The Econometrics Journal

Econometrics Journal (2022), volume **25**, pp. 277–300.
https://doi.org/10.1093/ectj/utac003

Causal mediation analysis with double machine learning

HELMUT FARBMACHER[†], MARTIN HUBER[‡], LUKÁŠ LAFFÉRS[‖], HENRIKA LANGEN[‡] AND MARTIN SPINDLER[¶]

[†]*TUM School of Management, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany.*
Email: farbmacher@tum.de

[‡]*Department of Economics, University of Fribourg, Bd de Pérolles 90, 1700 Fribourg, Switzerland.*
Email: martin.huber@unifr.ch, henrika.langen@unifr.ch

[‖]*Department of Mathematics, Matej Bel University, Tajovskeho 40, 97411 Banská Bystrica, Slovakia.*
Email: lukas.laffers@gmail.com

[¶]*Faculty of Business Administration, University of Hamburg, Moorweidenstr. 18, 20148 Hamburg, Germany.*
Email: martin.spindler@uni-hamburg.de

First version received: 9 July 2021; final version accepted: 10 December 2021.

Summary: This paper combines causal mediation analysis with double machine learning for a data-driven control of observed confounders in a high-dimensional setting. The average indirect effect of a binary treatment and the unmediated direct effect are estimated based on efficient score functions, which are robust with respect to 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. We demonstrate that the effect estimators are asymptotically normal and $n^{-1/2}$-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 US 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.

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

JEL codes: *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 control-

ling 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 et al. (2006), VanderWeele (2009), Hong (2010), Imai et al. (2010), Albert and Nelson (2011), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt et al. (2012), Imai and Yamamoto (2013), 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 et al. (2013), Huber (2015), Keele et al. (2015), Conti et al. (2016), Huber et al. (2017), Bellani and Bia (2018), Bijwaard and Jones (2018), and Huber et al. (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 with reference to (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.

To improve upon this practice, 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 et al. (2018) for a data-driven control of observed confounders to obtain valid inference under specific regularity conditions. In particular, 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.[1] 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 $n^{-1/2}$-consistency and asymptotic normality of the proposed effect estimators under specific regularity conditions by verifying that the general framework of Chernozhukov et al. (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 et al. (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. We show that the efficient score function of Tchetgen Tchetgen and Shpitser (2012) satisfies the so-called Neyman (1959) orthogonality discussed in Chernozhukov et al. (2018), which makes the

[1] Different from conventional semiparametric methods, the double machine learning framework does not require the set of potential confounders to be restricted by Donsker conditions, but permits the set to be unbounded and to grow with the sample size.

estimation of direct and indirect effects rather insensitive to (local) estimation errors in the plug-in estimates. We transform the score function of Tchetgen Tchetgen and Shpitser (2012) by an application of Bayes' Law in a way that avoids the estimation of the conditional mediator density, as discussed in Zheng and van der Laan (2012) and also adopted by Díaz and Hejazi (2020), and show it to be Neyman orthogonal. This appears particularly useful when the mediator is a vector of variables and/or continuous, making conditional mediator density estimation cumbersome. Further, we establish the score function required for estimating the controlled direct effect along with Neyman orthgonality.

Neyman orthgonality is key for the fruitful application of double machine learning, ensuring robustness in the estimation of the nuisance parameters which is crucial when applying modern machine learning methods. Random sample splitting—to estimate the parameters 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 effect 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 et al. (2018). Under specific regularity conditions, this convergence rate can attained by various machine learning algorithms including lasso regression, see Tibshirani (1996).

We investigate the estimators' finite sample behaviour based on the score function of Tchetgen Tchetgen and Shpitser (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) conducted by the Bureau of Labor Statistics at the US Department of Labor (2019), where a large set of potential control variables is available. We disentangle the short-term 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, though statistically insignificant, health-improving direct effect, the indirect effect is very close to zero. We therefore do not find evidence that health insurance coverage affects general health through routine checkups in the short run.

We note that basing estimation on efficient score functions is not the only framework satisfying the previously mentioned robustness w.r.t. estimation errors in plug-in parameters. This property is also satisfied by the targeted maximum likelihood estimation (TMLE) framework by van der Laan and Rubin (2006), see the discussion in Díaz (2020). TMLE relies on iteratively updating (or robustifying) an initial estimate of the parameter of interest based on regression steps that involve models for the plug-in parameters. Zheng and van der Laan (2012) have developed an estimation approach for natural direct and indirect effects using TMLE, where the plug-in parameters might by estimated by machine learners, e.g., the super learner, an ensemble method suggested by van der Laan et al. (2007). This iterative estimation approach is therefore an alternative to the double machine learning based approach suggested in this paper, for which we demonstrate $n^{-1/2}$-consistency under specific conditions.

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 $n^{-1/2}$-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 et al. (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.[2] 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:

$$\theta(d) = E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}. \tag{2.1}$$

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.

$$\delta(d) = E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}. \tag{2.2}$$

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

$$\begin{aligned}\Delta &= E[Y(1, M(1)) - Y(0, M(0))] \\ &= E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0) \\ &= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1). \end{aligned} \tag{2.3}$$

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 checkups (M). We note that a different approach to dealing with the interaction effects between D and M is a three-way decomposition of the ATE into the

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

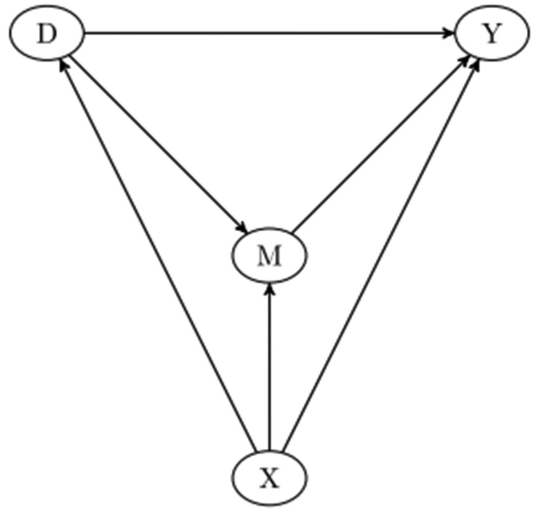

Figure 1. Causal paths under conditional exogeneity given pre-treatment covariates.

pure direct effect ($\theta(0)$), the pure indirect effect ($\delta(0)$) and the mediated interaction effect, see VanderWeele (2013).

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:

$$\gamma(m) = E[Y(1, m) - Y(0, m)], \qquad \text{for } m \text{ in the support of } M. \tag{2.4}$$

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 et al. (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 conditioning 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.

Formally, the first assumption invokes conditional independence of the treatment and potential mediators or outcomes given X. This restriction has been referred to as 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 nonexperimental data, the plausibility of this assumption critically hinges on the richness of X.

ASSUMPTION 3.1. (CONDITIONAL INDEPENDENCE OF THE TREATMENT) $\{Y(d', m), M(d)\} \perp D|X = x$ *for all* $d', d \in \{0, 1\}$ *and* m, x *in the support of* M, X, *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 3.2. (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 3.2 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 3.3. (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.

The common support assumption, also known as positivity or covariate overlap assumption, 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 treatment must not be deterministic in X, otherwise no comparable units in terms of X are available across treatment states. By Bayes' Law, Assumption 3.3 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 3.1 to 3.3 are standard in the causal mediation literature, see for instance Imai et al. (2010), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt et al. (2012), and Huber (2014), or also Pearl (2001), Petersen et al. (2006), and Hong (2010), for closely related restrictions.

We identify the counterfactual $E[Y(d, M(1 - d))]$ based on the following lemma proven by Tchetgen Tchetgen and Shpitser (2012).

LEMMA 3.1. *Under Assumptions 3.1, 3.2, and 3.3, the counterfactual* $E[Y(d, M(1-d))]$ *is identified by the following efficient score function:*

$$
\begin{aligned}
E[Y(d, M(1-d))] &= E[\psi_d], \\
\text{with } \psi_d &= \frac{I\{D=d\} \cdot f(M|1-d, X)}{p_d(X) \cdot f(M|d, X)} \cdot [Y - \mu(d, M, X)] \\
&\quad + \frac{I\{D=1-d\}}{1-p_d(X)} \cdot \Bigg[\mu(d, M, X) \\
&\quad - \int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X)\, dm\Bigg] \\
&\quad + \int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X)\, dm
\end{aligned} \tag{3.1}
$$

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*.*

(3.1) satisfies a multiple robustness property in the sense that estimation remains consistent even if one out of the three models for the plug-in parameters $f(M|D, X)$, $p_d(X)$, and $\mu(D, M, X)$ is misspecified.

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

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

where $f(M|X)$ is the conditional distribution of M given X and $p_d(X, M) = \Pr(D = d|X, M)$. Furthermore,

$$
\int \mu(d, m, X) \cdot f(m|1-d, X) dm = E\Big[\mu(d, M, X)\Big|D = 1-d, X\Big].
$$

As also noticed in Zheng and van der Laan (2012), the counterfactual can as well be identified based on an alternative multiply robust representation of (3.1), as provided in the following lemma.

LEMMA 3.2. *Under Assumptions 3.1, 3.2, and 3.3, the counterfactual* $E[Y(d, M(1-d))]$ *is identified by the following alternative efficient score function:*

$$E[Y(d, M(1-d))] = E[\psi_d^*],$$
$$\text{with } \psi_d^* = \frac{I\{D=d\}\cdot(1-p_d(M,X))}{p_d(M,X)\cdot(1-p_d(X))}\cdot[Y-\mu(d,M,X)]$$
$$+\frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Big[\mu(d,M,X)-E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]\Big]$$
$$+E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]. \tag{3.2}$$

Similarly, as the approaches based on inverse probability weighting (rather than efficient scores) in Huber (2014) and Tchetgen Tchetgen (2013), (3.2) avoids conditional mediator densities, which appears attractive if M is continuous and/or multidimensional. On the other hand, it requires the estimation of an additional parameter, namely the nested conditional mean $E[\mu(d, M, X)|D = 1-d, X]$, as similarly found in Miles et al. (2020), who suggest a multiply robust score function for assessing path-specific effects. Alternatively to rearranging the score function by Tchetgen Tchetgen and Shpitser (2012) as outlined above, ratios of conditional densities, as for instance appearing in the first component of (3.1), might be treated as additional nuisance parameters and estimated directly via density-ratio estimation, see e.g., Sugiyama et al. (2010) for density-ratio estimation in high-dimensional settings. Such methods based on directly estimating the density ratio without going through estimating the densities in numerator and denominator separately are shown in several studies to compare favourably with estimating the densities separately, see e.g., Kanamori et al. (2012).

Efficient score-based identification of $E[Y(d, M(d))]$ under $Y(d,m)\perp\{D, M\}|X = x$ (see Assumptions 3.1 and 3.2) has been established in the literature on doubly robust ATE estimation, see for instance Robins et al. (1994) and Hahn (1998):

LEMMA 3.3. *Under Assumptions 3.1, 3.2 and 3.3, the potential outcome* $E[Y(d, M(d))]$ *is identified by the following efficient score function:*

$$E[Y(d, M(d))] = E[\alpha_d] \text{ with } \alpha_d = \frac{I\{D=d\}\cdot[Y-\mu(d,X)]}{p_d(X)}+\mu(d,X), \tag{3.3}$$

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.

For identifying the controlled direct effect, we now assume that M is discrete (while this need not be the case in the context of natural direct and indirect effects) such that for all m in the support of M, it must hold that $\Pr(M = m) > 0$. As Assumptions 3.1 and 3.2 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 (3.3) by $I\{D = d, M = m\} = I\{M = m\}\cdot I\{D = d\}$ and $\Pr(D = d, M = m|X) = f(m|d, X)\cdot p_d(X)$:

LEMMA 3.4. *Under Assumptions 3.1, 3.2, and 3.3, the potential outcome* $E[Y(d, m)]$ *is identified by the following efficient score function:*

$$E[Y(d,m)] = E[\psi_{dm}]$$
$$\text{with } \psi_{dm} = \frac{I\{D=d\} \cdot I\{M=m\} \cdot [Y - \mu(d,m,X)]}{f(m|d,X) \cdot p_d(X)} + \mu(d,m,X). \tag{3.4}$$

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 $d \in \{0, 1\}$ based on the efficient score function by Tchetgen Tchetgen and Shpitser (2012) provided in (3.1) and show its $n^{-1/2}$-consistency under specific regularity conditions. To this end, let $\mathcal{W} = \{W_i | 1 \leq i \leq N\}$ with $W_i = (Y_i, M_i, D_i, X_i)$ for $i = 1, \ldots, n$ denote the set of observations in an i.i.d. sample of size n. η denotes the plug-in (or nuisance) parameters, i.e., the conditional mean outcome, mediator density and treatment probability. Their respective estimates 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 $\eta_0 = \{\mu_0(D, M, X), f_0(M|D, X), p_{d0}(X)\}$. Finally, $\psi_{d0} = E[Y(d, M(1-d))]$ denotes the true counterfactual.

We suggest estimating ψ_{d0} using the following algorithm that combines orthogonal score estimation with sample splitting and is $n^{-1/2}$-consistent under conditions outlined further below.

ALGORITHM 1: *Estimation of* $E[Y(d, M(1-d))]$ *based on equation (3.1)*

(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 $\mathcal{W}_k$.
(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 in $\mathcal{W}_k$, 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 ψ_d in (3.1)) for each observation i in $\mathcal{W}_k$, denoted by $\hat{\psi}_{d,i}^k$:

$$\begin{aligned}\hat{\psi}_{d,i}^k = &\frac{I\{D_i = d\} \cdot \hat{f}^k(M_i|1-d, X_i)}{\hat{p}_d^k(X_i) \cdot \hat{f}^k(M_i|d, X_i)} \cdot [Y_i - \hat{\mu}^k(d, M_i, X_i)] \\ &+ \frac{I\{D_i = 1-d\}}{1 - \hat{p}_d^k(X_i)} \cdot \Bigg[\hat{\mu}^k(d, M_i, X_i) \\ &- \int_{m \in \mathcal{M}} \hat{\mu}^k(d, m, X_i) \cdot \hat{f}^k(m|1-d, X_i) dm\Bigg] \\ &+ \int_{m \in \mathcal{M}} \hat{\mu}^k(d, m, X_i) \cdot \hat{f}^k(m|1-d, X_i) dm. \end{aligned} \tag{4.1}$$

(4) Average the estimated scores $\hat{\psi}_{d,i}^k$ over all observations across all K subsamples to obtain an estimate of $\psi_{d0} = E[Y(d, M(1-d))]$ in the total sample, denoted by $\hat{\psi}_d = 1/n \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\psi}_{d,i}^k$.

Algorithm 1 can be adapted to estimate the counterfactuals required for the controlled direct effect, see (3.4). To this end, denote by $\psi_{dm0} = E[Y(d, m)]$ the true counterfactual of interest, which is estimated by replacing ψ_d and ψ_{d0} by ψ_{dm} and ψ_{dm0}, respectively, everywhere in Algorithm 1.

In order to achieve $n^{-1/2}$-consistency for counterfactual estimation, we make specific assumptions about the prediction qualities of the machine learners for our plug-in estimates of the nuisance parameters. Closely following Chernozhukov et al. (2018), to this end we introduce some further notation. Let $(\delta_n)_{n=1}^{\infty}$ and $(\Delta_n)_{n=1}^{\infty}$ denote sequences of positive constants with $\lim_{n\to\infty} \delta_n = 0$ and $\lim_{n\to\infty} \Delta_n = 0$. Also, let c, ϵ, C, $\underline{f}$, $\overline{f}$ and q be positive constants such that $q > 2$, and let $K \geq 2$ be a fixed integer. Furthermore, for any random vector $Z = (Z_1, \ldots, Z_l)$, let $\|Z\|_q = \max_{1\leq j\leq l} \|Z_l\|_q$, where $\|Z_l\|_q = \left(E\left[|Z_l|^q\right]\right)^{1/q}$. For the sake of easing notation, we assume that n/K is an integer. For brevity, we omit the dependence of probability $\Pr_P(\cdot)$, expectation $E_P(\cdot)$, and norm $\|\cdot\|_{P,q}$ on the probability measure P.

ASSUMPTION 4.1. (REGULARITY CONDITIONS AND QUALITY OF PLUG-IN PARAMETER ESTIMATES) *For all probability laws $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of all possible probability laws, the following conditions hold for the random vector (Y, D, M, X) for $d \in \{0, 1\}$:*

(a) $\|Y\|_q \leq C$ *and* $\left\|E[Y^2|d, M, X]\right\|_\infty \leq C^2$,
(b) $\Pr(\epsilon \leq p_{d0}(X) \leq 1 - \epsilon) = 1$,
(c) $\Pr(\underline{f} \leq f(M|D, X) \leq \overline{f}) = 1$,
(d) $\|Y - \mu_0(d, M, X)\|_2 = E\Big[\,(Y - \mu_0(d, M, X))^2\,\Big]^{1/2} \geq c$
(e) Given a random subset $\mathcal{W}_k$ of size n/K, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0(\mathcal{W}_k^C)$ satisfies the following conditions. With P-probability no less than $1 - \Delta_n$:

$$\begin{aligned}
\|\hat{\eta}_0 - \eta_0\|_q &\leq C,\\
\|\hat{\eta}_0 - \eta_0\|_2 &\leq \delta_n,\\
\|\hat{p}_{d0}(X) - 1/2\|_\infty &\leq 1/2 - \epsilon,\\
\left\|\hat{f}_0(M|D, X) - (\underline{f} + \overline{f})/2\right\|_\infty &\leq (\overline{f} - \underline{f})/2,\\
\|\hat{\mu}_0(D, M, X) - \mu_0(D, M, X)\|_2 \times \|\hat{p}_{d0}(X) - p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2},\\
\|\hat{\mu}_0(D, M, X) - \mu_0(D, M, X)\|_2 \times \left\|\hat{f}_0(M|1 - D, X) - f_0(M|1 - D, X)\right\|_2 &\leq \delta_n n^{-1/2}.
\end{aligned}$$

For demonstrating $n^{-1/2}$-consistency of the proposed estimation strategy for the counterfactual, we heavily draw from Chernozhukov et al. (2018) by showing that our estimation strategy satisfies the requirements for their double machine learning framework.

LEMMA 4.1. (NEYMAN ORTHOGONALITY AND LINEARITY) *The following conditions are satisfied: (a) the moment condition $E[\psi_d(W, \eta_0, \psi_{d0})] = 0$ holds, (b) the score $\psi_d(W, \eta_0, \psi_{d0})$ is linear in ψ_{d0}, (c) the second Gateaux derivative of $\eta \mapsto E\Big[\psi_d(W, \hat{\eta}, \psi_{d0})\Big]$ is continuous, (d) the score function is Neyman orthogonal and (e) singular values of $E[\psi_d^a(W; \eta_0)]$ are bounded.*
The proof is provided in Online Appendix S2.1.1.

Then, as, e.g., $\psi_d(W, \eta, \psi_{d0})$ is smooth in (η, ψ_{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 nets. The rates for L_2-boosting were, for instance, derived in Luo and Spindler (2016).

THEOREM 4.1. *Under Assumptions 3.1–3.3 and 4.1, it holds for estimating* $E[Y(d, M(1-d))]$, $E[Y(d, m)]$ *based on Algorithm 1:*
$\sqrt{n}(\hat{\psi}_d - \psi_{d0}) \rightarrow N(0, \sigma^2_{\psi_d})$, *where* $\sigma^2_{\psi_d} = E[(\psi_d - \psi_{d0})^2]$.
$\sqrt{n}(\hat{\psi}_{dm} - \psi_{dm0}) \rightarrow N(0, \sigma^2_{\psi_{dm}})$, *where* $\sigma^2_{\psi_d} = E[(\psi_d - \psi_{dm0})^2]$.
The proof is provided in Online Appendix S2.1.

Analogous results follow for the estimation of $\Lambda = E[Y(d, M(d))]$ when replacing $\hat{\psi}_d$ in the algorithm above by an estimate of score function α_d from (3.3),

$$\hat{\alpha}_d = \frac{I\{D=d\} \cdot (Y_i - \hat{\mu}^k(d, X_i))}{\hat{p}_d{}^k(X_i)} + \hat{\mu}^k(d, X_i), \tag{4.2}$$

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 et al. (2017) and Chernozhukov et al. (2018). Denoting by $\hat{\Lambda}$ the estimate of Λ, it follows under Assumptions 3.1–3.3 and 4.1 that $\sqrt{n}\left(\hat{\Lambda}_d - \Lambda_d\right) \rightarrow N(0, \sigma^2_{\alpha_d})$, where $\sigma^2_{\alpha_d} = E[(\alpha_d - \Lambda_d)^2]$. Therefore, $n^{-1/2}$-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, $\hat{\Delta} = \hat{\Lambda}_1 - \hat{\Lambda}_0$, $\hat{\theta}(1) = \hat{\Lambda}_1 - \hat{\psi}_0$, $\hat{\theta}(0) = \hat{\psi}_1 - \hat{\Lambda}_0$, $\hat{\delta}(1) = \hat{\Lambda}_1 - \hat{\psi}_1$, and $\hat{\delta}(0) = \hat{\psi}_0 - \hat{\Lambda}_0$.

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 $Var(\hat{\theta}(1)) = Var(\alpha_1 - \psi_0)/n = (\sigma^2_{\alpha_1} + \sigma^2_{\psi_0} - 2Cov(\alpha_1, \psi_0))/n$.

Chernozhukov et al. (2018) show that under Assumptions 3.1–3.3 and 4.1, $\hat{\sigma}^2_{\psi_d}$ can be estimated as:

$$\hat{\sigma}^2_{\psi_d} = \frac{1}{K} \sum_{k=1}^{K} \left[1/n_k \sum_{i=1}^{n_k} \psi_d(W_i, \hat{\eta}_0^k, \hat{\psi}_d)^2 \right]. \tag{4.3}$$

The asymptotic variance of α_d can be estimated accordingly, with ψ_d and $\hat{\psi}_{d0}$ substituted by α_d and $\hat{\Lambda}_{d0}$.

We subsequently discuss estimation based on the score function ψ^*_d in expression (3.2). We note that, in this case, we have to estimate the nested nuisance parameter $E\Big[\mu(d, M, X)\Big| D = 1-d, X\Big]$, which we henceforth denote by $\omega(1-d, X)$. To avoid overfitting, the models for $\mu(d, M, X)$ and $\omega(1-d, X)$ are estimated in different subsamples. The plug-in estimates for the conditional mean outcome, the nested conditional mean outcome, mediator density, and treatment probability are referred to by $\hat{\eta}^* = \{\hat{\mu}(D, M, X), \hat{\omega}(D, X), \hat{p}_d(M, X), \hat{p}_d(X)\}$ and the true nuisance parameters by $\eta^*_0 = \{\mu_0(D, M, X), \omega_0(D, X), p_{d0}(M, X), p_{d0}(X)\}$.

ALGORITHM 2: *Estimation of* $E[Y(d, M(1-d))]$ *based on equation (3.2)*

(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 $\mathcal{W}_k$.
(2) For each k, use $\mathcal{W}_k^C$ to estimate the model parameters of $p_d(X)$ and $p_d(M, X)$. Split $\mathcal{W}_k^C$ into 2 nonoverlapping subsamples, estimate the model parameters of the conditional mean $\mu(d, M, X)$ in one subsample and use it for estimating the nested conditional

mean $\omega(1-d,X)=E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]$ in the other subsample. Predict the nuisance parameters in $\mathcal{W}_k$, where the predictions are denoted by $\hat{p}_d{}^k(X)$, $\hat{p}_d^k(M,X)$, $\hat{\mu}^k(D,M,X)$ and $\hat{\omega}(D,X)^k$.

(3) For each k, obtain an estimate of the efficient score function (see ψ_d^* in (3.2)) for each observation i in $\mathcal{W}_k$, denoted by $\hat{\psi}_{d,i}^{*k}$:

$$\begin{aligned}\hat{\psi}_{d,i}^{*k} = {} & \frac{I\{D_i=d\}\big(1-\hat{p}_d^k(M_i,X_i)\big)}{\hat{p}_d^k(M_i,X_i)\big(1-\hat{p}_d^k(X_i)\big)}\cdot[Y-\hat{\mu}^k(d,M_i,X_i)] \\ & + \frac{I\{D_i=1-d\}}{1-\hat{p}_d^k(X_i)}\cdot\Big[\hat{\mu}^k(d,M_i,X_i) \\ & - \hat{\omega}(1-d,X_i)^k\Big]+\hat{\omega}(1-d,X_i)^k. \end{aligned} \tag{4.4}$$

(4) Average the estimated scores $\hat{\psi}_{d,i}^{*k}$ over all observations across all K subsamples to obtain an estimate of $\psi_{d0}=E[Y(d,M(1-d))]$ in the total sample, denoted by $\hat{\psi}_d^*=\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\hat{\psi}_{d,i}^{*k}$.

Also this approach can be shown to be $n^{-1/2}$-consistent under specific regularity conditions outlined below.

ASSUMPTION 4.2. (REGULARITY CONDITIONS AND QUALITY OF PLUG-IN PARAMETER ESTIMATES) *For all probability laws $P\in\mathcal{P}$ the following conditions hold for the random vector (Y,D,M,X) for all $d\in\{0,1\}$:*

(a) $\|Y\|_q\le C$ *and* $\big\|E[Y^2|d,M,X]\big\|_\infty\le C^2$,
(b) $\Pr(\epsilon\le p_{d0}(X)\le 1-\epsilon)=1$,
(c) $\Pr(\epsilon\le p_{d0}(M,X)\le 1-\epsilon)=1$,
(d) $\|Y-\mu_0(d,M,X)\|_2=E\Big[(Y-\mu_0(d,M,X))^2\Big]^{1/2}\ge c$
(e) Given a random subset $\mathcal{W}_k$ of size n/K, the nuisance parameter estimator $\hat{\eta}_0^*=\hat{\eta}_0^*(\mathcal{W}_k^C)$ satisfies the following conditions. With P-probability no less than $1-\Delta_n$:

$$\begin{aligned}\big\|\hat{\eta}_0^*-\eta_0^*\big\|_q &\le C,\\ \big\|\hat{\eta}_0^*-\eta_0^*\big\|_2 &\le \delta_n,\\ \|\hat{p}_{d0}(X)-1/2\|_\infty &\le 1/2-\epsilon,\\ \|\hat{p}_{d0}(M,X)-1/2\|_\infty &\le 1/2-\epsilon,\\ \|\hat{\mu}_0(D,M,X)-\mu_0(D,M,X)\|_2\times\|\hat{p}_{d0}(X)-p_{d0}(X)\|_2 &\le \delta_n n^{-1/2},\\ \|\hat{\mu}_0(D,M,X)-\mu_0(D,M,X)\|_2\times\|\hat{p}_{d0}(M,X)-p_{d0}(M,X)\|_2 &\le \delta_n n^{-1/2},\\ \|\hat{\omega}_0(D,X)-\omega_0(D,X)\|_2\times\|\hat{p}_{d0}(X)-p_{d0}(X)\|_2 &\le \delta_n n^{-1/2}.\end{aligned}$$

THEOREM 4.2. *Under Assumptions 3.1–3.3 and 4.2, it holds for estimating $E[Y(d,M(1-d))]$ based on Algorithm 2:*

$\sqrt{n}\left(\hat{\psi}_d^* - \psi_{d0}^*\right) \to N(0, \sigma^2_{\psi_d^*})$, *where* $\sigma^2_{\psi_d^*} = E[(\psi_d^* - \psi_{d0}^*)^2]$.
The proof is provided in Online Appendix S2.2.

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:

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

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 Σ. The latter is defined based on setting the covariance of the ith and jth covariate in X to $\Sigma_{ij} = 0.5^{|i-j|}$.[3] Coefficients β 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 4.1 using the identification result in expression (3.1) derived by Tchetgen Tchetgen and Shpitser (2012) and (ii) Theorem 4.2 using the modified score function in expression (3.2), which avoids conditional mediator densities. The nuisance parameters are estimated by post-lasso regression based on the 'causalweight' package by Bodory and Huber (2018) for the statistical software 'R' (R Core Team, 2020), in which our estimation procedure is made available, 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 $\omega(1 - d, X)$. The estimation of direct and indirect effects is based on four-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%). Furthermore, we normalize the weights related to the inverse propensity scores in our estimators such that they sum up to one within treatment groups, as for instance advocated in Busso et al. (2009).

In our first simulation design, we set $p = 200$ and the ith element in the coefficient vector β to $0.3/i^2$ for $i = 1, \ldots, p$, 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 on Theorem 4.2 on average exhibits a slightly lower absolute bias ('abias') and standard deviation ('sd') than estimation based on Theorem 4.1 The behaviour of both approaches improves when increasing sample size to $n = 4000$, as the absolute bias is very close to zero for any effect estimate and standard deviation is roughly cut by half. Under the larger sample size, differences in terms of root mean

[3] The results presented below are hardly affected when setting Σ to the identity matrix (zero correlation across X).

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

	Coefficients given by $0.3/i^2$ for $i = 1, \ldots, p$							Coefficients given by $0.5/i^2$ for $i = 1, \ldots, p$						
	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 4.1													
$\hat{\Delta}$	0.01	0.08	0.08	0.00	0.04	0.04	1.02	0.00	0.10	0.10	0.01	0.04	0.05	1.00
$\hat{\theta}(1)$	0.01	0.08	0.08	0.00	0.04	0.04	0.84	0.01	0.09	0.09	0.01	0.04	0.04	0.83
$\hat{\theta}(0)$	0.00	0.08	0.08	0.00	0.04	0.04	0.75	0.00	0.10	0.10	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	0.07	0.07	0.00	0.03	0.03	0.25
$\hat{\delta}(0)$	0.01	0.05	0.05	0.00	0.02	0.02	0.18	0.01	0.05	0.05	0.00	0.02	0.02	0.17
trimmed		17.24			19.19				80.25			237.50		
	Double machine learning based on Theorem 4.2													
$\hat{\Delta}$	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
$\hat{\theta}(1)$	0.00	0.07	0.07	0.00	0.04	0.04	0.84	0.01	0.08	0.08	0.00	0.04	0.04	0.83
$\hat{\theta}(0)$	0.00	0.08	0.08	0.00	0.04	0.04	0.75	0.00	0.08	0.08	0.00	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
$\hat{\delta}(0)$	0.00	0.04	0.04	0.00	0.02	0.02	0.18	0.00	0.04	0.04	0.00	0.02	0.02	0.17
trimmed		1.20			0.11				16.76			25.45		

Notes: '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.

squared error ('rmse') between estimation based on Theorems 4.1 and 4.2 are very close to zero. By and large, the results suggest that the estimators converge to the true effects at rate $n^{-1/2}$.

In our second simulation, confounding is increased by setting β to $0.5/i^2$ for $i = 1, \ldots, p$. 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. Again, estimation based on Theorem 4.2 slightly dominates in terms of having a smaller absolute bias and standard deviation, in particular for $n = 1000$. However, in other settings, the two methods might compare differently in terms of finite sample performance. Both methods based on Theorems 4.1 and 4.2, respectively, appear to converge to the true effects at rate $n^{-1/2}$, and differences in terms of root mean squared errors are minor for $n = 4000$.

Online Appendix S1 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 conducted by the Bureau of Labor Statistics at the US Department of Labor (2019) following a US 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 cycle. 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 to identify 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 Baicker et al. (2013), Cardella and Depew (2014), Yörük (2016), Simon et al. (2017), and Sommers et al. (2017) for the US and King et al. (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), Burstin et al. (1998), Fowler-Brown et al. (2007), Press (2014)), others exploit natural experiments: Simon et al. (2017) estimate a difference-in-differences model comparing states that did and did not expand Medicaid to low-income adults in 2005, while Baicker et al. (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 et al. (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 et al. (2010) for a literature review, several experimental contributions also assess general health checkups. For instance, Rasmussen et al. (2007) conducted an experiment with individuals aged 30–49 in Denmark by randomly offering a set of health screenings, including advice on healthy living, and found a significant positive effect on life expectation. In a study on Japan's elderly population, Nakanishi et al. (1996) found a significantly negative correlation between the rate of attendance at health checkups 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 knowledge, not yet been investigated. A further distinction to most previous studies is that we consider comparably young individuals with an average age below thirty. 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 focuses 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 the 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 had gone for routine checkups since the 2006 interview. This information serves as binary mediator, measured post-treatment but pre-outcome.

To ensure that the control variables (X) are not influenced by the treatment, they come from the pre-treatment 2005 and earlier interview rounds. They cover demographic characteristics, family background and quality of the home environment during youth, education and training, labour 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 behaviour 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 individuals' 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 nonresponse 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 755 control variables, 593 of which are dummy variables (incl. 251 dummies for missing values).

After excluding 1,498 observations with either mediator or treatment status missing, we remain with 7,486 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

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	0.54	0.41	0.13	0	0.66	0.35	0.31	0
Age	22.5	22.54	22.44	0.1	0	22.54	22.46	0.08	0.02
Ethnicity									
Black	0.27	0.25	0.3	−0.04	0	0.32	0.22	0.1	0
Hispanic	0.21	0.19	0.25	−0.06	0	0.21	0.22	−0.01	0.58
Mixed	0.01	0.01	0.01	0	0.35	0.01	0.01	0	0.3
White or Other	0.51	0.55	0.44	0.11	0	0.46	0.55	−0.1	0
Relationship/marriage									
Not cohabiting	0.62	0.61	0.65	−0.03	0	0.61	0.64	−0.03	0.01
Cohabiting	0.17	0.16	0.18	−0.02	0.01	0.16	0.17	0	0.61
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	0.02	0.02	0.02	0	0.55
Missing	0	0	0	0	0.42	0	0	0	0.92
Urban	0.72	0.75	0.67	0.08	0	0.75	0.7	0.05	0
Missing	0.11	0.08	0.16	−0.08	0	0.09	0.14	−0.05	0
HH Income[4]	41,851	47,908	31,433	16,475	0	43,338	40,460	2,878	0.03
Missing	0.24	0.2	0.31	−0.11	0	0.21	0.26	−0.05	0
HH Size	2.99	3.05	2.89	0.16	0	3.1	2.89	0.21	0
Missing	0.09	0.06	0.14	−0.09	0	0.06	0.11	−0.05	0
HH Members under 18	0.67	0.65	0.69	−0.04	0.13	0.76	0.58	0.18	0
Missing	0.09	0.06	0.14	−0.09	0	0.07	0.11	−0.05	0
Biological children	0.48	0.47	0.5	−0.02	0.24	0.55	0.42	0.13	0

Table 2. Continued

n	overall 7,061	$D = 1$ 2,335	$D = 0$ 4,726	diff	*p*-val	$M = 1$ 3,612	$M = 0$ 3,449	diff	*p*-val
Highest grade	11.78	12.61	10.36	2.25	0	12.26	11.33	0.93	0
Missing	0.09	0.06	0.15	−0.09	0	0.07	0.11	−0.05	0
Employment									
Employed	0.71	0.73	0.68	0.05	0	0.7	0.72	−0.02	0.11
Unemployed	0.05	0.04	0.07	−0.03	0	0.05	0.06	−0.01	0.24
Out of labour force	0.2	0.19	0.23	−0.04	0	0.21	0.2	0	0.7
Military	0.03	0.04	0.01	0.02	0	0.04	0.02	0.02	0
Missing	0	0	0.01	0	0.01	0	0	0	0.59
Working hours (per week)	24.04	25.35	21.79	3.57	0	24.11	23.98	0.13	0.78
Missing	0.09	0.06	0.14	−0.09	0	0.06	0.11	−0.05	0
Weight (pounds)	152	156	145	11	0	152	152	1	0.72
Missing	0.11	0.08	0.17	−0.09	0	0.09	0.14	−0.05	0
Height (feet)	4.97	5.16	4.64	0.52	0	5.03	4.91	0.12	0
Missing	0.12	0.08	0.18	−0.1	0	0.09	0.14	−0.05	0
Days 5+ drinks (per month)	1.57	1.55	1.62	−0.07	0.44	1.22	1.9	−0.68	0
Missing	0.11	0.08	0.17	−0.09	0	0.09	0.14	−0.05	0
Days of exercise (per week)	2.37	2.42	2.3	0.11	0.05	2.33	2.41	−0.08	0.15
Missing	0.06	0.05	0.09	−0.05	0	0.05	0.08	−0.03	0
Depressed/down									
Never	0.3	0.31	0.28	0.03	0	0.29	0.31	−0.02	0.05
Sometimes	0.49	0.52	0.45	0.07	0	0.51	0.47	0.04	0
Mostly	0.09	0.09	0.09	0	0.68	0.1	0.08	0.02	0
Always	0.02	0.02	0.02	−0.01	0.03	0.02	0.02	0	0.41
Missing	0.1	0.07	0.16	−0.09	0	0.08	0.12	−0.04	0

Notes: 'overall', '$D = 1$', '$D = 0$', '$M = 1$', '$M = 0$' report the mean of the respective variable in the total sample, among treated, among nontreated, among mediated, and among nonmediated, 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.

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

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 4.1 and expression (3.1) derived by Tchetgen Tchetgen and Shpitser (2012) as well as Theorem 4.2 and expression (3.2). We estimate the nuisance parameters and treatment effects in the same way as outlined in Section 5 (i.e., post-lasso regression for modelling the nuisance parameters and three-fold cross-fitting for effect estimation) after augmenting the set of covariates with 380 selected interaction and higher order terms of covariates measuring demographic characteristics, health status, and health-related behaviour. The trimming threshold for discarding observations with too extreme propensity scores is set to 0.02 (2%), such that 777 and 54 observations are dropped when basing estimation on Theorems 4.1 and 4.2, respectively. As for the simulations, the propensity score-based weights in our estimators are normalized such that they sum up to one within treatment groups.

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 nontreatment for comparison ('$\hat{E}[Y(0, M(0))]$'). The ATEs of health insurance coverage on general health in the year 2008 (columns 2 and 8), estimated based on Theorems 4.1 or 4.2, are statistically significant at the 10% and 5% levels, respectively. As the outcome is measured on an ordinal scale ranging from 'excellent' to 'poor', the negative ATEs suggest a short-term health-improving effect of health coverage. The direct effects under treatment (columns 3 and 9) and under nontreatment (columns 4 and 10) mostly have a similar magnitude as the ATEs, even though they are not statistically significant in 3 out of 4 cases. The indirect effects under treatment (columns 5 and 11) and nontreatment (columns 6 and 12) are generally close to zero and not statistically significant in three out of four cases either. We therefore conclude that, in the short run, health insurance coverage does not seem to importantly affect general health of young adults in the US through routine checkups.

7. CONCLUSION

In this paper, we combined causal mediation analysis with double machine learning under selection-on-observables assumptions, which avoids *ad hoc* 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 $n^{-1/2}$-consistency and asymptotic normality of the effect estimators under specific regularity conditions. Furthermore, we investigated the finite sample behaviour 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 US National Longitudinal Survey of Youth 1997 and found a moderate short-term effect of health insurance coverage on general health, which was, however, not importantly mediated by routine checkups. The estimators considered in the simulation study and the application are available in the 'causalweight' package for the statistical software 'R'.

Table 3. Total, direct, and indirect effects on general health in 2008.

	Estimations based on Theorem 4.1						Estimations based on Theorem 4.2					
	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$	$\hat{E}[Y(0, M(0))]$	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$	$\hat{E}[Y(0, M(0))]$
effect	−0.05	−0.04	−0.04	−0.01	−0.01	2.27	−0.05	−0.03	−0.05	−0.00	−0.02	2.28
se	0.03	0.03	0.03	0.01	0.01	0.03	0.03	0.03	0.03	0.01	0.01	0.02
p-val	0.10	0.23	0.23	0.49	0.17	0.00	0.04	0.22	0.05	0.88	0.04	0.00

Notes: '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.

ACKNOWLEDGEMENTS

Lukáš Lafférs acknowledges support provided by the Slovak Research and Development Agency under contract no. APVV-17-0329 and VEGA-1/0692/20. Helmut Farbmacher and Martin Spindler acknowledge support from the German Research Foundation (DFG), Germany - 431701914.

REFERENCES

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine 27*, 1282–304.

Albert, J. M. and S. Nelson (2011). Generalized causal mediation analysis. *Biometrics 67*, 1028–38.

Baicker, K., S. L. Taubman, H. L. Allen, M. Bernstein, J. H. Gruber, J. P. Newhouse, 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), 1713–22.

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*, 1173–82.

Bellani, L. and M. Bia (2018). The long-run effect of childhood poverty and the mediating role of education. *Journal of the Royal Statistical Society: Series A (Statistics in Society).182*(1), 37–68.

Belloni, A., V. Chernozhukov, I. Fernández-Val and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*, 233–98.

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*, *57*, 1–47

Bodory, H. and M. Huber (2018). The causalweight package for causal inference in R. SES Working Paper, *493*, University of Fribourg.

Bureau of Labor Statistics at the US Department of Labor. (2019). National longitudinal survey of youth 1997 cohort, 1997–2017 (rounds 1–18). Produced and distributed by the Center for Human Resource Research (CHRR), The Ohio State University. https://www.nlsinfo.org

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*, *35*(4),389–97.

Busso, M., J. DiNardo and J. McCrary (2014). New evidence on the finite sample properties of propensity score matching and reweighting estimators. *Review of Economics and Statistics 96*(5), 885–97.

Cardella, E. and B. Depew (2014). The effect of health insurance coverage on the reported health of young adults. *Economics Letters 124*(3), 406–10.

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*, C1–68.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics 13*, 261–81.

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*, F28–65.

Díaz, I. (2020). Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics 21*(2), 353–8.

Díaz, I. and N. S. Hejazi (2020). Causal mediation analysis for stochastic interventions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(3), 661–83.

Faulkner, L. and H. Schauffler (1997). The effect of health insurance coverage on the appropriate use of recommended clinical preventive services. *American Journal of Preventive Medicine 13*(6), 453–58.

Flores, C. A. and A. Flores-Lagunes (2009). Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. *Institute of Labor Economics (IZA) 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), 502–7.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–31.

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*, 2052–86.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. In *Proceedings of the American Statistical Association, Biometrics Section*, 2401–15. Alexandria, VA, American Statistical Association.

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics 29*, 920–43.

Huber, M. (2015). Causal pitfalls in the decomposition of wage gaps. *Journal of Business and Economic Statistics 33*, 179–91.

Huber, M., M. Lechner and G. Mellace (2017). Why do tougher caseworkers increase employment? The role of program assignment as a causal mechanism. *Review of Economics and Statistics 99*, 180–3.

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*, 441–63.

Imai, K. and T. Yamamoto (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis 21*, 141–71.

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.

Judd, C. M. and D. A. Kenny (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review 5*, 602–19.

Kanamori, T., T. Suzuki and M. Sugiyama (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning 86(3)*, 335–67.

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*, 937–63.

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. *Lancet 373*(9673), 1447–54.

Luo, Y. and M. Spindler (2016). High-dimensional $l_2$boosting: Rate of convergence. *arXiv preprint arXiv:1602.08927*.

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), 1656–60.

Miles, C. H., I. Shpitser, P. Kanki, S. Meloni and E. J. Tchetgen Tchetgen (2020). On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika 107*(1), 159–72.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika 78*, 691–92.

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. In J. Wiley (Ed.), *Probability and Statistics*, 213–34.

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), 1359–69.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–20. Morgan Kaufman, San Francisco.

Petersen, M. L., S. E. Sinisi and M. J. van der Laan (2006). Estimation of direct causal effects. *Epidemiology 17*, 276–84.

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), 365–72.

R Core Team(2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology 3*, 143–55.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122–29.

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*, 846–66.

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), 1119–28.

Sugiyama, M., M. Kawanabe and P. L. Chui (2010). Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks 23(1)*, 44–59.

Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in Medicine 32*, 4567–80.

Tchetgen Tchetgen, E. J. and I. Shpitser (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics 40*, 1816–45.

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*, 926–34.

Tibshirani, R. (1996). Regresson shrinkage and selection via the lasso. *Journal of the Royal Statistical Society 58*, 267–88.

van der Laan, M. and D. Rubin (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics 2*, 1–38.

van der Laan, M. J., E. C. Polley and A. E. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*, Article 25.

VanderWeele, T. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology 24*, 224–32.

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*, 129–58.

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), 285–95.

Zheng, W. and M. J. van der Laan (2012). Targeted maximum likelihood estimation of natural direct effects. *International Journal of Biostatistics 8*, 1–40.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Victor Chernozhukov handled this manuscript.