Robust and flexible estimation of data-dependent stochastic mediation effects: a proposed method and example in a randomized trial setting

Kara E. Rudolph*[1], Oleg Sofrygin[2], Wenjing Zheng[3], and Mark J. van der Laan[2]

[1]Division of Epidemiology, School of Public Health, University of California, Berkeley, California
[2]Division of Biostatistics, School of Public Health, University of California, Berkeley, California
[3]Center for AIDS Research, University of California, San Francisco, California

Abstract

Background: Causal mediation analysis can improve understanding of the mechanisms underlying epidemiologic associations. However, the utility of natural direct and indirect effect estimation has been limited by the assumption of no confounder of the mediator-outcome relationship that is affected by prior exposure—an assumption frequently violated in practice.

Methods: We build on recent work that identified alternative estimands that do not require this assumption and propose a flexible and double robust semiparametric targeted minimum loss-based estimator for data-dependent stochastic direct and indirect effects. The proposed method treats the intermediate confounder affected by prior exposure as a time-varying confounder and intervenes stochastically on the mediator using a distribution which conditions on baseline covariates and marginalizes over the intermediate confounder. In addition, we assume the stochastic intervention is given, conditional on observed data, which results in a simpler estimator and weaker identification assumptions.

Results: We demonstrate the estimator's finite sample and robustness properties in a simple simulation study. We apply the method to an example from the Moving to Opportunity experiment. In this application, randomization to receive a housing voucher is the treatment/instrument that influenced moving to a low-poverty neighborhood, which is the intermediate confounder. We estimate the data-dependent stochastic direct effect of randomization to the voucher group on adolescent marijuana use not mediated by change in school district

*Corresponding author:
13B University Hall, Division of Epidemiology, School of Public Health, Berkeley, CA 94720
kara.rudolph@berkeley.edu
tel. +15107619404

and the stochastic indirect effect mediated by change in school district. We find no evidence of mediation.

Conclusions: Our estimator is easy to implement in standard statistical software, and we provide annotated R code to further lower implementation barriers.

Keywords: mediation; direct effect; indirect effect; double robust; targeted minimum loss-based estimation; targeted maximum likelihood estimation; data-dependent

1 Introduction

Mediation allows for an examination of the mechanisms driving a relationship. Much of epidemiology entails reporting exposure-outcome associations where the exposure may be multiple steps removed from the outcome. For example, risk-factor epidemiology demonstrated that obesity increases the risk of type 2 diabetes, but biochemical mediators linking the two have advanced our understanding of the causal relationship (15). Mediators have been similarly important in understanding how social exposures become embodied to influence health outcomes; for example, neighborhood poverty may influence an individual's physiologic stress response which may in turn influence mental health (12, 19, 20, 32).

Causal mediation analysis (13, 38, 47) (also called mediation analysis using the counterfactual framework (38)) shares similar goals with the standard mediation approaches, e.g., structural equation modeling and the widely used Baron and Kenny "product method" approach (4, 38). They all aim to test mechanisms and estimate direct and indirect effects. The primary advantage of causal mediation analysis is that it makes fewer restrictive parametric modeling assumptions. For example, in contrast to traditional approaches, causal mediation analysis 1) allows for interaction between the treatment and mediator (43), 2) allows for modeling nonlinear relationships between mediators and outcomes (43), and 3) allows for incorporation of data-adaptive machine learning methods and double robust estimation (35, 47).

However, despite these advantages, the assumptions required to estimate certain causal mediation effects may sometimes be untenable; for example, the assumption that there is no confounder of the mediator-outcome relationship that is affected by treatment (in the literature, such a confounder is referred to as confounding by a causal intermediate (27), a time-varying confounder affected by prior exposure (44), or time-dependent confounding by an intermediate covariate (39)). For brevity, we will refer to such a variable as an intermediate confounder. Zheng & van der Laan and VanderWeele & Tchetgen Tchetgen recently proposed causal mediation estimands, called randomized (i.e., stochastic) interventional direct effects and interventional indirect effects that do not require this assumption (44, 46, 48). We build on their work, proposing a robust and flexible semiparametric estimator to estimate data-dependent versions of these effects, which we call data-dependent stochastic direct effects (SDE) and stochastic indirect effects (SIE).

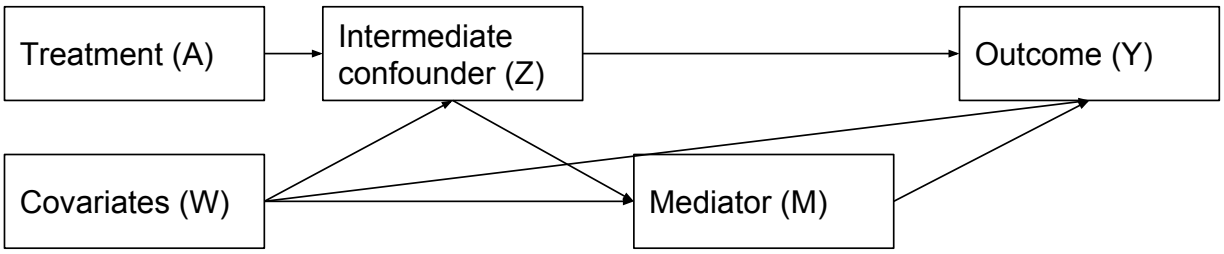
This paper is organized as follows. In the following section, we review and compare common causal mediation estimands, providing the assumptions necessary for their identification. Then, we describe our proposed estimator, its motivation, and its implementation in detail. Code to implement this method is provided in the Appendix. We then provide results from a limited simulation study demonstrating the estimator's finite sample performance and robustness properties. Lastly, we apply the method in a longitudinal, randomized trial setting.

2 Notation and Causal Mediation Estimands

Let observed data: $O = (W, A, Z, M, Y)$ with n i.i.d. copies $O_1, ..., O_n \sim P_0$, where W is a vector of pre-treatment covariates, A is the treatment, Z is the intermediate confounder

affected by A, M is the mediator, and Y is the outcome. For simplicity, we assume that $A, Z, M,$ and Y are binary. In our illustrative example, A is an instrument so it is reasonable to assume that M and Y are not affected by A except through its effect on Z. Mirroring the structural causal model (SCM) of our illustrative example, we assume that M is affected by $\{Z, W\}$ but not A, and that Y is affected by $\{M, Z, W\}$ but not A. We assume exogenous random errors: $(U_W, U_A, U_Z, U_M, U_Y)$. This SCM is represented in Figure 1. Note that this SCM (including that U_Y and U_M are not affected by A) puts the following assumptions on the probability distribution: $P(Y|M, Z, A, W) = P(Y|M, Z, W)$ and $P(M|Z, A, W) = P(M|Z, W)$. However, our approach generalizes to scenarios where A also affects M and Y as well as to scenarios where A is not random. We provide details and discuss these generalizations in the Appendix. We can factorize the likelihood for the SCM reflecting our illustrative example as follows: $P(O) = P(Y|M, Z, W)P(M|Z, W)P(Z|A, W)P(A)P(W)$.

Figure 1: Structural causal model reflecting the illustrative example.



Causal mediation analysis typically involves estimating one of two types of estimands: controlled direct effects (CDE) or natural direct and indirect effects (NDE, NIE). Controlled direct effects involve comparing expected outcomes under different values of the treatment and setting the value of the mediator for everyone in the sample. For example the CDE can be defined: $E(Y_{a,m} - Y_{a*,m})$, where Y is the counterfactual outcome setting treatment A equal to a or a^* (the two treatment values being compared) and setting mediator M equal to m. In contrast, the NDE can be defined: $E(Y_{a,M_{a^*}} - Y_{a*,M_{a^*}})$, where Y is the counterfactual outcome setting A equal to a or a^* but this time setting M to be the counterfactual value of the mediator had A been set to a (possibly contrary to fact). Similarly, the NIE can be defined: $E(Y_{a,M_a} - Y_{a,M_{a*}})$. Natural direct and indirect effects are frequently used in epidemiology and have the appealing property of adding to the total effect (25).

Although the NDE and NIE are popular estimands, their identification assumptions may sometimes be untenable. Broadly, identification of their causal effects relies on the sequential randomization assumption on intervention nodes A and M and positivity. Two specific ignorability assumptions are required to identify CDEs and NDE/NIEs: 1) $A \perp Y_{am}|W$ and 2) $M \perp Y_{am}|W, A$ (25). The positivity assumptions are: $P(M = m|A = a, W) > 0$ *a.e.* and $P(A = a|W) > 0$ *a.e.* Two additional ignorability assumptions are required to identify NDE/NIEs: 3) $A \perp M_a|W$ and 4) $M_{a*} \perp Y_{am}|W$ (25). This last assumption states that, conditional on W, knowledge of M in the absence of exposure A provides no information of the effect of A on Y (27). This assumption is violated when there is a confounder of the $M - Y$ relationship that is affected by A (2, 27, 44). This assumption is also problematic because it involves independence of counterfactuals under separate worlds (a and a^*) which

can never simultaneously exist.

This last assumption that there is no confounder of the mediator-outcome relationship affected by prior exposure is especially concerning for epidemiology studies where longitudinal cohort data may reflect a data structure in which a treatment affects an individual characteristic measured at follow-up that in turn affects both a mediating variable and the outcome variable (see (5, 10, 28) for some examples). It is also problematic for mediation analyses of randomized encouragement-design interventions where an instrument, A, encourages the exposure of interest, Z, which then may influence Y potentially through M. Such a design is present in the Moving to Opportunity (MTO) experiment that we will use as an illustrative example in this paper (16). In MTO, participating families living in public housing were randomized to receive a voucher that they could use to rent housing on the private market (A), which was the instrument that "encouraged" the exposure of interest, moving to a low-poverty neighborhood (Z, which we will call the intermediate confounder). In turn, Z may influence subsequent drug use among adolescent participants at follow-up (Y), possibly through a change the children's school environment (M).

There has been recent work to relax the assumption of no confounder of the mediator-outcome relationship affected by prior exposure (7, 14, 44, 46, 48). In this paper, we build on the approach described by VanderWeele & Tchetgen Tchetgen (2016) in which they redefined the NDE and NIE by using a stochastic intervention on the mediator (44). In other words, instead of formulating the individual counterfactual values of M_a or M_{a^*}, values are stochastically drawn from the distribution of M, conditional on covariates W but marginal over causal intermediates Z, setting $A = a$ or $A = a^*$, respectively. We refer to this stochastic distribution as: $g_{M|a,W}$ or $g_{M|a^*,W}$, where

$$g_{M|a^*,W}(W) = \sum_{z=0}^{1} P(M = 1|Z = z, W)P(Z = z|A = a^*, W). \tag{1}$$

The corresponding estimands of interest are the $SDE = E(Y_{a,g_{M|a^*,W}}) - E(Y_{a*,g_{M|a^*,W}})$, and $SIE = E(Y_{a,g_{M|a,W}}) - E(Y_{a,g_{M|a^*,W}})$. Others have taken a similar approach (9, 45, 46, 48). For example, others (48) formulate a stochastic intervention on M that is fully conditional on the past:

$$g_{M|Z,a^*,W}(Z, W) = P(M = 1|Z = z, W). \tag{2}$$

The corresponding estimands are the stochastic direct and indirect effects fully conditional on the past: $CSDE = E(Y_{a,g_{M|Z,a^*,W}}) - E(Y_{a*,g_{M|Z,a^*,W}})$, and $CSIE = E(Y_{a,g_{M|Z,a,W}}) - E(Y_{a,g_{M|Z,a^*,W}})$.

However, the formulation shown in Equation 2 is not useful for understanding mediation under the instrumental variable SCM we consider here, as there is no direct pathway from A to M or from A to Y under the instrumental variable SCM. Because of the restriction on our statistical model that $P(M|Z, A, W) = P(M|Z, W)$, $g_{M|Z,a^*,W}(Z, W) = g_{M|Z,a,W}(Z, W)$, so CSIE's under this model would equal 0. Thus, in this scenario, the NDE and CSDE are very different parameters. We note that it is also because of these restrictions on our statistical model stemming from the instrumental variable SCM that the sequential mediation analysis approach proposed by VanderWeele and Vansteelandt (2014) would also result in indirect effects equal to 0 (42). Because the CSIE and CSDE do not aid in understanding

the role of M as a potential mediator in this scenario, we focus instead on VanderWeele and Tchetgen Tchetgen's SDE and SIE that condition on W but marginalize over Z, thus completely blocking arrows into M. The SDE and SIE coincide with the NDE and NIE in the absence of causal intermediates (44). Assuming the sequential randomization assumption on intervention nodes A and M, the causal estimand $E(Y_{a,gM|a^*,W})$, for a particular a and $g_{M|a^*,W}$, can be identified from the observed data distribution using the g-computation formula as discussed by VanderWeele and Tchetgen Tchetgen (44).

However, the efficient influence curves (EIC) of these parameters are complicated due to the dependence of the unknown marginal stochastic intervention on the data distribution. The EIC would include an M component, the form of which would be more complex due to the distribution of M being marginalized over Z. No statistical tools for solving an EIC of that form currently exist. Therefore, we focus on data-dependent versions of the SDE and SIE estimands where we replace the true stochastic distribution $g_{M|a^*,W}$ with an estimated, data-dependent distribution $\hat{g}_{M|a^*,W}$. Similar to CDEs, the data-dependent SDE and SIE estimate the outcome under fixed values of the mediator but differ in allowing individual variation in those fixed values. Also similar to CDEs, only positivity and sequential randomization assumptions 1) $Y_{am} \perp A|W$ and 2) $Y_{am} \perp M|W, A = a, Z$ are necessary for the data-dependent version, $E(Y_{a,\hat{g}_{M|a^*,W}})$.

In addition to simplifying the EIC into a form for which statistical tools exist, a substantive reason for the data-dependent version of the SDE and SIE estimands to be of interest is that researchers may prefer to use the mediator distributions that are most relevant in their particular sample instead of in the underlying population. Although with increasingly large sample size the two will converge, in finite samples the observed data may differ from the true data distribution in meaningful ways just by chance. In the case of the MTO experiment that we use as an illustrative example, one reason to examine mediators is for their potential utility in explaining previously puzzling results in which the intervention improved many health-related outcomes for girls but negatively affected many of those same outcomes for boys (17, 18, 22–24, 33). Estimating mediating relationships using the estimated distribution $\hat{g}_{M|a^*,W}$ instead of the true distribution $g_{M|a^*,W}$ would be more relevant in achieving this goal.

The data-dependent, stochastic mediation estimand $E(Y_{a,\hat{g}_{M|a^*W}})$ can be identified via sequential regression, which provides the framework for our proposed estimator that follows. For intervention $(A = a, M = \hat{g}_{M|a^*,W})$, we have $\bar{Q}_Y^{\hat{g}}(Z, W) \equiv E_{\hat{g}_{M|a^*,W}}(E(Y|M, Z, W)|Z, W)$, where we integrate out M under our stochastic intervention $\hat{g}_{M|a^*,W}$. This is accomplished by evaluating the inner expectation at each m and multiplying it by the probability that $M = m$ under $\hat{g}_{M|a^*,W}$, summing over all m. We then integrate out Z and set $A = a$: $\bar{Q}_Z^a(W) \equiv E_Z(\bar{Q}_Y^{\hat{g}}(Z, W)|A = a, W)$. Taking the empirical mean gives the statistical parameter: $\Psi(P)(a, \hat{g}_{M|a^*,W}) = E_W(\bar{Q}_Z^a(W)|W)$.

3 Targeted minimum loss-based estimator

We propose a novel, robust, and flexible semiparametric estimator for the data-dependent SDE and SIE. Previously, a parametric marginal structural model estimator was proposed that can estimate non-data-dependent versions of these effects (44). Although such an es-

timator could also be used to estimate the data-dependent SDE and SIE, it has limited flexibility and may be sensitive to positivity violations given that multiple sets of weights are multiplied (26).

Our proposed estimator is designed explicitly to address these gaps, and responds to a call for additional, double robust estimators (44, p18). It uses TMLE (41), targeting the data-dependent, stochastic, counterfactual outcomes that comprise the SDE and SIE. TMLE is a substitution estimation method that solves the EIC estimating equation. It inherits the double robustness of the EIC, wherein if at least one of the Y or A model is correct, then one obtains a consistent estimator of the parameter. (Note that the estimand itself changes based on the observed distribution of the mediator, so robustness to misspecification of that model is not applicable). And with any approach that solves the EIC estimating equation, it is also straightforward to incorporate machine learning into modeling relationships with theory-based inference.

The estimator integrates two previously developed TMLEs: one for stochastic interventions (21) and one for multiple time-point interventions (39), which is built on the iterative/recursive g-computation approach (3). This TMLE is not efficient under the SCM considered here which puts the following restrictions on our statistical model: $P(Y|M,Z,A,W) = P(Y|M,Z,W)$ and $P(M|Z,A,W) = P(M|Z,W)$. However, it is still a consistent estimator if the restrictions on our model do not hold (i.e., $P(Y|M,Z,A,W) \neq P(Y|M,Z,W)$ and $P(M|Z,A,W) \neq P(M|Z,W)$)), because the targeting step adds dependence on A. The TMLE is constructed using the sequential regressions described in the above section with an additional targeting step after each regression. The TMLE solves the EIC, which has been described previously (39). The EIC for the parameter $\Psi(P)(a, \hat{g}_{M|a^*,W})$ is given by

$$
\begin{aligned}
D^*(a,\hat{g}_{M|a^*,W}) &= \sum_{k=0}^{2} D_k^*(a,\hat{g}_{M|a^*,W}), \text{ where} \\
D_0^*(a,\hat{g}_{M|a^*,W}) &= \bar{Q}_Z^a(W) - \Psi(P)(a,\hat{g}_{M|a^*,W}) \\
D_1^*(a,\hat{g}_{M|a^*,W}) &= \frac{I(A=a)}{P(A=a|W)}(\bar{Q}_Y^{\hat{g}}(Z,W) - \bar{Q}_Z^a(W)) \\
D_2^*(a,\hat{g}_{M|a^*,W}) &= \frac{I(A=a)\{I(M=1)\hat{g}_{M|a^*,W} + I(M=0)(1-\hat{g}_{M|a^*,W})\}}{P(A=a|W)\{I(M=1)g_{M|Z,W} + I(M=0)(1-g_{M|Z,W})\}} \\
&\quad \times (Y - \bar{Q}_Y^{\hat{g}}(Z,W)).
\end{aligned}
\tag{3}
$$

We now describe how to compute the TMLE. In doing so, we use parametric model/regression language for simplicity but data-dependent estimation approaches that incorporate machine learning (e.g., 40) may be substituted and may be preferable (we use such a data-dependent approach in the illustrative example analysis). We note that survey or censoring weights could be incorporated into this estimator as described previously (30).

First, one estimates $\hat{g}_{M|a^*,W}(W)$ as previously defined. Consider a binary Z. We estimate $g_{Z|a^*,W}(W) = P(Z=1|A=a^*,W)$. We then estimate $g_{M|z,W}(W) = P(M=1|Z=z,W)$ for $z \in \{0,1\}$. We use these quantities to calculate $\hat{g}_{M|a^*,W} = \hat{g}_{M|z=1,W}\hat{g}_{Z|a^*,W} + \hat{g}_{M|z=0,W}(1-\hat{g}_{Z|a^*,W})$. We can obtain $\hat{g}_{Z|a^*,W}(W)$ from a logistic regression of Z on A,W setting $A=a^*$, and $\hat{g}_{M|z,W}(W)$ from a logistic regression of M on Z,W, setting $Z=\{0,1\}$. We will then use

this data-dependent stochastic intervention in the TMLE, whose implementation is described as follows.

1. Let $\bar{Q}^{\hat{g}}_{Y,n}(Z,W)$ be an estimate of $\bar{Q}^{\hat{g}}_{Y}(Z,W) \equiv E_{\hat{g}_{M|a^*,W}}(E(Y|M,Z,W)|Z,W)$. To obtain $\bar{Q}^{\hat{g}}_{Y,n}(Z,W)$, first predict values of Y from a regression of Y on M,Z,W, setting $m=1$ and $m=0$, giving $\hat{Y}(m=1,z,w)$ and $\hat{Y}(m=0,z,w)$. Then, multiply the predicted outcomes by their probabilities under $\hat{g}_{M|a^*,W}(W)$ (for $a \in \{a,a^*\}$), and add them together (i.e., $\bar{Q}^{\hat{g}}_{Y,n}(Z,W) = \hat{Y}(m=1,z,w)\hat{g}_{M|a^*,W} + \hat{Y}(m=0,z,w)(1-\hat{g}_{M|a^*,W})$).

2. Estimate the weights to be used for the initial targeting step: $h_1(a) = \frac{I(A=a)\{I(M=1)\hat{g}_{M|a^*,W}+I(M=0)(1-\hat{g}_{M|a^*,W})\}}{P(A=a)\{I(M=1)g_{M|Z,W}+I(M=0)(1-g_{M|Z,W})\}}$, where $\hat{g}_{M|Z,W}$ are predicted probabilities from a logistic regression of $M=m$ on Z and W. Let $h_{1,n}(a)$ denote the estimate of $h_1(a)$.

3. Target the estimate of $\bar{Q}^{\hat{g}}_{Y,n}(Z,W)$ by considering a univariate parametric submodel $\{\bar{Q}^{\hat{g}}_{Y,n}(Z,W)(\epsilon) : \epsilon\}$ defined as: $logit(\bar{Q}^{\hat{g}}_{Y,n}(\epsilon)(Z,W)) = logit(\bar{Q}^{\hat{g}}_{Y,n}(Z,W)) + \epsilon$. Let ϵ_n be the MLE fit of ϵ. We obtain ϵ_n by setting ϵ as the intercept of a weighted logistic regression model of Y with $logit(\bar{Q}^{\hat{g}}_{Y,n}(Z,W))$ as an offset and weights $h_{1,n}(a)$. (Note that this is just one possible TMLE.) The update is given by $\bar{Q}^{\hat{g},*}_{Y,n}(Z,W) = \bar{Q}^{\hat{g}}_{Y,n}(\epsilon_n)(Z,W)$. Y can be bounded to the [0,1] scale as previously recommended (11).

4. We now fit a regression of $\bar{Q}^{\hat{g},*}_{Y,n}(Z,W)$ on W among those with $A=a$. We call the predicted values from this regression $\bar{Q}^{a}_{Z,n}(W)$. The empirical mean of these predicted values is the TMLE estimate of $\Psi(P)(a,\hat{g}_{M|a^*,W})$.

5. Repeat the above steps for each of the interventions. For example, for binary A, we would execute these steps to estimate: 1) $\Psi(P)(1,\hat{g}_{M|1,W})$, 2) $\Psi(P)(1,\hat{g}_{M|0W})$, and 3) $\Psi(P)(0,\hat{g}_{M|0,W})$.

6. The SDE can then be obtained by substituting estimates of parameters $\Psi(P)(a,\hat{g}_{M|a^*,W}) - \Psi(P)(a^*,\hat{g}_{M|a^*,W})$ and the SIE can be obtained by substituting estimates of parameters $\Psi(P)(a,\hat{g}_{M|a,W}) - \Psi(P)(a,\hat{g}_{M|a^*,W})$.

7. The variance of each estimate from Step 9 can be estimated as the sample variance of the EIC (defined above) divided by n. First, we estimate the EIC for each component of the SDE/SIE, which we call $EIC_{\Psi(P)(a,\hat{g}_{M|a^*,W})}$. Then we estimate the EIC for the estimand of interest by subtracting the EICs corresponding to the components of the estimand. For example $EIC_{SDE} = EIC_{\Psi(P)(a,\hat{g}_{M|a^*,W})} - EIC_{\Psi(P)(a^*,\hat{g}_{M|a^*,W})}$. The sample variance of this EIC divided by n is the influence curve-based variance of the estimator.

4 Simulation

4.1 Data generating mechanism

We conduct a simulation study to examine finite sample performance of the TMLE estimators for the SDE and SIE from the data-generating mechanism (DGM) shown in Table 1. Under this DGM, we have: $SDE = E(Y_{1,\hat{g}_{M|0,W}}) - E(Y_{0,\hat{g}_{M|0,W}})$, and $SIE = E(Y_{1,\hat{g}_{M|1,W}}) - E(Y_{1,\hat{g}_{M|0,W}})$.

We compare performance of the TMLE estimator to an inverse-probability weighted estimator (IPTW) and estimator that solves the EIC estimating equation (EE) but differs from TMLE in that it lacks the targeting steps and is not a substitution estimator. We show estimator performance in terms of absolute bias, percent bias, closeness to the efficiency bound (mean estimator standard error (SE, estimated from the sample variance of the EIC) $\times$ the square root of the number of observations), 95% confidence interval coverage, and mean squared error (MSE) across 1,000 simulations for a sample size of N=10,000. We evaluate performance under correct model specification and misspecification of the Y model that included a term for Z only.

Table 1: Simulation data-generating mechanism .

$W_1 \sim Ber(0.5)$
$W_2 \sim Ber(0.4 + 0.2W_1)$
$\Delta \sim Ber(-1 + log(4)W_1 + log(4)W_2$
$A = \Delta A^*$, where $A^* \sim Ber(0.5)$
$Z = \Delta Z^*$, where $Z^* \sim Ber(log(4)A - log(2)W_2)$
$M = \Delta M^*$, where $M^* \sim Ber(-log(3) + log(10)Z - log(1.4)W_2)$
$Y = \Delta Y^*$, where $Y^* \sim Ber(log(1.2) + log(3)Z + log(3)M - log(1.2)W_2 + log(1.2)ZW_2)$

4.2 Performance

As seen in Table 2, the TMLE and EE estimators perform similarly. They are both consistent when all models are correctly specified or when the outcome model is misspecified. The 95% CI for the TMLE and EE estimators results in coverage close to 95%, and they are both close to the efficiency bounds. The IPTW estimator shows small bias of about 1%. Its confidence intervals are conservative (estimated using the variance of its efficient influence curve). As expected, IPTW is less efficient than TMLE and EE.

Table 2: Simulation results under correct specification of all parametric models and 2) misspecification of the outcome model. Note that misspecification of the Z or M models result in a new estimand. N=10,000; 1,000 simulations. Estimation methods compared include targeted minimum loss-based estimation (TMLE), inverse probability weighting estimation (IPTW), and solving the estimating equation (EE). Bias and MSE values are averages across the simulations. The estimator standard error $\times\sqrt{n}$ should be compared to the efficiency bound, which is 1.11 for the SDE and 0.24 for the SIE.

Estimand	Bias	%Bias	SE$\times\sqrt{n}$	95%CI Cov	MSE
All models correctly specified					
TMLE					
SDE	1.75e-04	0.26	1.46	94.70	2.07e-04
SIE	1.24e-04	0.47	0.31	94.50	1.04e-05
IPTW					
SDE	2.25e-04	0.33	3.00	99.50	4.89e-04
SIE	3.82e-04	1.44	1.55	100.00	2.88e-05
EE					
SDE	1.87e-04	0.28	1.47	94.90	2.08e-04
SIE	1.32e-04	0.50	0.31	94.30	1.04e-05
Y model misspecified					
TMLE					
SDE	2.28e-04	0.34	1.49	95.00	2.13e-04
SIE	1.24e-04	0.47	0.31	94.30	1.05e-05
IPTW					
SDE	2.25e-04	0.33	3.00	99.50	4.89e-04
SIE	3.82e-04	1.44	1.55	100.00	2.88e-05
EE					
SDE	2.55e-04	0.38	1.49	95.00	2.15e-04
SIE	1.48e-04	0.56	0.32	94.60	1.08e-05

5 Empirical Illustration

5.1 Overview and set-up

We now apply our proposed estimator to a longitudinal, randomized trial. MTO is a housing policy experiment in which participating families living in public housing in 5 U.S. cities were randomized to receive a housing voucher that they could then use to rent housing on the private market with the goal being to move to a lower-poverty neighborhood (16). Thus, randomization to receive a housing voucher was the instrument (A) that influenced the intermediate confounder of moving to a low-poverty neighborhood (Z). We estimate the data-dependent SDE of being randomized to receive a housing voucher (A) on marijuana use (Y) not mediated by change in school district (M) and the data-dependent SIE mediated by M among adolescent boys in the Boston site.

We restrict to adolescents less than 18 years old who were present at interim follow-up,

as those participants had school data and were eligible to be asked about marijuana use. We restrict to boys in the Boston site as previous work has shown important quantitative and qualitative differences in MTO's effects by sex (17, 18, 22–24, 33) and by city (31). We choose to present results from a restricted analysis instead of a stratified analysis, as our goal is to illustrate the proposed method. A more thorough mediation analysis considering all sexes and sites is the subject of a future paper.

Marijuana use was self-reported by adolescents at the interim follow-up, which occurred 4-7 years after baseline, and was defined as ever versus never use. Change in school district was defined as the current school and school at randomization being in the same district. The instrument was binary and defined as randomization to receive a voucher to move versus to not receive a voucher, as has been done previously (23). Intermediate confounder is defined as living in a low-poverty neighborhood at follow-up, where neighborhood was defined as the 2000 Census tract of residence and a low-poverty neighborhood was defined as less than 25% of residents living at or below the federal poverty line. Numerous baseline characteristics included individual and family sociodemographics, motivation for participating in the study, neighborhood perceptions, and school-related characteristics of the adolescent.

We used machine learning to flexibly and data-adaptively model the following relationships: instrument to intermediate confounder, intermediate confounder to mediator, and mediator to outcome. Specifically, we use gradient boosted machines (6, 37) and choose the most predictive model over a high-dimensional grid of candidate models to model each relationship (e.g, predicting $P(Z = 1|A, W)$, $P(M = 1|Z, W)$, and $P(Y = 1|Z, M, W)$) to generate the predicted probabilities needed for the TMLE data-dependent stochastic mediation estimators.

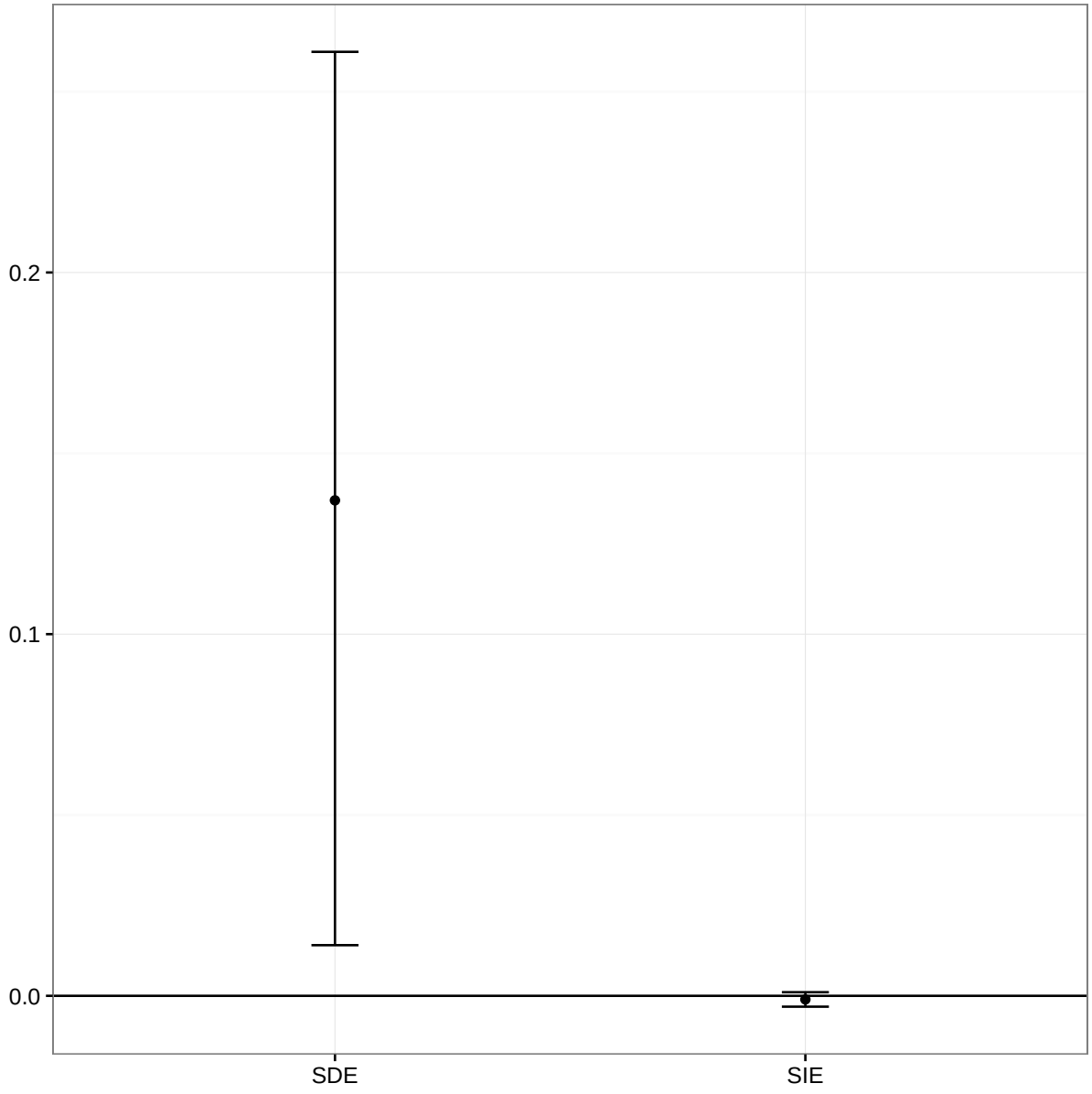
5.2 Results

Figure 2 shows the data-dependent SDE, and SIE estimates for boys in the Boston MTO site (N=228). We find no evidence that change in school district mediated the effect of being randomized to the voucher group on marijuana use. The direct effect of randomization to the housing voucher group on marijuana use is statistically significant, suggesting that boys who were randomized to this group were 14% more likely to use marijuana than boys in the control group (risk difference: 0.137, 95% CI: 0.014-0.261). In contrast, there is no indirect effect through change in school district (risk difference: -0.001, 95% CI: -0.003-0.001).

6 Discussion

We proposed robust, semi-parametric targeted minimum loss-based estimators for data-dependent versions of stochastic direct and indirect effects. These estimators build on previous work identifying and estimating the SDE and SIE (44). The SDE and SIE have the appealing properties of 1) relaxing the assumption of no intermediate confounder affected by prior exposure, and 2) utility in studying mediation in the context of instrumental variables that adhere to the exclusion restriction assumption (a common assumption of instrumental variables which states that there is no direct effect between A and Y or between A and M (1)) due to completely blocking arrows into the mediator by marginalizing over the inter-

Figure 2: Mediated effect estimates and 95% confidence intervals using interim follow-up data from adolescent boys in the Boston site of the Moving to Opportunity experiment. The data-dependent SDE is interpreted as the direct effect of being randomized to receive a housing voucher on risk of marijuana use that is not mediated through a change in school district. The data-dependent SIE is interpreted as the effect of being randomized to receive a housing voucher on marijuana use that is mediated by changing school districts.



mediate confounder, Z. Given the restrictions that this assumption places on the statistical model, several alternative estimands are not appropriate for understanding mediation in this context as the indirect effect would always equal zero (e.g., 36, 42, 48).

The data-dependent versions of the SDE and SIE that we proposed have additional

properties that may be appealing in some cases. First, because the stochastic intervention on the mediator reflects the observed distribution of the data, these estimands may be advantageous in cases where the research question relates to the observed sample instead of to the underlying population and when the sample data size is also small and may differ from the true underlying distribution by chance. We describe such a scenario in terms of our MTO illustrative example, above. A second reason these estimands may be appealing is that their causal interpretation is valid even if the stochastic intervention on M is not correctly estimated. In contrast, the interpretation of the non-data-dependent version relies on correctly specifying $g_{M|a^*,W}$. In addition, the data-dependent estimands rely on weaker identification assumptions (44). Practically, the data-dependent versions result in a simpler EIC with a form conducive to using existing statistical tools to solve its estimating equation. Because of this, we were able to develop a TMLE estimator that was simple to implement in standard statistical software and retained the desired properties of double robustness and flexibility in integrating machine learning. To our knowledge, this is the first estimator of data-dependent causal mediation effects.

TMLE, along with other estimators that solve the EIC estimating equation, has the advantage of giving theory-based inference when incorporating machine learning approaches. The ability to incorporate machine learning is a significant strength in this case; if using the parametric alternative, multiple models would need to be correctly specified (44). Another advantage of using TMLE is that it is doubly robust, so if one were to use parametric models, one could misspecify either the Y model or the A model and still obtain a consistent estimate. In addition, our proposed estimation strategy is less sensitive to positivity violations than weighting-based approaches. First, TMLE is usually less sensitive to these violations than weighting estimators, due in part to it being a substitution estimator, which means that its estimates lie within the global constraints of the statistical model. Second, we formulate our TMLE such that the targeting is done as a weighted regression, which may smooth highly variable weights (34). In addition, moving the targeting into the weights improves computation time (34).

However, there are also limitations to the proposed approach. We have currently only implemented it for a binary A and M, though extensions to multinomial or continuous versions of those variables are possible (8, 29). Extending the estimator to allow for a high-dimensional M is less straightforward, though it is of interest and an area for future work as allowing for high-dimensional M is a strength of other mediation approaches (36, 48).

We also plan to focus future work on developing a robust, semiparametric estimator for non-data-dependent versions of the SDE and SIE. We acknowledge that there exist many research questions for which the non-data-dependent SDE and SIE would be more appropriate than the data-dependent versions. These might include scenarios where the desired intervention would be applied to the true, underlying population instead of to the sample, and/or where the sample size is large and likely closely approximates the underlying population.

References

[1] Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 91, 444–

455.

[2] Avin, C., I. Shpitser, and J. Pearl (2005): "Identifiability of path-specific effects," in *Proceedings of the 19th international joint conference on Artificial intelligence*, Morgan Kaufmann Publishers Inc., 357–363.

[3] Bang, H. and J. M. Robins (2005): "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–973.

[4] 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.

[5] Bild, D. E., D. A. Bluemke, G. L. Burke, R. Detrano, A. V. D. Roux, A. R. Folsom, P. Greenland, D. R. JacobsJr, R. Kronmal, K. Liu, et al. (2002): "Multi-ethnic study of atherosclerosis: objectives and design," *American journal of epidemiology*, 156, 871–881.

[6] Click, C., J. Lanford, M. Malohlava, V. Parmar, and H. Roark (2015): *Gradient Boosted Models*, URL `http://h2o.ai/resources`.

[7] Daniel, R., B. De Stavola, S. Cousens, and S. Vansteelandt (2015): "Causal mediation analysis with multiple mediators," *Biometrics*, 71, 1–14.

[8] Díaz, I. and M. Rosenblum (2015): "Targeted maximum likelihood estimation using exponential families," *The international journal of biostatistics*, 11, 233–251.

[9] Didelez, V., A. P. Dawid, and S. Geneletti (2006): "Direct and indirect effects of sequential treatments," in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 138–146.

[10] Eaton, W. W. and L. G. Kessler (2012): *Epidemiologic field methods in psychiatry: the NIMH Epidemiologic Catchment Area Program*, Academic Press.

[11] Gruber, S. and M. J. van der Laan (2010): "A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome," *The International Journal of Biostatistics*, 6.

[12] Hackman, D. A., L. M. Betancourt, N. L. Brodsky, H. Hurt, and M. J. Farah (2012): "Neighborhood disadvantage and adolescent stress reactivity," *Frontiers in human neuroscience*, 6, 277.

[13] Imai, K., L. Keele, and D. Tingley (2010): "A general approach to causal mediation analysis." *Psychological methods*, 15, 309.

[14] Imai, K. and T. Yamamoto (2013): "Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments," *Political Analysis*, 141–171.

[15] Kahn, S. E., R. L. Hull, and K. M. Utzschneider (2006): "Mechanisms linking obesity to insulin resistance and type 2 diabetes," *Nature*, 444, 840–846.

[16] Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental analysis of neighborhood effects," *Econometrica*, 75, 83–119.

[17] Kling, J. R., J. Ludwig, and L. F. Katz (2005): "Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment," *The Quarterly Journal of Economics*, 87–130.

[18] Leventhal, T. and V. Dupéré (2011): "Moving to opportunity: Does long-term exposure to 'low-poverty'neighborhoods make a difference for adolescents?" *Social Science & Medicine*, 73, 737–743.

[19] Mair, C. F., A. V. D. Roux, and S. Galea (2008): "Are neighborhood characteristics associated with depressive symptoms? a critical review," *Journal of epidemiology and community health*, jech–2007.

[20] McEwen, B. S. (2007): "Physiology and neurobiology of stress and adaptation: central role of the brain," *Physiological reviews*, 87, 873–904.

[21] Muñoz, I. D. and M. van der Laan (2012): "Population intervention causal effects based on stochastic interventions," *Biometrics*, 68, 541–549.

[22] Orr, L., J. Feins, R. Jacob, E. Beecroft, L. Sanbonmatsu, L. F. Katz, J. B. Liebman, and J. R. Kling (2003): "Moving to opportunity: Interim impacts evaluation," .

[23] Osypuk, T. L., N. M. Schmidt, L. M. Bates, E. J. Tchetgen-Tchetgen, F. J. Earls, and M. M. Glymour (2012): "Gender and crime victimization modify neighborhood effects on adolescent mental health," *Pediatrics*, 130, 472–481.

[24] Osypuk, T. L., E. J. T. Tchetgen, D. Acevedo-Garcia, F. J. Earls, A. Lincoln, N. M. Schmidt, and M. M. Glymour (2012): "Differential mental health effects of neighborhood relocation among youth in vulnerable families: results from a randomized trial," *Archives of general psychiatry*, 69, 1284–1294.

[25] Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 411–420.

[26] Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan (2010): "Diagnosing and responding to violations in the positivity assumption," *Statistical methods in medical research*, 0962280210386207.

[27] Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006): "Estimation of direct causal effects," *Epidemiology*, 17, 276–284.

[28] Phair, J., L. Jacobson, R. Detels, C. Rinaldo, A. Saah, L. Schrager, and A. Muñoz (1992): "Acquired immune deficiency syndrome occurring within 5 years of infection with human immunodeficiency virus type-1: the multicenter aids cohort study." *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 5, 490–496.

[29] Rosenblum, M. and M. J. van der Laan (2010): "Targeted maximum likelihood estimation of the parameter of a marginal structural model," *The international journal of biostatistics*, 6, 19.

[30] Rudolph, K. E., I. Díaz, M. Rosenblum, and E. A. Stuart (2014): "Estimating population treatment effects from a survey subsample," *American journal of epidemiology*, kwu197.

[31] Rudolph, K. E., N. M. Schmidt, M. M. Glymour, R. E. Crowder, J. Galin, J. Ahern, and T. L. Osypuk (2017): "Composition or context: using transportability to understand drivers of site differences in a large-scale housing experiment," *Epidemiology*, In process.

[32] Rudolph, K. E., E. A. Stuart, T. A. Glass, A. H. Marques, R. Duncko, K. R. Merikangas, et al. (2014): "The association between cortisol and neighborhood disadvantage in a us population-based sample of adolescents," *Health & place*, 25, 68–77.

[33] Sanbonmatsu, L., J. Ludwig, L. F. Katz, L. A. Gennetian, G. J. Duncan, R. C. Kessler, E. Adam, T. W. McDade, and S. T. Lindau (2011): "Moving to opportunity for fair housing demonstration program–final impacts evaluation," .

[34] Stitelman, O. M., V. De Gruttola, and M. J. van der Laan (2012): "A general implementation of tmle for longitudinal data applied to causal inference in survival analysis," *The international journal of biostatistics*, 8.

[35] Tchetgen, E. T. and I. Shpitser (2014): "Estimation of a semiparametric natural direct effect model incorporating baseline covariates," *Biometrika*, 101, 849–864.

[36] Tchetgen Tchetgen, E. J. (2013): "Inverse odds ratio-weighted estimation for causal mediation analysis," *Statistics in medicine*, 32, 4567–4580.

[37] team, T. H. (2017): *H2O*, URL `http://h2o-release.s3.amazonaws.com/h2o/rel-turin/3/R`, 3.

[38] Valeri, L. and T. J. VanderWeele (2013): "Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with sas and spss macros." *Psychological methods*, 18, 137.

[39] van der Laan, M. J. and S. Gruber (2012): "Targeted minimum loss based estimation of causal effects of multiple time point interventions," *The International Journal of Biostatistics*, 8, article 9.

[40] van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007): "Super learner," *Statistical applications in genetics and molecular biology*, 6.

[41] van der Laan, M. J. and D. Rubin (2006): "Targeted maximum likelihood learning," *The International Journal of Biostatistics*, 2.

[42] VanderWeele, T. and S. Vansteelandt (2014): "Mediation analysis with multiple mediators," *Epidemiologic methods*, 2, 95–115.

[43] VanderWeele, T. J. (2009): "Marginal structural models for the estimation of direct and indirect effects," *Epidemiology*, 20, 18–26.

[44] VanderWeele, T. J. and E. J. Tchetgen Tchetgen (2016): "Mediation analysis with time varying exposures and mediators," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

[45] VanderWeele, T. J., S. Vansteelandt, and J. M. Robins (2014): "Effect decomposition in the presence of an exposure-induced mediator-outcome confounder," *Epidemiology (Cambridge, Mass.)*, 25, 300.

[46] Zheng, W. and M. J. van der Laan (2012): "Causal mediation in a survival setting with time-dependent mediators," .

[47] Zheng, W. and M. J. van der Laan (2012): "Targeted maximum likelihood estimation of natural direct effects," *The international journal of biostatistics*, 8, 1–40.

[48] Zheng, W. and M. J. van der Laan (2017): "Longitudinal mediation analysis with time-varying mediators and exposures: with application to survival outcomes," *Working Paper*.

A Generalizations to other structural causal models

A.1 Nonrandom Treatment

Let observed data: $O = (W, A, Z, M, Y)$ with n i.i.d. copies $O_1, ..., O_n \sim P_0$, where W is a vector of pre-treatment covariates, A is the treatment, Z is the intermediate confounder affected by A, M is the mediator, and Y is the outcome. For simplicity, we assume that $A, Z, M,$ and Y are binary. We assume that M and Y are not affected by A except through its effect on Z. We assume that A is affected by $\{W\}$, Z is affected by $\{A, W\}$, M is affected by $\{Z, W\}$ but not A, and that Y is affected by $\{M, Z, W\}$ but not A. We assume exogenous random errors: $(U_W, U_A, U_Z, U_M, U_Y)$. Note that this SCM (including that U_Y and U_M are not affected by A) puts the following assumptions on the probability distribution: $P(Y|M, Z, A, W) = P(Y|M, Z, W)$ and $P(M|Z, A, W) = P(M|Z, W)$. We can factorize the likelihood for this SCM as follows: $P(O) = P(Y|M, Z, W)P(M|Z, W)P(Z|A, W)P(A|W)P(W)$.

The data-dependent, stochastic mediation estimand $E(Y_{a,\hat{g}_{M|a^*W}})$ can be identified via sequential regression, which provides the framework for our proposed estimator that follows. For intervention $(A = a, M = \hat{g}_{M|a^*,W})$, we have
$\bar{Q}_Y^{\hat{g}}(Z, A, W) \equiv E_{\hat{g}_{M|a^*,W}}(E(Y|M, Z, A, W)|Z, A, W)$, where we integrate out M under our stochastic intervention $\hat{g}_{M|a^*,W}$. This is accomplished by evaluating the inner expectation at each m and multiplying it by the probability that $M = m$ under $\hat{g}_{M|a^*,W}$, summing over all m. We then integrate out Z and set $A = a$: $\bar{Q}_Z^a(W) \equiv E_Z(\bar{Q}_Y^{\hat{g}}(Z, A, W)|A = a, W)$. Taking the empirical mean gives the statistical parameter: $\Psi(P)(a, \hat{g}_{M|a^*,W}) = E_W(\bar{Q}_Z^a(W)|W)$.

The EIC for the parameter $\Psi(P)(a, \hat{g}_{M|a^*,W})$ is given by

$$
\begin{aligned}
D^*(a, \hat{g}_{M|a^*,W}) &= \sum_{k=0}^{2} D_k^*(a, \hat{g}_{M|a^*,W}), \text{ where} \\
D_0^*(a, \hat{g}_{M|a^*,W}) &= \bar{Q}_Z^a(W) - \Psi(P)(a, \hat{g}_{M|a^*,W}) \\
D_1^*(a, \hat{g}_{M|a^*,W}) &= \frac{I(A=a)}{P(A=a|W)}(\bar{Q}_Y^{\hat{g}}(Z,W) - \bar{Q}_Z^a(W)) \\
D_2^*(a, \hat{g}_{M|a^*,W}) &= \frac{I(A=a)\{I(M=1)\hat{g}_{M|a^*,W} + I(M=0)(1-\hat{g}_{M|a^*,W})\}}{P(A=a|W)\{I(M=1)g_{M|Z,W} + I(M=0)(1-g_{M|Z,W})\}}(Y - \bar{Q}_Y^{\hat{g}}(Z,W)).
\end{aligned} \tag{4}
$$

We now describe how to compute the TMLE. The estimation of $\hat{g}_{M|a^*,W}(W)$ does not differ from that described in the main text.

1. Let $\bar{Q}_{Y,n}^{\hat{g}}(Z,W)$ be an estimate of $\bar{Q}_Y^{\hat{g}}(Z,W) \equiv E_{\hat{g}_{M|a^*,W}}(E(Y|M,Z,W)|Z,W)$. To obtain $\bar{Q}_{Y,n}^{\hat{g}}(Z,W)$, first predict values of Y from a regression of Y on M, Z, W, setting $m = 1$ and $m = 0$, giving $\hat{Y}(m = 1, z, w)$ and $\hat{Y}(m = 0, z, w)$. Then, multiply the predicted outcomes by their probabilities under $\hat{g}_{M|a^*,W}(W)$ (for $a \in \{a, a^*\}$), and add them together (i.e., $\bar{Q}_{Y,n}^{\hat{g}}(Z,W) = \hat{Y}(m = 1, z, w)\hat{g}_{M|a^*,W} + \hat{Y}(m = 0, z, w)(1 - \hat{g}_{M|a^*,W})$).

2. Estimate the weights to be used for the initial targeting step: $h_1(a) = \frac{I(A=a)\{I(M=1)\hat{g}_{M|a^*,W}+I(M=0)(1-\hat{g}_{M|a^*,W})\}}{P(A=a)\{I(M=1)g_{M|Z,W}+I(M=0)(1-g_{M|Z,W})\}}$, where $\hat{g}_{M|Z,W}$ are predicted probabilities from a logistic regression of $M = m$ on Z and W. Let $h_{1,n}(a)$ denote the estimate of $h_1(a)$.

3. Target the estimate of $\bar{Q}_{Y,n}^{\hat{g}}(Z,W)$ by considering a univariate parametric submodel $\{\bar{Q}_{Y,n}^{\hat{g}}(Z,W)(\epsilon) : \epsilon\}$ defined as: $logit(\bar{Q}_{Y,n}^{\hat{g}}(\epsilon)(Z,W)) = logit(\bar{Q}_{Y,n}^{\hat{g}}(Z,W)) + \epsilon$. Let ϵ_n be the MLE fit of ϵ. We obtain ϵ_n by setting ϵ as the intercept of a weighted logistic regression model of Y with $logit(\bar{Q}_{Y,n}^{\hat{g}}(Z,W))$ as an offset and weights $h_{1,n}(a)$. (Note that this is just one possible TMLE.) The update is given by $\bar{Q}_{Y,n}^{\hat{g},*}(Z,W) = \bar{Q}_{Y,n}^{\hat{g}}(\epsilon_n)(Z,W)$. Y can be bounded to the [0,1] scale as previously recommended (11).

4. We now fit a regression of $\bar{Q}_{Y,n}^{\hat{g},*}(Z,W)$ on W among those with $A = a$. We call the predicted values from this regression $\bar{Q}_{Z,n}^a(W)$.

5. Complete a second targeting step: $logit(\bar{Q}_{Z,n}^a(\epsilon)(W)) = logit(\bar{Q}_{Z,n}^a(W)) + \epsilon h_{2,n}(a)$, where $h_{2,n}(a)$ is an estimate of $h_2(a) = \frac{I(A=a)}{P(A=a|W)}$ and $P(A = a|W)$ can be estimated from a logistic regression model of $A = a$ on W. Let ϵ_n again be the MLE fit of ϵ, which can be obtained by fitting an intercept-only weighted logistic regression model of $\bar{Q}_{Y,n}^{\hat{g},*}(Z,W)$ with $logit(\bar{Q}_{Z,n}^a(W))$ as an offset and weights $h_{2,n}(a)$. (Alternatively, we could fit an unweighted logistic regression model of $\bar{Q}_{Y,n}^{\hat{g},*}(Z,W)$ with $logit(\bar{Q}_{Z,n}^a(W))$ as an offset and $h_{2,n}(a)$ as a covariate, where ϵ_n is the fitted coefficient on $h_{2,n}(a)$.) The update is given by $\bar{Q}_{Z,n}^{a,*}(W) = \bar{Q}_{Z,n}^{a,a^*}(\epsilon_n)(A,W)$.

6. The TMLE of $\Psi(P)(a, \hat{g}_{M|a^*,W})$ is the empirical mean of $\bar{Q}^{a,*}_{Z,n}(W)$.

7. Repeat the above steps for each of the interventions. For example, for binary A, we would execute these steps to estimate: 1) $\Psi(P)(1, \hat{g}_{M|1,W})$, 2) $\Psi(P)(1, \hat{g}_{M|0W})$, and 3) $\Psi(P)(0, \hat{g}_{M|0,W})$.

8. The SDE can then be obtained by substituting estimates of parameters $\Psi(P)(a, \hat{g}_{M|a^*,W}) - \Psi(P)(a^*, \hat{g}_{M|a^*,W})$ and the SIE can be obtained by substituting estimates of parameters $\Psi(P)(a, \hat{g}_{M|a,W}) - \Psi(P)(a, \hat{g}_{M|a^*,W})$.

9. The variance of each estimate can be estimated as the sample variance of the EIC (defined above) divided by n. First, we estimate the EIC for each component of the SDE/SIE, which we call $EIC_{\Psi(P)(a,\hat{g}_{M|a^*,W})}$. Then we estimate the EIC for the estimand of interest by subtracting the EICs corresponding to the components of the estimand. For example $EIC_{SDE} = EIC_{\Psi(P)(a,\hat{g}_{M|a^*,W})} - EIC_{\Psi(P)(a^*,\hat{g}_{M|a^*,W})}$. The sample variance of this EIC divided by n is the influence curve-based variance of the estimator.

A.2 There exist direct effects between A and M and between A and Y

Let observed data: $O = (W, A, Z, M, Y)$ with n i.i.d. copies $O_1, ..., O_n \sim P_0$, where W is a vector of pre-treatment covariates, A is the treatment, Z is the intermediate confounder affected by A, M is the mediator, and Y is the outcome. For simplicity, we assume that $A, Z, M,$ and Y are binary. We assume that A is exogenous, Z is affected by $\{A, W\}$, M is affected by $\{A, Z, W\}$, and that Y is affected by $\{A, M, Z, W\}$. We assume exogenous random errors: $(U_W, U_A, U_Z, U_M, U_Y)$. We can factorize the likelihood for this SCM as follows: $P(O) = P(Y|M, Z, A, W)P(M|Z, A, W)P(Z|A, W)P(A)P(W)$.

Under this SCM, our proposed TMLE is efficient. The sequential regression used to identify the data-dependent, stochastic mediation estimands does not change from the that given in the main text for this SCM.

The EIC for the parameter $\Psi(P)(a, \hat{g}_{M|a^*,W})$ is given by

$$
\begin{aligned}
D^*(a, \hat{g}_{M|a^*,W}) &= \sum_{k=0}^{2} D^*_k(a, \hat{g}_{M|a^*,W}), \text{ where} \\
D^*_0(a, \hat{g}_{M|a^*,W}) &= \bar{Q}^a_Z(W) - \Psi(P)(a, \hat{g}_{M|a^*,W}) \\
D^*_1(a, \hat{g}_{M|a^*,W}) &= \frac{I(A=a)}{P(A=a)}(\bar{Q}^{\hat{g}}_Y(Z, A, W) - \bar{Q}^a_Z(W)) \\
D^*_2(a, \hat{g}_{M|a^*,W}) &= \frac{I(A=a)\{I(M=1)\hat{g}_{M|a^*,W} + I(M=0)(1-\hat{g}_{M|a^*,W})\}}{P(A=a)\{I(M=1)g_{M|Z,A,W} + I(M=0)(1-g_{M|Z,A,W})\}}(Y - \bar{Q}^{\hat{g}}_Y(Z, A, W)).
\end{aligned} \tag{5}
$$

We now describe how to compute the TMLE. First, one estimates $\hat{g}_{M|a^*,W}(W) = \sum_{z=0}^{1} P(M = 1|Z = z, A, W)P(Z = z|A = a^*, W)$. Consider a binary Z. We first estimate $g_{Z|a^*,W}(W) = P(Z = 1|A = a^*, W)$. We then estimate $g_{M|z,A,W}(W) = P(M = 1|Z = z, A, W)$ for $z \in$

$\{0,1\}$. We use these quantities to calculate $\hat{g}_{M|a^*,W} = (\hat{g}_{M|z=1,A,W} \times \hat{g}_{Z|a^*,W}) + (\hat{g}_{M|z=0,A,W} \times (1 - \hat{g}_{Z|a^*,W}))$. We can obtain $\hat{g}_{Z|a^*,W}(W)$ from a logistic regression of Z on A, W setting $A = a^*$, and $\hat{g}_{M|z,A,W}(W)$ from a logistic regression of M on Z, A, W, setting $Z = \{0,1\}$. We will then use this data-dependent stochastic intervention in the TMLE, whose implementation is described as follows.

1. Let $\bar{Q}^{\hat{g}}_{Y,n}(Z,A,W)$ be an estimate of $\bar{Q}^{\hat{g}}_{Y}(Z,A,W) \equiv E_{\hat{g}_{M|a^*,W}}(E(Y|M,Z,A,W)|Z,A,W)$. To obtain $\bar{Q}^{\hat{g}}_{Y,n}(Z,A,W)$, first predict values of Y from a regression of Y on M, Z, A, W, setting $m = 1$ and $m = 0$, giving $\hat{Y}(m = 1, z, a, w)$ and $\hat{Y}(m = 0, z, a, w)$. Then, multiply the predicted outcomes by their probabilities under $\hat{g}_{M|a^*,W}(W)$ (for $a \in \{a, a^*\}$), and add them together (i.e., $\bar{Q}^{\hat{g}}_{Y,n}(Z,A,W) = \hat{Y}(m = 1, z, a, w)\hat{g}_{M|a^*,W} + \hat{Y}(m = 0, z, a, w)(1 - \hat{g}_{M|a^*,W})$).

2. Estimate the weights to be used for the initial targeting step: $h_1(a) = \frac{I(A=a)\{I(M=1)\hat{g}_{M|a^*,W} + I(M=0)(1-\hat{g}_{M|a^*,W})\}}{P(A=a)\{I(M=1)g_{M|Z,A,W} + I(M=0)(1-g_{M|Z,A,W})\}}$, where $\hat{g}_{M|Z,A,W}$ are predicted probabilities from a logistic regression of $M = m$ on Z, A, and W. Let $h_{1,n}(a)$ denote the estimate of $h_1(a)$.

3. Target the estimate of $\bar{Q}^{\hat{g}}_{Y,n}(Z,A,W)$ by considering a univariate parametric submodel $\{\bar{Q}^{\hat{g}}_{Y,n}(Z,A,W)(\epsilon) : \epsilon\}$ defined as: $logit(\bar{Q}^{\hat{g}}_{Y,n}(\epsilon)(Z,A,W)) = logit(\bar{Q}^{\hat{g}}_{Y,n}(Z,A,W)) + \epsilon$. Let ϵ_n be the MLE fit of ϵ. We obtain ϵ_n by setting ϵ as the intercept of a weighted logistic regression model of Y with $logit(\bar{Q}^{\hat{g}}_{Y,n}(Z,A,W))$ as an offset and weights $h_{1,n}(a)$. (Note that this is just one possible TMLE.) The update is given by $\bar{Q}^{\hat{g},*}_{Y,n}(Z,A,W) = \bar{Q}^{\hat{g}}_{Y,n}(\epsilon_n)(Z,A,W)$.

4. We now fit a regression of $\bar{Q}^{\hat{g},*}_{Y,n}(Z,A,W)$ on W among those with $A = a$. We call the predicted values from this regression $\bar{Q}^{a}_{Z,n}(W)$. The empirical mean of these predicted values is the TMLE estimate of $\Psi(P)(a, \hat{g}_{M|a^*,W})$.

5. Repeat the above steps for each of the interventions. For example, for binary A, we would execute these steps to estimate: 1) $\Psi(P)(1, \hat{g}_{M|1,W})$, 2) $\Psi(P)(1, \hat{g}_{M|0W})$, and 3) $\Psi(P)(0, \hat{g}_{M|0,W})$.

6. The SDE can then be obtained by substituting estimates of parameters $\Psi(P)(a, \hat{g}_{M|a^*,W}) - \Psi(P)(a^*, \hat{g}_{M|a^*,W})$ and the SIE can be obtained by substituting estimates of parameters $\Psi(P)(a, \hat{g}_{M|a,W}) - \Psi(P)(a, \hat{g}_{M|a^*,W})$.

7. The variance of each estimate can be estimated as the sample variance of the EIC (defined above) divided by n. First, we estimate the EIC for each component of the SDE/SIE, which we call $EIC_{\Psi(P)(a,\hat{g}_{M|a^*,W})}$. Then we estimate the EIC for the estimand of interest by subtracting the EICs corresponding to the components of the estimand. For example $EIC_{SDE} = EIC_{\Psi(P)(a,\hat{g}_{M|a^*,W})} - EIC_{\Psi(P)(a^*,\hat{g}_{M|a^*,W})}$. The sample variance of this EIC divided by n is the influence curve-based variance of the estimator.

B Function code

```
## A is the randomization/treatment variable. It needs to be named "a" and have values 0/1.
## Z is the causal intermediate variable. It needs to be named "z" and have values 0/1.
## M is the mediator variable. It needs to be named "m" and have values 0/1.
## Y is the outcome variable. It needs to be named "y" and have values 0/1.
## W are the covariate variables. They need to be in a dataframe named "w" with names "w1", ... , "wn".
## Weights need to be numeric vector named "svywt".
## The parametric model for Z is named zmodel.
## The parametric model for M is named mmodel.
## The parametric model for Y is named ymodel.
## The parametric model for Q is named qmodel.

medtmle<-function(a, z, m, y, w, svywt, zmodel, mmodel, ymodel, qmodel){
  datw<-w

  #make gm
  za0<-predict(glm(formula=zmodel, family="binomial", data=data.frame(cbind(datw, a=a, z=z))), newdata=data.frame(cbind(datw, a=0)), type="response")
  za1<-predict(glm(formula=zmodel, family="binomial", data=data.frame(cbind(datw, a=a, z=z))), newdata=data.frame(cbind(datw, a=1)), type="response")

  mz1<-predict(glm(formula=mmodel, family="binomial", data=data.frame(cbind(datw, z=z, m=m))), newdata=data.frame(cbind(datw, z=1)), type="response")
  mz0<-predict(glm(formula=mmodel, family="binomial", data=data.frame(cbind(datw, z=z, m=m))), newdata=data.frame(cbind(datw, z=0)), type="response")

  gm<-(mz1*za0) + (mz0*(1-za0))
  gma1<-(mz1*za1) + (mz0*(1-za1))

  tmpdat<-data.frame(cbind(datw, a=a))

  tmpdat$q1<-(predict(glm(formula=ymodel, family="binomial", data=data.frame(cbind(datw, z=z, m=m, y=y))), newdata=data.frame(cbind(datw, z=z, m=1)), type="response")*gm) + (predict(glm(formula=ymodel, family="binomial", data=data.frame(cbind(datw, z=z, m=m, y=y))), newdata=data.frame(cbind(datw, z=z, m=0)), type="response")*(1-gm))

  qfit<-glm(formula=paste("q1", qmodel, sep="~"), data=tmpdat[tmpdat$a==1,], family="quasibinomial")
  q2preda1<-predict(qfit, newdata=tmpdat, type="response")
  qfita0<-glm(formula=paste("q1", qmodel, sep="~"), data=tmpdat[tmpdat$a==0,], family="quasibinomial")
  q2preda0<-predict(qfita0, newdata=tmpdat, type="response")

  psa1<-I(a==1)/mean(a)
```

```
35  psa0<-I(a==0)/mean(1-a)
36  mz<-predict(glm(formula=mmodel, family="binomial", data=data.frame(cbind(
        datw, z=z, m=m))), newdata=data.frame(cbind(datw, z=z)), type="response"
        )
37  psm<-(mz*m) + ((1-mz)*(1-m))
38
39  #get E(Y_{1, g_0})
40  #clever covariate
41  tmpdat$h<-((m*gm + (1-m)*(1-gm))/psm) * psa1 * svywt
42  tmpdat$y<-y
43
44  epsilon<-coef(glm(y ~ 1 , weights=tmpdat$h, offset=(qlogis(q1)), family="
        quasibinomial", data=tmpdat))
45
46  tmpdat$q1up<-plogis(qlogis(tmpdat$q1) + epsilon)
47  eic1<-tmpdat$h * (tmpdat$y - tmpdat$q1up)
48
49  q2predfit<-glm(formula=paste("q1up", qmodel, sep="~"), data=tmpdat[tmpdat$a
        ==1,], family="quasibinomial")
50  q2pred<-predict(q2predfit, type="response", newdata=tmpdat)
51
52  tmlea1m0<-sum(q2pred*svywt)/sum(svywt)
53
54  epsilon2<-coef(glm(q1up~ 1 , weights=psa1*svywt, offset=qlogis(q2pred),
        family="quasibinomial", data=tmpdat))
55  q2up<-plogis(qlogis(q2pred) + epsilon2)
56
57  eic2<-psa1*svywt*(tmpdat$q1up - q2up)
58  eic<-eic1 + eic2
59
60  #get E(Y_{0, g_0})
61  #clever covariate
62  tmpdat$ha0<-((m*gm + (1-m)*(1-gm))/psm) * psa0 * svywt
63  epsilona0<-coef(glm(y ~ 1 , weights=tmpdat$ha0, offset=(qlogis(q1)), family
        ="quasibinomial", data=tmpdat))
64
65  tmpdat$q1upa0<-plogis(qlogis(tmpdat$q1) + epsilona0)
66  eic1a0<-tmpdat$ha0 * (tmpdat$y - tmpdat$q1upa0)
67
68  q2predfita0<-glm(formula=paste("q1upa0", qmodel, sep="~"), data=tmpdat[
        tmpdat$a==0,], family="quasibinomial")
69  q2preda0<-predict(q2predfita0, type="response", newdata=tmpdat)
70
71  tmlea0m0<-sum(q2preda0*svywt)/sum(svywt)
72
73  epsilon2a0<-coef(glm(q1upa0~ 1 , weights=psa0*svywt, offset=qlogis(q2preda0)
        , family="quasibinomial", data=tmpdat))
74  q2upa0<-plogis(qlogis(q2preda0) + epsilon2a0)
75
76  eic2a0<-psa0*svywt*(tmpdat$q1upa0 - q2upa0)
77  eica0<-eic1a0 + eic2a0
78
79  ndeeic<-eic - eica0
80  vareic<-var(ndeeic)/nrow(tmpdat)
```

```
81
82    #get E(Y_{1, g_1})
83    tmpdat$hma1<-((m*gma1 + (1-m)*(1-gma1))/psm) * psa1 * svywt
84
85    tmpdat$q1ma1<-(predict(glm(formula=ymodel, family="binomial", data=data.
          frame(cbind(datw, z=z, m=m, y=y))), newdata=data.frame(cbind(datw, z=z,
          m=1)), type="response")*gma1) + (predict(glm(formula=ymodel, family="
          binomial", data=data.frame(cbind(datw, z=z, m=m, y=y))), newdata=data.
          frame(cbind(datw, z=z, m=0)), type="response")*(1-gma1))
86
87    epsilonma1<-coef(glm(y ~ 1 , weights=tmpdat$hma1, offset=(qlogis(q1ma1)),
          family="quasibinomial", data=tmpdat))
88
89    tmpdat$q1upma1<-plogis(qlogis(tmpdat$q1ma1) + epsilonma1)
90    eic1ma1<-tmpdat$hma1 * (tmpdat$y - tmpdat$q1upma1)
91
92    q2predfitma1<-glm(formula=paste("q1upma1", qmodel, sep="~"), data=tmpdat[
          tmpdat$a==1,], family="quasibinomial")
93    q2predma1<-predict(q2predfitma1, type="response", newdata=tmpdat)
94
95    tmlea1m1<-sum(q2predma1*svywt)/sum(svywt)
96
97    epsilon2ma1<-coef(glm(q1upma1~ 1 , weights=psa1*svywt, offset=qlogis(
          q2predma1), family="quasibinomial", data=tmpdat))
98    q2upma1<-plogis(qlogis(q2predma1) + epsilon2ma1)
99
100   eic2ma1<-psa1*svywt*(tmpdat$q1upma1 - q2upma1)
101   eicma1<-eic1ma1 + eic2ma1
102
103   nieeic<-eicma1 - eic
104   varnieeic<-var(nieeic)/nrow(tmpdat)
105
106   return(list("a1m1"=tmlea1m1, "a1m0"=tmlea1m0, "a0m0"=tmlea0m0, "nde"=
          tmlea1m0-tmlea0m0, "ndevar"=vareic, "nie"=tmlea1m1-tmlea1m0, "nievar"=
          varnieeic))
107 }
108
109
110 set.seed(21230)
111 n<-10000
112
113 w0<-rbinom(n, 1, .5)
114 w1<-rbinom(n,1, .4 + (.2*w0))
115
116 probsel<-plogis(-1+ log(4)*race + log(4)*site)
117 psel<-rbinom(n, 1, probsel)
118 svywt<-mean(probsel)/probsel
119
120 #instrument
121 a<-rbinom(n, 1, .5)
122
123 #exposure
124 z0<-rbinom(n,1,plogis(       - log(2)*w1))
125 z1<-rbinom(n,1,plogis(log(4) - log(2)*w1))
```

```
z<-ifelse(a==1, z1, z0)

#mediator
m0<-rbinom(n, 1, plogis(-log(3)              - log(1.4)*w1))
m1<-rbinom(n, 1, plogis(-log(3) + log(10)- log(1.4)*w1))
m<-ifelse(z==1, m1, m0)

#outcomes
y<-rbinom(n,1, plogis(log(1.2)  + (log(3)*z)  + log(3)*m - log(1.2)*w1 + log
    (1.2)*w1*z) )

dat<-data.frame(w1=w1, a=a, z=z, m=m, y=y, psel=psel, svywt=svywt, radid_
    person=seq(1,n,1))
obsdat<-dat[dat$psel==1,]

zmodel<-"z ~ a + w1 "
mmodel<-"m ~ z + w1"
ymodel<-"y ~ m + z*w1"
qmodel<-"w1"

medtmle(a=obsdat$a, z=obsdat$z, m=obsdat$m, y=obsdat$y, w=data.frame(w1=obsdat
    $w1), svywt=obsdat$svywt, zmodel=zmodel, mmodel=mmodel, ymodel=ymodel,
    qmodel=qmodel)
```

exampcode.R