# Double Machine Learning and Automated Confounder Selection — A Cautionary Tale

Paul Hünermund[†] Beyers Louw[‡] Itamar Caspi[*]

[†] *Copenhagen Business School, Kilevej 14A, Frederiksberg, 2000, DK.*
E-mail: phu.si@cbs.dk

[‡] *Maastricht University, Tongersestraat 53, 6211 LM Maastricht, NL.*
E-mail: jb.louw@maastrichtuniversity.nl

[*] *Bank of Israel, P.O.Box 780, 91007, Jerusalem, IL*
E-mail: itamar.caspi@boi.org.il

This version: February 9, 2022
First version: August 26, 2021

**Summary** Double machine learning (DML) is becoming an increasingly popular tool for automatic model selection in high-dimensional settings. These approaches rely on the assumption of conditional independence, which may not hold in big-data settings where the covariate space is large. This paper shows that DML is very sensitive to the inclusion of even a few "bad controls" in the covariate space. The resulting bias varies with the nature of the causal model, which raises concerns about the feasibility of selecting control variables in a data-driven way.

**Keywords**: *Double/Debiased Machine Learning, Directed Acyclic Graphs, Structural Causal Models, Backdoor Adjustment, Collider Bias*

> *"No causes in, no causes out."*
> — Nancy Cartwright

## 1. INTRODUCTION

Machine learning approaches for selecting suitable control variables to establish identification of causal parameters in high-dimensional settings are gaining increasing attention in economics (Belloni et al., 2014b; Chernozhukov et al., 2018; Angrist and Frandsen, 2019; Wüthrich and Zhu, 2021). This rising popularity can be explained by two potential advantages that these methods offer. First, a mostly data-driven, automated procedure of model selection allows to systematize the research process and make it more transparent (Athey, 2019). And second, the ability to consider a large number of covariates — possibly larger than the sample size — renders selection-on-observables types of identification assumptions more plausible (Belloni et al., 2014a, p. 640).

Consider the following system of partially linear equations

$$y = \theta_0 d + g_0(x) + u, \tag{1.1}$$

$$d = m_0(x) + v, \tag{1.2}$$

with primary interest in the causal effect $\theta_0$ of a treatment $D$ on outcome $Y$. The vector $X = (X_1, \ldots, X_p)$ consists of a set of exogenous covariates and $(U, V)$ are two disturbances with 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, regularized regression techniques such as LASSO (Belloni et al., 2014a) or $l_2$-boosting (Bühlmann and Yu, 2003) can be applied to automatically select the most suitable among a large set of potential control variables.

Yet, a naive application of regularization to equation (1.1) will fail, because it only selects variables that are highly correlated with the outcome $Y$, but not with the treatment $D$, which can lead to substantial omitted variable bias. The naive approach therefore generally does not result in a root-$N$ consistent estimator for the structural parameter $\theta_0$ (Chernozhukov et al., 2018). Instead, the analysis should be based on orthogonal, or doubly robust, moment conditions that are insensitive to approximation errors stemming from regularization (Belloni et al., 2017). In particular, let the moment condition be denoted by

$$E[\psi(W;\theta_0,\eta_0)] = 0, \tag{1.3}$$

with i.i.d. data $W$, nuisance parameter $\eta_0$, and moment function $\psi$. Then, (1.3) needs to fulfill the Neyman orthogonality property

$$\partial_\eta E[\psi(W;\theta_0,\eta)]|_{\eta=\eta_0} = 0, \tag{1.4}$$

where $\partial_\eta$ denotes the pathwise Gateaux derivative operator, for root-$N$ consistent estimation.

Two main solutions to the variable selection problem are proposed in the literature: (a) partialling out, and (b) double selection, which both take into account the strength of association between $D$ and $X$. The former uses regularization to estimate the residuals $\rho^y = y - x'\pi_0^y$ and $\rho^d = d - x'\pi_0^d$, and then, following the Frisch-Waugh-Lovell theorem, finds $\theta_0$ by regressing $\rho^y$ on $\rho^d$. The latter 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 of theses approaches ensure Neyman orthogonality (Chernozhukov et al., 2018).

However, a key assumption for applying the double machine learning framework in these settings is *unconfoundedness* (Imbens, 2004; Belloni et al., 2014b). Given the high-dimensional vector of control variables, treatment status is required to be conditionally independent of potential outcomes

$$Y_{D=d} \perp\!\!\!\perp D|X. \tag{1.5}$$

This assumption can easily be violated, if $X$ includes variables that are not fully exogenous. In this paper, we explore the consequences of violations of unconfoundedness for DML due to the presence of *bad controls* in the conditioning set (Angrist and Pischke, 2009). We focus on the LASSO case, which was one of the main motivations for developing these methods and has received the most attention by economists so far (Belloni et al., 2014b; Angrist and Frandsen, 2019; Knaus, 2021), because of its appealing combination of interpretability and accuracy. However, as we will show, our argument applies more broadly, also to the use of other machine learning algorithms for automated model selection in a causal inference setting.

In a first step, we make precise the notion of bad controls in regression analyses by building on the *backdoor criterion* from the graphical causal models literature (Pearl, 2009; Cinelli et al., 2020). We then show in simulation studies that DML is very sensitive to the inclusion of bad controls. Depending on the exact source of endogeneity, which will be further specified in more detail, the advantage of DML over naive LASSO vanishes completely. This is because bad controls, although they do not necessarily exert a causal influence, are often highly correlated with the treatment or the outcome (since they are related to unobservables that affect $D$ or $Y$). Therefore, bad controls are very likely to be picked 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 (2017). 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.

Taken as a whole, these results highlight significant pitfalls of automated, data-driven model selection in high-dimensional settings. In particular, if numerous potential controls are considered, in an attempt to justify a selection-on-observables assumption, with little economic intuition to guide the choice, the likelihood that some bad controls are included in the conditioning set might be high. We show that this problem is not only prevalent for post-treatment variables, so that researchers cannot rely on simple rules of thumb for variable inclusion. Instead, each potential control requires its own careful identification argument based on economic theory, which is difficult to provide if the feature space is large and ultimately undermines the purpose of automated model selection.[1]

We stress, however, that DML has broader applications, e.g., for the estimation of high-dimensional instrumental variable models (Belloni et al., 2017) and arbitrary do-calculus objects (Jung et al., 2021), as well as for data-splitting to reduce over-fitting. Our argument therefore specifically applies to the case when machine learning tools are used for the purpose of confounder detection.

## 2. A THEORY OF BAD CONTROLS

In order to make precise the notion of bad controls, we begin by defining structural causal models (SCM) and directed acyclic graphs (DAG). An SCM is a 4-tuple $\langle V, U, F, P(u)\rangle$, where $V = \{V_1, \ldots, V_m\}$ 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 $\{f_1, \ldots, f_m\}$ that assign values to the corresponding $V_i \in V$, such that $v_i \leftarrow f_i(pa_i, u_i)$, for $i = 1, \ldots, m$, and $PA_i \subseteq V \setminus V_i$.[2] Finally, $P(u)$ is a probability function defined over the domain of $U$.

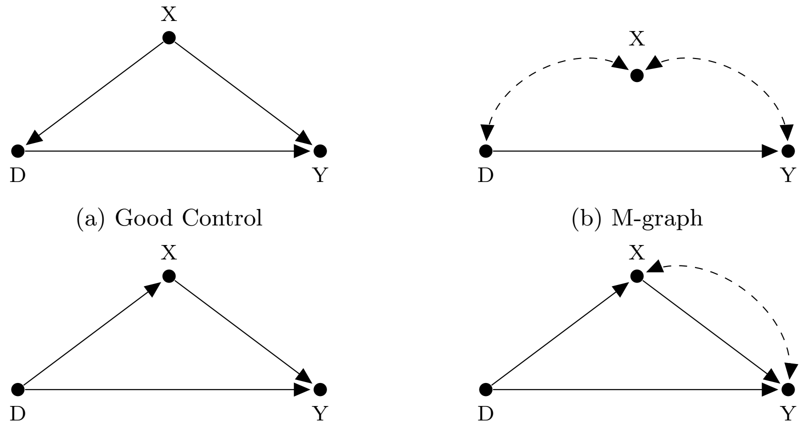
Every SCM defines a directed graph $\mathcal{G} \equiv (V, E)$, where $V$ is the set of endogenous variables, denoted as nodes (vertices) in the graph, and $E$ is a set of edges (links) pointing from $PA_i$ (the set of parent nodes) to $V_i$. An example is given by Fig. 1a, which corresponds to the SCM

$$\begin{aligned} x &\leftarrow f_1(u_1), \\ d &\leftarrow f_2(x, u_2), \\ y &\leftarrow f_3(d, x, u_3). \end{aligned} \tag{2.6}$$

Unobserved parents nodes induce a correlation between background factors in $U$. This

[1] Koopmans (1947): "Without resort to theory [...] conclusions relevant to the guidance of economic policies cannot be drawn."

[2] The SCM literature uses assignment operators instead of equations to capture the asymmetric nature of causal relationships (Hünermund and Bareinboim, 2021).



(a) Good Control

(b) M-graph

(c) Mediator

(d) Confounded Mediator

**Figure 1**: Directed acyclic graphs representing different structural causal models.

is depicted by bidirected dashed arcs in the graph, which render the causal model *semi-Markovian* (Pearl, 2009, p. 30). Fig. 1b 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 (Strotz and Wold, 1960), directed paths represent the direction of causal influence in the graph. Due to the notion of causality being asymmetric (Woodward, 2003; Cartwright, 2007), 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 to decompose the distribution of the observed variables according to the factorization: $P(v) = \sum_u \prod_i P(v_i|pa_i, u_i)P(u)$ (Pearl, 2009). 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 (Pearl, 1988). Consider three disjoint sets of variables, $X$, $Y$, and $Z$ in a 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 $Y$. By contrast, in the collider structure, $X$ and $Y$ are independent from the outset, $X \perp\!\!\!\perp Y|\emptyset$, whereas conditioning on $Z$ (or a descendant of $Z$; see Pearl, 2009, def. 1.2.3) would unblock the path, such that $X \not\perp\!\!\!\perp Y|Z$.[3]

The conditional independence relations that can be read off the graph via d-separation are of particular relevance for causal inference. Causal effects are defined in terms of inter-

[3] 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 $P(u)$ (Hünermund and Bareinboim, 2021).

ventions in the SCM, denoted by the $do(\cdot)$-operator (Haavelmo, 1943; Strotz and Wold, 1960; Pearl, 1995). For example, the intervention $do(D = d')$ in eq. (2.6) entails to delete the function $f_2(\cdot)$, which normally assigns values to $D$, from the model and to replace it with the constant value $d'$. The target is then to estimate the post-intervention distribution of the outcome variable, $P(Y = y|do(D = d'))$, that results from this manipulation. Other quantities, such as the average causal effect (ACE) of a discrete change in treatment from $d'$ to $d''$, which frequently is the parameter of interest in econometric studies, can then be computed as $E(y|do(D = d'')) - E(y|do(D = d'))$. However, since $P(y|do(d))$ 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 (Bareinboim and Pearl, 2016; Hünermund and Bareinboim, 2021). This constitutes the *identification* step in the graphical causal models literature (Koopmans, 1950; Pearl, 2009).

One particularly common 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* (Pearl, 1995).

Definition 2.1. *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 $\mathcal{G}_{\underline{D}}$, which is formed by deleting all edges from $\mathcal{G}$ that are emitted by $D$.*

Deleting edges emitted by $D$ ensures that all directed, causal paths between $D$ and $Y$ are removed from $\mathcal{G}$. The remaining paths are non-causal and thus create a spurious correlation between the treatment and outcome.[4] 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 (Pearl, 2009)

$$P(y|do(d)) = \sum_x P(y|d, x)P(x). \tag{2.7}$$

Since the right-hand side expression does not contain any do-operator, it can be estimated from observational 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 Fig. 1 as an example. In 1a there is one causal path, $D \to Y$, and one backdoor path, $D \leftarrow X \to Y$ (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 Fig. 1b, which is known under the name of *m-graph* in the epidemiology literature (Greenland, 2003), $X$ exerts no causal influence on any variable in the graph. Still, there are unobserved confounders that result in a backdoor path, $D \dashleftarrow\dashrightarrow X \dashleftarrow\dashrightarrow Y$, 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

[4] Since these paths point into $D$, they are said to "enter through the backdoor".

(or any of its desecendants) would unblock the path and therefore produce a spurious correlation. By contrast, $X$ does not lie on a backdoor path in Fig. 1c, 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, $D \rightarrow X \rightarrow Y$, (Imai et al., 2010). However, this direct effect is generally different from the ACE, which has to be kept in mind for interpretation of results.[5] Moreover, such an approach is risky, because if there are unobserved confounders between $X$ and $Y$, as depicted in Fig. 1d, $X$ becomes a collider on the path $D \rightarrow X \dashleftarrow\dashrightarrow Y$ and would thus lead to bias if conditioned on.[6]

## 3. SIMULATION RESULTS

In the following, we present a variety of simulations results to assess the magnitude of the bias introduced by including bad controls in the DML. 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 naive 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 $n = 100$ and number of covariates at $p = 100$. To introduce sparsity, only $q = 10$ out of these variables are specified to have non-zero coefficients. The treatment effect $\theta_0$ is constant and set equal to one.

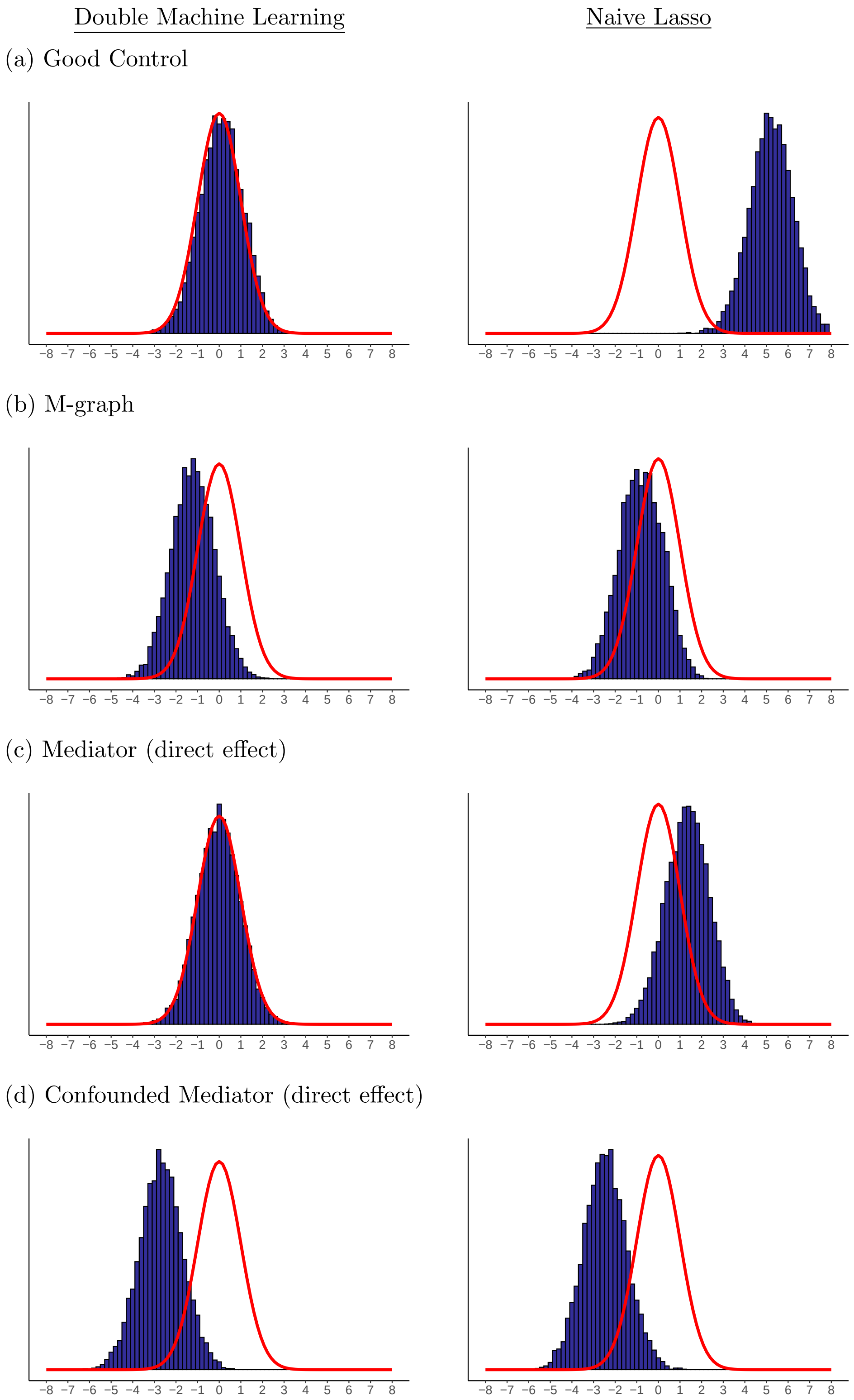
We then simulate data according to the four structural causal models depicted in Fig. 1. All exogenous nodes (which do not receive any incoming arrows) are specified as standard normal. In the baseline, parameters are chosen in such a way that the strength, measured as the product of structural coefficients, of each path connecting the (non-zero) 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 $b_2 = 0.2$.[7] We then test the performance of DML across these different setups by running 10,000 simulations each. Furthermore, we compare DML to naive LASSO, by applying $l_1$-regularization to the outcome equation alone and then re-estimating the structural parameters based on the selected control variables (i.e, post-LASSO).

Fig. 2 shows results using centered and studentized quantities, next to their theoretical distributions. In panel (a) we observe the familiar picture from Belloni et al. (2014b). DML is able to reliably filter out the good controls from irrelevant covariates, which leads to a distribution that closely matches the theoretical one. By contrast, naive 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

[5] Additionally, following Imai et al. (2010), 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.

[6] See Cinelli et al. (2020) for a more comprehensive discussion of bad controls in graphical causal models that goes beyond the scope of this paper.

[7] For the exact parameterization used in the data generating process, see Fig. S1 in the online supplement.



Double Machine Learning

Naive Lasso

(a) Good Control

(b) M-graph

(c) Mediator (direct effect)

(d) Confounded Mediator (direct effect)

**Figure 2**: Performance of DML compared to naive LASSO for different causal models.

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

| $(b_1, b_2) =$ | $(0.8, 0.2)$ | $(0.6, 0.4)$ | $(0.5, 0.5)$ | $(0.4, 0.6)$ | $(0.2, 0.8)$ |
|---|---|---|---|---|---|
| Good Control | 0.012 | 0.074 | 0.097 | 0.109 | 0.076 |
| M-graph | -0.126 | -0.178 | -0.187 | -0.185 | -0.149 |
| Mediator | -0.010 | -0.010 | -0.009 | -0.010 | -0.012 |
| Confounded Mediator | -0.536 | -0.482 | -0.418 | -0.344 | -0.178 |
| $q =$ | 1 | 5 | 10 | 20 | 50 |
| Good Control | -0.001 | -0.006 | 0.012 | 0.199 | 0.240 |
| M-graph | -0.050 | -0.106 | -0.126 | -0.141 | -0.150 |
| Mediator | -0.004 | -0.010 | -0.010 | -0.003 | 0.005 |
| Confounded Mediator | -0.136 | -0.403 | -0.536 | -0.642 | -0.738 |

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 naive LASSO in (a) vanishes completely ($bias^{DML} = -0.126$, and $bias^{LASSO} = -0.089$). Given the chosen parameterization with only a moderately high correlation between the covariates and the outcome, the naive approach selects fewer bad controls, which is the reason why it performs better in this setting.

In panel (c) we investigate the mediator case. Now, the covariates are post-treatment variables, which nonetheless end up getting selected as controls by the DML. According to the discussion in Sec. 2, this allows 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 the DML is unable to consistently estimate the total effect of treatment. Moreover, once we introduce a confounded mediator in panel (d), both DML and LASSO perform equally poorly. The direct effect cannot be consistently estimated in this model, as neither controlling for the mediators as well as 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 confounded mediator, substantial bias arises regardless of the chosen parameterization. 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 larger when the strength of association is stronger with the outcome than the treatment.

In the bottom panel of Table 1, we vary the number of covariates with non-zero 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 $p = 100$.[8] Interestingly, a noticeable

[8] 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.

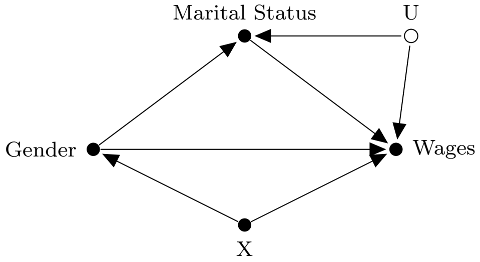


**Figure 3**: Causal diagram for the gender wage gap study in Blau and Kahn (2017). Unobserved variables are depicted by hollow nodes, $X$ denotes a vector of observed control variables.

bias (around 5 percent for the m-graph and 14 percent 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 case of good controls, DML performs well up until $q = 10$, after which performance deteriorates substantially. This result can be explained by a breakdown of the approximate sparsity regularity condition (Belloni et al., 2014b, Sec. 3.1) when $q \geq 20$ and the sample size is equal to $n = 100$.[9]

## 4. APPLICATION

For an application to real-world data, we take the *Panel Study of Income Dynamics* (PSID) microdata provided by Blau and Kahn (2017). 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, as well as regional and industry characteristics. However, Blau and Kahn deliberately decide 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 3).

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

[9] This interpretation is supported by the simulation results with $n = 1,000$ presented in Table S3 in the online supplement, which exhibit much smaller bias for the good control case, while the bias for bad controls remains large.

**Table 2**: Effect of gender on log wages using PSID data from Blau and Kahn (2017). Standard errors in parentheses.

| Wave = | 1981 | 1990 | 1999 | 2007 | 2009 | 2011 |
|---|---|---|---|---|---|---|
| OLS | -0.249 | -0.137 | -0.158 | -0.168 | -0.157 | -0.145 |
| | (0.016) | (0.014) | (0.016) | (0.015) | (0.015) | (0.016) |
| DML | -0.268 | -0.139 | -0.158 | -0.164 | -0.157 | -0.136 |
| | (0.017) | (0.015) | (0.016) | (0.016) | (0.016) | (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) |

range from $(1 - \exp(-0.249)) \approx 22$ percentage points in 1981 to approximately 13.5 p.p. in 2011. Most of the convergence between male and female wages happens in the 1980s, which coincides with the results in Blau and Kahn (2017). Although the DML relies on a much larger set of covariates, the results are very similar to OLS. We find greater discrepancies, however, 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 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 being married, induced by unobservables, that gets activated when conditioning on marital status as a collider. 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.[10]

## 5. DISCUSSION

In this paper, we demonstrated the sensitivity of automated confounder selection using double machine learning 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 naive LASSO. 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 feature space.

We believe that these results ultimately limit the usefulness of DML for automated confounder selection. In high-dimensional settings, with a large number of covariates, unintentionally including bad controls in the estimation could happen quite frequently. Moreover, simple rules of thumb, such as only considering pre-treatment variables for the conditioning set, do not offer adequate safeguards against this problem. Indeed, as Figure 1b shows, our results are not limited to the case of the inclusion of post-treatment controls. The intricacies of the backdoor criterion (recall, e.g., the implications of subtle differences between Figures 1c and 1d) imply that a vague economic intuition about the appropriateness of certain variables as controls will likely not be sufficient to ensure identification.

[10] 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.

Because DML already assumes unconfounded covariates (Chernozhukov et al., 2018, sec. 5), 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 (1989)'s familiar adage: "no causes in, no causes out." Since this is difficult to achieve in high-dimensional settings, however, 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 (Spirtes et al., 2000; Peters et al., 2017). 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 (Bareinboim et al., 2020). The same applies to DML, which is a highly effective tool within a particular domain, but needs to be considered in a broader context of possible causal structures that can occur in applied empirical settings.

## ACKNOWLEDGEMENTS

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

## REFERENCES

Angrist, J. and B. Frandsen (2019). Machine labor. Technical Report 26584, NBER Working Paper Series.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Athey, S. (2019). The impact of machine learning on economics. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*. Chicago, IL, USA: The University of Chicago Press.

Bareinboim, E., J. D. Correa, D. Ibeling, and T. Icard (2020). On Pearl's hierarchy and the foundations of causal inference. Technical Report R-60, Columbia University. Forthcoming in ACM special volume in honor of Judea Pearl.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences 113*, 7345–7352.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*, 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*, 608–650.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*, 789–865.

Bühlmann, P. and B. Yu (2003). Boosting with the $\mathrm{l}_2$ loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–339.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement.* Oxford, UK: Clarendon Press.

Cartwright, N. (2007). *Hunting Causes and Using Them.* Cambridge, UK: Cambridge University Press.

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

Cinelli, C., A. Forney, and J. Pearl (2020, March). A crash course in good and bad controls. Technical Report R-493, UCLA.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology 14*, 300–306.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica 11*, 1–12.

Hünermund, P. and E. Bareinboim (2021). Causal inference and data fusion in econometrics. Technical Report R-51, Causal Artificial Intelligence Lab, Columbia University.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*, 51–71.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86*, 4–29.

Jung, Y., J. Tian, and E. Bareinboim (2021). Estimating identifiable causal effects through double machine learning. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Knaus, M. C. (2021). A double machine learning approach to estimate efffects of musical practice on student's skills. *Journal of the Royal Statistical Society: Series A 184*(1), 282–300.

Koopmans, T. C. (1947). Measurement without theory. *The Review of Economics and Statistics 29*(3), 161–172.

Koopmans, T. C. (1950). *Cowles Foundation Monograph 10: Statistical Inference in Dynamic Economic Models.* John Wiley & Sons.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA, USA: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–709.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). New York, NY, USA: Cambridge University Press.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference.* Cambridge, MA, USA: MIT Press.

Spirtes, P., C. N. Glymour, R. Scheines, and D. Heckerman (2000). *Causation, Prediction, and Search.* Cambridge, MA, USA: MIT Press.

Strotz, R. H. and H. O. A. Wold (1960). Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica 28*, 417–427.

Woodward, J. (2003). *Making Things Happen.* Oxford Studies in Philosophy of Science. Oxford, UK: Oxford University Press.

Wüthrich, K. and Y. Zhu (2021). Omitted variable bias of lasso-based inference methods: A finite sample analysis. *The Review of Economics and Statistics*. forthcoming.