UNI FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

# WORKING PAPERS SES

## Causal mediation analysis with double machine learning

Helmut Farbmacher,
Martin Huber,
Henrika Langen,
and
Martin Spindler

N. 515
V.2020

Faculté des sciences économiques et sociales et du management
Wirtschafts- und sozialwissenschaftliche Fakultät

# Causal mediation analysis with double machine learning

Helmut Farbmacher+, Martin Huber*, Henrika Langen*, Martin Spindler**

+Max Planck Society, Munich Center for the Economics of Aging

*University of Fribourg, Dept. of Economics

**University of Hamburg, Faculty of Business Sciences

**Abstract:** This paper combines causal mediation analysis with double machine learning to control for observed confounders in a data-driven way under a selection-on-observables assumption in a high-dimensional setting. We consider the average indirect effect of a binary treatment operating through an intermediate variable (or mediator) on the causal path between the treatment and the outcome, as well as the unmediated direct effect. Estimation is based on efficient score functions, which possess a multiple robustness property w.r.t. misspecifications of the outcome, mediator, and treatment models. This property is key for selecting these models by double machine learning, which is combined with data splitting to prevent overfitting in the estimation of the effects of interest. We demonstrate that the direct and indirect effect estimators are asymptotically normal and root-n consistent under specific regularity conditions and investigate the finite sample properties of the suggested methods in a simulation study when considering lasso as machine learner. We also provide an empirical application to the U.S. National Longitudinal Survey of Youth, assessing the indirect effect of health insurance coverage on general health operating via routine checkups as mediator, as well as the direct effect. We find a moderate short term effect of health insurance coverage on general health which is, however, not mediated by routine checkups.

**Keywords:** mediation, direct and indirect effects, causal mechanisms, double machine learning, efficient score.

**JEL classification: C21**.

# 1 Introduction

Causal mediation analysis aims at decomposing the causal effect of a treatment on an outcome of interest into an indirect effect operating through a mediator (or intermediate outcome) and a direct effect comprising any causal mechanisms not operating through that mediator. Even if the treatment is random, direct and indirect effects are generally not identified by naively controlling for the mediator without accounting for its likely endogeneity, see Robins and Greenland (1992). While much of the earlier literature either neglected endogeneity issues or relied on restrictive linear models, see for instance Cochran (1957), Judd and Kenny (1981), and Baron and Kenny (1986), more recent contributions consider more general identification approaches using the potential outcome framework. Some of the numerous examples are Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), VanderWeele (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt, Bekaert, and Lange (2012), and Huber (2014). Using the denomination of Pearl (2001), the literature distinguishes between natural direct and indirect effects, where mediators are set to their potential values 'naturally' occurring under a specific treatment assignment, and the controlled direct effect, where the mediator is set to a 'prescribed' value.

The vast majority of identification strategies relies on selection-on-observable-type assumptions implying that the treatment and the mediator are conditionally exogenous when controlling for observed covariates. Empirical examples in economics and policy evaluation include Flores and Flores-Lagunes (2009), Heckman, Pinto, and Savelyev (2013), Keele, Tingley, and Yamamoto (2015), Conti, Heckman, and Pinto (2016), Huber (2015), Huber, Lechner, and Mellace (2017), Bellani and Bia (2018), Bijwaard and Jones (2018), and Huber, Lechner, and Strittmatter (2018). Such studies typically rely on the (implicit) assumption that the covariates to be controlled for can be unambiguously preselected by the researcher, for instance based on institutional knowledge or theoretical considerations. This assumes away uncertainty related to model selection w.r.t. covariates to be included and entails incorrect inference under the common practice of choosing and refining the choice of covariates based on their predictive power.

For this reason, this paper combines causal mediation analysis based on efficient score functions, see Tchetgen Tchetgen and Shpitser (2012), with double machine learning as outlined

in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) for a data-driven control of observed confounders to obtain valid inference under specific regularity conditions. Bluntly speaking, one important condition is that the number of important confounders (that make the selection-on-observables assumptions to hold approximately) is not too large relative to the sample size. However, the set of these important confounders need not be known a priori and the set of potential confounders can be even larger than the sample size. This is particularly useful in high dimensional data with a vast number of covariates that could potentially serve as control variables, which can render researcher-based covariate selection complicated if not infeasible. We demonstrate root-n consistency and asymptotic normality of the proposed effect estimators under specific regularity conditions by verifying that the general framework of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) for well-behaved double machine learning is satisfied in our context.

Tchetgen Tchetgen and Shpitser (2012) suggest estimating natural direct and indirect effects based on the efficient score functions of the potential outcomes, which requires plug-in estimates for the conditional mean outcome, mediator density, and treatment probability. Analogous to doubly robust estimation of average treatment effects, see Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995), the resulting estimators are semiparametrically efficient if all models of the plug-in estimates are correctly specified and remain consistent even if one model is misspecified. Our first contribution is to show that the efficient score function of Tchetgen Tchetgen and Shpitser (2012) satisfies the so-called Neyman (1959) orthogonality discussed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), which makes the estimation of direct and indirect effects rather insensitive to estimation errors in the plug-in estimates. Second, we show that by an application of Bayes' Theorem, the score function of Tchetgen Tchetgen and Shpitser (2012) can be transformed in a way that avoids estimation of the conditional mediator density and show it to be Neyman orthogonal. This appears particularly useful when the mediator is a vector of variables and/or continuous. Third, we establish the score function required for estimating the direct controlled effect along with Neyman orthgonality.

Neyman orthgonality is key for the fruitful application of double machine learning, allowing for robustness in the estimation of the nuisance parameters which is crucial when applying modern machine learning methods. Random sample splitting – to estimate the param-

eters of the plug-in models in one part of the data, while predicting the score function and estimating the direct and indirect effects in the other part – avoids overfitting the plug-in models (e.g. by controlling for too many covariates). It increases the variance by only using part of the data for effect estimation. This is avoided by cross-fitting which consists of swapping the roles of the data parts for estimating the plug-in models and the treatment effects to ultimately average over the effects estimates in either part. When combining efficient score-based effect estimation with sample splitting, $n^{-1/2}$-convergence of treatment effect estimation can be obtained under a substantially slower convergence of $n^{-1/4}$ for the plug-in estimates, see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). Under specific regularity conditions, this convergence rate is attained by many machine learning algorithms like lasso regression, see Tibshirani (1996).

We investigate the estimators' finite sample behavior based on the score function of Tchetgen Tchetgen and S (2012) and the alternative score suggested in this paper when using post-lasso regression as machine learner for the plug-in estimates. Furthermore, we apply our method to data from the National Longitudinal Survey of Youth 1997 (NLSY97), where a large set of potential control variables is available. We disentangle the short run effect of health insurance coverage on general health into an indirect effect which operates via the incidence of a routine checkup in the last year and a direct effect covering any other causal mechanisms. While we find a moderate health-improving direct effect, the indirect effect is very close to zero and never statistically insignificant. We therefore do not find evidence that health insurance coverage affects general health through routine checkups in the short run.

This paper proceeds as follows. Section 2 introduces the concepts of direct and indirect effect identification in the potential outcome framework. In Section 3, we present the identifying assumptions and discuss identification based on efficient score functions. Section 4 proposes an estimation procedure based on double machine learning and shows root-n consistency and asymptotic normality under specific conditions. Section 5 provides a simulation study. Section 6 presents an empirical application to data from the NLSY97. Section 7 concludes.

## 2 Definition of direct and indirect effects

We aim at decomposing the average treatment effect (ATE) of a binary treatment, denoted by $D$, on an outcome of interest, $Y$, into an indirect effect operating through a discrete mediator, $M$, and a direct effect that comprises any causal mechanisms other than through $M$. We use the potential outcome framework, see for instance Rubin (1974), to define the direct and indirect effects of interest, see also Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007) and Albert (2008) for further examples in the context of mediation. $M(d)$ denotes the potential mediator under treatment value $d \in \{0, 1\}$, while $Y(d, m)$ denotes the potential outcome as a function of both the treatment and some value $m$ of the mediator $M$.[1] The observed outcome and mediator correspond to the respective potential variables associated with the actual treatment assignment, i.e. $Y = D \cdot Y(1, M(1)) + (1-D) \cdot Y(0, M(0))$ and $M = D \cdot M(1) + (1-D) \cdot M(0)$, implying that any other potential outcomes or mediators are a priori (i.e. without further statistical assumptions) unknown.

We denote the ATE by $\Delta = E[Y(1, M(1)) - Y(0, M(0))]$, which comprises both direct and indirect effects. To decompose the latter, note that the average direct effect, denoted by $\theta(d)$, equals the difference in mean potential outcomes when switching the treatment while keeping the potential mediator fixed, which blocks the causal mechanism via $M$:

$$\theta(d) \quad = \quad E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}. \tag{1}$$

The (average) indirect effect, $\delta(d)$, equals the difference in mean potential outcomes when switching the potential mediator values while keeping the treatment fixed to block the direct effect.

$$\delta(d) \quad = \quad E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}. \tag{2}$$

Robins and Greenland (1992) and Robins (2003) referred to these parameters as pure/total direct and indirect effects, Flores and Flores-Lagunes (2009) as net and mechanism average treatment effects, and Pearl (2001) as natural direct and indirect effects, which is the denomination used in the remainder of this paper.

The ATE is the sum of the natural direct and indirect effects defined upon opposite treatment

[1] Throughout this paper, capital letters denote random variables and small letters specific values of random variables.

states $d$, which can be easily seen from adding and subtracting the counterfactual outcomes $E[Y(0, M(1))]$ and $E[Y(1, M(0))]$:

$$\begin{aligned}
\Delta &= E[Y(1, M(1)) - Y(0, M(0))] \\
&= E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0) \\
&= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1). \qquad (3)
\end{aligned}$$

The distinction between $\theta(1)$ and $\theta(0)$ as well as $\delta(1)$ and $\delta(0)$ hints to the possibility of heterogeneous effects across treatment states $d$ due to interaction effects between $D$ and $M$. For instance, the direct effect of health insurance coverage ($D$) on general health ($Y$) might depend on whether or not a person underwent routine check-ups ($M$).

The so-called controlled direct effect, denoted by $\gamma(m)$, is a further parameter that received much attention in the mediation literature. It corresponds to the difference in mean potential outcomes when switching the treatment and fixing the mediator at some value $m$:

$$\gamma(m) = E[Y(1, m) - Y(0, m)], \qquad \text{for } m \text{ in the support of } M. \qquad (4)$$

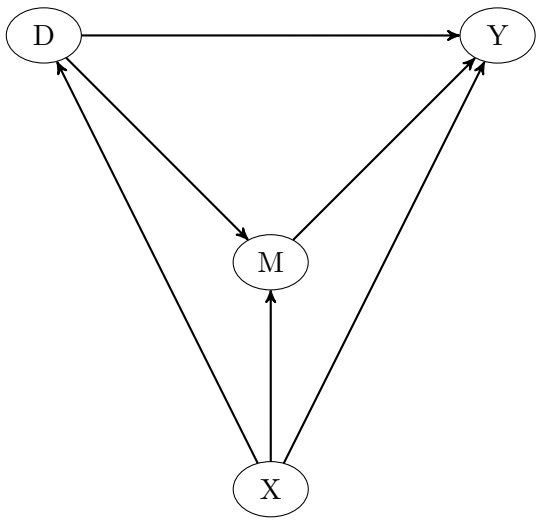
In contrast to $\theta(d)$, which is conditional on the potential mediator value 'naturally' realized for treatment $d$ which may differ across subjects, $\gamma(m)$ is conditional on enforcing the same mediator state in the entire population. The two parameters are only equivalent in the absence of an interaction between $D$ and $M$. Whether the natural or controlled direct effect is more relevant depends on the feasibility and desirability to intervene on or prescribe the mediator, see Pearl (2001) for a discussion of the 'descriptive' and 'prescriptive' natures of natural and controlled effects. There is no indirect effect parameter matching the controlled direct effect, implying that the difference between the total effect and the controlled direct effect does in general not correspond to the indirect effect, unless there is no interaction between $D$ and $M$, see e.g. Kaufman, MacLehose, and Kaufman (2004).

## 3 Assumptions and identification

Our identification strategy is based on the assumption that confounding of the treatment-outcome, treatment-mediator, and mediator-outcome relations can be controlled for by con-

ditioning on observed covariates, denoted by $X$. The latter must not contain variables that are influenced by the treatment, such that $X$ is typically evaluated prior to treatment assignment. Figure 1 provides a graphical illustration using a directed acyclic graph, with arrows representing causal effects. Each of $D$, $M$, and $Y$ might be causally affected by distinct and statistically independent sets of unobservables not displayed in Figure 1, but none of these unobservables may jointly affect two or all three elements $(D, M, Y)$ conditional on $X$.

Figure 1: Causal paths under conditional exogeneity given pre-treatment covariates




Formally, the first assumption invokes conditional independence of the treatment and potential mediators or outcomes given $X$. This restriction has been referred to conditional independence, selection on observables, or exogeneity in the treatment evaluation literature, see e.g. Imbens (2004). This rules out confounders jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand conditional on $X$. In non-experimental data, the plausibility of this assumption critically hinges on the richness of $X$.

**Assumption 1 (conditional independence of the treatment):**

$\{Y(d', m), M(d)\} \perp D|X$ for all $d', d \in \{0, 1\}$ and $m$ in the support of $M$,

where '$\perp$' denotes statistical independence. The second assumption requires the mediator to be conditionally independent of the potential outcomes given the treatment and the covariates.

**Assumption 2 (conditional independence of the mediator):**

$Y(d', m) \perp M|D = d, X = x$ for all $d', d \in \{0, 1\}$ and $m, x$ in the support of $M, X$.

Assumption 2 rules out confounders jointly affecting the mediator and the outcome conditional

on $D$ and $X$. If $X$ is pre-treatment (as is common to avoid controlling for variables potentially affected by the treatment), this implies the absence of post-treatment confounders of the mediator-outcome relation. Such a restriction needs to be rigorously scrutinized and appears for instance less plausible if the time window between the measurement of the treatment and the mediator is large in a world of time-varying variables.

The third assumption imposes common support on the conditional treatment probability across treatment states.

**Assumption 3 (common support):**

$\Pr(D = d|M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and $m, x$ in the support of $M, X$.

Assumption 3 restricts the conditional probability to be or not be treated given $M, X$, henceforth referred to as propensity score, to be larger than zero. It implies the weaker condition that $\Pr(D = d|X = x) > 0$ such that the the treatment must not be deterministic in $X$, otherwise no comparable units in terms of $X$ are available across treatment states. By Bayes' theorem, Assumption 3 also implies that $\Pr(M = m|D = d, X = x) > 0$ if $M$ is discrete or that the conditional density of $M$ given $D, X$ is larger than zero if $M$ is continuous. Conditional on $X$, the mediator state must not be deterministic in the treatment, otherwise no comparable units in terms of the treatment are available across mediator states. Assumptions 1 to 3 are standard in the causal mediation literature, see for instance Imai, Keele, and Yamamoto (2010), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt, Bekaert, and Lange (2012), and Huber (2014), or also Pearl (2001), Petersen, Sinisi, and van der Laan (2006), and Hong (2010), for closely related restrictions.

Tchetgen Tchetgen and Shpitser (2012) discuss identification of the counterfactual $E[d, M(1-d)]$ based on the efficient score function:

$$\begin{aligned} E[Y(d, M(1-d))] &= E[\psi_d], \\ \text{with } \psi_d &= \frac{I\{D = d\} \cdot f(M|1-d, X)}{p_d(X) \cdot f(M|d, X)} \cdot [Y - \mu(d, M, X)] \\ &\quad + \frac{I\{D = 1-d\}}{1 - p_d(X)} \cdot \left[\mu(d, M, X) - \int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X)\ dm\right] \\ &\quad + \int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X)\ dm \end{aligned} \tag{5}$$

where $f(M|D, X)$ denotes the conditional density of $M$ given $D$ and $X$ (if $M$ is discrete, this

is a conditional probability and integrals need to be replaced by sums), $p_d(X) = Pr(D = d|X)$ the probability of treatment $D = d$ given $X$, and $\mu(D, M, X) = E(Y|D, M, X)$ the conditional expectation of outcome $Y$ given $D$, $M$, and $X$.

To derive an alternative expression for identification, note that by Bayes' Law,

$$\frac{f(M|1-d,X)}{p_d(X)\cdot f(M|d,X)} = \frac{\big(1-p_d(M,X)\big)\, f(M|X)}{1-p_d(X)} \cdot \frac{p_d(X)}{p_d(M,X)\cdot f(M|X)\cdot p_d(X)} = \frac{1-p_d(M,X)}{p_d(M,X)\,\big(1-p_d(X)\big)}$$

where $f(M|X)$ is the conditional distribution of $M$ given $X$. Furthermore,

$$\begin{aligned} \int \mu(d,m,X)\cdot f(m|1-d,X)dm &= \int \mu(d,m,X)\cdot \frac{\big(1-p_d(m,X)\big)\, f(m|X)}{1-p_d(X)}dm \\ = \quad E\Big[\mu(d,M,X)\cdot \frac{1-p_d(M,X)}{1-p_d(X)}\Big|X\Big]. \end{aligned}$$

Therefore, an alternative representation of (5) is

$$\begin{aligned} E[Y(d,M(1-d))] &= E[\psi_d^*], \\ \text{with } \psi_d^* &= \frac{I\{D=d\}(1-p_d(M,X))}{p_d(M,X)\cdot 1-p_d(X)}\cdot[Y-\mu(d,M,X)] \\ &+ \frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Big[\mu(d,M,X) - \frac{1}{1-p_d(X)}E\Big[\mu(d,M,X)\,(1-p_d(M,X))\Big|X\Big]\Big] \\ &+ \frac{1}{1-p_d(X)}E\Big[\mu(d,M,X)\,\big(1-p_d(M,X)\big)\Big|X\Big]. \end{aligned} \tag{6}$$

This representation avoids conditional mediator densities which appears attractive if $M$ is continuous and/or multidimensional. On the other hand, it requires estimation of an additional parameter, namely $E\Big[\mu(d,M,X)\cdot(1-p_d(M,X))\Big|X\Big]$.

Efficient score-based identification of $E[Y(d,M(d))]$ under $Y(d,m)\perp\{D,M\}|X = x$ (see Assumption 1) has been established in the literature on doubly robust ATE estimation, see for instance Robins, Rotnitzky, and Zhao (1994) and Hahn (1998):

$$E[Y(d,(M(d))] = E[\alpha_d] \text{ with } \alpha_d = \frac{I\{D=d\}\cdot[Y-\mu(d,X)]}{p_d(X)} + \mu(d,X) \tag{7}$$

where $\mu(D,X) = E(Y|D,M(D),X) = E(Y|D,X)$ is the conditional expectation of outcome $Y$ given $D$ and $X$.

Assume that $M$ is discrete. As Assumptions 1 and 2 imply $Y(d,m)\perp\{D,M\}|X = x$, doubly

robust identification of the potential outcome $E[Y(d,m)]$, which is required for the controlled direct effect, follows from replacing $I\{D=d\}$ and $p_d(X)$ in (7) by $I\{D=d, M=m\} = I\{M=m\} \cdot I\{D=d\}$ and $\Pr(D=d, M=m|X) = f(m|d,X) \cdot p_d(m,X)$:

$$E[Y(d,m)] = E[\psi_{dm}] \text{ with } \psi_{dm} = \frac{I\{D=d\} \cdot I\{M=m\} \cdot [Y - \mu(d,m,X)]}{f(m|d,X) \cdot p_d(m,X)} + \mu(d,m,X). \quad (8)$$

## 4 Estimation of the counterfactual with K-fold Cross-Fitting

We subsequently propose an estimation strategy for the counterfactual $E[Y(d, M(1-d))]$ with $d \in \{1,0\}$ and show its root-n consistency under specific regularity conditions. The strategy will rely on the above mentioned score function from Tchetgen Tchetgen and Shpitser (2012):

$$\begin{aligned}\psi_d(W,\eta,\Psi_{d0}) \quad = \quad & \frac{I\{D=d\} \cdot f(M|1-d,X)}{p_d(X) \cdot f(M|d,X)} \cdot [Y - \mu(d,M,X)] \\ & + \frac{I\{D=1-d\}}{1-p_d(X)} \cdot \Big[\mu(d,M,X) - \int_{m \in \mathcal{M}} \mu(d,m,X) \cdot f(m|1-d,X) dm\Big] \\ & + \int_{m \in \mathcal{M}} \mu(d,m,X) \cdot f(m|1-d,X) dm \quad - \quad \Psi_{d0} \qquad (9)\end{aligned}$$

where $\mathcal{W} = \{W_i | 1 \leq i \leq N\}$ with $W_i = (Y_i, M_i, D_i, X_i)$ for $i = 1, \ldots, N$ denotes the set of observations. The plug-in estimates for the conditional mean outcome, mediator density and treatment probability are referred to by $\hat{\eta} = \{\hat{\mu}(D,M,X), \hat{f}(M|D,X), \hat{p_d}(X)\}$ and the true nuisance parameters by $\eta_0 = \{\mu_0(D,M,X), f_0(M|D,X), p_{d0}(X)\}$. $\Psi_{d0} = E[Y(d, M(1-d))]$ denotes the true counterfactual.

We suggest estimating the $\Psi_{d0}$ using the following algorithm that combines orthogonal score estimation and sample splitting. Further below we will outline the conditions under which this estimation strategy leads to root-n consistent estimates for the counterfactual.

**Algorithm 1: Estimation of $E[Y(d, M(1-d))]$ based on equation (5)**

1. Split $\mathcal{W}$ in $K$ subsamples. For each subsample $k$, let $n_k$ denote its size, $\mathcal{W}_k$ the set of observations in the sample and $\mathcal{W}_k^C$ the complement set of all observations not in $k$.

2. For each $k$, use $\mathcal{W}_k^C$ to estimate the model parameters of $p_d(X)$, $f(M|D,X)$, and $\mu(D,M,X)$ in order to predict these models on $\mathcal{W}_k$, where the predictions are denoted by $\hat{p_d}^k(X)$, $\hat{f}^k(M|D,X)$, and $\hat{\mu}^k(D,M,X)$.

3. For each $k$, obtain an estimate of the efficient score function (see $\psi_d$ in (5)) for each observation $i$ in $\mathcal{W}_k$, denoted by $\hat{\psi}^k_{d,i}$ :

$$\begin{aligned}\hat{\psi}^k_{d,i} &= \frac{I\{D_i = d\} \cdot \hat{f}^k(M_i|1-d, X_i)}{\hat{p_d}^k(X_i) \cdot \hat{f}^k(M_i|d, X_i)} \cdot [Y_i - \hat{\mu}^k(d, M_i, X_i)] \\ &+ \frac{I\{D_i = 1-d\}}{1-\hat{p_d}^k(X_i)} \cdot \left[\hat{\mu}^k(d, M_i, X_i) - \int_{m\in\mathcal{M}} \hat{\mu}^k(d, m, X_i) \cdot \hat{f}^k(m|1-d, X_i)dm\right] \\ &+ \int_{m\in\mathcal{M}} \hat{\mu}^k(d, m, X_i) \cdot \hat{f}^k(m|1-d, X_i)dm. \end{aligned} \tag{10}$$

4. Average the estimated scores $\hat{\psi}^k_{d,i}$ over all observations across all $K$ subsamples to obtain an estimate of $\Psi_d = E[Y(d, M(1-d))]$ in the total sample, denoted by $\hat{\Psi}_d = 1/n \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\psi}^k_{d,i}$.

Algorithm 1 can be adapted to estimate the counterfactuals required for the controlled direct effect, see (8). To this end, denote by $\Psi_{dm0} = E[Y(d, m)]$ the true counterfactual of interest and define the score function

$$\psi_{dm}(W, \eta, \Psi_{dm0}) = \frac{I\{D = d\} \cdot I\{M = m\} \cdot [Y - \mu(d, m, X)]}{f(m|d, X) \cdot p_d(m, X)} + \mu(d, m, X) - \Psi_{dm0}. \tag{11}$$

$E[Y(d, m)]$ can then be estimated by replacing $\psi_d$ and $\Psi_d$ by $\psi_{dm}$ and $\Psi_{dm0}$, respectively, everywhere in Algorithm 1.

In order to achieve root-n consistency for counterfactual estimation, we make the following assumption on the prediction qualities of the machine learners for our plug-in estimates of the nuisance parameters .

**Assumption 4 (quality of plug-in parameter estimates):**
Given a random subset $k$, the plug-in estimators $\hat{\eta}_k$ estimated in this subset belong to the shrinking neighborhoods of the true parameters $\eta_0$, denoted by $\mathcal{T}_k = \{\eta^* \in \mathbb{R} \times (0,1) \times (0,1) \times \mathbb{R} \mid \ \|\eta_0 - \hat{\eta}_k\|_2 \leq o(n_k^{-1/4}\}$, with probability of at least $1 - o(1)$.

For demonstrating root-n consistency of the proposed estimation strategy for the counterfactual, we heavily draw from Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). We show that our estimation strategy satisfies the requirements for their Double Ma-

chine Learning framework by first verifying linearity and Neyman orthogonality of the score (see Appendix A). Then, as e.g. $\psi_d(W, \eta, \Psi_{d0})$ is smooth in $(\eta, \Psi_{d0})$, the plug-in estimators must converge with rate $n^{-1/4}$ in order to achieve $n^{-1/2}$-convergence for the estimation of $\hat{\Psi}_d$. This convergence rate of $n^{-1/4}$ is achievable for many commonly used machine learners such as lasso, random forest, boosting and neural net estimation. The rates for $L_2$-boosting were, for instance, derived in Luo and Spindler (2016).

**Theorem 1**

Under Assumptions 1-4, it holds for estimating $E[Y(d, M(1-d))]$, $E[Y(d, m)]$ based on Algorithm 1:

$\sqrt{n}\Big(\hat{\Psi}_d - \Psi_{d0}\Big) \to N(0, \sigma_{\psi_d})$, where $\sigma_{\psi_d} = E[(\psi_d - \Phi_{d0})^2]$.

$\sqrt{n}\Big(\hat{\Psi}_{dm} - \Psi_{dm0}\Big) \to N(0, \sigma_{\psi_{dm}})$, where $\sigma_{\psi_d} = E[(\psi_d - \Phi_{dm0})^2]$.

For the proofs of score linearity an Neyman orthogonality, see Appendices A and B.

Analogous results follow for the estimation of the counterfactual $\Lambda = E[Y(d, M(d))]$ when replacing $\hat{\psi}_d$ in the algorithm above by an estimate of score function $\alpha_d$ in (7),

$$\hat{\alpha}_d = \frac{I\{D=d\}\cdot(Y_i - \hat{\mu}^k(d, X_i))}{\hat{p_d}^k(X_i)} + \hat{\mu}^k(d, X_i), \tag{12}$$

where $\hat{\mu}^k(d, x)$ is an estimate of $\mu(d, x)$. This approach has been discussed in literature on ATE estimation based on double machine learning, see for instance Belloni, Chernozhukov, Fernández-Val, and Hanse (2017) and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). Denoting by $\hat{\Lambda}$ the estimate of $\Lambda$, it follows under Assumptions 1-4 that $\sqrt{n}\Big(\hat{\Lambda}_d - \Lambda_d\Big) \to N(0, \sigma_{\alpha_d})$, where $\sigma_{\alpha_d} = E[(\alpha_d - \Lambda_d)^2]$. Therefore, root-n-consistent estimates of the total as well as the direct and indirect effects are obtained as difference of the estimated potential outcomes, which we denote by $\hat{\Delta}$, $\hat{\theta}(d)$, and $\hat{\delta}(d)$.That is, $\hat{\Delta} = \hat{\Lambda}_1 - \hat{\Lambda}_0$, $\hat{\theta}(1) = \hat{\Lambda}_1 - \hat{\Phi}_0$, $\hat{\theta}(0) = \hat{\Phi}_1 - \hat{\Lambda}_0$, $\hat{\delta}(1) = \hat{\Lambda}_1 - \hat{\Phi}_1$, and $\hat{\delta}(0) = \hat{\Phi}_0 - \hat{\Lambda}_0$.

Naturally, the asymptotic variance of any effect is obtained based on the variance of the difference in the score functions of the potential outcomes required for the respective effect. For instance, the asymptotic variance of $\hat{\theta}(1)$ is given by $Var(\hat{\theta}(1)) = Var(\alpha_1 - \psi_0)/n = (\sigma_{\alpha_1} + \sigma_{\psi_0} - 2Cov(\alpha_1, \psi_0))/n$.

We subsequently discuss estimation based on the score function $\psi_d^*$ in expression (6). We

note that in this case, one needs to estimate a nested nuisance parameter $E\Big[\mu(d,M,X)\,(1-p_d(M,X))\Big|X\Big]$. To avoid overfitting, the models for $\mu(d,M,X)$, $1-p_d(M,X)$, and $E\Big[\mu(d,M,X)\,(1-p_d(M,X))\Big|X\Big]$ are therefore estimated in different subsamples.

**Algorithm for estimating $E[Y(d,M(1-d))]$ based on equation (6)**

1. Split $\mathcal{W}$ in $K$ subsamples. For each subsample $k$, let $n_k$ denote its size, $\mathcal{W}_k$ the set of observations in the sample and $\mathcal{W}_k^C$ the complement set of all observations not in $k$.

2. For each $k$, use $\mathcal{W}_k^C$ to estimate the model parameters of $p_d(X)$. Split $\mathcal{W}_k^C$ into 3 nonoverlapping subsamples and estimate the model parameters of $\mu(d,M,X)$, $p_d(M,X)$, and $E\Big[\mu(d,M,X)\,(1-p_d(M,X))\Big|X\Big]$ in the distinct subsamples. Predict the models among $\mathcal{W}_k$, where the predictions are denoted by $\hat{p_d}^k(X)$, $\hat{p_d}^k(M,X)$, $\hat{\mu}^k(D,M,X)$ and $\hat{E}\Big[\mu(d,M,X)\,(1-p_d(M,X))\Big|X\Big]^k$.

3. For each $k$, obtain an estimate of the efficient score function (see $\psi_d^*$ in (6)) for each observation $i$ in $\mathcal{W}_k$, denoted by $\hat{\psi}_{d,i}^k$ :

$$\begin{aligned}\hat{\psi}_{d,i}^{*k} \;=\;& \frac{I\{D_i=d\}\big(1-\hat{p_d}^k(M_i,X_i)\big)}{\hat{p_d}^k(M_i,X_i)\,\big(1-\hat{p_d}^k(X_i)\big)}\cdot[Y-\hat{\mu}^k(d,M_i,X_i)]\\ &+\frac{I\{D_i=1-d\}}{1-\hat{p_d}^k(X_i)}\cdot\Big[\hat{\mu}^k(d,M_i,X_i)-\frac{1}{1-\hat{p_d}^k(X_i)}\cdot\hat{E}\Big[\hat{\mu}^k(d,M_i,X_i)\,\big(1-\hat{p_d}^k(M_i,X_i)\big)\Big|X_i\Big]\Big]\\ &+\frac{1}{1-\hat{p_d}^k(X_i)}E\Big[\hat{\mu}^k(d,M_i,X_i)\,\big(1-\hat{p_d}^k(M_i,X_i)\big)\Big|X\Big]. \qquad (13)\end{aligned}$$

4. Average the estimated scores $\hat{\psi}_{d,i}^{*k}$ over all observations across all $K$ subsamples to obtain an estimate of $\Psi_d = E[Y(d,M(1-d))]$ in the total sample, denoted by $\hat{\Psi}_d^* = 1/n\sum_{k=1}^{K}\sum_{i=1}^{n_k}\hat{\psi}_{d,i}^{*k}$.

Also this approach can be shown to be root-n- consistent under our assumptions.

**Theorem 2**

Under Assumptions 1-4, it holds for estimating $E[Y(d,M(1-d))]$ based on Algorithm 2:
$\sqrt{n}\Big(\hat{\Psi}_d^*-\Psi_{d0}^*\Big)\to N(0,\sigma_{\psi_d^*})$, where $\sigma_{\psi_d^*}=E[(\psi_d^*-\Phi_{d0}^*)^2]$.
For the proofs of score linearity an Neyman orthogonality, see Appendix C.

## 5 Simulation study

This section provides a simulation study to investigate the finite sample behaviour of the proposed methods based on the following data generating process:

$$\begin{aligned} Y &= 0.5D + 0.5M + 0.5DM + X'\beta + U, \\ M &= I\{0.5D + X'\beta + V > 0\}, \quad D = I\{X'\beta + W > 0\}, \\ X &\sim N(0, \Sigma), \quad U, V, W \sim N(0,1) \text{ independently of each other and } X. \end{aligned}$$

Outcome $Y$ is a function of the observed variables $D, M, X$, including an interaction between the mediator and the treatment, and an unobserved term $U$. The binary mediator $M$ is a function of $D, X$ and the unobservable $V$, while the binary treatment $D$ is determined by $X$ and the unobservable $W$. $X$ is a vector of covariates of dimension $p$, which is drawn from a multivariate normal distribution with zero mean and covariance matrix $\Sigma$. The latter is defined based on setting the covariance of the $i$th and $j$th covariate in $X$ to $\Sigma_{ij} = 0.5^{|i-j|}$.[2] Coefficients $\beta$ gauge the impact of $X$ on $Y$, $M$, and $D$, respectively, and thus, the strength of confounding. $U, V, W$ are random and standard normally distributed scalar unobservables. We consider two sample sizes of $n = 1000, 4000$ and run 1000 simulations per data generating process.

We investigate the performance of effect estimation based on (i) Theorem 1 using the identification result in expression (5) derived by Tchetgen Tchetgen and Shpitser (2012) as well as (ii) Theorem 2 using the modified score function in expression (6) which avoids conditional mediator densities. The nuisance parameters are estimated by post-lasso regression based on the 'hdm' package by Spindler, Chernozhukov, and Hansen (2016) for the statistical software 'R' with its default options, using logit specifications for $p_d(X)$, $p_d(M, X)$, and $f(M|D, X)$ and linear specifications for $\mu(D, M, X)$ and $E\Big[\mu(d, M, X) \cdot (1 - p_d(M, X))\Big|X\Big]$. The estimation of direct and indirect effects is based on 4-fold cross-fitting. For all methods investigated, we drop observations whose (products of) estimated conditional probabilities in the denominator of any potential outcome expression are close to zero, namely smaller than a trimming threshold of 0.05 (or 5%).

In our first simulation design, we set $p = 200$ and the $i$th element in the coefficient vector

[2] The results presented below are hardly affected when setting $\Sigma$ to the identity matrix (zero correlation across $X$).

$\beta$ to $0.3/i^2$ for $i = 1, ..., p$, meaning a quadratic decay of covariate importance in terms of confounding. This specification implies that the $R^2$ of $X$ when predicting $Y$ amounts to 0.22 in large samples, while the Nagelkerke (1991) pseudo-$R^2$ of $X$ when predicting $D$ and $M$ by probit models amounts to 0.10 and 0.13, respectively. The left panel of Table 1 reports the results for either sample size. For $n = 1000$, double machine learning based Theorem 1 generally exhibits a lower absolute bias ('abias') and also a slightly smaller standard deviation ('sd') than based on Theorem 2. The behavior of both estimation approaches improve when increasing sample size to $n = 4000$, as the bias and standard deviation go down. Under the larger sample size, differences in terms of root mean squared error ('rmse') between estimation based on Theorems 1 and 2 are relatively small. By and large, the results suggest that the estimators converge to the true effects at root-$n$ rate.

In our second simulation, confounding is increased by setting $\beta$ to $0.5/i^2$ for $i = 1, ..., p$. This specification implies that the $R^2$ of $X$ when predicting $Y$ amounts to 0.42, while the Nagelkerke (1991) pseudo-$R^2$ of $X$ when predicting $D$ and $M$ amounts to 0.23 and 0.28, respectively. The results are displayed in the right panel of Table 1. The increase of confounding entails a non-negligible increase in the standard deviation of estimation based of Theorem 2, while estimation based on Theorem 1 is hardly affected. Therefore, the latter method dominates in terms of a lower mean squared error under either sample size. However, both estimators convergence to the true effects as the sample size increases, and differences of root mean squared errors across methods become smaller.

Table 1: Simulation results for effect estimates ($p = 200$)

| | Coefficients given by $0.3/i^2$ for $i = 1, ..., p$ | | | | | | | Coefficients given by $0.5/i^2$ for $i = 1, ..., p$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | abias | sd | rmse | abias | sd | rmse | true | abias | sd | rmse | abias | sd | rmse | true |
| | $n$=1000 | | | $n$=4000 | | | | $n$=1000 | | | $n$=4000 | | | |
| | Double machine learning based on Theorem 1 | | | | | | | | | | | | | |
| $\hat{\Delta}$ | 0.01 | 0.08 | 0.08 | 0.00 | 0.04 | 0.04 | 1.02 | 0.02 | 0.09 | 0.09 | 0.02 | 0.04 | 0.05 | 1.00 |
| $\hat{\theta}(1)$ | 0.00 | 0.09 | 0.09 | 0.00 | 0.04 | 0.04 | 0.84 | 0.01 | 0.09 | 0.09 | 0.01 | 0.04 | 0.04 | 0.83 |
| $\hat{\theta}(0)$ | 0.01 | 0.08 | 0.08 | 0.00 | 0.04 | 0.04 | 0.75 | 0.02 | 0.08 | 0.09 | 0.01 | 0.04 | 0.04 | 0.75 |
| $\hat{\delta}(1)$ | 0.00 | 0.06 | 0.06 | 0.00 | 0.03 | 0.03 | 0.27 | 0.00 | 0.06 | 0.06 | 0.00 | 0.03 | 0.03 | 0.25 |
| $\hat{\delta}(0)$ | 0.01 | 0.06 | 0.06 | 0.00 | 0.02 | 0.02 | 0.18 | 0.01 | 0.06 | 0.06 | 0.00 | 0.02 | 0.02 | 0.17 |
| trimmed | | 17.12 | | | 19.27 | | | | 80.14 | | | 237.63 | | |
| | Double machine learning based on Theorem 2 | | | | | | | | | | | | | |
| $\hat{\Delta}$ | 0.01 | 0.08 | 0.08 | 0.00 | 0.04 | 0.04 | 1.02 | 0.01 | 0.09 | 0.09 | 0.01 | 0.04 | 0.04 | 1.00 |
| $\hat{\theta}(1)$ | 0.03 | 0.10 | 0.11 | 0.01 | 0.04 | 0.04 | 0.84 | 0.01 | 0.15 | 0.15 | 0.01 | 0.07 | 0.07 | 0.83 |
| $\hat{\theta}(0)$ | 0.10 | 0.13 | 0.16 | 0.02 | 0.04 | 0.04 | 0.75 | 0.11 | 0.18 | 0.21 | 0.02 | 0.07 | 0.07 | 0.75 |
| $\hat{\delta}(1)$ | 0.11 | 0.11 | 0.16 | 0.02 | 0.03 | 0.04 | 0.27 | 0.12 | 0.18 | 0.21 | 0.02 | 0.07 | 0.07 | 0.25 |
| $\hat{\delta}(0)$ | 0.02 | 0.08 | 0.08 | 0.00 | 0.03 | 0.03 | 0.18 | 0.00 | 0.13 | 0.13 | 0.01 | 0.06 | 0.06 | 0.17 |
| trimmed | | 34.78 | | | 2.12 | | | | 69.47 | | | 48.70 | | |

Note: 'abias', 'sd', and 'rmse' denote the absolute bias, standard deviation and root mean squared error of the respective effect estimate. 'true' provides the true effect. 'trimmed' is the average number of trimmed observations per simulation. The propensity score-based trimming threshold is set to 0.05.

Appendix D reports the simulation results (namely the absolute bias, standard deviation, and root mean squared error) for the standard errors obtained by an asymptotic approximation based on the estimated variance of the score functions. The results suggest that the asymptotic standard errors decently estimate the actual standard deviation of the point estimators.

# 6 Application

In this section, we apply our method to data from the National Longitudinal Survey of Youth 1997 (NLSY97), a survey following a U.S. nationally representative sample of 8,984 individuals born in the years 1980-84. Since 1997, the participants have been interviewed on a wide range of demographic, socioeconomic, and health-related topics in a one- to two-year circle. We investigate the causal effect of health insurance coverage ($D$) on general health ($Y$) and decompose

it into an indirect pathway via the incidence of a regular medical checkup ($M$) and a direct effect entailing any other causal mechanisms. Whether or not an individual undergoes routine checkups appears to be an interesting mediator, as it is likely to be affected by health insurance coverage and may itself have an impact on the individual's health, because checkups can help identifying medical conditions before they get serious to prevent them from affecting a person's general health state.

The effect of health insurance coverage on self-reported health has been investigated in different countries with no compulsory medical insurance and no publicly provided universal health coverage, see for example Simon, Soni, and Cawley (2017), Sommers, Maylone, Blendon, Orav, and Epstein (2017), Baicker, Taubman, Allen, Bernstein, Gruber, Newhouse, Schneider, Wright, Zaslavsky, and Finkelstein (2013), Yörük (2016) and Cardella and Depew (2014) for the U.S. and King, Gakidou, Imai, Lakin, Moore, Nall (2009) for Mexico). Most of these studies find a significant positive effect of insurance coverage on self-reported health. The impact of insurance coverage on the utilization of preventive care measures, particularly routine checkups like cancer, diabetes and cardiovascular screenings, is also extensively covered in public health literature. Most studies find that health insurance coverage increases the odds of attending routine checkups. While some contributions include selected demographic, socioeconomic and health-related control variables to account for the endogeneity of health insurance status (see e.g. Faulkner and Schauffler (1997), Press (2014), Burstin, Swartz, O'Neil, Orav, and Brennan (1998), Fowler-Brown, Corbie-Smith, Garrett, and Lurie (2007)), others exploit natural experiments: Simon, Soni, and Cawley (2017) estimate a difference-in-differences model comparing states which did and did not expand Medicaid to low-income adults in 2005, while Baicker, Taubman, Allen, Bernstein, Gruber, Newhouse, Schneider, Wright, Zaslavsky, an (2013) exploit that the state of Oregon expanded Medicaid based on lottery drawings from a waiting list. The results of both studies suggest that the Medicaid expansions increased use of certain forms of preventive care. In a study on Mexican adults, Pagán, Puig, and Soldo (2007) use self-employment and commission pay as instruments for insurance coverage and also find a more frequent use of some types of preventive care by individuals with health insurance coverage.

While the bulk of studies investigating checkups focus on one particular type of screening (rather than general health checkups), see Maciosek, Coffield, Flottemesch, Edwards, and Solberg (2010) for a literature review, several experimental contributions also assess general health checkups. For instance, Rasmussen, Thomsen, Kilsmark, Hvenegaard, Engberg, Lauritzen, and Sogaard

(2007) conduct an experiment with individuals aged 30 to 49 in Denmark by randomly offering a set of health screenings, including advice on healthy living and find a significant positive effect on life expectation. In a study on Japan's elderly population, Nakanishi, Tatara, and Fujiwara (1996) find a significantly negative correlation between the rate of attendance at health check-ups and hospital admission rates. Despite the effects of health insurance coverage and routine checkups being extensively covered in the public health literature, the indirect effect of insurance on general health operating via routine checkups as mediator has to the best of our best knowledge not yet been investigated. A further distinction to most previous studies is that we consider comparably young individuals with an average age below 30. For this population, the relative importance of different health screenings might differ from that for other age groups. We also point out that our application focusses on short term health effects.

We consider a binary indicator for health insurance coverage, equal to one if an individual reports to have any kind of health insurance when interviewed in 2006 and zero otherwise. The outcome, self-reported general health, is obtained from 2008 interview and measured with an ordinal variable, taking on the values 'excellent', 'very good', 'good', 'fair' and 'poor'. In the 2007 interview, participants were asked whether they have gone for routine checkups since the 2006 interview. This information serves as binary mediator, measured post-treatment but pre-outcome.

The control variables ($X$) come from the 2005 and earlier interview rounds. They cover demographic characteristics, family background and quality of the home environment during youth, education and training, labor market status, income and work experience, marital status and fertility, household characteristics, received monetary transfers, attitudes and expectations, state of physical and mental health as well as health-related behavior regarding e.g. nutrition and physical activity. For some variables, we only consider measurements from 2005 or from the initial interview round covering demographics and family related topics. For other variables we include measurements from both the indiviuals' youth and 2005 in order to capture their social, emotional and physical development. Treatment and mediator state in the pre-treatment period (2005) are also considered as potential control variables. Item non-response in control variables is dealt with by including missing dummies for each control variable and setting the respective missing values to zero. In total, we end up with a set of 770 control variables, 601 of which are dummy variables (incl. 252 dummies for missing values).

After excluding all observations with either mediator or treatment status missing, we remain with 7,061 observations. Table 2 presents some descriptive statistics for a selection of control variables. It shows that the group of individuals with and without health insurance coverage differ substantially. There are significant differences with respect to most of the control variables listed in the table. Females are significantly more likely to have health insurance coverage. Education and household income also show a significant positive correlation with health insurance coverage while the number of household members for example is negatively correlated with insurance coverage. Regarding the mediator, we find a similar pattern as for the treatment. With respect to many of the considered variables, the group of individuals who went for medical checkup differs substantially from those who did not. Further, we see that the correlation between many control variables and the treatment appear to have the same sign as that with the mediator.

In order to assess the direct and indirect effect of health insurance coverage on general health, we consider estimation based on Theorem 1 and expression (5) derived by Tchetgen Tchetgen and Shpitser (2012). We estimate the nuisance parameters and treatment effects in the same way as outlined in Section 5 (i.e. post-lasso regression for modeling the nuisance parameters and 4-fold cross fitting for effect estimation). The trimming threshold for discarding observations with too extreme propensity scores is set to 0.02 (2%), leading to 814 observations being dropped.

Table 3 provides the estimated effects along with the standard error ('se') and p-value ('p-val') and also provides the estimated mean potential outcome under non-treatment for comparison ('$\hat{E}[Y(0, M(0))]$'). Estimation based on Theorem 1 yields a moderately negative ATE of health insurance coverage on general health in the year 2008 (column 2) that is statistically significant at the 10% level. As the outcome is measured on an ordinal scale ranging from 'excellent' to 'poor', this suggests a short term health improving effect of health coverage. The direct effects under treatment (column 3) and non-treatment (column 4) are very similar to the ATE, while the indirect effects under treatment (column 5) and non-treatment (column 6) are close to zero and never significant. Thus, health insurance coverage does not seem to affect general health of young adults in the U.S. through routine checkups in the short run, but rather through other mechanisms.

[3]The HH income variable is the sum of several variables measuring HH income components (different sources and receivers). These variables are capped but only a total of 11 observations are in critical cap categories

Table 2: Descriptive Statistics

| | **overall** | $D = 1$ | $D = 0$ | **diff** | **p-val** | $M = 1$ | $M = 0$ | **diff** | **p-val** |
|---|---|---|---|---|---|---|---|---|---|
| N | 7,061 | 2,335 | 4,726 | | | 3,612 | 3,449 | | |
| Female | 0.5 | 0.55 | 0.42 | 0.13 | 0 | 0.66 | 0.34 | 0.32 | 0 |
| Age | 28.51 | 28.47 | 28.59 | -0.12 | 0 | 28.46 | 28.55 | -0.09 | 0.01 |
| Ethnicity | | | | | | | | | |
| *Black* | 0.27 | 0.25 | 0.3 | -0.05 | 0 | 0.32 | 0.21 | 0.11 | 0 |
| *Hispanic* | 0.21 | 0.19 | 0.26 | -0.07 | 0 | 0.21 | 0.21 | 0 | 0.76 |
| *Mixed* | 0.01 | 0.01 | 0.01 | 0 | 0.34 | 0.01 | 0.01 | 0 | 0.42 |
| *White or Other* | 0.51 | 0.55 | 0.43 | 0.12 | 0 | 0.46 | 0.56 | -0.1 | 0 |
| Relationship/Marriage | | | | | | | | | |
| *Not Cohabiting* | 0.62 | 0.61 | 0.65 | -0.03 | 0.01 | 0.61 | 0.64 | -0.03 | 0.01 |
| *Cohabiting* | 0.17 | 0.16 | 0.18 | -0.03 | 0.01 | 0.16 | 0.17 | 0 | 0.72 |
| *Married* | 0.18 | 0.21 | 0.14 | 0.07 | 0 | 0.2 | 0.17 | 0.03 | 0 |
| *Separated/ Widowed* | 0.02 | 0.02 | 0.03 | -0.01 | 0.03 | 0.02 | 0.02 | 0 | 0.62 |
| *Missing* | 0 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0.79 |
| Urban | 1.75 | 1.75 | 1.73 | 0.02 | 0.03 | 1.76 | 1.73 | 0.03 | 0.01 |
| *Missing* | 0.08 | 0.08 | 0.09 | -0.01 | 0.16 | 0.08 | 0.09 | -0.01 | 0.14 |
| HH Income [3] | 43,406 | 48,388 | 33,322 | 15,066 | 0 | 44,217 | 42,556 | 1,661 | 0.24 |
| *Missing* | 0.21 | 0.19 | 0.24 | -0.05 | 0 | 0.2 | 0.22 | -0.01 | 0.2 |
| HH Size | 3.09 | 3.06 | 3.15 | -0.1 | 0.04 | 3.13 | 3.05 | 0.08 | 0.07 |
| *Missing* | 0.06 | 0.05 | 0.07 | -0.02 | 0.01 | 0.05 | 0.07 | -0.01 | 0.03 |
| HH Members under 18 | 0.69 | 0.65 | 0.76 | -0.11 | 0 | 0.77 | 0.6 | 0.17 | 0 |
| *Missing* | 0.06 | 0.06 | 0.07 | -0.02 | 0.01 | 0.06 | 0.07 | -0.01 | 0.04 |
| Biological Children | 0.49 | 0.47 | 0.54 | -0.07 | 0 | 0.56 | 0.43 | 0.13 | 0 |
| Highest Grade | 12.17 | 12.65 | 11.21 | 1.44 | 0 | 12.41 | 11.93 | 0.48 | 0 |
| *Missing* | 0.06 | 0.06 | 0.07 | -0.02 | 0 | 0.06 | 0.07 | -0.01 | 0.04 |
| Employment | | | | | | | | | |
| *Employed* | 0.71 | 0.73 | 0.68 | 0.05 | 0 | 0.7 | 0.72 | -0.02 | 0.05 |
| *Unemployed* | 0.05 | 0.04 | 0.08 | -0.03 | 0 | 0.05 | 0.06 | -0.01 | 0.17 |
| *Out of Labor Force* | 0.21 | 0.19 | 0.24 | -0.04 | 0 | 0.21 | 0.2 | 0.01 | 0.4 |
| *Military* | 0.02 | 0.03 | 0.01 | 0.03 | 0 | 0.03 | 0.01 | 0.02 | 0 |
| *Missing* | 0 | 0 | 0 | 0 | 0.13 | 0 | 0 | 0 | 0.45 |
| Working Hours (per week) | 24.83 | 25.47 | 23.53 | 1.94 | 0 | 24.44 | 25.24 | -0.81 | 0.09 |
| *Missing* | 0.06 | 0.05 | 0.07 | -0.02 | 0.01 | 0.05 | 0.07 | -0.01 | 0.04 |
| Weight (pounds) | 157 | 157 | 157 | -1 | 0.64 | 154 | 160 | -6 | 0 |
| *Missing* | 0.08 | 0.08 | 0.1 | -0.02 | 0.01 | 0.08 | 0.09 | -0.01 | 0.09 |
| Height (feet) | 5.12 | 5.18 | 5.01 | 0.17 | 0 | 5.08 | 5.17 | -0.09 | 0.02 |
| *Missing* | 0.09 | 0.08 | 0.11 | -0.04 | 0 | 0.08 | 0.09 | -0.01 | 0.13 |
| Days 5+ drinks (per month) | 1.64 | 1.56 | 1.8 | -0.25 | 0.02 | 1.24 | 2.06 | -0.82 | 0 |
| *Missing* | 0.09 | 0.08 | 0.1 | -0.03 | 0 | 0.08 | 0.09 | -0.02 | 0.01 |
| Days of Exercise (per week) | 2.39 | 2.41 | 2.36 | 0.05 | 0.42 | 2.32 | 2.46 | -0.14 | 0.01 |
| *Missing* | 0.05 | 0.05 | 0.06 | -0.01 | 0.03 | 0.04 | 0.06 | -0.01 | 0.03 |
| Depressed/ Down | | | | | | | | | |
| *Never* | 0.31 | 0.32 | 0.29 | 0.02 | 0.05 | 0.29 | 0.32 | -0.03 | 0.01 |
| *Sometimes* | 0.51 | 0.52 | 0.48 | 0.03 | 0.01 | 0.52 | 0.49 | 0.02 | 0.06 |
| *Mostly* | 0.09 | 0.09 | 0.1 | -0.01 | 0.14 | 0.1 | 0.08 | 0.02 | 0 |
| *Always* | 0.02 | 0.02 | 0.03 | -0.01 | 0 | 0.02 | 0.02 | 0 | 0.68 |
| *Missing* | 0.08 | 0.07 | 0.1 | -0.03 | 0 | 0.07 | 0.09 | -0.01 | 0.02 |

Note: 'overall', '$D = 1$', '$D = 0$', '$M = 1$', '$M = 0$' report the mean of the respective vaiable in the total sample, among treated, among non-treated, among mediated, and among non-mediated, respectively. 'diff' and 'p-val' provide the mean difference (across treatment or mediator states) and the p-value of a two-sample t-test, respectively.

Table 3: Effects on general health in 2008 based on Theorem 1

| | $\hat{\Delta}$ | $\hat{\theta}(1)$ | $\hat{\theta}(0)$ | $\hat{\delta}(1)$ | $\hat{\delta}(0)$ | $\hat{E}[Y(0, M(0))]$ |
|---|---|---|---|---|---|---|
| effect | -0.05 | -0.05 | -0.05 | 0.00 | -0.00 | 2.33 |
| se | 0.03 | 0.04 | 0.03 | 0.01 | 0.02 | 0.02 |
| p-val | 0.07 | 0.26 | 0.05 | 0.76 | 0.83 | 0.00 |

Note: 'effect', 'se', and 'p-val' report the respective effect estimate, standard error and p-value. Lasso regression is used for the estimation of nuisance parameters. The propensity score-based trimming threshold is set to 0.02.

# 7 Conclusion

In this paper, we combined causal mediation analysis with double machine learning under selection-on-observables assumptions which avoids adhoc pre-selection of control variables. Thus, this approach appears particularly fruitful in high-dimensional data with many potential control variables. We proposed estimators for natural direct and indirect effects as well as the controlled direct effect exploiting efficient score functions, sample splitting, and machine learning-based plug-in estimates for conditional outcome means, mediator densities, and/or treatment propensity scores. We demonstrated the root-n consistency and asymptotic normality of the effect estimators under specific regularity conditions. Furthermore, we investigated the finite sample behavior of the proposed estimators in a simulation study and found the performance to be decent in samples with several thousand observations. Finally, we applied our method to data from the U.S. National Longitudinal Survey of Youth 1997 and found a moderate short term effect of health insurance coverage on general health, which was, however, not mediated by routine checkups.

# References


Albert, J. M. (2008): "Mediation analysis via potential outcomes models," *Statistics in Medicine*, 27, 1282–1304.

Albert, J. M., and S. Nelson (2011): "Generalized causal mediation analysis," *Biometrics*, 67, 1028–1038.

Baicker, K., S. L. Taubman, H. L. Allen, M. Bernstein, J. H. Gruber, J. P. New-

HOUSE, E. C. SCHNEIDER, B. J. WRIGHT, A. M. ZASLAVSKY, AND A. N. FINKELSTEIN (2013): "The Oregon experiment—effects of Medicaid on clinical outcomes," *New England Journal of Medicine*, 368(18), 1713–1722.

BARON, R. M., AND D. A. KENNY (1986): "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51, 1173–1182.

BELLANI, L., AND M. BIA (2018): "The long-run effect of childhood poverty and the mediating role of education," *forthcoming in the Journal of the Royal Statistical Society: Series A (Statistics in Society).*

BELLONI, A., V. CHERNOZHUKOV, I. FERNÁNDEZ-VAL, AND C. HANSEN (2017): "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, 85, 233–298.

BIJWAARD, G. E., AND A. M. JONES (2018): "An IPW estimator for mediation effects in hazard models: with an application to schooling, cognitive ability and mortality," *Empirical Economics*, pp. 1–47.

BURSTIN, H. R., K. SWARTZ, A. C. O'NEIL, E. J. ORAV, AND T. A. BRENNAN (1998): "The effect of change of health insurance on access to care," *Inquiry*, pp. 389–397.

CARDELLA, E., AND B. DEPEW (2014): "The effect of health insurance coverage on the reported health of young adults," *Economics Letters*, 124(3), 406–410.

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," *The Econometrics Journal*, 21, C1–C68.

COCHRAN, W. G. (1957): "Analysis of Covariance: Its Nature and Uses," *Biometrics*, 13, 261–281.

CONTI, G., J. J. HECKMAN, AND R. PINTO (2016): "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour," *The Economic Journal*, 126, F28–F65.

Faulkner, L., and H. Schauffler (1997): "The Effect of Health Insurance Coverage on the Appropriate Use f Recommended Clinical Preventive Services," *American journal of preventive medicine*, 13(6), 453–458.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA DP No. 4237*.

Fowler-Brown, A., G. Corbie-Smith, J. Garrett, and N. Lurie (2007): "Risk of cardiovascular events and death—does insurance matter?," *Journal of general internal medicine*, 22(4), 502–507.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103, 2052–2086.

Hong, G. (2010): "Ratio of mediator probability weighting for estimating natural direct and indirect effects," in *Proceedings of the American Statistical Association, Biometrics Section*, p. 2401–2415. Alexandria, VA: American Statistical Association.

Huber, M. (2014): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *Journal of Applied Econometrics*, 29, 920–943.

Huber, M. (2015): "Causal pitfalls in the decomposition of wage gaps," *Journal of Business and Economic Statistics*, 33, 179–191.

Huber, M., M. Lechner, and G. Mellace (2017): "Why Do Tougher Caseworkers Increase Employment? The Role of Program Assignment as a Causal Mechanism," *The Review of Economics and Statistics*, 99, 180–183.

Huber, M., M. Lechner, and A. Strittmatter (2018): "Direct and indirect effects of training vouchers for the unemployed," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 441–463.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imai, K., and T. Yamamoto (2013): "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," *Political Analysis*, 21, 141–171.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Judd, C. M., and D. A. Kenny (1981): "Process Analysis: Estimating Mediation in Treatment Evaluations," *Evaluation Review*, 5, 602–619.

Kaufman, J. S., R. F. MacLehose, and S. Kaufman (2004): "A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation," *Epidemiologic Perspectives & Innovations*, 1, 4.

Keele, L., D. Tingley, and T. Yamamoto (2015): "Identifying mechanisms behind policy interventions via causal mediation analysis," *Journal of Policy Analysis and Management*, 34, 937–963.

King, G., E. Gakidou, K. Imai, J. Lakin, R. T. Moore, C. Nall, N. Ravishankar, M. Vargas, M. M. Tellez-Rojo, J. E. H. Avila, et al. (2009): "Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme," *The lancet*, 373(9673), 1447–1454.

Luo, Y., and M. Spindler (2016): "High-Dimensional $L_2$Boosting: Rate of Convergence," .

Maciosek, M. V., A. B. Coffield, T. J. Flottemesch, N. M. Edwards, and L. I. Solberg (2010): "Greater use of preventive services in US health care could save lives at little or no cost," *Health Affairs*, 29(9), 1656–1660.

Nagelkerke, N. J. D. (1991): "A note on a general definition of the coefficient of determination," *Biometrika*, 78, 691–692.

Nakanishi, N., K. Tatara, and H. Fujiwara (1996): "Do preventive health services reduce eventual demand for medical care?," *Social Science & Medicine*, 43(6), 999–1005.

Neyman, J. (1959): *Optimal asymptotic tests of composite statistical hypotheses*p. 416–444. Wiley.

Pagán, J. A., A. Puig, and B. J. Soldo (2007): "Health insurance coverage and the use of preventive services by Mexican adults," *Health economics*, 16(12), 1359–1369.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006): "Estimation of Direct Causal Effects," *Epidemiology*, 17, 276–284.

Press, R. (2014): "Insurance Coverage and Preventive Care Among Adults," .

Rasmussen, S. R., J. L. Thomsen, J. Kilsmark, A. Hvenegaard, M. Engberg, T. Lauritzen, and J. Sogaard (2007): "Preventive health screenings and health consultations in primary care increase life expectancy without increasing costs," *Scandinavian journal of public health*, 35(4), 365–372.

Robins, J. M. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects," in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.

Robins, J. M., and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are not Always Observed," *Journal of the American Statistical Association*, 90, 846–866.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Simon, K., A. Soni, and J. Cawley (2017): "The impact of health insurance on preventive care and health behaviors: evidence from the first two years of the ACA Medicaid expansions," *Journal of Policy Analysis and Management*, 36(2), 390–417.

Sommers, B. D., B. Maylone, R. J. Blendon, E. J. Orav, and A. M. Epstein (2017): "Three-year impacts of the Affordable Care Act: improved medical care and health among low-income adults," *Health Affairs*, 36(6), 1119–1128.

Spindler, M., V. Chernozhukov, and C. Hansen (2016): "High-Dimensional Metrics," *arXiv:1608.00354*.

Tchetgen Tchetgen, E. J., and I. Shpitser (2012): "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis," *The Annals of Statistics*, 40, 1816–1845.

Ten Have, T. R., M. M. Joffe, K. G. Lynch, G. K. Brown, S. A. Maisto, and A. T. Beck (2007): "Causal mediation analyses with rank preserving models," *Biometrics*, 63, 926–934.

Tibshirani, R. (1996): "Regresson shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society*, 58, 267–288.

VanderWeele, T. J. (2009): "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26.

Vansteelandt, S., M. Bekaert, and T. Lange (2012): "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects," *Epidemiologic Methods*, 1, 129–158.

Yörük, B. K. (2016): "Health insurance coverage and self-reported health: new estimates from the NLSY97," *International journal of health economics and management*, 16(3), 285–295.

# Appendices

## A Score linearity & Neyman orthogonality for the score in (5)

The score function for the counterfactual proposed by Tchetgen Tchetgen and Shpitser (2012) for $\Psi_{d0} = E[d, M(1-d)]$ is given by:

$$\begin{aligned}
\psi_d(W, \eta, \Psi_{d0}) \quad = \quad & \frac{I\{D=d\} \cdot f(M|1-d, X)}{p_d(X) \cdot f(M|d, X)} \cdot [Y - \mu(d, M, X)] \\
& + \frac{I\{D=1-d\}}{1-p_d(X)} \cdot \Big[\mu(d, M, X) - \overbrace{\int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X) dm}^{=:\nu(1-d,X)}\Big] \\
& + \underbrace{\int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X) dm}_{=:\nu(1-d,X)} \quad - \quad \Psi_{d0}
\end{aligned}$$

**Linearity:** The score $\psi_d(W, \eta_0, \Psi_{d0})$ is linear in $\Psi_{d0}$ as it can be written as: $\psi_d(W, \eta_0, \Psi_{d0}) = \psi_d^a(W, \eta_0) \cdot \Psi_0 + \psi_d^b(W, \eta_0)$ with $\psi_d^a(W, \eta_0) = -1$ and

$$\psi_d^b(W, \eta_0) = \frac{I\{D=d\} \cdot f_0(M|1-d, X)}{p_{d0}(X) \cdot f_0(M|d, X)}[Y - \mu_0(d, M, X)] + \frac{I\{D=1-d\}}{1-p_{d0}(X)}\Big[\mu_0(d, M, X) - \nu_0(1-d, X)\Big] + \nu_0(1-d, X)$$

**Moment Condition:** The moment condition $E\Big[\psi_d(W, \hat{\eta}, \Psi_{d0})\Big] = 0$ is satisfied:

$$\begin{aligned}
E\Big[\psi_d(W, \hat{\eta}, \Psi_{d0})\Big] \quad = \quad & E\Bigg[\frac{I\{D=d\} \cdot \hat{f}(M|1-d, X)}{\hat{p_d}(X) \cdot \hat{f}(M|d, X)} \cdot \overbrace{E\Big[Y - \hat{\mu}(d, M, X)\Big| D=d, M, X\Big]}^{=0}\Bigg] \\
& + E\Bigg[\overbrace{E\Big[E\Big[\frac{I\{D=1-d\}}{1-\hat{p_d}(X)} \cdot [\hat{\mu}(d, M, X) - \hat{\nu}(1-d, X)]\Big| M, X\Big]\Big| X\Big]}^{E\left[\frac{1-\hat{p_d}(M,X)}{1-\hat{p_d}(X)} \cdot [\hat{\mu}(d,M,X) - \hat{\nu}(1-d,X)]\Big| X\right] = E\left[\frac{1-\hat{p_d}(M,X)}{1-\hat{p_d}(X)} \cdot \hat{\mu}(d,M,X)|X\right] - E[\hat{\nu}(1-d,X)|X] = 0}\Bigg] \\
& + E[\hat{\nu}(1-d, X)] \quad - \quad \Psi_{d0} \\
= \quad & \Psi_{d0} \quad - \quad \Psi_{d0} \quad = 0
\end{aligned}$$

**Neyman Orthogonality**: The Gateaux derivative in the direction $\eta - \eta_0 = (f(M|D, X) - f_0(M|D, X), p_d(X) - p_{d0}(X), \mu(d, M, X) - \mu_0(d, M, X))$ is given by:

$$
\begin{aligned}
&\partial E\big[\psi_d(W,\eta,\Psi_d)\big]\big[\eta-\eta_0\big] \\
&= E\left[\frac{\big[f(M|1-d,X)-f_0(M|1-d,X)\big]\cdot f_0(M|d,X)-\big[f(M|d,X)-f_0(M|d,X)\big]\cdot f_0(M|1-d,X)}{f_0(M|d,X)^2}\cdot\frac{I\{D=d\}}{p_{d0}(X)}\cdot\overbrace{\Big(Y-\mu_0(d,M,X)\Big)}^{E[\cdot|D=d,M,X]=0}\right] \\
&\underbrace{- E\left[\underbrace{\frac{I\{D=1-d\}}{1-p_{d0}(X)}}_{E[\cdot|X]=1}\cdot\partial E[\nu_0(1-d,X)][f(M|1-d,X)-f_0(M|1-d,X)]\right]+\partial E[\nu_0(1-d,X)][f(M|1-d,X)-f_0(M|1-d,X)]}_{=0} \\
&- E\left[\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_{d0}^2(X)\cdot f_0(M|d,X)}\cdot\underbrace{\Big(Y-\mu_0(d,M,X)\Big)}_{E[\cdot|D=d,M,X]=0}\cdot\Big[p_d(X)-p_{d0}(X)\Big]\right] \\
&+ E\left[\underbrace{\frac{I\{D=1-d\}}{(1-p_{d0}(X))}\cdot\Big(\mu_0(d,M,X)-\nu_0(1-d,X)\Big)}_{E[E[\cdot|M,X]|X]=E\left[\frac{1-p_d(M,X)}{1-p_d(X)}\cdot\left(\mu_0(d,M,X)-\nu_0(1-d,X)\right)\Big|X\right]=E\left[\mu_0(d,M,X)-\nu_0(1-d,X)\Big|D=1-d,X\right]=0}\cdot\frac{1}{(1-p_{d0}(X))}\cdot\Big[p_d(X)-p_{d0}(X)\Big]\right] \\
&- \underbrace{E\left[\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}\cdot\Big[\mu(d,M,X)-\mu_0(d,M,X)\Big]\right]}_{E[\cdot]=E[E[\cdot|M,X]]=E\left[\frac{p_{d0}(M,X)\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}\cdot[\mu(d,M,X)-\mu_0(d,M,X)]\right]} \qquad (*) \\
&+ \underbrace{E\left[\frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot\Big[\mu(d,M,X)-\mu_0(d,M,X)\Big]\right]}_{E[\cdot]=E[E[\cdot|M,X]]=E\left[\frac{1-p_{d0}(M,X)}{1-p_{d0}(X)}\cdot[\mu(d,M,X)-\mu_0(d,M,X)]\right]} \qquad (**) \\
&\underbrace{- E\left[\underbrace{\frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot\partial E[\nu_0(1-d,X)][\mu(d,M,X)-\mu_0(d,M,X)]}_{E[\cdot|X]=\frac{1-p_{d0}(X)}{1-p_{d0}(X)}\cdot\partial E[\nu_0(1-d,X)][\mu(d,M,X)-\mu_0(d,M,X)]}\right]+\partial E[\nu_0(1-d,X)][\mu(d,M,X)-\mu_0(d,M,X)],}_{=0}
\end{aligned}
$$

where terms $(*)$ and $(**)$ cancel out by Bayes' Law, $\frac{p_{d0}(M,X)\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}=\frac{p_{d0}(M,X)\cdot(1-p_{d0}(M,X))}{p_{d0}(M,X)\cdot(1-p_{d0}(X))}=\frac{1-p_{d0}(M,X)}{1-p_{d0}(X)}$. Thus, it follows that:

$$
\partial E\big[\psi_d(W,\eta,\Psi_d)\big]\big[\eta-\eta_0\big]=0
$$

proving that the score function is orthogonal.

# B Score linearity & Neyman orthogonality for the score in (8)

The score for the controlled direct effect is given by:

$$
E\big[\psi_d(W,\eta,\Psi_{dm0})\big]=E\left[\frac{I\{D=d\}\cdot I\{M=m\}\cdot[Y-\mu(d,m,X)]}{f(m|d,X)\cdot p_d(m,X)}+\mu(d,m,X)-\Psi_{dm0}\right]
$$

**Linearity:** The score $\psi_d(W, \eta_0, \Psi_{dm0})$ is linear in $\Psi_{dm0}$ as it can be written as: $\psi_d(W, \eta_0, \Psi_{dm0}) = \psi_d^a(W, \eta_0) \cdot \Psi_{dm0} + \psi_d^b(W, \eta_0)$ with $\psi_d^a(W, \eta_0) = -1$ and

$$\psi_d^b(W, \eta_0) = \frac{I\{D = d\} \cdot I\{M = m\} \cdot [Y - \mu(d, m, X)]}{f(m|d, X) \cdot p_d(m, X)} + \mu(d, m, X)$$

**Moment condition:** The moment condition $E\Big[\psi_d(W, \hat{\eta}, \Psi_{dm0})\Big] = 0$ is satisfied:

$$\begin{aligned}
E\Big[\psi_{dm}(W, \hat{\eta}, \Psi_{dm0})\Big] &= E\left[\frac{I\{D = d\} \cdot I\{M = m\} \cdot [Y - \hat{\mu}(d, m, X)]}{\hat{f}(m|d, X) \cdot \hat{p_d}(m, X)} + \hat{\mu}(d, m, X) - \Psi_{dm0}\right] \\
&= E\left[\frac{I\{D = d\} \cdot I\{M = m\} \cdot \overbrace{E[Y - \hat{\mu}(d, m, X)|d, m, X]}^{=0}}{\hat{f}(m|d, X) \cdot \hat{p_d}(m, X)}\right] + E\Big[\hat{\mu}(d, m, X)\Big] - \Psi_{dm0} \\
&= \Psi_{dm0} - \Psi_{dm0} = 0
\end{aligned}$$

**Neyman orthogonality**: The Gateaux derivative in the direction $\eta - \eta_0 = (f(M|D, X) - f_0(M|D, X), p_d(X) - p_{d0}(X), \mu_d(d, M, X) - \mu_0(d, M, X))$ is given by:

$$\begin{aligned}
&\partial E\big[\psi_{dm}(W, \eta, \Psi_{dm})\big]\big[\eta - \eta_0\big] \\
&= \overbrace{-E\left[\underbrace{\frac{I\{D = d\} \cdot I\{M = m\}}{f_0(m|d, X) \cdot p_{d0}(m, X)}}_{E[\cdot|X] = \frac{\Pr(D=d, M=m|X)}{\Pr(D=d, M=m|X)} = 1} \cdot \Big[\mu(d, m, X) - \mu_0(d, m, X)\Big]\right] + E\Big[\mu(d, m, X) - \mu_0(d, m, X)\Big]}^{=0} \\
&\quad - E\left[\frac{I\{D = d\} \cdot I\{M = m\} \cdot \overbrace{[Y - \mu_0(d, m, X)]}^{E[\cdot|D=d, M=m, X]=0} \cdot \Big[f(m|d, X) - f_0(m|d, X)\Big]}{f_0(m|d, X)^2 \cdot p_{d0}(m, X)}\right] \\
&\quad - E\left[\frac{I\{D = d\} \cdot I\{M = m\} \cdot \overbrace{[Y - \mu_0(d, m, X)]}^{E[\cdot|D=d, M=m, X]=0} \cdot \Big[p_d(m, X) - p_{d0}(m, X)\Big]}{f_0(m|d, X) \cdot p_{d0}(m, X)^2}\right].
\end{aligned}$$

Thus, it follows that:

$$\partial E\big[\psi_{dm}(W, \eta, \Psi_{dm})\big]\big[\eta - \eta_0\big] = 0$$

proving that the score function is orthogonal.

## C Score linearity & Neyman orthogonality for the score in (6)

The alternative score for the counterfactual based on (6) is given by:

$$\begin{aligned}\psi_d^*(W,\eta,\Psi_{d0}) = & \frac{I\{D=d\}\cdot(1-p_d(M,X))}{p_d(M,X)\cdot(1-p_d(X))}\cdot\Big[Y-\mu(d,M,X)\Big] \\ & + \frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Bigg[\mu(d,M,X) - \overbrace{E\Big[\mu(d,M,X)\cdot\frac{1-p_d(M,X)}{1-p_d(X)}\,\Big|X\Big]}^{=:\omega(1-d,X)}\Bigg] \\ & + \overbrace{E\Big[\mu(d,M,X)\cdot\frac{1-p_d(M,X)}{1-p_d(X)}\,\Big|X\Big]}^{=:\omega(1-d,X)}\Bigg] - \Psi_{d0}\end{aligned}$$

with $\eta = (p_d(M,X), p_d(X), \mu(D,M,X), \omega(D,X))$.

**Linearity:** The score $\psi_d^*(W,\eta_0,\Psi_{d0})$ is linear in $\Psi_{d0}$ as it can be written as: $\psi_d^*(W,\eta_0,\Psi_{d0}) = \psi_d^a(W,\Psi_{d0})\cdot\Psi_{d0} + \psi_d^b(W,\eta_0)$ with $\psi_d^a(W,\eta_0) = -1$ and

$$\psi_d^b(W,\eta_0) = \frac{I\{D=d\}\cdot(1-p_{d0}(M,X))}{p_{d0}(M,X)\cdot(1-p_{d0}(X))}\cdot\Big[Y-\mu_0(d,M,X)\Big] + \frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot\Big[\mu_0(d,M,X)-\omega_0(1-d,X)\Big] + \omega_0(1-d,$$

**Moment condition:** The moment condition $E\Big[\psi_d(W,\hat{\eta},\Psi_{d0})\Big] = 0$ is satisfied:

$$\begin{aligned}E\Big[\psi_d^*(W,\hat{\eta},\Psi_{d0})\Big] &= E\Bigg[\frac{I\{D=d\}\cdot(1-\hat{p_d}(M,X))}{\hat{p_d}(M,X)\cdot(1-\hat{p_d}(X))}\cdot\overbrace{E\Big[Y-\hat{\mu}(d,M,X)\Big|D=d,M,X]\Big]}^{=0}\Bigg] \\ &\quad + E\Bigg[\overbrace{E\Big[E\Big[\frac{I\{D=1-d\}}{1-\hat{p_d}(X)}\cdot[\hat{\mu}(d,M,X)-\hat{\omega}(1-d,X)]\Big|M,X\Big]\Big|X\Big]}^{E\left[\frac{1-\hat{p_d}(M,X)}{1-\hat{p_d}(X)}\cdot[\hat{\mu}(d,M,X)-\hat{\omega}(1-d,X)]|X\right]=E[\frac{1-\hat{p_d}(M,X)}{1-\hat{p_d}(X)}\cdot\hat{\mu}(d,M,X)|X]-E[\hat{\omega}(1-d,X)|X]=0}\Bigg] \\ &\quad + E[\hat{\omega}(1-d,X)] - \Psi_{d0} \\ &= \Psi_{d0} - \Psi_{d0} = 0\end{aligned}$$

**Neyman orthogonality**:

The Gateaux derivative in the direction $\eta-\eta_0 = (p_d(M.X)-p_{d0}(M,X), p_d(X)-p_{d0}(X), \mu_d(d,M,X)-$

$\mu_0(d, M, X), \omega(1-d, X) - \omega_0(1-d, X))$ is given by:

$$
\begin{aligned}
&\partial E\big[\psi_d^*(W, \eta, \Psi_d)\big]\big[\eta - \eta_0\big] \\
&= E\left[\frac{-p_d(M,X) - p_{d0}(M,X)}{p_d(M,X)^2} \cdot \frac{I\{D=d\}}{1-p_d(X)} \cdot \overbrace{\big(Y - \mu(d,M,X)\big)}^{E[\cdot|D=d,M,X]=0}\right] \\
&+ E\left[\frac{I\{D=d\} \cdot (1-p_d(M,X))}{p_d(M,X)} \cdot \overbrace{\big(Y - \mu(d,M,X)\big)}^{E[\cdot|D=d,M,X]=0} \cdot \frac{p_d(X) - p_{d0}(X)}{(1-p_d(X))^2}\right] \\
&+ E\Bigg[\underbrace{\frac{I\{D=1-d\}}{(1-p_d(X))} \cdot \big(\mu_0(d,M,X) - \omega_0(1-d,X)\big)}_{E[E[\cdot|M,X]|X] = E\left[\frac{1-p_d(M,X)}{1-p_d(X)} \cdot \left(\mu_0(d,M,X) - \omega_0(1-d,X)\right)\middle| X\right] = E\left[\mu_0(d,M,X) - \omega_0(1-d,X)\middle| D=1-d, X\right] = 0} \cdot \frac{p_d(X) - p_{d0}(X)}{(1-p_d(X))}\Bigg] \\
&\quad \underbrace{-E\left[\underbrace{\frac{I\{D=d\}}{p_d(M,X)}}_{E[\cdot|M,X]=1} \cdot \frac{(1-p_d(M,X))}{(1-p_d(X))} \cdot \Big[\mu(d,M,X) - \mu_0(d,M,X)\Big]\right] + E\left[\underbrace{\frac{I\{D=1-d\}}{1-p_d(X)}}_{E[\cdot|M,X] = \frac{1-p_d(M,X)}{1-p_d(X)}} \cdot \Big[\mu(d,M,X) - \mu_0(d,M,X)\Big]\right]}_{=0} \\
&\quad \underbrace{-E\left[\underbrace{\frac{I\{D=1-d\}}{1-p_d(X)}}_{E[\cdot|X]=1} \cdot \Big[\omega(1-d,X) - \omega_0(1-d,X)\Big] + \Big[\omega(1-d,X) - \omega_0(1-d,X)\Big]\right]}_{=0}.
\end{aligned}
$$

Thus, it follows that:

$$
\partial E\big[\psi_d^*(W, \eta, \Psi_{d0})\big]\big[\eta - \eta_0\big] = 0
$$

proving that the score function is orthogonal.

# D Simulation results for standard errors

Table D.1: Simulation results for standard errors ($p = 200$)

| | Coefficients given by $0.3/i^2$ for $i = 1, ..., p$ | | | | | | | Coefficients given by $0.5/i^2$ for $i = 1, ..., p$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | abias | sd | rmse | abias | sd | rmse | true | abias | sd | rmse | abias | sd | rmse | true |
| | $n$=1000 | | | $n$=4000 | | | | $n$=1000 | | | $n$=4000 | | | |
| | Double machine learning based on Theorem 1 | | | | | | | | | | | | | |
| $se(\hat{\Delta})$ | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.04 |
| $se(\hat{\theta}(1))$ | 0.02 | 0.01 | 0.02 | 0.01 | 0.00 | 0.01 | 0.04 | 0.02 | 0.01 | 0.02 | 0.00 | 0.00 | 0.01 | 0.04 |
| $se(\hat{\theta}(0))$ | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.04 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.04 |
| $se(\hat{\delta}(1))$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.03 |
| $se(\hat{\delta}(0))$ | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.02 |
| | Double machine learning based on Theorem 2 | | | | | | | | | | | | | |
| $se(\hat{\Delta})$ | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.04 |
| $se(\hat{\theta}(1))$ | 0.01 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.04 | 0.04 | 0.00 | 0.01 | 0.01 | 0.07 |
| $se(\hat{\theta}(0))$ | 0.00 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.04 | 0.01 | 0.05 | 0.05 | 0.00 | 0.01 | 0.01 | 0.07 |
| $se(\hat{\delta}(1))$ | 0.01 | 0.02 | 0.03 | 0.00 | 0.00 | 0.00 | 0.03 | 0.01 | 0.05 | 0.06 | 0.00 | 0.01 | 0.01 | 0.07 |
| $se(\hat{\delta}(0))$ | 0.00 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.03 | 0.01 | 0.04 | 0.05 | 0.00 | 0.01 | 0.01 | 0.06 |

Note: 'abias', 'sd', and 'rmse' denote the absolute bias, standard deviation and root mean squared error of the respective standard error ('se'). 'true' provides the true standard deviation.

Working Paper 515 may 2020

## Authors

Helmut FARBMACHER
Max Planck Society, MEA, Amalienstr. 33, 80799 Munich GERMANY
Phone: +49 (0)89/38602-330, Email: farbmacher@mpisoc.mpg.de; Website: www.farbmacher.de

Martin HUBER
University of Fribourg, Faculty of Economics and Social Sciences, Chair of Applied Econometrics - Evaluation of Public Policies, Bd. de Pérolles 90, 1700 Fribourg, Switzerland.
Phone: +41 26 300 8255; Email: martin.huber@unifr.ch; Website: http://www.unifr.ch/appecon/en/

Henrika LANGEN
University of Fribourg, Department of Economics, Bd. de Pérolles 90, 1700 Fribourg, Switzerland. Phone: +41 26 300 8276; Email: henrika.langen@unifr.ch; Website: https://www3.unifr.ch/appecon/en/chair/team/henrika-langen

Martin SPINDLER
Universität Hamburg, Fakultät für Betriebswirtschaft, Moorweidenstr. 18, 20148 Hamburg, Germany
Phone: +49 40 42838-1544, Email: martin.spindler@uni-hamburg.de; Website: https://www.bwl.uni-hamburg.de/matstat/team/

## Abstract

This paper combines causal mediation analysis with double machine learning to control for observed confounders in a data-driven way under a selection-on-observables assumption in a high-dimensional setting. We consider the average indirect effect of a binary treatment operating through an intermediate variable (or mediator) on the causal path between the treatment and the outcome, as well as the unmediated direct effect. Estimation is based on efficient score functions, which possess a multiple robustness property w.r.t. misspecifications of the outcome, mediator, and treatment models. This property is key for selecting these models by double machine learning, which is combined with data splitting to prevent overfitting in the estimation of the effects of interest. We demonstrate that the direct and indirect effect estimators are asymptotically normal and root-n consistent under specific regularity conditions and investigate the finite sample properties of the suggested methods in a simulation study when considering lasso as machine learner. We also provide an empirical application to the U.S. National Longitudinal Survey of Youth, assessing the indirect effect of health insurance coverage on general health operating via routine checkups as mediator, as well as the direct effect. We find a moderate short term effect of health insurance coverage on general health which is, however, not mediated by routine checkups.

## Citation proposal

Helmut Farbmacher, Martin Huber, Henrika Langen, Martin Spindler. 2020. «Causal mediation analysis with double machine learning». Working Papers SES 515, Faculty of Management, Economics and Social Sciences, University of Fribourg (Switzerland)

## Jel Classification

C21

## Keywords

Mediation; direct and indirect e□ects; causal mechanisms; double machine learning; effcient score.

## Working Papers SES collection

### Last published

509 Songane M.: Patenting Government Funded Innovations as a Strategy to Increase Budget for Academic Vaccine R&D; 2019
510 Songane M., Grossmann V.: The Patent Buyout Price for Human Papilloma Virus (HPV) Vaccine and the Ratio of R&D Costs to the Patent Value; 2019
511 Ducret R., Isakov D.: The Korea discount and chaebols; 2019
512 Wu J., Morschett D.: M&A by Chinese POEs in Developed Countries. Acquiring and Bundling Strategic Assets; 2020
513 Wallimann H., Imhof D., Huber M.: A Machine Learning Approach for Flagging Incomplete Bid-rigging Cartels; 2020
514 Huber M., Lafférs L.: Bounds on direct and indirect effects under treatment/mediator endogeneity and outcome attrition; 2020

### Catalogue and download links

http://www.unifr.ch/ses/wp
http://doc.rero.ch/collection/WORKING_PAPERS_SES

### Publisher

Université de Fribourg, Suisse, Faculté des sciences économiques et sociales et du management
Universität Freiburg, Schweiz, Wirtschafts- und sozialwissenschaftliche Fakultät
University of Fribourg, Switzerland, Faculty of Management, Economics and Social Sciences

Bd de Pérolles 90
CH-1700 Fribourg
Tél.: +41 26 300 8200
decanat-ses@unifr.ch