

## Research Article

Paul Hünermund\*, Beyers Louw, and Itamar Caspi

# Double machine learning and automated confounder selection: A cautionary tale

<https://doi.org/10.1515/jci-2022-0078>

received November 28, 2022; accepted March 22, 2023

**Abstract:** Double machine learning (DML) has become an increasingly popular tool for automated variable selection in high-dimensional settings. Even though the ability to deal with a large number of potential covariates can render selection-on-observables assumptions more plausible, there is at the same time a growing risk that endogenous variables are included, which would lead to the violation of conditional independence. This article demonstrates that DML is very sensitive to the inclusion of only a few “bad controls” in the covariate space. The resulting bias varies with the nature of the theoretical causal model, which raises concerns about the feasibility of selecting control variables in a data-driven way.

**Keywords:** double/debiased machine learning, bad controls, backdoor adjustment, collider bias, causal hierarchy

**MSC 2020:** 62D20

## 1 Introduction

Machine learning approaches for selecting suitable control variables to establish causal identification in high-dimensional settings are gaining increasing attention [1,2]. Besides the evident benefits of automation for the analysis of high-dimensional data, this rising popularity can be explained by two specific advantages that applied researchers attribute to these methods. First, a mostly data-driven, automated procedure of model selection allows us to systematize the research process and make it more transparent [3]. Second, the ability to consider a large number of covariates – possibly larger than the sample size – could render selection-on-observables types of identification assumptions more plausible [4]. For these reasons, automated variable selection has seen several recent applications in economics [5–7], finance [8], political science [9,10], and organizational studies [11], as well as the introduction of dedicated open-source software libraries in *R* and *Python* [12,13].<sup>1</sup>

Double/debiased machine learning (DML) is a method developed to use regularized regression techniques, such as least absolute shrinkage and selection operator (LASSO) [14] or  $l_2$ -boosting [15], for variable selection in a high-dimensional causal inference setting [4]. Compared to standard regularization on a single outcome equation, it seeks variables that are highly correlated with both treatment *and* outcome,

---

<sup>1</sup> See, for example, the vignette of the *R*-package *hdm*, which presents automated variable selection as a main application for illustrating the usefulness of DML approaches.

---

\* Corresponding author: Paul Hünermund, Copenhagen Business School, Kilevej 14A, Frederiksberg, 2000, Denmark, e-mail: phu.si@cbs.dk

Beyers Louw: Maastricht University, Tongersestraat 53, 6211 LM Maastricht, Netherlands, e-mail: jb.louw@maastrichtuniversity.nl

Itamar Caspi: Bank of Israel, P.O. Box 780, 91007, Jerusalem, Israel, e-mail: itamar.caspi@boi.org.il

which immunizes the procedure against small approximation errors that inevitably arise when selecting among a large set of covariates. Consider the following system of partially linear equations

$$y = \theta_0 d + g_0(x) + u, \quad (1)$$

$$d = m_0(x) + v, \quad (2)$$

with a primary interest in the causal effect  $\theta_0$  of a treatment  $D$  on outcome  $Y$ . The vector  $X = (X_1, \dots, X_p)$  consists of a set of covariates and  $(U, V)$  are two disturbances with a zero conditional mean. In settings where  $X$  is high-dimensional and  $g_0(\cdot)$  and  $m_0(\cdot)$  are approximately linear and sparse, meaning that only a few elements of  $X$  are important for predicting the treatment and outcome, regularization can be applied to automatically select the most suitable among a large set of potential control variables.

Yet, a naïve application of regularization to equation (1) can lead to substantial omitted variable bias (OVB), as it only selects variables that are highly correlated with the outcome  $Y$  but not with the treatment  $D$ . The naïve approach, therefore, generally does not result in a root- $N$  consistent estimator for the structural parameter  $\theta_0$  [2]. Two main solutions to this problem are proposed in the literature: (a) partialling out and (b) double selection, both of which consider the strength of association between  $D$  and  $X$ . The former uses regularization to estimate the residuals of the outcome equation,  $\rho^y = y - x'\pi_0^y$ , and treatment equation,  $\rho^d = d - x'\pi_0^d$ , with  $\pi_0^y$  and  $\pi_0^d$  being the respective coefficient vectors. It then finds the causal effect of interest  $\hat{\theta}$  by regressing  $\rho^y$  on  $\rho^d$  [16]. The latter solution first determines suitable predictors for  $Y$ , then similarly finds predictors for  $D$ , and finally regresses  $Y$  on the union of the selected controls. It can be shown that both approaches rely on doubly robust moment conditions and are thus insensitive to approximation errors stemming from regularization [2,17].

To causal inference scholars, it is generally well known that model-free covariate selection is a theoretical impossibility – a fact that was conceptualized by Pearl and Mackenzie under the rubric of the *ladder of causation* [18] and recently proven by Bareinboim *et al.* [19]. From this vantage point, the DML research program appears puzzling. If the starting point is a standard textbook regression equation in which each variable  $X_k$  is exogenous and the number of parameters  $p$  is allowed to grow large, then the variable selection is obviously feasible. Identification is achieved by assumption, and the only task left for the machine learning algorithm is to pick the covariates with nonzero coefficients. But this ignores the problem that, in reality, not all covariates will be suitable controls.

The key identification assumption within the DML framework is *ignorability* [1,20]. Given the high-dimensional vector of control variables, treatment status is required to be conditionally independent of potential outcomes

$$Y_{D=d} \perp\!\!\!\perp D|X, \quad (3)$$

with  $Y_{D=d}$  denoting the potential outcome of  $Y$  given treatment status  $D = d$ . This assumption can easily be violated if  $X$  includes variables that are not fully exogenous. In the following, we explore the consequences of violations of ignorability due to the presence of *bad controls* in the conditioning set of the DML algorithm [21,22]. We focus on the LASSO case, which has received the most attention so far [5,7,11,23], presumably because of its appealing combination of interpretability and accuracy. However, as we will show, our arguments also apply more broadly to the use of other machine learning algorithms for automated variable selection in a causal inference setting.

In the first step, we make precise the notion of bad controls in regression analyses by building on the *backdoor criterion* from the graphical causal model literature [22,24]. We then show in simulations that DML is very sensitive to minor violations of the ignorability assumption. Depending on the exact source of endogeneity, the advantage of DML over naïve LASSO – which was one of the main motivations for developing the method – vanishes completely. This is because bad controls, although they do not necessarily exert a causal influence, are often highly correlated with the treatment or outcome (since they are related to unobservables that affect  $D$  or  $Y$ ). Therefore, bad controls are very likely to be picked up by DML, which has quantitative implications even if only a few endogenous variables are present in the conditioning set. We demonstrate this in an application of DML to the estimation of the gender wage gap using the data

provided by Blau and Kahn [25]. We find that the estimation results obtained by the original study differ in non-negligible ways compared to when marital status, which the literature identifies as being likely endogenous with respect to women's labor-force decisions, is included in the covariate space.

Our study is related to a growing literature studying the performance of DML under various practically relevant data-generating processes, of which most work has been focused on the OVB case. Wüthrich and Zhu [26] show that double selection LASSO can exhibit substantial OVB as a result of variable under-selection in finite samples, even in favorable settings such as – most relevant for this manuscript – with uncorrelated, exogenous controls. Their findings render an application of the asymptotic distribution of  $\sqrt{n}(\hat{\theta} - \theta_0)$  derived in the study by Belloni et al. [1] potentially problematic. Moreover, Chernozhukov et al. [27] derived sharp bounds on the OVB in the presence of unobserved confounders, which can be used to perform sensitivity analysis. Instead, we focus on the case with endogenous, bad controls in the conditioning set.

Our results highlight the significant pitfalls of automated, data-driven variable selection in high-dimensional settings. In particular, if numerous potential controls are considered in an attempt to justify selection-on-observables, without theoretical background knowledge to guide the choice, the likelihood that some bad controls are accidentally included in the algorithm is high. Therefore, dealing with a large covariate space in an automated fashion might do little to approximate the ignorability assumption and is instead more useful to determine a suitable functional-form specification for a *small* set of covariates, e.g., by considering higher-order polynomial terms [1,28]. We show that this problem is not prevalent for posttreatment variables or variables that are themselves considered outcome variables [11] so researchers cannot rely on simple rules of thumb for variable inclusion. By contrast, each potential control requires its own careful identification argument based on domain knowledge<sup>2</sup>, which is difficult to provide if the feature space is large and ultimately undermines the purpose of automated variable selection. We stress, however, that DML has broader applications, e.g., for the estimation of high-dimensional instrumental variable models [17] and arbitrary do-calculus objects [30], as well as for data splitting to reduce over fitting. Therefore, our argument specifically applies to the case when machine learning tools are used for the purpose of confounder selection.

## 2 Preliminaries

An SCM is a 4-tuple  $\langle V, U, F, P(u) \rangle$ , where  $V = \{V_1, \dots, V_m\}$  is a set of endogenous variables that are determined in the model and  $U$  denotes a set of (exogenous) background factors.  $F$  is a set of functions  $\{f_1, \dots, f_m\}$  that assign values to the corresponding  $V_i \in V$  such that  $v_i \leftarrow f_i(pa_i, u_i)$ , for  $i = 1, \dots, m$ , and  $PA_i \subseteq V \setminus V_i$ .<sup>3</sup> Finally,  $P(u)$  is a probability function defined over the domain  $U$ .

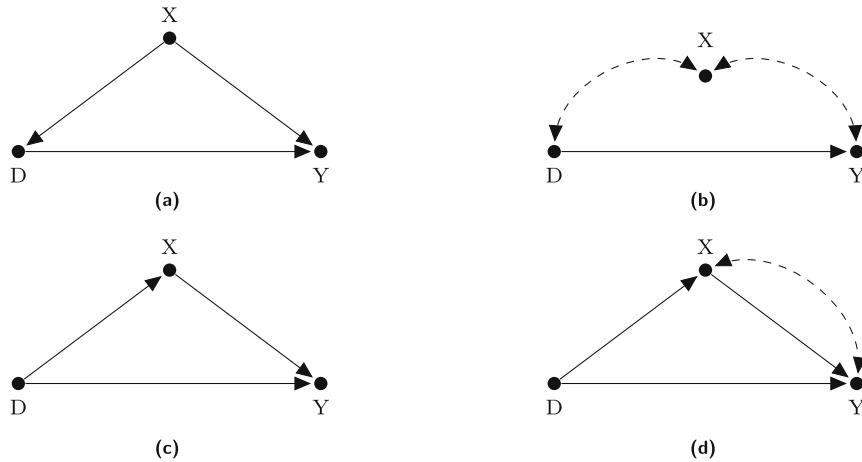
Every SCM defines a directed graph  $\mathcal{G} \equiv (V, E)$ , where  $V$  is the set of endogenous variables, denoted as nodes (vertices) in the graph, and  $E$  is a set of edges (links) pointing from  $PA_i$  (the set of parent nodes) to  $V_i$ . An example is shown in Figure 1(a), which corresponds to the SCM

$$\begin{aligned} x &\leftarrow f_1(u_1), \\ d &\leftarrow f_2(x, u_2), \\ y &\leftarrow f_3(d, x, u_3). \end{aligned} \tag{4}$$

Unobserved parents nodes induce a correlation between background factors in  $U$ . This is depicted by bidirected dashed arcs in the graph, which render the causal model *semi-Markovian* [32, p. 30]. Figure

<sup>2</sup> Early econometricians such as Tjalling Koopmans were of course well aware of this fact: "Without resort to theory [...] conclusions relevant to the guidance of economic policies cannot be drawn" [29].

<sup>3</sup> The SCM literature use assignment operators instead of equations to capture the asymmetric nature of causal relationships [31].



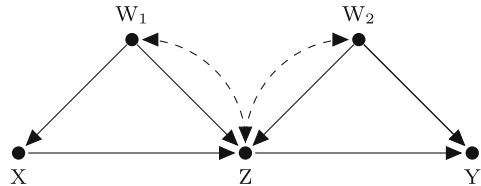
**Figure 1:** Directed acyclic graphs representing different structural causal models. (a) Good control, (b) M-graph, (c) mediator, and (d) confounded mediator.

1(b) depicts an example where the background factors of  $X$  and  $D$ , as well as  $X$  and  $Y$ , are correlated due to the presence of common influence factors that remain unobservable to the analyst.

A sequence of edges connecting two nodes in  $\mathcal{G}$  is called a *path*. Paths can be either undirected or directed (i.e., following the direction of arrowheads). Since edges correspond to stimulus response relations between variables in the underlying SCM [33], directed paths represent the direction of causal influence in the graph. Due to the notion of causality being asymmetric [34,35], directed cycles (i.e., loops from a node back to itself) are excluded to rule out that a variable can be an instantaneous cause of itself. This assumption renders  $\mathcal{G}$  acyclic.

A semi-Markovian causal graph  $\mathcal{G}$  allows us to decompose the distribution of the observed variables according to the factorization:  $P(v) = \sum_u \prod_i P(v_i|pa_i, u_i)P(u)$  [32]. The close connection between the topology of  $\mathcal{G}$  and the probabilistic relationships – in particular conditional independence relations – between the variables that represent its nodes is further exemplified by the *d-separation* criterion [36]. Consider three disjoint sets of variables,  $X$ ,  $Y$ , and  $Z$ , in a directed acyclic graphs (DAG). These sets can either be connected via a (causal) chain,  $X \rightarrow Z \rightarrow Y$ , or a fork,  $X \leftarrow Z \rightarrow Y$ , where  $Z$  acts as a common parent of  $X$  and  $Y$ . A third possible configuration is the collider,  $X \rightarrow Z \leftarrow Y$ . In a chain and fork, conditioning on  $Z$  renders  $X$  and  $Y$  conditionally independent such that  $X \perp\!\!\!\perp Y|Z$ .  $Z$  is then said to “d-separate” or “block the path between”  $X$  and  $Y$ . In contrast, in the collider structure,  $X$  and  $Y$  are independent of the outset,  $X \perp\!\!\!\perp Y|\emptyset$ , whereas conditioning on  $Z$  (or a descendant of  $Z$ ; see [32]) would unblock the path such that  $X \not\perp\!\!\!\perp Y|Z$ .<sup>4</sup>

D-separation gives rise to testable implications of graphical causal models [37]. Consider the following DAG:



This graph implies four d-separation relations between observed variables in the model:  $W_1 \perp\!\!\!\perp W_2$ ,  $X \perp\!\!\!\perp W_2$ ,  $X \perp\!\!\!\perp Y|W_2, Z$ , and  $Y \perp\!\!\!\perp W_1|W_2, Z$ . They can be tested in the data with the help of a suitable conditional independence test, and if rejected, the hypothesized causal model can be discarded and refined.

<sup>4</sup> Note that these d-separation relations hold for any distribution  $P(v)$  over the variables in the model, in particular, irrespective of any specific functional-form assumptions for  $f_i$  and any distributional assumptions for  $P(u)$  [31].

Causal effects are defined in terms of interventions in the SCM, denoted by the  $\text{do}(\cdot)$ -operator [24,33,38]. For example, the intervention  $\text{do}(D = d')$  in equation (4) entails deleting the function  $f_2(\cdot)$ , which normally assigns values to  $D$ , from the model and replacing it with the constant value  $d'$ . The target is then to estimate the post-intervention distribution of the outcome variable,  $P(Y = y|\text{do}(D = d'))$ , which results from this manipulation. Other quantities, such as the average causal effect (ACE) of a discrete change in treatment from  $d'$  to  $d''$ , can then be computed by taking the difference in expected values:  $E(Y|\text{do}(D = d'')) - E(Y|\text{do}(D = d'))$ . However, since  $P(y|\text{do}(d))$  is not directly observable in non-experimental data, it first needs to be transformed into a probability object that does not contain any do-operator before estimation can proceed [31,39]. This constitutes the *identification* step in the graphical causal model literature [32,40].

## 2.1 Backdoor adjustment

A popular strategy to identify the ACE is to control for confounding influence factors via covariate adjustment. This strategy can be rationalized with the help of the *backdoor criterion* [24].

**Definition 1.** Given an ordered pair of treatment and outcome variables  $(D, Y)$  in a causal graph  $\mathcal{G}$ , a set  $X$  is backdoor admissible if it blocks (in the d-separation sense) every path between  $D$  and  $Y$  in the subgraph  $\mathcal{G}_D$ , which is formed by deleting all edges from  $\mathcal{G}$  that are emitted by  $D$ .

Deleting edges emitted by  $D$  from  $\mathcal{G}$  ensures that all directed causal paths between  $D$  and  $Y$  are kept open. The remaining paths are non-causal and thus create a spurious correlation between the treatment and outcome.<sup>5</sup> Consequently, a backdoor admissible set  $X$  blocks all non-causal paths between  $D$  and  $Y$ , while leaving the causal paths intact. The post-intervention distribution is then identifiable via the adjustment formula [32]

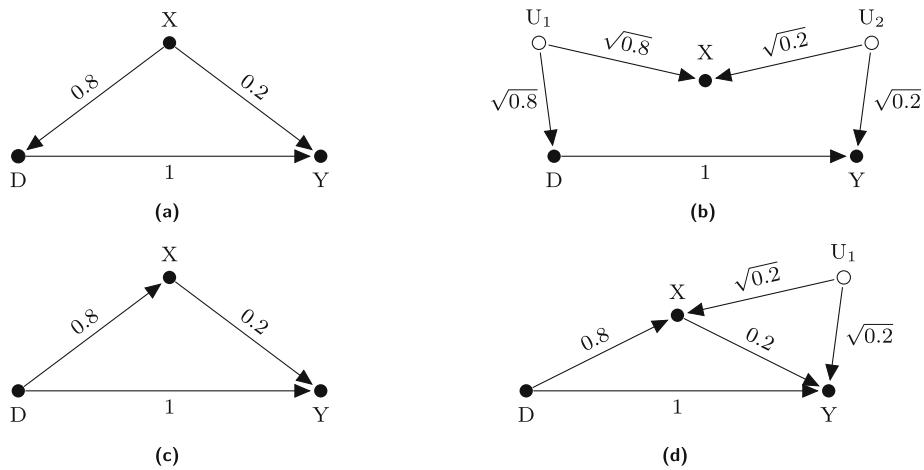
$$P(y|\text{do}(d)) = \sum_x P(y|d, x)P(x). \quad (5)$$

Since the right-hand side expression does not contain any do-operator, it can be estimated from observational data either by nonparametric methods, such as matching and inverse probability weighting, or, under additional functional-form assumptions, by parametric regression methods such as OLS.

However, following the d-separation criterion, correctly blocking backdoor paths via covariate adjustment can be intricate. Take Figure 1 as an example. In Figure 1(a), there is one causal path,  $D \rightarrow Y$ , and one backdoor path,  $D \leftarrow X \rightarrow Y$  (with  $X$  being possibly vector-valued). Following the d-separation criterion, the backdoor path can be blocked by conditioning on  $X$  so that only the causal influence of  $D$  remains. By contrast, in the other depicted cases, controlling for  $X$  would induce rather than reduce bias, thus, rendering  $X$  a *bad control* in these models. In Figure 1(b), which is known under the name of *m-graph* in the epidemiology literature [41],  $X$  exerts no causal influence on any variable in the graph. Still, there are unobserved confounders that result in a backdoor path,  $D \leftarrow \dots \rightarrow X \leftarrow \dots \rightarrow Y$ , which is already blocked however, since  $X$  acts as a collider on this path. At the same time, since  $X$  is a collider, conditioning on it (or any of its descendants) would unblock the path and therefore produce a spurious correlation. By contrast,  $X$  does not lie on a backdoor path in Figure 1(c), but acts as a mediator between  $D$  and  $Y$ . Controlling for  $X$  would allow to filter out the direct effect of the treatment,  $D \rightarrow Y$ , from its mediated portion,  $D \rightarrow X \rightarrow Y$ , [42]. However, this direct effect is generally different from the ACE, which has to be kept in mind for interpretation of results.<sup>6</sup> Moreover, such an approach is risky, because if there are

<sup>5</sup> Since these paths point into  $D$ , they are said to “enter through the backdoor.”

<sup>6</sup> In addition, following [42], identifying direct and indirect effects in a mediation setting requires the assumption of sequential ignorability, which is fulfilled in linear models with constant effects, but does not need to hold for every SCM.



**Figure 2:** Baseline parametrization of the simulations. (a) Good control, (b) M-graph, (v) Mediator, and (d) Confounded mediator.

unobserved confounders between  $X$  and  $Y$ , as depicted in Figure 1(d),  $X$  becomes a collider on the path  $D \rightarrow X \leftarrow \dots \rightarrow Y$  and would thus lead to bias if conditioned on.<sup>7</sup>

### 3 Simulation results

In the following, we present a variety of simulation results to assess the magnitude of the bias introduced by including bad controls in the DML algorithm. We focus on the high-dimensional linear setting and apply double selection DML based on  $l_1$ -regularization to automatically select covariates. However, our argument is not specific to the LASSO case. In the online supplement, we present additional simulation results using  $l_2$ -boosting, which show very similar patterns.

Since DML is specifically designed to spot variables that are mainly correlated with the treatment, which is the reason for its superior performance compared to naïve LASSO, for our baseline specification, we set a higher correlation between the controls and the treatment than with the outcome. We fix the sample size at  $n = 1,000$  and the number of covariates at  $p = 100$ . To introduce sparsity, only  $q = 10$  out of these variables are specified as having nonzero coefficients. The treatment effect  $\theta_0$  is constant and set equal to 1. All exogenous nodes (which do not receive any incoming arrows) are specified as standard normal. In the baseline, parameters are chosen in such a way that the strength, measured as the product of structural coefficients, of each path connecting the (nonzero) covariates and the treatment is equal to  $b_1 = 0.8$ . Similarly, the strength of paths connecting the covariates and the outcome is set to  $b_2 = 0.2$  (Figure 2 depicts the baseline parametrization in the form of a path diagram with associated coefficients as edge labels; hollow circles indicate unobserved variables).

Following the double selection method, we then regress  $Y$  on  $X$  using LASSO and record the variables with estimated nonzero coefficients. We do the same for a LASSO regression of  $D$  on  $X$ . Finally, we regress  $Y$  on the union of variables in  $X$  that have been picked in the preceding two LASSO regressions, this time using standard OLS. We record the estimated coefficients for the treatment effect of interest  $\hat{\theta}$  across 10,000 simulation runs. In addition, we compare double selection with the naïve (post)LASSO method, in which we repeat the previous protocol but without the second step of regressing  $D$  on  $X$ . That is in naïve LASSO variables are only selected once for the outcome regression. To summarize the estimation algorithms:

<sup>7</sup> See ref. [22] for a more comprehensive discussion of bad controls in graphical causal models that go beyond the scope of this article.

– **DML (double selection)**

- (1) Regress  $Y$  on  $X$  via LASSO and record all  $X_k$  with nonzero coefficients
- (2) Regress  $D$  on  $X$  via LASSO and record all  $X_{k'}$  with nonzero coefficients
- (3) Regress  $Y$  on the union of all  $X_k$  and  $X_{k'}$

– **Naïve (post-)LASSO**

- (1) Regress  $Y$  on  $X$  via LASSO and record all  $X_k$  with nonzero coefficients
- (2) Regress  $Y$  on all  $X_k$

Figure 3 shows simulation results using centered and studentized quantities, next to their theoretical (standard normal) distribution. In panel (a), we observe the familiar picture from [1]. DML is able to reliably filter out the good controls from irrelevant covariates, which leads to a distribution that closely matches the theoretical one. In contrast, naïve LASSO fails to pick relevant control variables that are only weakly correlated with the outcome, translating into substantial bias. However, this result reverses for the m-graph in panel (b). Here, the covariates are bad controls due to the collider structure and should not be included in the regression. They are nonetheless highly correlated with the treatment and thus get picked by the DML, leading to biased causal effect estimates. In fact, the advantage that DML had over naïve LASSO in (a) vanishes completely ( $\text{bias}^{\text{DML}} = -0.120$ , and  $\text{bias}^{\text{LASSO}} = -0.119$ ). Interestingly, given the chosen parameterization with only a moderately high correlation between the covariates and the outcome, the naïve approach consistently selects fewer bad controls than DML. The mode of the number of controls selected across simulations is 5 for the naïve LASSO and 10 for DML.

In panel (c), we investigate the mediator case. Now, the covariates are posttreatment variables, which nonetheless end up getting selected as controls by both the naïve LASSO and DML. According to the discussion in Section 2, this allows us to consistently estimate the direct effect of the treatment. However, the researcher needs to keep this change of target parameter in mind for interpretation, since both naïve LASSO and DML are unable to consistently estimate the total effect of treatment. Moreover, once we introduce a confounded mediator in panel (d), both DML and naïve LASSO perform equally poorly. The direct effect cannot be consistently estimated in this model, as neither controlling for the mediators nor leaving them out would be sufficient for identification. The total effect of treatment is likewise not estimable via DML (but would be by a simple regression of  $Y$  on  $D$ ).

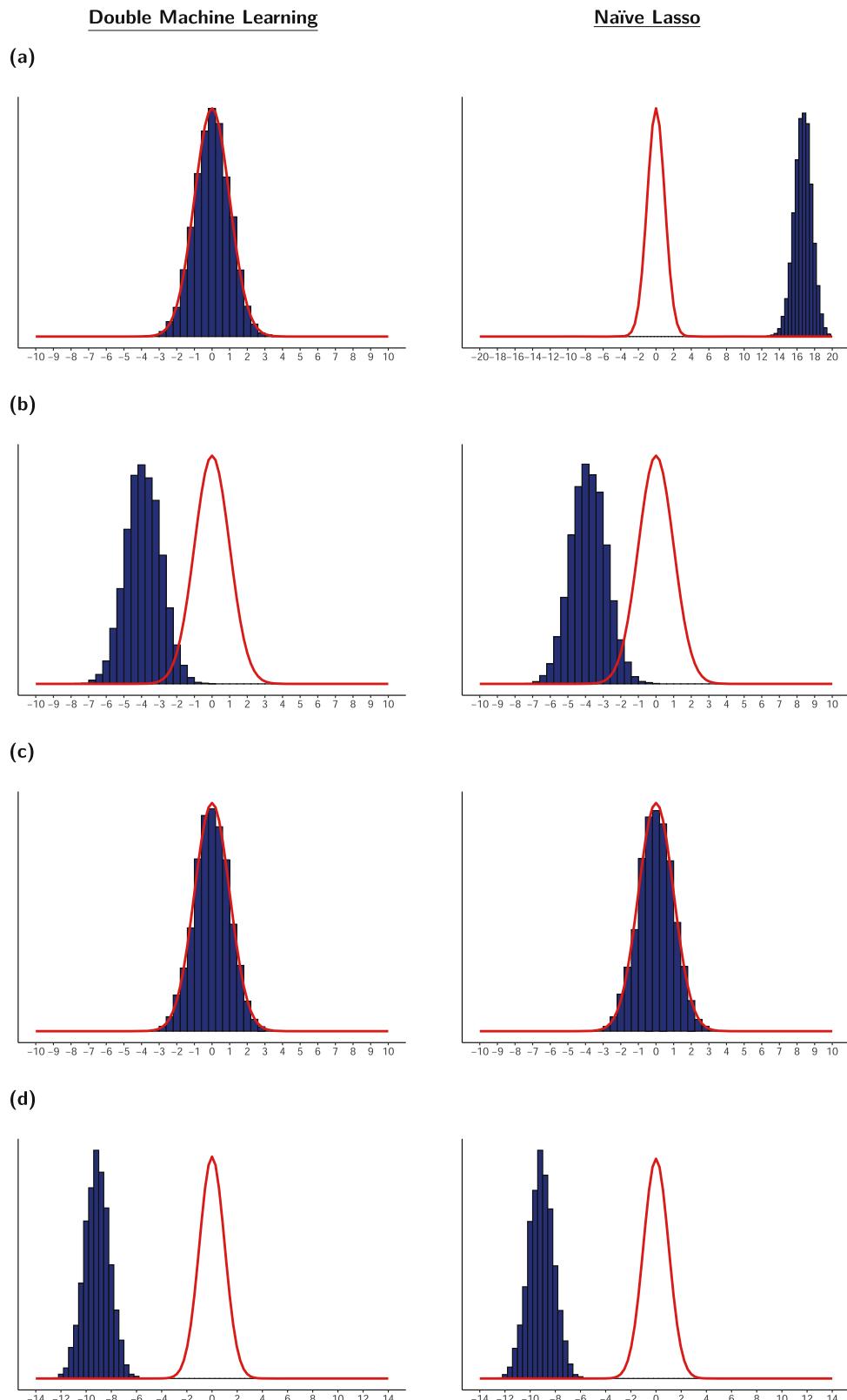
Table 1 depicts the bias obtained from DML for varying parameter constellations. In the top panel, we study performance depending on whether there is a higher strength of association between the covariates and the treatment or the outcome. For the two bad control cases, i.e., the m-graph and the confounded mediator, substantial bias arises regardless of the chosen parametrization. When taking into account the change of target parameter from the total to the direct effect, bias is low for the simple mediator model across all setups. Moreover, the DML generally performs well in the good control case, although bias becomes slightly larger when the strength of association is stronger with the outcome than the treatment (see also the  $n = 100$  case in the supplemental material).

In the bottom panel of Table 1, we vary the number of covariates with nonzero coefficients  $q$  (with  $b_1 = 0.8$ ,  $b_2 = 0.2$ , as before), while the total number of variables considered in the conditioning set remains fixed at  $p = 100$ .<sup>8</sup> Interestingly, a noticeable bias (around 5% for the m-graph and 13% for the confounded mediator) arises already with one bad control out of a hundred and increases monotonically in  $q$ . The bias for the direct effect remains low for the simple mediator model.

In the online supplement, we present additional simulations with varying  $p$  and  $n$ . In particular, we explore the case of  $n = 100$  since DML is often proposed as a technique for dealing with a large number of predictors in relatively small data sets ( $p \gg n$ ). We find results that are in line with the ones presented here in the main text.

---

<sup>8</sup> In unreported analyses, we find similar results if bad controls are mixed with good controls instead of irrelevant (zero-coefficient) ones. The two cases are conceptually similar since the DML either picks the good or leaves out the irrelevant controls, resulting in a zero-bias baseline, which then gets distorted by the selected bad controls.



**Figure 3:** Performance of DML compared to naïve LASSO for different causal models. (a) Good control, (b) M-graph, (c) Mediator (direct effect), and (d) Confounded mediator (direct effect).

**Table 1:** Bias obtained from DML under various parameter constellations ( $\theta_0 = 1$ )

$(b_1, b_2) =$	<b>(0.8, 0.2)</b>	<b>(0.6, 0.4)</b>	<b>(0.5, 0.5)</b>	<b>(0.4, 0.6)</b>	<b>(0.2, 0.8)</b>
Good control	0.000	0.000	0.000	0.000	0.000
M-graph	-0.120	-0.172	-0.179	-0.174	-0.126
Mediator	-0.001	-0.001	-0.001	0.000	0.000
Confounded mediator	-0.534	-0.480	-0.417	-0.343	-0.178
$q =$	<b>1</b>	<b>5</b>	<b>10</b>	<b>20</b>	<b>50</b>
Good control	0.000	0.000	0.000	0.000	0.000
M-graph	-0.054	-0.105	-0.120	-0.128	-0.134
Mediator	0.000	-0.001	-0.001	-0.001	-0.001
Confounded mediator	-0.134	-0.401	-0.534	-0.641	-0.728

## 4 Application

For an application to real-world data, we make use of the *Panel Study of Income Dynamics* (PSID) microdata provided by Blau and Kahn [25]. They estimate the extent of the gender wage gap in six waves of the PSID between 1981 and 2011. For their full specification, they employ a rich set of 50 control variables (as described in Section IV of their online appendix), including individual-level information on education, experience, race, occupation, unionization, and regional and industrial characteristics. However, Blau and Kahn deliberately decided to exclude marital status and number of children from their regressions, because these variables “are likely to be endogenous with respect to women’s labor-force decisions” (p. 797). Although the source of this endogeneity is not further discussed, we find it plausible that marital status acts as a confounded mediator, since it is likely influenced by the same unobserved background factors that also affect wages (Figure 4).

From the PSID data, we can infer a woman’s marital status based on whether she is recorded as “legally married wife” in her relation to the household head (men are by default indicated as household heads). Our goal is to test the sensitivity of the estimated (adjusted) gender wage gap to the inclusion of this potentially bad control. As a benchmark, we regress log wages on a female dummy and the original set of controls for each wave separately. We then employ DML using the double selection method, which allows us to include all interactions of the control variables up to degree 2. In a last step, we add marital status and its interactions to the model matrix in the DML.

Results are shown in Table 2. The estimated gender wage gaps in the OLS specifications range from  $(1 - \exp(-0.249)) \approx 22$  percentage points in 1981 to approximately 13.5 pp. in 2011. Most of the convergence between male and female wages happened in the 1980s, which coincides with the results obtained in the study by Blau and Kahn [25]. Although the DML relies on a much larger set of covariates, the results are very similar to those of OLS. However, we found greater discrepancies when marital status is included in the feature space. Across all six waves, marital status (as well as several interactions) ends up getting picked as control by the double selection DML. This has a non-negligible impact on the estimated gender wage gaps, which are 10.6% larger on average, in absolute terms, compared to the benchmark OLS. Under the assumption that marital status is a confounded mediator, larger gaps might be the result of a negative correlation between wages and the decision to get married, induced by unobservables. The respective path gets activated when marital status, as a collider, is conditioned on. Thus, the example demonstrates how having only one endogenous control within a large covariate space, paired with a flexible DML approach, can substantially affect the quantitative conclusions drawn from a study.<sup>9</sup>

<sup>9</sup> We find even larger differences if marital status is included as a single regressor, without interacting it with other covariates; see Table S4 in the online supplemental material.

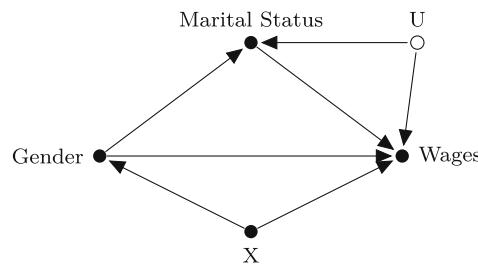


Figure 4: Causal diagram for the gender wage gap study in [25].

Table 2: Effect of gender on log wages using PSID data from [25] (standard errors in parentheses)

Wave	1981	1990	1999	2007	2009	2011
OLS	−0.249 (0.016)	−0.137 (0.014)	−0.158 (0.016)	−0.168 (0.015)	−0.157 (0.015)	−0.145 (0.016)
DML	−0.268 (0.017)	−0.139 (0.015)	−0.158 (0.016)	−0.164 (0.016)	−0.157 (0.016)	−0.136 (0.017)
DML incl.	−0.270	−0.154	−0.173	−0.190	−0.179	−0.163
Marital status	(0.022)	(0.019)	(0.020)	(0.019)	(0.020)	(0.021)

## 5 Discussion

In this article, we demonstrate the sensitivity of automated confounder selection using DML approaches to the inclusion of bad controls in the conditioning set. In our simulations, only when covariates are strictly exogenous, DML shows superior performance to naïve LASSO. In all other cases, it performs equally poorly or worse. Furthermore, our empirical application illustrates that a non-negligible bias can already occur with a small number of endogenous variables in an otherwise much larger covariate space.

These results highlight why it may be problematic to use machine learning techniques for the automatic selection of control variables in regression settings. While the ability to deal with a large set of potential controls in an automated fashion can add to the plausibility of selection-on-observable assumptions, there is an increasing chance that bad controls might be included unintentionally if the covariate space grows large. Automated approaches thus turn out to be a double-edged sword, in particular, if the number of control variables becomes so large that the researcher is unable to provide a sufficient theoretical discussion for each of them. We show that simple rules of thumb, such as restricting the conditioning set to only pretreatment variables, do not offer adequate safeguards against this problem. Indeed, as Figure 1(b) shows, our results are not limited to posttreatment variables. The intricacies of the backdoor criterion (recall, e.g., the implications of subtle differences between Figure 1(c) and (d)) imply that a vague intuition, without the guidance of a causal model, will likely be insufficient to ensure causal identification.

Because DML already assumes unconfounded covariates [2], using its ability to handle a large feature space in order to justify unconfoundedness ultimately leads to a circular argument. As long as causal inference is the goal, the analyst needs to provide a theoretical justification for the exogeneity of each of the considered control variables individually, which echoes Cartwright's [43]'s familiar adage: "no causes in, no causes out." Since this is difficult to achieve in high-dimensional settings, from a practical standpoint, smaller models that focus only on the most relevant covariates for a given context might actually be preferable.

For the purpose of automated model selection, causal discovery algorithms from the artificial intelligence literature could represent a viable alternative [44,45]. These methods do not rely on unconfoundedness and clarify the possibilities for data-driven causal learning based on a minimal set of assumptions. A key insight from this literature is that causal structures can only be learned up to a certain equivalence class

from data. As a result, the ultimate justification for a particular causal model needs to come from theoretical background knowledge [19]. The same applies to DML, which is a highly effective tool, e.g., for selecting suitable functional specifications involving a small set of controls in a data-driven way. In big data settings with a large number of potential covariates, however, DML needs to be applied carefully to avoid bad controls and ensure robust results.

**Acknowledgments:** The authors are grateful to Elias Bareinboim, Victor Chernozhukov, Jevgenij Gamper, Daniel Millimet, Judea Pearl, and seminar participants at Booking.com, Microsoft, RWTH Aachen, and Vinterd for useful comments and suggestions.

**Conflict of interest:** Prof. Paul Hünermund is an Editorial Board member of the Journal of Causal Inference but was not involved in the review process of this article.

## References

- [1] Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection among high-dimensional controls. *Rev Econ Stud.* 2014;81:608–50.
- [2] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J.* 2018;21:C1–68.
- [3] Athey S. The impact of machine learning on economics. In: Agrawal A, Gans J, Goldfarb A, editors. *The Economics of artificial intelligence: an Agenda.* Chicago, IL, USA: The University of Chicago Press; 2019.
- [4] Belloni A, Chernozhukov V, Hansen C. High-dimensional methods and inference on structural and treatment effects. *J Econ Perspect.* 2014;28(2):29–50.
- [5] Jones D, Molitor D, Reif J. What do workplace wellness programs do? Evidence from the Illinois workplace wellness study. *Quart J Econ.* 2019;134(4):1747–91.
- [6] Chang NC. Double/debiased machine learning for difference-in-differences models. *Econom J.* 2020;23(2):177–91.
- [7] Angrist JD, Frandsen B. Machine labor. *J Labor Econ.* 2022;40(S1):S97–140.
- [8] Feng G, Giglio S, Xiu D. Taming the factor zoo: a test of new factors. *J Financ.* 2020;75(3):1327–70.
- [9] Dutt P, Tsetlin I. Income distribution and economic development: Insights from machine learning. *Econ Polit.* 2018;33(1):1–36.
- [10] Blackwell M, Olson MP. Reducing model misspecification and bias in the estimation of interactions. *Polit Anal.* 2021;30(4):495–514.
- [11] Vanneste BS, Gulati R. Generalized trust, external sourcing, and firm performance in economic downturns. *Organ Sci.* 2021;33(4):1251–699.
- [12] Chernozhukov V, Hansen C, Spindler M. High-dimensional metrics in R; 2019. <https://CRAN.R-project.org/package=hdm>.
- [13] Bach P, Chernozhukov V, Kurz MS, Spindler M. DoubleML - an object-oriented implementation of double machine learning in python. *J Mach Learn Res.* 2022;23(53):1–6.
- [14] Tibshirani R. Regression shrinkage and selection via LASSO. *J R Stat Soc Series B Stat Methodol.* 1996;58(1):267–88.
- [15] Bühlmann P, Yu B. Boosting with the L2 loss: regression and classification. *J Am Stat Assoc.* 2003;98:324–39.
- [16] Robinson PM. Root-N-consistent semiparametric regression. *Econometrica.* 1988;56(4):931–54.
- [17] Belloni A, Chernozhukov V, Fernández-Val I, Hansen C. Program evaluation and causal inference with high-dimensional data. *Econometrica.* 2017;85:233–98.
- [18] Pearl J, Mackenzie D. *The book of why: the new science of cause and effect.* New York: Basic Books; 2018.
- [19] Bareinboim E, Correa JD, Ibeling D, Icard T. On Pearl’s hierarchy and the foundations of causal inference. *Probabilistic and Causal Inference: The Works of Judea Pearl.* New York, NY, USA: Association for Computing Machinery; 2022 Feb. p. 507–56.
- [20] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat.* 2004;86:4–29.
- [21] Angrist JD, Pischke JS. *Mostly harmless econometrics: an empiricist’s companion.* Princeton University Press; 2009. <https://doi.org/10.2307/j.ctvcm4j72>.
- [22] Cinelli C, Forney A, Pearl J. A crash course in good and bad controls. *Sociol Method Res.* 2022. [https://ftp.cs.ucla.edu/pub/stat\\_ser/r493.pdf](https://ftp.cs.ucla.edu/pub/stat_ser/r493.pdf).
- [23] Knaus MC. A double machine learning approach to estimate effects of musical practice on student’s skills. *J R Stat Soc Ser A Stat Soc.* 2021;184(1):282–300.

- [24] Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669–709.
- [25] Blau FD, Kahn LM. The gender wage gap: extent, trends, and explanations. *J Econ Lit*. 2017;55:789–865.
- [26] Wüthrich K, Zhu Y. Omitted variable bias of LASSO-based inference methods: a finite sample analysis. *Rev Econ Stat*. 2021. <http://dx.doi.org/10.2139/ssrn.3379123>.
- [27] Chernozhukov V, Cinelli C, Newey W, Sharma A, Syrgkanis V. Long story short: omitted variable bias in causal machine learning; 2022. doi: 10.48550/arXiv.2112.13398.
- [28] Athey S, Imbens GW. The state of applied econometrics: causality and policy evaluation. *J Econ Perspect*. 2017;31(2):3–32.
- [29] Koopmans TC. Measurement without theory. *Rev Econ Stat*. 1947;29(3):161–72.
- [30] Jung Y, Tian J, Bareinboim E. Estimating identifiable causal effects through double machine learning. *Proc AAAI Conf Artif Intell*. 2021;35:12113–22.
- [31] Hünermund P, Bareinboim E. Causal inference and data fusion in econometrics. *Econom J*. 2023. <https://doi.org/10.48550/arXiv.1912.09104>.
- [32] Pearl J. *Causality: models, reasoning, and inference*. 2nd ed. New York, NY, USA: Cambridge University Press; 2009.
- [33] Strotz RH, Wold HOA.. Recursive vs. nonrecursive systems: an attempt at synthesis (part I of a Triptych on causal chain systems). *Econometrica*. 1960;28:417–27.
- [34] Woodward J. *Making things happen*. Oxford Studies in Philosophy of Science. Oxford, UK: Oxford University Press; 2003.
- [35] Cartwright N. *Hunting causes and using them*. Cambridge, UK: Cambridge University Press; 2007.
- [36] Pearl J. *Probabilistic reasoning in intelligent systems*. San Mateo, CA, USA: Morgan Kaufmann; 1988.
- [37] Pearl J. *Causality: models, reasoning, and inference*. 1st ed. New York, NY, USA: Cambridge University Press; 2000.
- [38] Haavelmo T. The statistical implications of a system of simultaneous equations. *Econometrica*. 1943;11:1–12.
- [39] Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci*. 2016;113:7345–52.
- [40] Koopmans TC. Cowles foundation monograph 10: statistical inference in dynamic economic models. Hoboken, New Jersey: John Wiley & Sons; 1950.
- [41] Greenland S. Quantifying biases in causal models: classical confounding vs Collider-Stratification bias. *Epidemiology*. 2003;14:300–6.
- [42] Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci*. 2010;25:51–71.
- [43] Cartwright N. *Nature's capacities and their measurement*. Oxford, UK: Clarendon Press; 1989.
- [44] Spirtes P, Glymour CN, Scheines R, Heckerman D. *Causation, prediction, and search*. Cambridge, MA, USA: MIT Press; 2000.
- [45] Peters J, Janzing D, Schölkopf B. *Elements of causal inference*. Cambridge, MA, USA: MIT Press; 2017.