

Research Article

Niloofer Moosavi, Tetiana Gorbach, and Xavier de Luna*

Valid causal inference with unobserved confounding in high-dimensional settings

<https://doi.org/10.1515/jci-2023-0069>

received October 11, 2023; accepted May 26, 2025

Abstract: Various methods have recently been proposed to estimate causal effects with confidence intervals that are uniformly valid over a set of data-generating processes when high-dimensional nuisance models are estimated by post-model-selection or machine learning estimators. These methods typically require that all the confounders are observed to ensure identification of the effects. We contribute by showing how valid semi-parametric inference can be obtained in the presence of unobserved confounders and high-dimensional nuisance models. We propose uncertainty intervals that allow for unobserved confounding, and show that the resulting inference is valid when the amount of unobserved confounding is not arbitrarily large; the latter is formalized in terms of convergence rates. Simulation experiments illustrate the finite sample properties of the proposed intervals. Finally, a case study on the effect of smoking during pregnancy on birth weight is used to illustrate the use of the methods introduced to perform an informed sensitivity analysis to the presence of unobserved confounding.

Keywords: average causal effect, double robust estimator, inverse probability weighting, sensitivity analysis

MSC 2020: 62D20, 62G05, 62J07

1 Introduction

In contrast to randomized experiments, observational studies are prone to the presence of confounding variables that are not balanced among treated and control individuals [1]. In such studies, all confounders are often assumed to be observed for identifiability of the causal parameter of interest [2,3]. Efforts to make this assumption plausible and the use of flexible nuisance models often yield a high-dimensional setting, where the number of nuisance parameters to be fitted may be at least of the order of the sample size. Nonetheless, some confounders may still be unobserved. Therefore, well-conducted observational studies should investigate the sensitivity of the inference to the assumption of no unmeasured confounding [4].

In this article, we study sensitivity analysis for semiparametric inference on the average causal effect of a binary treatment in high-dimensional observational studies. In this context, under certain conditions, augmented inverse probability weighting (AIPW) [5] and targeted learning [6] estimators yield uniformly valid inferences even when nuisance models are fitted with flexible machine learning algorithms (see, e.g., Farrell [7], Van der Laan and Gruber [8], Chernozhukov et al. [9] and Belloni et al. [10], Moosavi et al. [11] for a recent review). Roughly, uniformly valid inference means that the finite sample behavior of the estimator can be well approximated by its asymptotic distribution, even if preliminary analysis, such as variable selection, has been performed on the nuisance models prior to final estimation. Farrell [7] studied uniformly valid

* **Corresponding author: Xavier de Luna**, Department of Statistics, USBE, Umeå University, 901 87, Umeå, Sweden, e-mail: xavier.de.luna@umu.se

Niloofer Moosavi: Department of Statistics, USBE, Umeå University, 901 87, Umeå, Sweden, e-mail: moosavi.n0@gmail.com

Tetiana Gorbach: Department of Statistics, USBE, Umeå University, 901 87, Umeå, Sweden, e-mail: tetiana.gorbach@umu.se

inference associated with the AIPW estimator when flexible estimators of nuisance models are weakly consistent and fulfill multiplicative rate conditions (Assumption 2). Here, we build on these results to allow for unobserved confounding and study the uniform validity of the resulting inference. For this, we first specify the confounding bias of the AIPW estimator of the causal effect as a function of unobserved confounding. We then propose uncertainty intervals for the causal effect that account for such bias [12,13]. Our suggested inference on the causal effect using the uncertainty intervals ignores the finite sample bias and randomness in the estimation of the confounding bias. It provides uniformly valid inference given assumptions on the amount of unobserved confounding relative to the sample size; the latter is formalized in terms of convergence rate. In particular, unobserved confounding cannot be arbitrarily large. While this may seem like a restriction, our simulation experiments show that valid inference is obtained with relevant amounts of unobserved confounding.

This work contributes to the sensitivity analysis literature that originated in Cornfield *et al.* [4] and has been studied further by many others (to cite only a few: [14–20]). More specifically, we contribute to research that considers the sensitivity parameter to be the correlation between the potential outcomes and the treatment assignment given the observed covariates induced by unmeasured confounders [13,21–23]. Our contribution is to consider post-model-selection inference, including high-dimensional situations, and inference following machine learning fits of the nuisance models. In this respect, our work is related to Nabi *et al.* [24], which also considers inference on a causal parameter based on flexible estimation of nuisance models and allows for unobserved confounders via sensitivity analysis. Their approach differs in how the confounding bias is parameterized. We contrast the methods using a case study in Section 4.

The rest of this article is organized as follows. Section 2 describes the context and states our result on the uniform validity of the inference for a post-model-selection estimator of a causal parameter, for a given amount of unobserved confounding described by a correlation parameter. In practice, this correlation is unknown and a plausible range of correlation values may be considered as a sensitivity analysis. In Section 3, simulations are provided to illustrate the relevance of the theory for finite samples, in both low- and high-dimensional settings. Section 4 presents a case study of the effect of maternal smoking on birth weight. This illustrates the use of the herein proposed methods implemented in the R-package `hdim.ui` (<https://github.com/stat4reg/hdim.ui>). Section 5 concludes this article. All proofs are given in the Appendix.

2 Theory and method

We aim to study the causal effect of a binary treatment T on a continuous outcome of interest Y . Let $Y(t)$, $t \in \{0, 1\}$, be the potential outcomes that would have been observed under a corresponding treatment level, and assume the observed outcome is $Y = TY(1) + (1 - T)Y(0)$, for all units in the study. The average causal effect of the treatment is defined as $E(Y(1) - Y(0))$. Without loss of generality, we focus on the parameter $\tau = E(Y(1))$; (see the Appendix for other parameters of interest). We denote the vector of pre-treatment covariates and their transformations by $X = (1, X^{(1)}, \dots, X^{(p-1)})$. For simplicity, we call this the set of covariates. The dimension p of this vector is considered to increase with the sample size n .

Consider $O_{i,n} = \{(X_{i,n}, Y_{i,n}, T_{i,n})\}_{i=1}^n$ to be an i.i.d. sample that follows a distribution P_n , which is allowed to vary with the sample size n . We drop the subscript n , which implies dependence on n , when it is clear from the context. Moreover, let n_t denote the number of individuals with observed treatment $T = 1$. We use the notation $E_n[W_i] = n^{-1}\sum_{i=1}^n W_i$, $E_n[W_i]^q = (n^{-1}\sum_{i=1}^n W_i^q)^q$, and $E_{n_t}[W_i] = n_t^{-1}\sum_{i=1}^{n_t} W_i$.

We consider the augmented inverse propensity weighting estimator (AIPW; see Robins *et al.* [5] and Scharfstein *et al.* [25]):

$$\hat{\tau}_{\text{AIPW}} = E_n \left[\frac{T_i(Y_i - \hat{m}(X_i))}{\hat{e}(X_i)} + \hat{m}(X_i) \right], \quad (1)$$

where $\hat{m}(X)$ and $\hat{e}(X)$ are the arbitrary estimators of $m(X) = E(Y(1)|X, T = 1)$ and $e(X) = E(T|X)$, respectively. Let us consider the following assumptions.

Assumption 1. $e(X) > 0$.

Assumption 2. Assume $\hat{e}(X)$ is constant conditional on the observed value of $\{X_i, T_{i|j=1}^n\}$ (assumption of “no additional randomness,” [26], revision of Farrel [7]) and

- a. $E_n[(\hat{e}(X_i) - e(X_i))^2] = o_p(1)$ and $E_n[(\hat{m}(X_i) - m(X_i))^2] = o_p(1)$.
- b. $E_n[(\hat{m}(X_i) - m(X_i))^2]^{1/2} E_n[(\hat{e}(X_i) - e(X_i))^2]^{1/2} = o_p(n^{-1/2})$.
- c. $E_n[(\hat{m}(X_i) - m(X_i))(1 - T_i/e(X_i))] = o_p(n^{-1/2})$.

The following theorem gives the asymptotic behaviour of the AIPW estimator and its asymptotic target parameter. The theorem generalizes Theorem 3 in Farrell [26] by avoiding the assumption $E(Y(1)|X = x, T = 1) = E(Y(1)|X = x)$, i.e., allowing for unobserved confounders.

Theorem 1. *Let Assumptions 1, 2, and Assumption A1 in the Appendix hold. The AIPW estimator (1) is asymptotically linear as follows:*

$$\sqrt{n}(\hat{\tau}_{\text{AIPW}} - \tau^-) = \sum_{i=1}^n \Psi_i / \sqrt{n} + o_p(1),$$

where $\tau^- = E(m(X))$ and $\Psi_i = \frac{T_i(Y_i - m(X_i))}{e(X_i)} + m(X_i) - E(m(X_i))$.

Assumption 1 can be checked empirically. Assumption 2 is central. It allows data-adaptive fits of the propensity score and outcome model with rates lower than $n^{1/2}$. For instance, Assumption 2b is fulfilled if both nuisance functions are fitted at the $n^{1/4}$ rate. Such rates are attainable by machine learning algorithms [27], e.g., Lasso under sparsity conditions [e.g., 7,11], Deep neural networks under smoothness conditions [26,28], and other nonparametric smoothers and combinations thereof via super-learners [8]. Thus, Theorem 1 ensures a parametric rate of convergence for AIPW, e.g., in sparse high-dimensional situations.

Following Genbäck and de Luna [13], we study the consequences of the presence of unobserved confounders of the treatment–outcome relationship on inference about τ by postulating a sensitivity model:

Assumption 3. Let $Y(1) = E(Y(1)|X) + \xi$, $T^* = g(X) + \eta$, $T = \mathbf{I}(T^* > 0)$, where $\eta \sim \mathcal{N}(0, 1)$ (i.e., a probit link is used), $\eta \perp\!\!\!\perp X$, and \mathbf{I} represents an indicator function. Moreover, let $\xi = \rho\sigma\eta + \varepsilon$, where $\rho = \text{corr}(\xi, \eta)$, $\sigma^2 = \text{var}(\xi) < \infty$, with ε satisfying $\varepsilon \perp\!\!\!\perp (X, \eta)$ and $E(\varepsilon) = 0$.

It is generally agreed upon that a sensitivity model should make as few restrictions as possible on the observed data distribution [29,30]. Assumption 3 makes one such slight restriction, i.e., assuming the normality of η (probit model for the propensity score). This restriction is, however, rather weak since it is typically the case that changing the link function, e.g., from probit to logit (corresponding to the logistic distribution for η) does not substantially change the fitted probabilities. We study the effect of misspecifying the distribution of η in our simulation experiments.

The case $\rho = 0$ corresponds to the inclusion of all confounders in X . Assumption 3, therefore, considers a departure from the commonly used identification assumption $E(Y(1)|X = x, T = 1) = E(Y(1)|X = x)$ (no unobserved confounders) in order to perform a sensitivity analysis. Assumption 3 allows us to consider a potential amount of unobserved confounding through an interpretable parameter ρ , as well as to compute the size of the confounding bias of the AIPW estimator.

Theorem 2. *Suppose that for some $g^0(X)$ and $m^0(X)$, we have $E_n[|\hat{g}(X_i) - g^0(X_i)|] = o_p(1)$ and $E_n[(\hat{m}(X_i) - m^0(X_i))^2] = o_p(1)$. Suppose Assumptions 1 and 3 also hold and Assumption A1 holds for $m^0(X)$. If $g^0(X) = g(X)$ or $m^0(X) = m(X)$, the (asymptotic) confounding bias of the AIPW estimator (1), which we denote by $b = \tau^- - \tau$, is equal to*

$$b = \rho\sigma E(\lambda(g(X))),$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$, and ϕ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively.

Note that under the conditions of Theorem 1, the AIPW estimator (1) is a semiparametric efficient estimator of τ^- [see, 31, Example 5]. However, if confounders of the treatment–outcome relationship are not observed (not included in X), ρ and consequently, b is nonzero, and hence, the AIPW estimate is not a consistent estimate of τ . It can be shown that for given values of ρ and σ , the efficient influence function of the bias term has the form $\rho\sigma E(\lambda(g(X))) - b$. This motivates us to estimate the confounding bias b empirically by plugging in the estimated $g(X_i)$ values. For a more realistic scenario, the parameter σ must also be estimated. An intuitive choice is to use $\hat{\sigma} = E_{n_i}[(Y(1)_i - \hat{m}(X_i))^2]^{1/2}$. This estimator of the variance, $\hat{\sigma}^2$, is biased due to both high dimensionality/variable selection and unobserved confounders. Theorem A1 in the Appendix provides the asymptotic properties of $\hat{\tau}_{\text{AIPW}} - \hat{b}$ as an estimator of τ , assuming approximate sparsity and known ρ .

Theorem 3 introduces a corrected estimator of σ . Given the true value of the sensitivity parameter, this corrected estimator of the variance gives us a consistent estimate of the causal parameter τ .

Theorem 3. *Let $E_n[|\hat{g}(X_i) - g(X_i)|] = o_p(1)$ and $E_n[(\hat{m}(X_i) - m(X_i))^2] = o_p(1)$. Assume, further, that Assumptions 1, 3, A1, and A2 hold. Let*

$$\begin{aligned}\hat{\sigma}_c^2 &= E_{n_i}[(Y(1)_i - \hat{m}(X_i))^2] / (1 - \rho^2 E_{n_i}[\hat{g}(X_i)\lambda(\hat{g}(X_i))] - \rho^2 E_{n_i}[\lambda^2(\hat{g}(X_i))]), \\ \hat{b}_c &= \rho \hat{\sigma}_c E_n[\lambda(\hat{g}(X_i))].\end{aligned}$$

Then, we have

$$\begin{aligned}\hat{\sigma}_c^2 &\xrightarrow{P} \sigma^2, \\ \hat{\tau}_{\text{AIPW}} - \hat{b}_c &\xrightarrow{P} \tau.\end{aligned}$$

Theorem 4 provides further asymptotic properties of $\hat{\tau}_{\text{AIPW}} - \hat{b}_c$ as an estimator of the average causal effect τ .

Theorem 4. *Let Assumptions 1–3 hold. Moreover, let Assumptions A1, and A4 hold and assume $\rho = o(1)$, $\sqrt{n} \rho E_n[\hat{g}(X_i) - g(X_i)] = o_p(1)$, and $\sqrt{n} \rho E_{n_i}[(m(X_i) - \hat{m}(X_i))^2]^{1/2} = o_p(1)$. We have*

- a. $\sqrt{n}((\hat{\tau}_{\text{AIPW}} - \hat{b}_c) - \tau) = \sum_{i=1}^n \Psi_i / \sqrt{n} + o_p(1)$,
- b. $V^{-1/2} \sqrt{n}((\hat{\tau}_{\text{AIPW}} - \hat{b}_c) - \tau) \rightarrow_d \mathcal{N}(0, 1)$,
- c. $\hat{V} - V = o_p(1)$,

where $V = E(\Psi_i^2)$, and $\hat{V} = E_n\left[\frac{T_i(Y_i - \hat{m}(X_i))^2}{\hat{e}(X_i)^2}\right] + E_n[(\hat{m}(X_i) - \hat{\tau}_{\text{AIPW}})^2]$.

As a corollary, the following $(1 - \alpha)\%$ confidence interval for τ given ρ is uniformly valid.

Corollary 1. *For each n , let \mathcal{P}_n be the set of distributions obeying the assumptions of Theorem 4. Then, we have:*

$$\sup_{P \in \mathcal{P}_n} |\Pr_P(\tau \in \{(\hat{\tau}_{\text{AIPW}} - \hat{b}_c) \pm c_\alpha \sqrt{\hat{V}/n}\} - (1 - \alpha))| \rightarrow 0,$$

where $c_\alpha = \Phi^{-1}(1 - \alpha/2)$.

Remark 1. The assumptions on ρ imply that the amount of unobserved confounding cannot be arbitrarily large for any finite sample. Two arguments can be made for the relevance of such restriction in applications: first, in well-designed studies, where we believe we have observed all major confounders based on subject-matter knowledge, we may expect only small amount of unobserved confounding left. If we do not believe in that we have observed all major confounders based on subject-matter reasoning, one may argue that the causal interpretability of the analysis is questionable anyways. Second, with very large observational databases, one may be able to take into account larger number of background characteristics with increasing sample sizes.

In practice, ρ is not known and a sensitivity analysis is obtained by considering the 95% uncertainty interval constructed as the union of all 95% confidence intervals obtained by varying $\rho \in (\rho_{\min}, \rho_{\max})$,

a plausible interval for ρ . If the latter interval contains the true value of ρ , then this uncertainty interval covers the parameter τ with at least 95% probability [32, Corollary 4.1.1].

Remark 2. While confidence intervals provide inference on τ in a classical sense under the absence of unobserved confounding, uncertainty intervals are a tool to bound τ given prior knowledge on ρ ($\rho \in (\rho_{\min}, \rho_{\max})$) (see Section 4 for an illustration). Thus, if the confidence interval assuming $\rho = 0$ does not contain zero (i.e., the inference suggests a non-zero causal effect) and the uncertainty interval contains zero then the $\rho = 0$ analysis is said to be sensitive to unobserved confounding. A common alternative strategy is to increase $\rho_{\max} = \rho_{\min}$ until the uncertainty interval contains zero. Then, this ρ_{\max} value can be reported and discussed as being the amount of unobserved confounding necessary to invalidate the conclusion based on $\rho = 0$.

When no parametric models are assumed and/or the number of covariates is greater than the number of observations, lasso regression [33] can be used in the first step to select low-dimensional sets of variables for fitting linear and probit nuisance models. In other words, a linear (in the parameters) regression can be fitted using variables selected by a preliminary lasso regression. A probit regression for the treatment can be fitted using variables selected by a preliminary probit-lasso regression. When it comes to the rate conditions in Assumption 2, one can use a post-selection linear model fit for the outcome under common sparsity assumptions on the true data-generating process (for post-lasso regression, see [7], Corollary 5). We are unaware of any similar result for an estimator in a sparse probit model. However, we investigate the performance of a post-lasso probit regression in the simulation section.

In the following sections, we use the R package `hdim.ui` (<https://github.com/stat4reg/hdim.ui>), where the estimators mentioned previously have been implemented and can be used to obtain uncertainty intervals and perform sensitivity analysis in high-dimensional situations. The package builds on code from the `ui` package (<https://cran.r-project.org/web/packages/ui/index.html>) [13].

3 Simulation study

The simulation study in this section aims to illustrate the finite sample properties of the asymptotic approximations obtained in the previous section. Here, we consider a high-dimensional setting, while the results for low-dimensional settings are reported in the Supplementary Material. Data were generated according to the model in Assumption 3. All the covariates are generated to be independently and normally distributed with mean 0 and variance 1. The error terms are generated using $(\eta, \xi) \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. We let $p = n$ and simulate outcome and treatment as

$$\begin{aligned} Y(1) &= 2 + X\beta - \rho\lambda(X\gamma) + \xi, & T^* &= X\gamma + \eta, \\ \beta &= 0.6(1, 1/2, 1/3, 1/4, 1/5, 1, 1/2, 1/3, 1/4, 1/5, 0, \dots, 0)', \\ \gamma &= 0.3(1, 1/2, 1/3, 1/4, 1/5, 1, 1, 1, 1, 1, 0, \dots, 0)', \end{aligned}$$

where both models include covariates that are weakly associated with the dependent variable. Such covariates are likely to be missed in a variable selection step. To study the method behavior in a case where the probit link for the treatment model is misspecified, we also simulate the error η from a logistic distribution with mean 0 and variance 1, while keeping the correlation with ξ equal to ρ .

In order to investigate the performance of the inference under variable selection, we use $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}^{\text{refit}}$ (using $\hat{\sigma}$, Theorem A1) and $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}_c^{\text{refit}}$ (using $\hat{\sigma}_c$, Theorem 4). In these estimators, $\hat{m}(X)$ is found by refitting the linear regression using variables selected by preliminary linear-lasso regression and $\hat{g}(X)$ is found by refitting the probit model using variables selected by preliminary probit-lasso regression. The linear-lasso

regression is implemented using the `hdm` package in R [34,35]. The probit-lasso is implemented using the `glmnet` package [36], where the regularization parameter is found by cross-validation. Then, 95% confidence intervals are constructed for a range of values of the sensitivity parameter ρ according to Corollary 1 using the influence curve-based standard error estimator. If empirical coverages are close to nominal, then the corresponding uncertainty intervals will be conservative by construction, as mentioned previously [see 13,32]. We are particularly interested in investigating coverages of confidence intervals when varying ρ as sample size increases ($n = 500, 1,000, 1,500$). This is because Theorem A1 assumes $\sqrt{n}\rho^3 = o(1)$ when using the biased estimator of outcome error variance, while the weaker condition $\rho = o(1)$ is used in Theorem 4 when the corrected variance estimator is used. The results of the simulation study are based on 500 Monte-Carlo replications.

Table 1 reports, respectively, biases and empirical coverages of the 95% confidence intervals for τ using $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - b^*$, $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}^{\text{refit}}$ and $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}_c^{\text{refit}}$. Here, b^* is an approximation of the confounding bias b using the true values of σ and ρ and the Monte Carlo estimate of $E(\lambda(g(X)))$.

Biases are as expected negligible when correcting with the true bias b^* . Also as expected, when the bias b is estimated, biases in the estimation of τ are largest for the larger values of ρ . The confidence intervals constructed using $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - b^*$ have minimum empirical coverage of 0.91. When a plug-in estimator of the confounding bias is used low empirical coverage may arise when ρ is too large (relative to the sample size). However, because of the weaker condition on ρ in Theorem 4, the coverage for the estimator $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}_c^{\text{refit}}$ is, as expected, closer to nominal level compared to the estimator $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}^{\text{refit}}$, and this improvement is more pronounced for larger sample sizes.

4 Case study: Effect of maternal smoking on birth weight

We re-examine a study that aims to assess the effect of smoking during pregnancy on birth weight [24,37,38]. The data come from an open-access sample of 4996 individuals, a sub-sample of approximately 500000 singleton births in Pennsylvania between 1989 and 1991 (see Nabi *et al.* [24] and resources therein for more details on the data). Weight at birth in grams is the outcome variable. The treated group consists of pregnant women who smoked during pregnancy, and the control group consists of women who did not smoke.

Table 1: Biases and empirical coverages of 95% confidence intervals for corrected estimators of τ ; high-dimensional scenario

	$n \setminus \rho$	$\hat{\tau}_{\text{AIPW}}^{\text{refit}} - b^*$				$\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}^{\text{refit}}$				$\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}_c^{\text{refit}}$			
		0.8	0.6	0.4	0.3	0.8	0.6	0.4	0.3	0.8	0.6	0.4	0.3
Bias													
$\eta \sim N(0,1)$	500	-0.01	0.03	0.03	0.02	0.03	0.02	-0.00	-0.01	-0.07	-0.04	-0.03	0.00
	1,000	0.01	0.02	0.02	0.01	0.09	0.03	0.00	-0.00	-0.05	-0.03	-0.01	-0.01
	1,500	0.00	0.01	0.01	0.01	0.10	0.03	0.01	-0.00	-0.02	-0.02	-0.01	-0.01
$\eta \sim \text{Logistic}$	500	0.01	0.03	0.03	0.01	0.03	-0.00	-0.02	-0.02	-0.10	-0.05	-0.06	-0.01
	1,000	-0.04	0.01	-0.03	0.00	0.03	0.01	-0.05	-0.02	-0.06	-0.06	-0.02	-0.02
	1,500	0.00	0.00	-0.00	0.00	0.08	0.02	-0.01	-0.01	-0.04	-0.04	-0.02	-0.03
Coverage													
$\eta \sim N(0,1)$	500	0.92	0.91	0.91	0.91	0.80	0.90	0.92	0.93	0.82	0.92	0.93	0.91
	1,000	0.96	0.92	0.94	0.93	0.60	0.87	0.94	0.93	0.82	0.91	0.95	0.94
	1,500	0.94	0.92	0.93	0.95	0.40	0.83	0.92	0.95	0.87	0.93	0.93	0.93
$\eta \sim \text{Logistic}$	500	0.91	0.91	0.91	0.91	0.80	0.90	0.93	0.94	0.81	0.92	0.95	0.93
	1,000	0.94	0.93	0.97	0.93	0.71	0.89	0.96	0.92	0.84	0.89	0.94	0.94
	1,500	0.96	0.94	0.96	0.92	0.56	0.90	0.95	0.93	0.81	0.90	0.95	0.95

We use the same set of covariates as in Nabi et al. [24]. As they argue, sensitivity analysis is necessary to account for potential unobserved confounders, such as genetic factors not observed. The observed covariates that we consider are maternal data including numerical (age and the number of prenatal visits) and categorical variables (education -less than high school, high school, more than high school- and birth order -one, two, larger than two) and binary variables including white, hispanic, married, foreign and alcohol use. Higher-order and interaction terms of the numerical variables of order up to three are also considered. Finally, all the second-order interactions with categorical variables are added, which gives a total of 80 terms.

The analysis in Almond et al. [37] estimates that smoking has an average effect of -203.2 g on birth weight, determined by a regression model without taking into account the effect of unobserved confounding. In Nabi et al. [24], a semiparametric approach yields an estimate of -223 g (95% confidence interval $[-274, -172]$). However, after accounting for unobserved confounding in a sensitivity analysis, the latter study suggests that the average effect is not more extreme than -200 g (see (2) for details on the clinical assumptions used).

In our analysis, we use the AIPW estimator of average causal effect, $E(Y(1)) - E(Y(0))$. The results presented in this article are directly applicable to $E(Y(0))$ as well, and, therefore, to $E(Y(1)) - E(Y(0))$ under corresponding assumptions. In particular, there are now two sensitivity parameters through the sensitivity model of Assumption 3, for both $Y(1)$ and $Y(0)$, denoted, respectively, ρ_1 and ρ_0 . We use `subset = 'refit'` in function `ui.causal` to exploit the built-in variable selection function (`ui.causal` included in the R-package `hdim.ui`). The resulting AIPW estimation gives an estimate of the average causal effect of -221 g (95% confidence interval: $[-273, -168]$).

In their sensitivity analysis, Nabi et al. [24] argued that the following inequalities make clinical sense:

$$\begin{aligned} E(Y(1)|T = 1) &< E(Y(0)|T = 1) < E(Y(0)|T = 0), \\ E(Y(1)|T = 1) &< E(Y(1)|T = 0) < E(Y(0)|T = 0). \end{aligned} \quad (2)$$

For instance from the second line, under smoking, non-smokers are expected to have higher average birth weight than smokers, $E(Y(1)|T = 0) > E(Y(1)|T = 1)$. We choose to also use these prior clinical assumptions, which yield bounds for our sensitivity parameters ρ_1 and ρ_0 . That is because the terms on the right- and left-hand sides of the inequalities in (2) can be unbiasedly estimated using sample averages since the data are fully observed, whereas the terms in the middle are functions of the sensitivity parameter. Appendix provides a description of how they are estimated. According to the range of sensitivity parameters obtained from the above restrictions (Figure 1), we have derived estimates and the uniformly valid confidence intervals along with a resulting uncertainty interval for $E(Y(0))$ and $E(Y(1))$ (Figure 2). The resulting uncertainty interval for

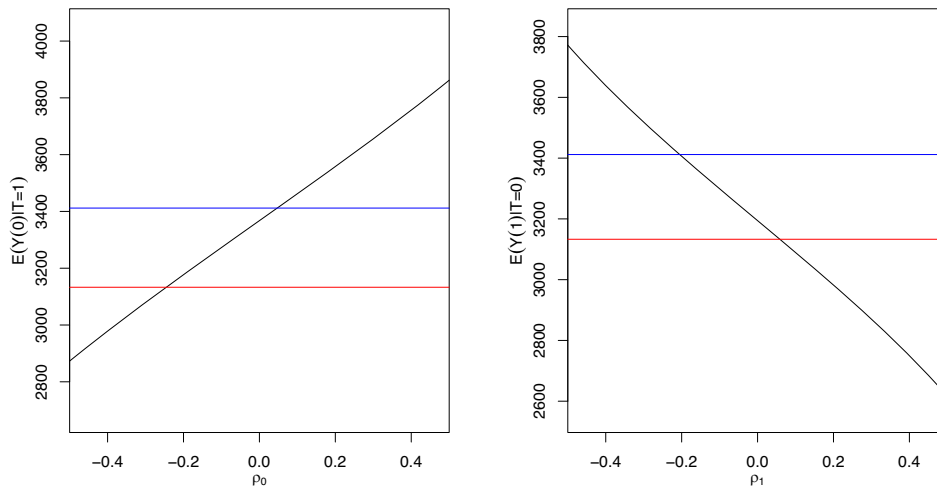


Figure 1: Left plot: estimated expectation of birth weight for smokers under not smoking, $E(Y(0)|T = 1)$. This estimate is constrained by the sample average estimations of $E(Y(1)|T = 1)$ in red and $E(Y(0)|T = 0)$ in blue, as (2) suggests. Right plot: the estimation of $E(Y(1)|T = 0)$. See Appendix for a description of the estimators. From these results, plausible values for the sensitivity parameters ρ_0 and ρ_1 are $(-0.25, 0.05)$ and $(-0.2, 0.06)$ respectively; i.e., where crossing with the red and blue bounds occur in respective plots.

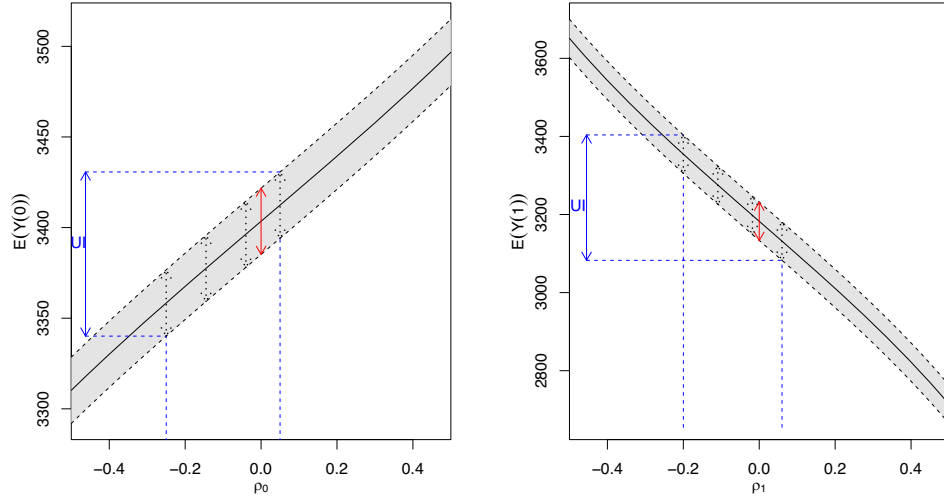


Figure 2: Estimates of average birth weight under smoking and non-smoking during pregnancy are given in the right and left plots, respectively; using output from `ui.causal` function (`hdim.ui` package). Solid black lines for bias-corrected AIPW estimates, $\hat{\tau}_{\text{AIPW}}^{\text{refit}} - \hat{b}^{\text{refit}}$, and 95% confidence intervals (gray area). Red intervals are confidence intervals assuming no unobserved confounding ($\rho_t = 0, t = 0, 1$), and blue intervals are 95% uncertainty intervals (using $-0.25 \leq \rho_0 \leq 0.05$ and $-0.20 \leq \rho_1 \leq 0.06$).

the average causal effect of maternal smoking on birth weight is $[-333, 49]$. Thus, while our AIPW estimation and 95% confidence interval under the assumption that all confounders are observed are similar to those of Nabi et al. [24], the sensitivity analysis is quite different, since they suggest that the average effect is not more extreme than -200 g while our analysis does not discard a potentially much larger adverse effect (up to 333 g weight loss) of maternal smoking depending on the strength of confounding.

The difference in conclusions may be due to the fact that Nabi et al. [24] analysis excludes the unconfoundedness situation as one of the possible scenarios while ours do not. This discrepancy might be due to different modeling assumptions and thereby different influence functions and corresponding estimation procedures. For instance, Nabi et al. [24] sensitivity model establishes a link between the observed and unobserved potential outcome densities, which necessitates the requirement of common supports; a condition stating that the support of each of the missing potential outcomes must be a subset of the support of the corresponding observed potential outcome. This assumption may affect the validity of the results (see Section 7 in [16]). Furthermore, our method only uses clinical assumptions for bounding sensitivity parameters, whereas Nabi et al. [24] uses those assumptions for both selecting a valid tilting function (defining the sensitivity model) and bounding sensitivity parameters. The effect on their analysis of using alternative tilting functions is unclear to us.

5 Discussion

Unobserved confounding cannot be discarded nor empirically investigated in observational studies, and therefore, sensitivity analysis to the unconfoundedness assumption should be common practice. Moreover, high-dimensional settings are typical in observational studies using large data sets and machine learning for nuisance models. We have presented here a novel method to conduct sensitivity analysis in such situations using uniformly valid estimators, which is essential in high-dimensional setting. In particular, the sensitivity analysis proposed is based on the construction of an uncertainty interval for the causal effect of interest which we show has uniformly valid coverage. Finite sample experiments confirm the asymptotic results.

We use a sensitivity model with a sensitivity parameter, a correlation, which is easy to interpret and discuss with subject-matter scientists [23]. As all sensitivity models, ours describes potential departures from the unconfoundedness assumption. If sensitivity is detected as is the case in the presented application on the

effect of smoking on birth weight, then this is important information. If no sensitivity is detected, then one might argue that this does not preclude the analysis to be sensitive to other departures from the unconfoundedness assumption. Note that our results show that we need to let ρ tend to zero with increasing sample size unless we can estimate bias due to unobserved confounding, and hence the propensity score, with a root- n convergence rate. An alternative to bias estimation is to fix the bias itself as the sensitivity parameter, at the expense of interpretability. Finally, future directions for research include generalizing our results from binary to continuous treatments, from continuous to binary outcome, as well as considering ultra-high dimensional situations ($p \gg n$), which have their own challenges [39].

6 Supplementary material

Supplementary material contains additional simulation results and is available online.

Acknowledgments: We are grateful to Minna Genbäck, Mohammad Ghasempour, and anonymous reviewers for their helpful comments.

Funding information: Funding from the Marianne and Marcus Wallenberg Foundation and the Swedish Research Council for Health, Working Life and Welfare is acknowledged.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results, and approved the final version of the manuscript. The study and manuscript have received significant contributions from all authors, with NM having the larger contribution overall.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The data analysed in this paper is available in the online supplementary material for Nabi et al. [24]. The code for the simulation study is available at https://github.com/mousavin0/simulations_validui.

References

- [1] Fisher R. Cigarettes, cancer, and statistics. *Centen Rev Arts Sci.* 1958;2:151–66.
- [2] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educat Psychol.* 1974;66(5):688–701.
- [3] Rubin DB. Formal mode of statistical inference for causal effects. *J Stat Plan Inference.* 1990;25(3):279–92.
- [4] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: Recent evidence and a discussion of some questions. *J Nat Cancer Inst.* 1959;22(1):173–203.
- [5] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Amer Stat Assoc.* 1994;89(427):846–66.
- [6] Van der Laan MJ. Targeted learning: causal inference for observational and experimental data. New York: Springer Science & Business Media; 2011.
- [7] Farrell MH. Robust inference on average treatment effects with possibly more covariates than observations. *J Econ.* 2015;189(1):1–23.
- [8] Van der Laan MJ, Gruber S. Collaborative double robust targeted maximum likelihood estimation. *Int J Biostat.* 2010;6(1):17.
- [9] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. *Econ J.* 2018;21(1):C1–68.
- [10] Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection among high-dimensional controls. *Rev Econ Stud.* 2014;81(2):608–50.
- [11] Moosavi N, Häggström J, de Luna X. The costs and benefits of uniformly valid causal inference with high-dimensional nuisance parameters. *Stat Sci.* 2023;38(1):1–12.

- [12] Vansteelandt S, Goetghebeur E, Kenward MG, Molenberghs G. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat Sin.* 2006;16(3):953–79.
- [13] Genbäck M, de Luna X. Causal inference accounting for unobserved confounding after outcome regression and doubly robust estimation. *Biometrics.* 2019;75(2):506–15.
- [14] Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika.* 1987;74(1):13–26.
- [15] Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass).* 2016;27(3):368–77.
- [16] Franks A, D’Amour A, Feller A. Flexible sensitivity analysis for observational studies without observable implications. *J Am Stat Assoc.* 2020;115(532):1730–46.
- [17] Bonvini M, Kennedy EH. Sensitivity analysis via the proportion of unmeasured confounding. *J Amer Stat Assoc.* 2022;117(539):1540–50.
- [18] Zhang B, Tchetgen Tchetgen EJ. A semi-parametric approach to model-based sensitivity analysis in observational studies. *J R Stat Soc Ser A.* 2022;185(S2):668–91.
- [19] Zhao Q, Small DS, Bhattacharya BB. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J R Stat Soc Ser B.* 2019;81(4):735–61.
- [20] Gabriel EE, Sjölander A, Sachs MC. Nonparametric bounds for causal effects in imperfect randomized experiments. *J Amer Stat Assoc.* 2023;118(541):684–92.
- [21] Copas JB, Li HG. Inference for non-random samples. *J R Stat Soc Ser B (Stat Methodol).* 1997;59(1):55–95.
- [22] Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods.* 2010;15(4):309–34.
- [23] Cinelli C, Hazlett C. Making sense of sensitivity: extending omitted variable bias. *J R Stat Soc Ser B Stat Methodol.* 2019;82(1):39–67.
- [24] Nabi R, Bonvini M, Kennedy EH, Huang MY, Smid M, Scharfstein DO. Semiparametric sensitivity analysis: unmeasured confounding in observational studies. *Biometrics.* 2024;80(4):ujae106.
- [25] Scharfstein D, Rotnitzky A, Robins J. Rejoinder to comments on “adjusting for non-ignorable drop-out using semiparametric non-response models?”. *J Amer Stat Assoc.* 1999;94:1121–46.
- [26] Farrell MH. Robust inference on average treatment effects with possibly more covariates than observations. 2018. arXiv:13094686v3.
- [27] Kennedy EH. Semiparametric theory and empirical processes in causal inference. In: He H, Wu P, Chen D-G, editors. *Statistical causal inferences and their applications in public health research.* Cham: Springer; 2016. p. 141–67.
- [28] Ghasempour M, Moosavi N, de Luna X. Convolutional neural networks for valid and efficient causal inference. *J Comput Graph Stat.* 2024;33(2):714–23.
- [29] Luedtke AR, Diaz I, van der Laan MJ. The statistics of sensitivity analyses. UC Berkeley Division of Biostatistics Working Paper Series Working Paper 341. 2015.
- [30] Gustafson P, McCandless LC. When is a sensitivity parameter exactly that? *Stat Sci.* 2018;33(1):86–95.
- [31] Hines O, Dukes O, Diaz-Ordaz K, Vansteelandt S. Demystifying statistical learning based on efficient influence functions. *Am Stat.* 2022;76(3):292–304.
- [32] Gorbach T, de Luna X. Inference for partial correlation when data are missing not at random. *Stat Probab Lett.* 2018;141:82–9.
- [33] Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodological).* 1996;58(1):267–88.
- [34] R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2019. Available from: <https://www.R-project.org/>.
- [35] Chernozhukov V, Hansen C, Spindler M. hdm: high-dimensional metrics. *R J.* 2016;8(2):185–99.
- [36] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Soft.* 2010;33(1):1–22.
- [37] Almond D, Chay KY, Lee DS. The costs of low birth weight. *Q J Econ.* 2005;120(3):1031–83.
- [38] Cattaneo MD. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *J Econ.* 2010;155(2):138–54.
- [39] Tang D, Kong D, Pan W, Wang L. Ultra-high dimensional variable selection for doubly robust causal inference. *Biometrics.* 2023;79(2):903–14.
- [40] Bishop YM, Fienberg SE, Holland PW. *Discrete multivariate analysis: theory and practice.* New York: Springer Science & Business Media; 2007.

Appendix

We use the notations $f(X) = E(Y(1)|X)$, $\tilde{\sigma}^2 = E_{n_i}[(Y(1)_i - m(X_i))^2]^{1/2}$ and $v = E((Y(1) - m(X))^2|T = 1)$. Furthermore, we use $a \lesssim_p b$ to denote $a = O_p(b)$.

Assumption A1. (Part of Assumption 2 in Farrell [26]) Let $U = Y(1) - m(X)$. P_n obey the following conditions, with bounds uniform in n .

- $E[|U|^4|X] \leq \mathcal{U}^4$.
- For some $r > 0$: $E[|m(X_i)|^{2+2r}]$ and $E[|u_i|^{4+r}]$ are bounded.

Assumption A2. We have $E(\lambda^4(\hat{g}(X))) < \infty$, $E(g^4(X)) < \infty$, $E(g^2(X)\lambda^2(g(X))) < \infty$, $E(\lambda^4(g(X))) < \infty$, and $E(\lambda^2(\hat{g}(X))\lambda^2(g(X))) < \infty$.

Assumption A3. Assume $E((Y(1) - f(X))^4) < \infty$, $E(\lambda^2(g(X))) < \infty$, $E((Y(1) - \hat{m}(X))^4) < \infty$, $E(g(X)\lambda(g(X))|T = 1) < \infty$, and $E(\lambda^2(g(X))|T = 1) < \infty$.

Assumption A4. Assume $E((Y(1) - f(X))^4) < \infty$, $E((Y(1) - \hat{m}(X))^4) < \infty$, $E(g^2(X)\lambda^2(g(X))) = O_p(1)$, $E(\lambda^4(g(X))) = O_p(1)$, $\lambda(\hat{g}(X)) = O_p(1)$, and $g(X) = O_p(1)$.

We frequently use the following lemma, which is a direct result of Bishop et al. [40, Theorem 14.1–1]. The lemma is used to translate some regularity conditions in the form of order in probability to moment conditions.

Lemma A1. Assume that $\text{var}(A_i) < C$ across i and n , then $E_n[A_i] - E(A_i) = O_p(n^{-1/2})$.

A.1 Proof of Theorem 1

Suppose that Assumptions 1, A1, and 2 hold. By Farrell [7, Theorem 3(1)], we have

$$\sqrt{n} \left[E_n \left[\frac{T_i(Y_i - \hat{m}(X_i))}{\hat{e}(X_i)} + \hat{m}(X_i) \right] - \tau^- \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\frac{T_i(Y_i - m(X_i))}{e(X_i)} + m(X_i) - \tau^- \right] + o_p(1).$$

In the aforementioned representation, unlike [7], the expectation $m(X) = E(Y(1)|X, T = 1)$ is not necessarily equal to $E(Y(1)|X)$ and, therefore, the asymptotic expectation of the AIPW estimator, denoted by τ^- , can be different from τ .

A.2 Proof of Theorem 2

The steps in the proof follow Genbäck and de Luna [13]. However, the parametric modelling assumptions are dropped here. By Farrell [7, Theorem 2], we have that under the consistency of one of the nuisance models and other regularity conditions specified $\hat{\tau}_{\text{AIPW}} \xrightarrow{P} E(m(X))$. Also,

$$\begin{aligned} E(\xi|X, T = 1) &= E(\rho\sigma\eta + \varepsilon|X, T = 1) \\ &= \rho\sigma E(\eta|\eta > -g(X)) \\ &= \rho\sigma\lambda(g(X)). \end{aligned} \tag{A1}$$

Using (A1) and Assumption 3, $\tau^- = E(E(Y(1)|X, T = 1)) = \tau + \rho\sigma E(\lambda(g(X)))$.

A.3 Asymptotic results for estimation using uncorrected variance estimator

Theorem A1. *Let Assumptions 1–3 hold. Moreover, let Assumptions A1 and A3 hold and assume $\sqrt{n}\rho^3 = o(1)$, $\sqrt{n}\rho E_n[\hat{g}(X_i) - g(X_i)] = o_p(1)$, and $\sqrt{n}\rho E_n[(m(X_i) - \hat{m}(X_i))^2]^{1/2} = o_p(1)$. We have:*

- $\sqrt{n}((\hat{\tau}_{AIPW} - \hat{b}) - \tau) = \sum_{i=1}^n \Psi_i / \sqrt{n} + o_p(1)$,
- $V^{-1/2} \sqrt{n}((\hat{\tau}_{AIPW} - \hat{b}) - \tau) \rightarrow_d \mathcal{N}(0, 1)$,
- $\hat{V} - V = o_p(1)$,

where $V = E(\Psi_i^2)$ and $\hat{V} = E_n\left[\frac{T_i(X_i - \hat{m}(X_i))^2}{\hat{e}(X_i)^2}\right] + E_n[(\hat{m}(X_i) - \hat{\tau}_{AIPW})^2]$.

Proof. If we show that $\sqrt{n}(\hat{b} - b) = o_p(1)$, the asymptotic linearity result is a direct result of Theorem 1. We have

$$\sqrt{n}(\hat{b} - b) = \sqrt{n}\rho\hat{\sigma}E_n[\lambda(\hat{g}(X_i)) - \lambda(g(X_i))] + \sqrt{n}\rho E_n[\lambda(g(X_i))](\hat{\sigma} - \sigma) + \sqrt{n}\rho\sigma\{E_n[\lambda(g(X_i))] - E(\lambda(g(X_i)))\}.$$

For the first term, we have

$$\sqrt{n}\rho\hat{\sigma}E_n[\lambda(\hat{g}(X_i)) - \lambda(g(X_i))] \leq_p \sqrt{n}\rho E_n[\hat{g}(X_i) - g(X_i)] = o_p(1),$$

where the inequality in probability is derived using Assumption A3, Lemma A1, Lipschitz continuity of inverse Mills ratio, while the equality follows from assumptions on ρ in Theorem A1. For the third term, using Lemma A1 and Assumption A3, one can show that $\sqrt{n}\rho\sigma\{E_n[\lambda(g(X_i))] - E(\lambda(g(X_i)))\} = o_p(1)$. This shows that the first and third terms in the aforementioned decomposition are negligible. Therefore, using some moment conditions in Assumption A3, we have

$$\begin{aligned} \sqrt{n}(\hat{b} - b) &\leq_p \sqrt{n}\rho E_n[\lambda(g(X_i))](\hat{\sigma} - \sigma) \\ &\leq_p \sqrt{n}\rho(\hat{\sigma}^2 - \sigma^2) \\ &\leq_p \sqrt{n}\rho(\hat{\sigma}^2 - \tilde{\sigma}^2)(= \mathbf{R}_{21}) + \sqrt{n}\rho(\tilde{\sigma}^2 - v)(= \mathbf{R}_{22}) + \sqrt{n}\rho(v - \sigma^2)(= \mathbf{R}_{23}). \end{aligned}$$

For \mathbf{R}_{21} , using Assumption A3 and a condition on ρ stated, we have

$$\begin{aligned} \sqrt{n}\rho(\hat{\sigma}^2 - \tilde{\sigma}^2) &\leq_p \sqrt{n}\rho(\hat{\sigma} - \tilde{\sigma}) \\ &\leq_p \sqrt{n}\rho E_n[(Y(1)_i - m(X_i))^2]^{1/2} + \sqrt{n}\rho E_n[(m(X_i) - \hat{m}(X_i))^2]^{1/2} \\ &\quad - \sqrt{n}\rho E_n[(Y(1)_i - m(X_i))^2]^{1/2} = o_p(1), \end{aligned}$$

where the second inequality in probability holds by Minkowski inequality.

For \mathbf{R}_{22} , using $\rho = o(1)$, Assumptions 1 and A3, we have

$$\begin{aligned} \sqrt{n}\rho(\tilde{\sigma}^2 - v) &= \sqrt{n}\rho\left(\frac{n}{n_t}E_n[T_i(Y(1)_i - m(X_i))^2] - v\right) \\ &\leq_p \sqrt{n}\rho\left(\frac{n}{n_t}E_n[T_i(Y(1)_i - m(X_i))^2] - P(T=1)^{-1}E_n[T_i(Y(1)_i - m(X_i))^2]\right) \\ &\quad + \sqrt{n}\rho(P(T=1)^{-1}E_n[T_i(Y(1)_i - m(X_i))^2] - P(T=1)^{-1}E(T(Y(1) - m(X))^2)) \\ &\leq_p \sqrt{n}\rho(E_n[T_i]^{-1} - P(T=1)^{-1}) \\ &\quad + \sqrt{n}\rho(E_n[T_i(Y(1)_i - m(X_i))^2] - E(T(Y(1) - m(X))^2)) = o_p(1). \end{aligned}$$

For \mathbf{R}_{23} , we have

$$\begin{aligned} E((Y(1) - m(X))^2|T=1) &= E((Y(1) - f(X) - \rho\sigma\lambda(g(X)))^2|T=1) \\ &= E((Y(1) - f(X))^2 + \rho^2\sigma^2\lambda^2(g(X)) - 2\rho\sigma(Y(1) - f(X))\lambda(g(X))|T=1) \\ &= E((Y(1) - f(X))^2|T=1) + \rho^2\sigma^2E(\lambda^2(g(X))|T=1) \\ &\quad - 2\rho\sigma E((Y(1) - f(X))\lambda(g(X))|T=1), \end{aligned}$$

where

$$E((Y(1) - f(X))^2|T = 1) = -\sigma^2\rho^2E(g(X)\lambda(g(X))|T = 1) + \sigma^2,$$

using equation (A.2) in [32] and

$$\begin{aligned} 2\rho\sigma E((Y(1) - f(X))\lambda(g(X))|T = 1) &= 2\rho\sigma E(\xi\lambda(g(X))|T = 1) \\ &= 2\rho\sigma E((\rho\sigma\eta + \varepsilon)\lambda(g(X))|T = 1) \\ &= 2\rho^2\sigma^2 E(\eta\lambda(g(X))|T = 1) + 2\rho\sigma E(\varepsilon\lambda(g(X))|T = 1) \\ &= 2\rho^2\sigma^2 E(\lambda(g(X))E(\eta|X, T = 1)|T = 1) + 2\rho\sigma E(\lambda(g(X))E(\varepsilon|X, T = 1)|T = 1) \\ &= 2\rho^2\sigma^2 E(\lambda^2(g(X))|T = 1). \end{aligned}$$

Therefore,

$$E((Y(1) - m(X))^2|T = 1) = \sigma^2(1 - \rho^2E(g(X)\lambda(g(X))|T = 1) - \rho^2E(\lambda^2(g(X))|T = 1)).$$

Finally, using assumptions on ρ and Assumption A3, we have the following for \mathbf{R}_{23} :

$$\sqrt{n}\rho(v - \sigma^2) = \sqrt{n}\rho^3E(g(X)\lambda(g(X))|T = 1) - \sqrt{n}\rho^3\sigma^2E(\lambda^2(g(X))|T = 1) = o_p(1).$$

Theorem A1 (b) is a direct result of Theorem A1 (a) and the moment condition in Assumption 2 (c).

Theorem A1 holds based on the proof of Theorem 3.3 in Farrell [7], Assumptions 1 and 2, and Theorem A1 (b). \square

As a corollary, the following 95% confidence interval for τ given ρ is uniformly valid.

Corollary A1. *For each n , let \mathcal{P}_n be the set of distributions obeying the assumptions of Theorem A1. Then, we have:*

$$\sup_{P \in \mathcal{P}_n} |\text{pr}_P(\tau \in \{(\hat{\tau}_{\text{AIPW}} - \hat{b}) \pm c_\alpha \sqrt{\hat{V}/n}\}) - (1 - \alpha)| \rightarrow 0,$$

where $c_\alpha = \Phi^{-1}(1 - \alpha/2)$.

Proof. The corollary follows from Theorem A1 [see 7, Corollary 2]. \square

A.4 Proof of Theorem 3

Step 1. To prove the consistency of the variance estimator, we first find the limit of the term in the numerator of the variance estimator. By the triangle inequality, we have

$$\begin{aligned} &|E_{n_t}^{1/2}[(Y(1)_i - m(X_i))^2] - E_{n_t}^{1/2}[(m(X_i) - \hat{m}(X_i))^2]| \\ &\leq E_{n_t}^{1/2}[(Y(1)_i - \hat{m}(X_i))^2] \\ &\leq E_{n_t}^{1/2}[(Y(1)_i - m(X_i))^2] + E_{n_t}^{1/2}[(m(X_i) - \hat{m}(X_i))^2], \end{aligned}$$

where using Assumption 1 (which implies $n/n_t = O_p(1)$) and consistency of $\hat{m}(X)$ we have

$$E_{n_t}^{1/2}[(m(X_i) - \hat{m}(X_i))^2] \leq \frac{n}{n_t} E_n^{1/2}[T_i(m(X_i) - \hat{m}(X_i))^2] \xrightarrow{P} 0.$$

To bound $E_{n_t}^{1/2}[(Y(1)_i - \hat{m}(X_i))^2]$ by the squeeze theorem, we just need to find the limit of $E_{n_t}^{1/2}[(Y(1)_i - m(X_i))^2]$. We have

$$\begin{aligned} E_{n_t}^{1/2}[(Y(1)_i - m(X_i))^2] &= \frac{n}{n_t} E_n[T_i(Y(1)_i - m(X_i))^2] \\ &\xrightarrow{P} (P(T = 1))^{-1} E((Y(1) - m(X))^2|T = 1) P(T = 1) \\ &= E((Y(1) - m(X))^2|T = 1), \end{aligned}$$

where the convergence is the result of Lemma A1 and Assumption A2.

Note that based on the proof of Theorem A1 for \mathbf{R}_{23} , the limit found earlier has the following relationship with the true parameter σ^2 :

$$E((Y(1) - m(X))^2|T = 1) = \sigma^2(1 - \rho^2 E(g(X)\lambda(g(X))|T = 1) - \rho^2 E(\lambda^2(g(X))|T = 1)).$$

Step 2. It now remains to show that

$$E_{n_t}[\hat{g}(X_i)\lambda(\hat{g}(X_i))] - E_{n_t}[\lambda^2(\hat{g}(X_i))] \xrightarrow{p} E(g(X)\lambda(g(X))|T = 1) - E(\lambda^2(g(X))|T = 1).$$

We have

$$\begin{aligned} & E_{n_t}[\hat{g}(X_i)\lambda(\hat{g}(X_i))] - E(g(X)\lambda(g(X))|T = 1) \\ &= \frac{n}{n_t} E_n[T\hat{g}(X_i)\lambda(\hat{g}(X_i))] - \frac{n}{n_t} E_n[Tg(X_i)\lambda(\hat{g}(X_i))] \\ & \quad + \frac{n}{n_t} E_n[Tg(X_i)\lambda(\hat{g}(X_i))] - \frac{n}{n_t} E_n[Tg(X_i)\lambda(g(X_i))] \\ & \quad + \frac{n}{n_t} E_n[Tg(X_i)\lambda(g(X_i))] - E(g(X)\lambda(g(X))|T = 1) \\ &\leq \frac{n}{n_t} E_n^{1/2}[\lambda^2(\hat{g}(X_i))] E_n^{1/2}[T(\hat{g}(X_i) - g(X_i))^2] \\ & \quad + \frac{n}{n_t} E_n^{1/2}[g^2(X_i)] E_n^{1/2}[T(\lambda(\hat{g}(X_i)) - \lambda(g(X_i)))^2] \\ & \quad + \frac{n}{n_t} E_n[Tg(X_i)\lambda(g(X_i))] - E(g(X)\lambda(g(X))|T = 1) \xrightarrow{p} 0, \end{aligned}$$

where the inequality is due to the Cauchy–Schwarz inequality and the convergences follows from Assumption 1, consistency assumption for $\hat{g}(X)$, Lipschitz continuity of the inverse Mills ratio $\lambda(\cdot)$, Lemma A1 and Assumption A2. Similarly, we have

$$\begin{aligned} & E_{n_t}[\lambda^2(\hat{g}(X_i))] - E(\lambda^2(g(X))|T = 1) \\ &= \frac{n}{n_t} (E_n[\lambda^2(\hat{g}(X_i))] - E_n[\lambda^2(g(X_i))] + E_n[\lambda^2(g(X_i))]) - E(\lambda^2(g(X))|T = 1) \\ &= \frac{n}{n_t} E_n[(\lambda(\hat{g}(X_i)) - \lambda(g(X_i)))(\lambda(\hat{g}(X_i)) + \lambda(g(X_i)))] + \frac{n}{n_t} E_n[\lambda^2(g(X_i))] - E(\lambda^2(g(X))|T = 1) \\ &\leq \frac{n}{n_t} E_n^{1/2}[(\lambda(\hat{g}(X_i)) - \lambda(g(X_i)))^2] E_n^{1/2}[(\lambda(\hat{g}(X_i)) + \lambda(g(X_i)))^2] + \frac{n}{n_t} E_n[\lambda^2(g(X_i))] - E(\lambda^2(g(X))|T = 1) \xrightarrow{p} 0. \end{aligned}$$

Step 3. The consistency of the causal parameter estimator can be shown by the consistency of the variance estimator $\hat{\sigma}_c^2$, Slutsky's theorem and $E_n[\lambda(\hat{g}(X_i))] \xrightarrow{p} E(\lambda(g(X_i)))$. The latter holds based on consistency of $\hat{g}(X)$, Lipschitz continuity of $\lambda(\cdot)$, Lemma A1 and Assumption A2.

A.5 Proof of Theorem 4

We have

$$\begin{aligned} \sqrt{n}(\hat{b}_c - b) &\leq_p \sqrt{n}\rho E_n[\lambda(g(X_i))](\hat{\sigma}_c - \sigma) \leq_p \sqrt{n}\rho(\hat{\sigma}_c^2 - \sigma^2) \\ &\leq_p \sqrt{n}\rho(\hat{\sigma}^2 - \bar{\sigma}^2)/(1 - \rho^2 E_n[\hat{g}(X_i)\lambda(\hat{g}(X_i))] - \rho^2 E_n[\lambda^2(\hat{g}(X_i))]) (= \mathbf{R}_{21}) \\ & \quad + \sqrt{n}\rho(\bar{\sigma}^2 - v)/(1 - \rho^2 E_n[\hat{g}(X_i)\lambda(\hat{g}(X_i))] - \rho^2 E_n[\lambda^2(\hat{g}(X_i))]) (= \mathbf{R}_{22}) \\ & \quad + \sqrt{n}\rho(v/(1 - \rho^2 E_n[\hat{g}(X_i)\lambda(\hat{g}(X_i))] - \rho^2 E_n[\lambda^2(\hat{g}(X_i))]) - \sigma^2) (= \mathbf{R}_{23}), \end{aligned}$$

where the first inequality can be shown from the first lines of the proof of Theorem A1. Moreover, both \mathbf{R}_{21} and \mathbf{R}_{22} can be shown to be $o_p(1)$ from the proof of Theorem 3 regarding the terms with the same name and under some extra moment conditions in Assumption A4. From the proof concerning the term named \mathbf{R}_{23} in Theorem

A1, we have the following. If Assumption 3 holds, we have $v/(1 - \rho^2 E(g(X_i)\lambda(g(X_i))) - \rho^2 E(\lambda^2(g(X_i)))) = \sigma^2$. To complete the proof, we have to show that

$$\sqrt{n}\rho(v/(1 - \rho^2 E_{n_i}[\hat{g}(X_i)\lambda(\hat{g}(X_i))] - \rho^2 E_{n_i}[\lambda^2(\hat{g}(X_i))]) - v/(1 - \rho^2 E(g(X_i)\lambda(g(X_i))) - \rho^2 E(\lambda^2(g(X_i)))) = o_p(1).$$

Using Assumption A4 and assumptions on ρ stated in the theorem, we have

$$\begin{aligned} \mathbf{R}_{23} &\leq_p \sqrt{n}\rho^3(E_{n_i}[\hat{g}(X_i)\lambda(\hat{g}(X_i))] + E_{n_i}[\lambda^2(\hat{g}(X_i))]) - \sqrt{n}\rho^3(E(g(X_i)\lambda(g(X_i))) - E(\lambda^2(g(X_i)))) \\ &\leq_p \sqrt{n}\rho^3 E_{n_i}[\hat{g}(X_i)\lambda(\hat{g}(X_i)) - g(X_i)\lambda(\hat{g}(X_i))] + \sqrt{n}\rho^3 E_{n_i}[g(X_i)\lambda(\hat{g}(X_i)) - g(X_i)\lambda(g(X_i))] \\ &\quad + \sqrt{n}\rho^3 E_{n_i}[\lambda^2(\hat{g}(X_i)) - \lambda^2(g(X_i))] + \sqrt{n}\rho^3(E_{n_i}[g(X_i)\lambda(g(X_i))] - E(g(X_i)\lambda(g(X_i)))) \\ &\quad + \sqrt{n}\rho^3(E_{n_i}[\lambda^2(g(X_i))] - E(\lambda^2(g(X_i)))) = o_p(1), \end{aligned}$$

which completes the proof.

A.6 Proof of Corollary 1

The corollary follows from Theorem 4 [see 7, Corollary 2].

A.7 Other parameters of interest

Theorem 1 in the article only concerns the causal parameter $E(Y(1))$. The asymptotic linearity of the AIPW estimators of the causal parameters $E(Y(0))$, $E(Y(1)|T=0)$ and $E(Y(0)|T=1)$ is shown under similar regularity conditions Farrell [26, Theorem 3–4]. Under asymptotic normality and using Assumption 3, the asymptotic confounding bias of these estimators can be found similar to the one found for $E(Y(1))$ in Theorem 2. Here, we show the proof for the parameter $\tau_{10} = E(Y(1)|T=0)$. We have

$$\begin{aligned} &\frac{1}{P(T=0)} \left[E(E((1-T)m(X)|X)) + E \left[E \left(\frac{T(Y-m(X))(1-e(X))}{e(X)} \middle| X \right) \right] \right] - \tau_{10} \\ &= \frac{1}{P(T=0)} \left[E((1-e(X))m(X)) + E \left[\frac{1-e(X)}{e(X)} E(T(Y-m(X))|X) \right] \right] - \tau_{10} \\ &= \frac{1}{P(T=0)} \left[E((1-e(X))m(X)) + E \left[\frac{1-e(X)}{e(X)} E(Y(1)-m(X)|X, T=1) \Pr(T=1|X) \right] \right] - \tau_{10} \\ &= \frac{1}{P(T=0)} [E((1-\Pr(T=1|X))E(Y(1)|X, T=1))] - \tau_{10} \\ &= \frac{1}{P(T=0)} [E(E(Y(1)|X, T=1) - \Pr(T=1|X)E(Y(1)|X, T=1))] - \tau_{10} \\ &= \frac{1}{P(T=0)} [E(E(Y(1)|X) + \rho\sigma\lambda(g(X)) - \Pr(T=1|X)E(Y(1)|X, T=1))] - \tau_{10} \\ &= \frac{1}{P(T=0)} [\rho\sigma E(\lambda(g(X))) + \Pr(T=0|X)E(Y(1)|X, T=0)] - \tau_{10} \\ &= \frac{\rho\sigma E(\lambda(g(X)))}{P(T=0)} + \frac{E(\Pr(T=0|X)E(Y(1)|X, T=0))}{P(T=0)} - \tau_{10} \\ &= \frac{\rho\sigma E(\lambda(g(X)))}{P(T=0)} + \frac{E(E((1-T)Y(1)|X))}{P(T=0)} - \tau_{10} \\ &= \frac{\rho\sigma E(\lambda(g(X)))}{P(T=0)} + \frac{E((1-T)Y(1))}{P(T=0)} - \tau_{10} \\ &= \frac{\rho\sigma E(\lambda(g(X)))}{P(T=0)} + \frac{P(T=0)E(Y(1)|T=0)}{P(T=0)} - \tau_{10} \frac{\rho\sigma E(\lambda(g(X)))}{P(T=0)}. \end{aligned}$$

For estimating the confounding bias at a given ρ value, we use $\rho\hat{\sigma}_c E_n[\lambda(\hat{g}(X_i))]/E_n[T_i]$.