition detailed in Section 5 of the Supporting Information (Table S24)], and we considered the two confounder sets analogous with those that motivated the simulation study design [detailed in Section 3 of the Supporting Information (Tables S2–S4)]. Henceforth, we use the following terminology: partially adjusted refers to estimates obtained using a confounder set that contains demographic and background confounders only ($p = 14$), and fully adjusted refers

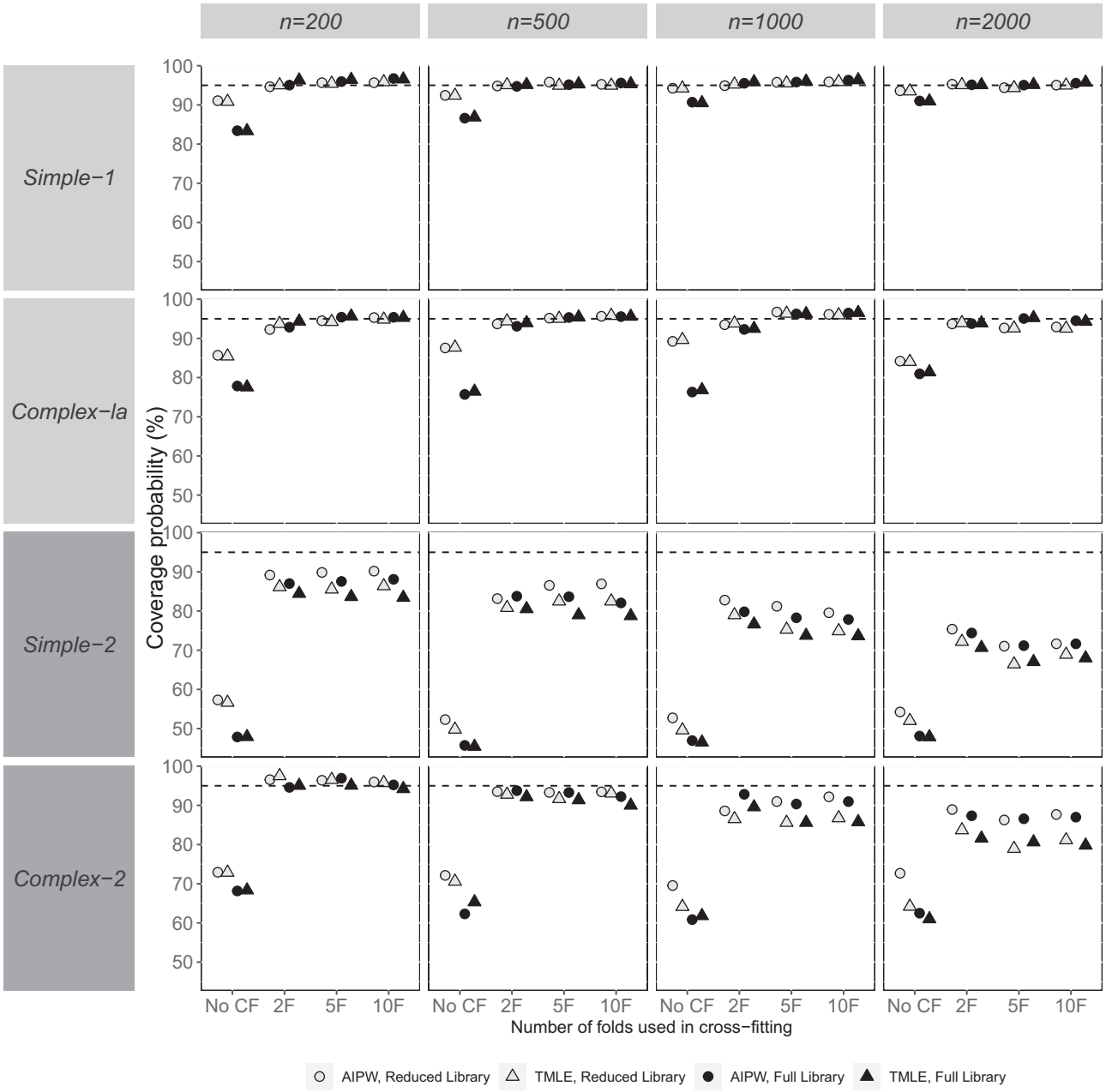

FIGURE 5 | Simulation study results for the coverage probability of the 95% CI for AIPW and TMLE with varying use of cross-fitting, by scenario (data generating mechanism and sample size). Results from a small number of datasets were excluded for AIPW (but not TMLE) due to the large effect that they had on some calculated performance measures (refer to Sections 7.1 and 5 of the Supporting Information for more details).

to an estimate adjusted for a set that contains demographic, background, and metabolomic confounders ($p = 87$).

8.1 | Results

The results are presented in Figure 6. Full results are available in Section 8 of the Supporting Information (Table S35, and the distribution of estimated propensity scores by exposure group is provided in Figures S1 and S2).

In general, estimated effect sizes were small for both methods, with and without CF and regardless of library used. A sensible fully adjusted estimate for the ACE was unable to be obtained for AIPW, with a full library and 10-fold CF, with the large and implausible estimate (−3.7; 95% CI [−17.2, 9.8]) reminiscent of unstable results reported in the simulation study for AIPW in some datasets. In addition, fully adjusted point estimates when using the reduced library appeared larger than those obtained when using the full library, potentially explained by the larger bias in point estimates seen in the simulation study with the reduced library in the most complex scenario.

Otherwise, point estimates and the width of CIs were similar for both methods. One exception was for AIPW with partial

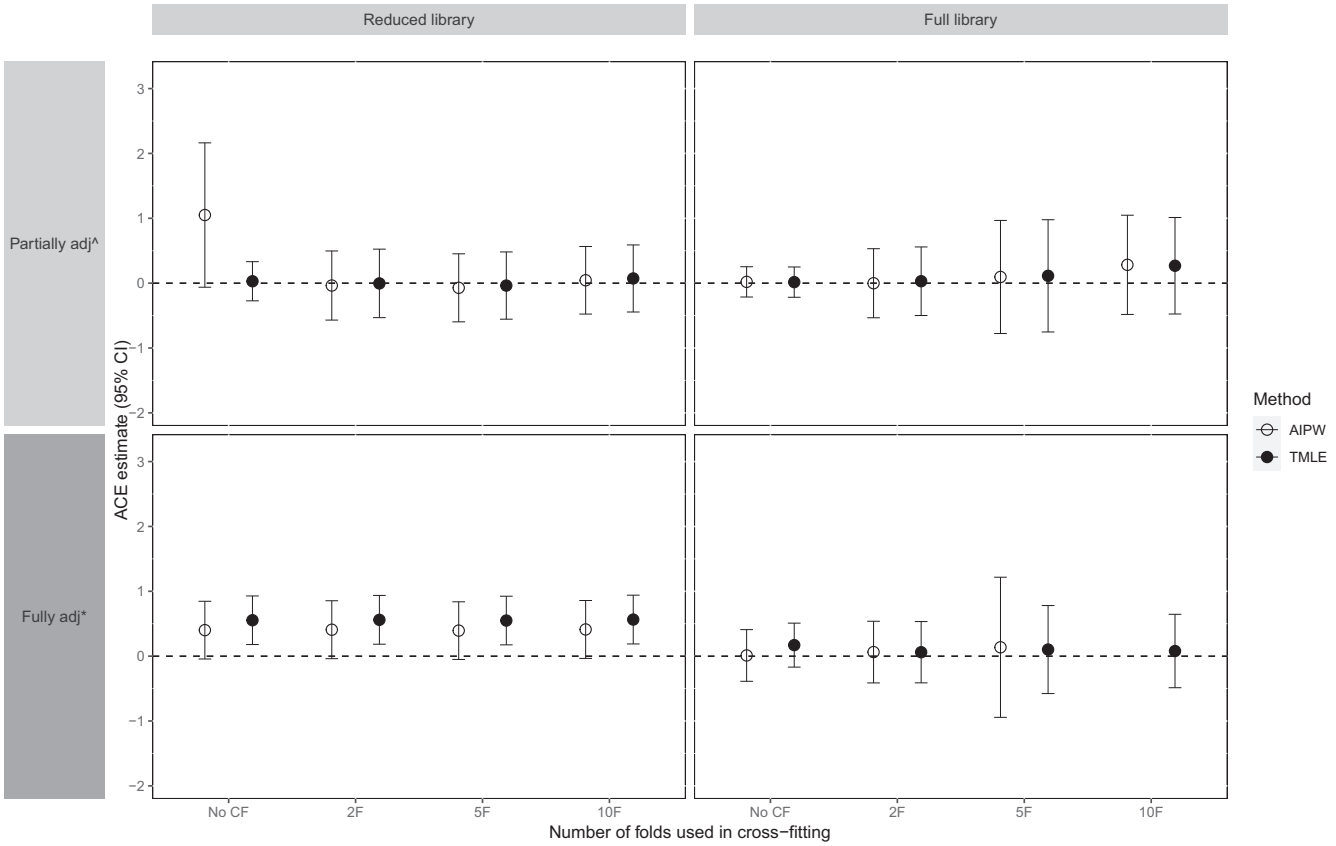

FIGURE 6 | Estimated average causal effect (ACE) with accompanying 95% CI, of inflammation (GlycA) in 1-year old infants on Pulse Wave velocity (PWV) at 4 years of age (standardized), obtained by applying the methods to the BIS motivating example. ^ demographic and background confounders only; * demographic, background, and metabolomic confounders. Estimate for AIPW using fully adjusted confounder set and full library in SL with 10-fold cross-fitting not shown in figure as too large to be sensible (refer to Table S35 in Section 8 of the Supporting Information).

adjustment when using the reduced library in SL without CF, where the estimate was larger and less precise than for TMLE (possibly due to the instability of AIPW mentioned above). Another exception was observed with full adjustment, where the CIs accompanying estimates (where available) were slightly wider for AIPW compared to TMLE, with this observation being more marked when the full library was used in SL.

In general, estimates were of similar magnitude with and without CF, but CIs were wider when implemented with CF compared to when no CF was applied. When CF was used and sensible estimates were able to be obtained, point estimates were similar regardless of the number of folds used. The width of CIs was similar when 5 or 10 folds were used with CF. With the full library, but not the reduced library, CI widths were larger with 5 or 10 folds than when 2 folds were used with CF.

8.2 | Insights From the Simulation Study

Next, we consider which of the several approaches used in the example would be considered most reliable in light of the simulation study results. Specifically, we consider simulation results for the complex scenarios because, in this example, we think it is highly plausible that there are complex relationships among variables. With this in mind, simulation results suggest that for a larger confounder set (full adjustment) and the small sample size in this motivating example, the use of the full library and CF may produce the most trustworthy and credible estimates. The findings of the simulation study suggest a slight preference for TMLE over AIPW due to stability. In contrast, for the small confounder set (partial adjustment), we found in the simulation study that method performance was best when using the reduced library rather than the full library and when CF was applied.

9 | Discussion

We conducted a simulation study that was motivated by a realistic study to answer key questions regarding the relative performance of AIPW and TMLE, the use of CF, and the number of folds used within CF when estimating the ACE in the presence of high-dimensional confounding. We found that, in general, in this setting, AIPW and TMLE performed similarly, but AIPW exhibited some stability issues. CF did not greatly affect the bias of point estimates but was important to reduce the bias in the model-based SE and for improving the coverage of the methods. In general, there was no substantial difference between using 2, 5, or 10 folds when using CF with the CML methods. In complex scenarios, the full, more flexible library was important for bias in point estimates, especially at larger sample sizes, and for empirical SE.

In the high-dimensional setting, our findings suggest that, although either AIPW or TMLE may be sensible choices of method to estimate the ACE, some caution is warranted with AIPW. Occasional instability of AIPW is likely explained by limitations of the method, where estimates produced may not always be within the natural bounds of the parameter space, in contrast to TMLE where this is not an issue because it is a substitution estimator [6]. With a complex scenario (regardless of the number of confounders), the full library appeared important, and this is as we would expect, with the more flexible algorithms needed to capture the complexity. Regardless of sample size or the number of confounders used, CF should be used with CML methods, and this is even more important if a full library is used. These results are in line with expectations that CF helps reduce the bias in variance estimation arising from the use of flexible ML algorithms. Interestingly, the number of folds used in CF did not have much impact on method performance, with other factors appearing more important (e.g., method and library choice).

Our findings align with previous methodological studies evaluating and comparing these CML methods with and without CF, although, unlike our study, most have considered low-dimensional confounding only (with very few or binary confounders only), with little focus on strong high-dimensional confounding [1, 3] or motivation from realistic case studies. One exception is a recent study by Meng and Huang [24] which considered both a realistic setting and high-dimensional confounding (≈1000 observations, 331 covariates), and like us, observed that AIPW and TMLE with CF performed similarly in general. Li et al. [53] evaluated CML methods, finding that both AIPW and TMLE with CF are favorable for robust inferences, as we found, but their study was in the context of a randomized experiment, with sample sizes in general much larger than we considered. In the literature it has been suggested that the number of folds considered with CF should be higher for smaller sample sizes than for larger sample sizes [12]. In simple settings without high-dimensional confounding, Chernozhukov et al. [22] report that 4 or 5 may work better than 2 folds. However, until our study, there has been no existing guidance on the number of folds to use in CF in realistic settings that feature high-dimensional confounding. We found that in general, CML method performance was similar regardless of the number of folds. In line with previous recommendations, other factors (library, method) appeared more important than number of folds used across the full range of performance indicators.

Meng and Huang [24] found that parametric learners generally performed similarly to or outperformed non-parametric learners in a realistic setting and had lower computational times. Our findings in simple scenarios are consistent with that study, but in complex scenarios (e.g., *complex-2*), we found that the use of the full, more flexible library was important to reduce bias in point estimates and variance, particularly for larger sample sizes, with variation depending on the number of confounders used. An explanation for this could be that our full library considered learners that extended beyond tree-based methods, with the importance of doing so highlighted in a recent commentary [12]. It could also be due to differences in the design of our simulation study, such as our data-generating mechanisms, because conclusions in regard to the use of parametric versus non-parametric learners will likely be very specific to the chosen data-generating mechanisms. Another possible explanation is the variant of CF used (single CF) as opposed to other variants of CF available, including what is referred to as double CF [3]. Phillips et al. [54] provide guidance regarding the choices that need to be made when using SL, including library construction. They state the importance of using a diverse library in SL, though emphasize that the number of learners included might need to be limited in the high-dimensional confounding setting. Our study adds to this important literature by re-emphasizing the need to carefully consider the setting-specific constraints (e.g., sample size and number of confounders) when choosing a library.

Our research strongly suggests that caution is warranted when applying these methods, with careful consideration of SL library composition required. We found indications that performance does not necessarily improve with increasing sample sizes. For example, in scenarios where we considered the large confounder set and used CF, we found increasing bias in model-SE and decreasing coverage, alongside decreasing bias in point estimates, as larger sample sizes were considered. For these scenarios (simple-2, complex-2), we noted larger SL coefficients were assigned to the glmnet learner at smaller sample sizes than at larger sample sizes (Supporting Information: Section 9, Figures S3 and S4). A possible explanation for this observation could be that simpler algorithms like glmnet are prioritized at smaller sample sizes, which are in turn less prone to overfitting or less likely to give extreme propensity score estimates. In keeping with this suggestion, we observed that the estimated propensity scores were more variable and had a much smaller minimum (closer to zero) for the larger sample sizes than for the smaller sample sizes (Supporting Information: Section 9, Figures S5 and S6). A possible explanation for this observation is that at the lower sample size, the screening within glmnet reduces the variable set, inducing bias in point estimates but dampening variability in the propensity scores. At larger sample sizes, the retention of variables helps reduce bias in point estimates but induces greater variability, with estimated propensity scores closer to zero resulting in biased model-SE's and poorer coverage. This phenomenon suggests that careful selection of the learners within SL is important in the sample sizes commonly encountered in modern observational studies. Future research should examine the inclusion of ridge regression or high-dimensional regression [55] as learners, which are expected to be better at handling a large number of variables with non-null effects. It is important to note that in practice, estimated propensity scores close to zero or one may indicate a lack of overlap, which raises conceptual issues, and therefore it is important for researchers to always inspect estimated propensity scores.

Strengths of our simulation study include that it was informed by a realistic case study, where we attempted to simulate confounders with dependencies consistent with associations observed between them in our realistic dataset. We considered several scenarios, representing both simple and complex data generating mechanisms, and simultaneously evaluated the performance of the methods across a range of modest sample sizes that are likely to be encountered in observational studies. However, although we considered a wide range of scenarios, further data generating mechanisms could be considered to control the heterogeneity of the ACE across scenarios. Indeed, in the complex scenarios, for the outcome generation models,

while the main effect coefficient value for the exposure was modified depending on sample size to control power (and thus, ensure comparability of coverage probability estimates across scenarios), coefficients for the interaction terms were not modified depending on sample size. This could have induced greater heterogeneity in the causal effect across strata in some scenarios, making it more challenging to compare the bias of point and SE estimates across scenarios. Further aspects of SL implementation that could potentially affect CML method performance [56] were not fully explored. For example, we did not consider the tuning of hyperparameters for the learners and did not compare the use of screening vs. no screening in SL implementations as we only considered libraries without separate explicit screening.

A further limitation of our study is that we did not examine the method for CF that aggregates over different random seeds to do the split, suggested by Chernozhukov et al. [22], which could be more robust than the methods we assessed that relied on a single seed/split. Recent studies have highlighted challenges in the practical implementation of DR methods with CF due to dependence on particular splits [28, 57, 58], and shown that increasing the number of folds in CF can potentially reduce the dependence on the seed [28]. We examined this in additional simulations, focusing on the most complex scenario (complex-2), at both the smallest ($n = 200$) and largest sample sizes ($n = 2000$). Specifically, we ran simulations under these settings to obtain estimates aggregated over multiple splits and examined the sensitivity of our simulation conclusions to this factor. Results are provided in Supporting Information, Section 10, Figures S7 and S8, and although there was some variation in results for these scenarios (e.g., for both sample sizes when the reduced library was used, the aggregate method had greater bias in model-SE and lower coverage than that observed in the original simulation study), our main conclusions regarding the impact of increasing number of folds held. We also compared the results for the case-study application with results obtained by aggregating estimates over multiple splits (e.g., applying Chernozhukov et al. [22] suggested approach) (Supporting Information, Section 10, Figure S9). Conclusions were similar in this case regardless of approach, but confidence intervals for the aggregated estimates were wider. We note that recent literature has suggested that an alternative way of reducing seed dependence is to increase folds in CF to a sensible number dependent on the dimensionality of the data [59]. This is less computationally demanding than aggregating. Careful consideration to the dependence on splits should be given in the practical application of these methods, with further research in this area warranted.

Finally, the simulation was based on one realistic motivating example, which may limit the guidance that is able to be provided from this study. Further, this case study had its own limitations. The exposure was dichotomized at an arbitrary cut-off. An alternative approach to tackle such settings is to consider the exposure as continuous, which would require defining the estimand in a different way (see below). In addition, we acknowledge that there are many interventions that may bring GlycA concentration below the 75th percentile, and each of these could have different effects on cardiovascular health; that is, the intervention remains ill-defined, which would challenge the causal consistency assumption. While these complexities arise often in discovery-phase studies of this kind, they do not justify ignoring confounding bias.

Future research directions could explore the CML methods in applications with a rare treatment or a binary, possibly rare outcome, or a large number of categorical confounders, all of which could bring about challenges with the application of CF. Exploring the impact of the CF variant [e.g., double CF compared to single CF (the variant of CF that we used)] on the performance of the methods would also provide insight. Double CF may possibly result in weaker rates of convergence requirements for the nuisance parameters compared to single CF [24]. However, Zivich and Breskin [3] discuss the potential challenge of using double CF with small sample sizes and suggest that single CF may overcome some of the difficulty. Limited sample sizes are common in realistic settings, and it remains for future research to determine whether double CF may be preferable to single CF in these situations. We have shown that further research is required into the composition of SL libraries, particularly as there is a need to balance the requirement for sufficient diversity of learners with performance across an extensive range of data-generating mechanisms, with or without sparsity. In our simulations, we observed that SE estimation was more sensitive to bias when using a large confounder set, reflecting a scenario that may be too complex. Therefore, it would also be helpful to explore any potential benefits of applying explicit screening within SL itself in a high-dimensional confounding setting, which potentially could assist with trimming down the size of the confounder set.

Further possibilities include extending this work beyond dichotomous exposures (e.g., continuous exposures) as they are often of interest. Possibilities for continuous exposures include exploring a shifting exposure distribution [60–63] or estimation of standard regression parameters via an influence curve approach [64]. The latter approach focuses on a summary of the exposure effect with reduced modeling assumptions, rather than ascertaining how the outcome changes with step increases or changes in exposure. It will also be important to explore challenges posed by multivariable missing data, which require consideration of the missingness mechanism as well as the identifiability or "recoverability" of the estimand [65–68]. In the high-dimensional confounding setting, evaluation and development of the CML methods in the presence of missing data is important. This is a growing area of research [69].

To conclude, in a high-dimensional confounding setting, we found that on most occasions AIPW and TMLE performed similarly for estimating the ACE, with a slight preference toward the use of TMLE, given its greater stability. In cases where complex confounding mechanisms are suspected, a more diverse and flexible SL library may be beneficial to reduce bias in point estimates and variance. CF is important for the estimation of SE and should be used, as without it we observed underestimation of model-SE and undercoverage, particularly with a more diverse library. Our study suggests that in the high-dimensional confounding setting, the number of folds used with CF is less important than the use of CF itself.

Author Contributions

All authors contributed to the planning and design of the simulation study. S.E. conducted the simulation study and performed the analyses and prepared the first draft of the manuscript. J.B.C., S.V., and M.M.-B. reviewed the manuscript, and S.E. revised the manuscript accordingly. All authors reviewed, sighted, and approved the final version of the manuscript.

Acknowledgments

We thank the BIS investigator group for providing access to the case-study data for illustrative purposes in this work. In addition, the authors thank David Burgner and Toby Mansell for sharing their expertise in the growing cardiovascular research area. Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. A. I. Naimi, A. E. Mishler, and E. H. Kennedy, "Challenges in Obtaining Valid Causal Effect Estimates With Machine Learning Algorithms," *American Journal of Epidemiology* 192, no. 9 (2023): 1536–1544, https://doi.org/10.1093/AJE/KWAB201.

2. M. S. Schuler and S. Rose, "Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies," *American Journal of Epidemiology* 185, no. 1 (2017): 65–73, https://doi.org/10.1093/aje/kww165.

3. P. N. Zivich and A. Breskin, "Machine Learning for Causal Inference: On the Use of Cross-Fit Estimators," *Epidemiology* 32, no. 3 (2021): 393–401, https://doi.org/10.1097/EDE.0000000000001332.

4. O. Dukes and S. Vansteelandt, "How to Obtain Valid Tests and Confidence Intervals After Propensity Score Variable Selection?," *Statistical Methods in Medical Research* 29, no. 3 (2020): 677–694, https://doi.org/10.1177/0962280219862005.

5. H. Leeb and B. M. Pöscher, "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?," *Annals of Statistics* 34, no. 5 (2006): 2554–2591, https://doi.org/10.1214/009053606000000821.

6. I. Díaz, "Machine Learning in the Estimation of Causal Effects: Targeted Minimum Loss-Based Estimation and Double/Debiased Machine Learning," *Biostatistics* 21, no. 2 (2019): 353–358, https://doi.org/10.1093/biostatistics/kxz042.

7. T. J. Vander Weele, "Principles of Confounder Selection," *European Journal of Epidemiology* 34, no. 3 (2019): 211–219, https://doi.org/10.1007/s10654-019-00494-6.

8. R. H. H. Groenwold, E. Hak, and A. W. Hoes, "Quantitative Assessment of Unobserved Confounding Is Mandatory in Nonrandomized Intervention Studies," *Journal of Clinical Epidemiology* 62, no. 1 (2009): 22–28, https://doi.org/10.1016/J.JCLINEPI.2008.02.011.

9. S. Rose and M. J. van der Laan, *Targeted Learning: Causal Inference for Observational and Experimental Data* (Springer, 2011).

10. M. J. van der Laan and D. Rubin, "Targeted maximum likelihood learning," *International Journal of Biostatistics* 2, no. 1 (2006): 1–38, https://doi.org/10.2202/1557-4679.1043.

11. V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey, "MACHINE LEARNING IN ECONOMETRICS Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review* 107, no. 5 (2017): 261–265, https://doi.org/10.1257/aer.p20171038.

12. L. B. Balzer and T. Westling, "Invited Commentary: Demystifying Statistical Inference When Using Machine Learning in Causal Research," *American Journal of Epidemiology* 192, no. 9 (2023): 1545–1549, https://doi.org/10.1093/aje/kwab200.

13. P. J. Bickel, F. Götze, and W. R. van Zwet, "Resampling Fewer Than n Observations: Gains, Losses, and Remedies," *Institute of Statistical Science, Academia Sinica* 7, no. 1 (1997): 1–31, https://doi.org/10.1007/978-1-4614-1314-1_17.

14. A. van der Vaart, "Higher Order Tangent Spaces and Influence Functions," *Statistical Science* 29, no. 4 (2014): 679–686, https://doi.org/10.1214/14-STS478.

15. H. Bang and J. M. Robins, "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics* 61, no. 4 (2005): 962–973, https://doi.org/10.1111/j.1541-0420.2005.00377.x.

16. J. M. Robins, A. Rotnitzky, P. L. Zhao, and P. L. Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association* 89, no. 427 (1994): 846–866, https://doi.org/10.1080/01621459.1994.10476818.

17. A. N. Glynn and K. M. Quinn, "An Introduction to the Augmented Inverse Propensity Weighted Estimator," *Political Analysis* 18, no. 1 (2010): 36–56, https://doi.org/10.1093/pan/mpp036.

18. S. D. Lendle, B. Fireman, and M. J. van der Laan, "Targeted Maximum Likelihood Estimation in Safety Analysis," *Journal of Clinical Epidemiology* 66, no. 8 SUPPL.8 (2013): S91–S98, https://doi.org/10.1016/j.jclinepi.2013.02.017.

19. S. Rose and D. Rizopoulos, "Machine Learning for Causal Inference in Biostatistics," *Biostatistics* 21, no. 2 (2019): 336–338, https://doi.org/10.1093/biostatistics/kxz045.

20. M. A. Luque-Fernandez, M. Schomaker, B. Rachet, and M. E. Schnitzer, "Targeted Maximum Likelihood Estimation for a Binary Treatment: A Tutorial," *Statistics in Medicine* 37, no. 16 (2018): 2530–2546, https://doi.org/10.1002/sim.7628.

21. J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky, "Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable," *Statistical Science* 22, no. 4 (2007): 523–539.

22. V. Chernozhukov, D. Chetverikov, M. Demirer, et al., "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal* 21, no. 1 (2018): C1–C68, https://doi.org/10.1111/ectj.12097.

23. W. Zheng and M. J. van der Laan, "Cross-Validated Targeted Minimum-Loss-Based Estimation," in *Targeted Learning. Springer Series in Statistics* (Springer, 2011).

24. X. Meng and J. Huang, "REFINE2: A Tool to Evaluate Real-World Performance of Machine-Learning Based Effect Estimators for Molecular and Clinical Studies," *ArXiv* (2022), https://arxiv.org/abs/2105.13148.

25. N. Kreif and K. DiazOrdaz, "Machine Learning in Policy Evaluation: New Tools for Causal Inference," in *Oxford Research Encyclopedia of Economics and Finance* (Oxford University Press, 2019).

26. A. Decruyenaere, J. Steen, K. Colpaert, D. D. Benoit, J. Decruyenaere, and S. Vansteelandt, "The Obesity Paradox in Critically Ill Patients: A Causal Learning Approach to a Casual Finding," *Critical Care (London, England)* 24, no. 1 (2020): 485, https://doi.org/10.1186/s13054-020-03199-5.

27. R. Herrera, U. Berger, S. Ehrenstein vO, et al., "Estimating the Causal Impact of Proximity to Gold and Copper Mines on Respiratory Diseases

in Chilean Children: An Application of Targeted Maximum Likelihood Estimation," *International Journal of Environmental Research and Public Health* 15, no. 1 (2018): 15, https://doi.org/10.3390/ijerph15010039.

28. P. N. Zivich, "Commentary the Seedy Side of Causal Effect Estimation With Machine Learning," *Epidemiology* 35, no. 6 (2024): 787–790, https://doi.org/10.1097/EDE.0000000000001783.

29. P. Vuillermin, R. Saffery, K. J. Allen, et al., "Cohort Profile: The Barwon Infant Study," *International Journal of Epidemiology* 44, no. 4 (2015): 1148–1160, https://doi.org/10.1093/IJE/DYV026.

30. D. Sorriento and G. Iaccarino, "Inflammation and Cardiovascular Diseases: The Most Recent Findings," *International Journal of Molecular Sciences* 20, no. 16 (2019): 5–8, https://doi.org/10.3390/ijms20163879.

31. S. DeWeerdt, "Inflammation in Heart Disease: Do Researchers Know Enough?," *Nature* 594, no. 7862 (2021): S8–S9, https://doi.org/10.1038/d41586-021-01453-6.

32. S. T. Chiesa, M. Charakida, G. Georgiopoulos, et al., "Glycoprotein Acetyls: A Novel Inflammatory Biomarker of Early Cardiovascular Risk in the Young," *Journal of the American Heart Association* 11, no. 4 (2022): e024380, https://doi.org/10.1161/JAHA.121.024380.

33. K. Sutton-Tyrrell, S. S. Najjar, R. M. Boudreau, et al., "Elevated Aortic Pulse Wave Velocity, a Marker of Arterial Stiffness, Predicts Cardiovascular Events in Well-Functioning Older Adults," *Circulation* 111, no. 25 (2005): 3384–3390, https://doi.org/10.1161/CIRCULATIONAHA.104.483628.

34. I. Sequí-Domínguez, I. Cavero-Redondo, C. Álvarez-Bueno, D. Pozuelo-Carrascosa, N. Arenas-Arroyo, and M.-V. dS, "Accuracy of Pulse Wave Velocity Predicting Cardiovascular and All-Cause Mortality. A Systematic Review and Meta-Analysis," *Journal of Clinical Medicine* 9, no. 7 (2020): 2080, https://doi.org/10.3390/jcm9072080.

35. M. A. Hernan, "A Definition of Causal Effect for Epidemiological Research," *Journal of Epidemiology and Community Health* 58, no. 4 (2004): 265–271, https://doi.org/10.1136/jech.2002.006361.

36. D. B. Rubin, "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Source: Journal of the American Statistical Association* 100, no. 469 (2005): 322–331, https://doi.org/10.1198/016214504000001880.

37. M. A. Hernán and J. M. Robins, *Causal Inference: What if* (Chapman & Hall/CRC, 2023).

38. R. M. Daniel, *Double Robustness. In: Wiley StatsRef: Statistics Reference Online, Chichester* (John Wiley & Sons, Ltd, 2018), 1–14.

39. E. H. Kennedy, "Semiparametric Theory and Empirical Processes in Causal Inference," in *Statistical Causal Inferences and Their Applications in Public Health Research*, ed. H. He, P. Wu, and D. G. D. Chen (Springer International Publishing, 2016), 141–167.

40. O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt, "Demystifying Statistical Learning Based on Efficient Influence Functions," *American Statistician* 76, no. 3 (2022): 292–304, https://doi.org/10.1080/00031305.2021.2021984.

41. D. Benkeser, M. Carone, M. J. van der Laan, and P. B. Gilbert, "Doubly Robust Nonparametric Inference on the Average Treatment Effect," *Biometrika* 104, no. 4 (2017): 863–880, https://doi.org/10.1093/biomet/asx053.

42. S. Wager, "STATS 361: Causal Inference. tech. rep., Stanford University," 2022.

43. R. Tibshirani, "Regression Shrinkage and Selection via the Lasso. Tech. Rep. 1," 1996.

44. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer, 2001).

45. L. Breiman, "Random Forests," *Machine Learning* 45 (2001): 5–32.

46. M. Kuhn and K. Johnson, *Applied Predictive Modeling* (Springer, 2013).

47. Z. H. Zhou, *Ensemble Methods, Foundations and Algorithms* (Chapman and Hall/CRC, 2012).

48. D. Jacob, "Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects," *ArXiv* (2020), https://arxiv.org/abs/2007.02852.

49. V. Chernozhukov, J. C. Escanciano, W. K. Newey, and J. M. Robins, "Locally Robust Semiparametric Estimation," Tech. Rep., 2020.

50. L. M. Montoya, M. J. van der Laan, A. R. Luedtke, and J. L. Skeem, "The Optimal Dynamic Treatment Rule Superlearner: Considerations, Performance, and Application to Criminal Justice," *International Journal of Biostatistics* 19, no. 1 (2023): 1–22, https://doi.org/10.1515/ijb-2020-0127.

51. J. Levy, "An Easy Implementation of CV-TMLE," *ArXiv* (2018), https://arxiv.org/abs/1811.04573.

52. T. P. Morris, I. R. White, and M. J. Crowther, "Using Simulation Studies to Evaluate Statistical Methods," *Statistics in Medicine* 38, no. 11 (2019): 2074–2102, https://doi.org/10.1002/sim.8086.

53. H. Li, S. Rosete, J. Coyle, et al., "Evaluating the Robustness of Targeted Maximum Likelihood Estimators via Realistic Simulations in Nutrition Intervention Trials," *Statistics in Medicine* 41, no. 12 (2022): 2132–2165, https://doi.org/10.1002/sim.9348.

54. R. V. Phillips, M. J. van der Laan, H. Lee, and S. Gruber, "Practical Considerations for Specifying a Super Learner," *International Journal of Epidemiology* 52, no. 4 (2023): 1276–1285, https://doi.org/10.1093/ije/dyad023.

55. P. Sur and E. J. Candès, "A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression," *Proceedings of the National Academy of Sciences of the United States of America* 116, no. 29 (2019): 14516–14525, https://doi.org/10.1073/pnas.1810420116.

56. P. Bach, O. Schacht, V. Chernozhukov, S. Klaassen, and M. Spindler, "Hyperparameter Tuning for Causal Inference With Double Machine Learning: A Simulation Study," *ArXiv* (2024), https://arxiv.org/abs/2402.04674.

57. L. Schader, W. Song, R. Kempker, and D. Benkeser, "Don't Let Your Analysis Go to Seed: On the Impact of Random Seed on Machine Learning-Based Causal Inference," *Epidemiology* 35, no. 6 (2024): 764–778, https://doi.org/10.1097/EDE.0000000000001782.

58. A. I. Naimi, Y. H. Yu, and L. M. Bodnar, "Pseudo-Random Number Generator Influences on Average Treatment Effect Estimates Obtained With Machine Learning," *Epidemiology* 35, no. 6 (2024): 779–786, https://doi.org/10.1097/EDE.0000000000001785.

59. N. T. Williams, A. Hung, and K. E. Rudolph, "Re: Don't Let Your Analysis Go to Seed: On the Impact of Random Seed on Machine Learning-Based Causal Inference," *Epidemiology* 36, no. 4 (2025): e12–e13, https://doi.org/10.1097/EDE.0000000000001860.

60. I. D. Muñoz and M. van der Laan, "Population Intervention Causal Effects Based on Stochastic Interventions," *Biometrics* 68, no. 2 (2012): 541–549, https://doi.org/10.1111/J.1541-0420.2011.01685.X.

61. I. Díaz and M. J. van der Laan, "Targeted Data Adaptive Estimation of the Causal Dose–Response Curve," *Journal of Causal Inference* 1, no. 2 (2013): 171–192, https://doi.org/10.1515/jci-2012-0005.

62. N. S. Hejazi and D. Benkeser, "'Txshift': Efficient Estimation of the Causal Effects of Stochastic Interventions in 'R'," *Journal of Open Source Software* 5, no. 54 (2020): 2447, https://doi.org/10.21105/JOSS.02447.

63. N. S. Hejazi, M. J. van der Laan, H. E. Janes, P. B. Gilbert, and D. C. Benkeser, "Efficient Nonparametric Inference on the Effects of Stochastic Interventions Under Two-Phase Sampling, With Applications to Vaccine Efficacy Trials," *Biometrics* 77, no. 4 (2021): 1241–1253, https://doi.org/10.1111/BIOM.13375.

64. S. Vansteelandt and O. Dukes, "Assumption-Lean Inference for Generalised Linear Model Parameters," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 84, no. 3 (2022): 657–685, https://doi.org/10.1111/rssb.12504.

65. K. J. Lee, J. B. Carlin, J. A. Simpson, and M. Moreno-Betancur, "Assumptions and Analysis Planning in Studies With Missing Data in Multiple Variables: Moving Beyond the MCAR/MAR/MNAR Classification," *International Journal of Epidemiology* 52, no. 4 (2023): 1268–1275, https://doi.org/10.1093/ije/dyad008.

66. K. Mohan and J. Pearl, "Graphical Models for Processing Missing Data," *Journal of the American Statistical Association* 116, no. 534 (2021): 1023–1037, https://doi.org/10.1080/01621459.2021.1874961.

67. M. Moreno-Betancur, K. J. Lee, F. P. Leacy, I. R. White, J. A. Simpson, and J. B. Carlin, "Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies," *American Journal of Epidemiology* 187, no. 12 (2018): 2705–2715, https://doi.org/10.1093/aje/kwy173.

68. K. Mohan and J. Pearl, "Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data," Tech. Rep., 2014.

69. S. G. Dashti, K. J. Lee, J. A. Simpson, I. R. White, J. B. Carlin, and M. Moreno-Betancur, "Handling Missing Data When Estimating Causal Effects With Targeted Maximum Likelihood Estimation," *American Journal of Epidemiology* 193, no. 7 (2024): 1019–1030, https://doi.org/10.1093/aje/kwae012.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.