Causal mediation analysis for stochastic interventions

Iván Díaz  
Weill Cornell Medicine, New York, USA  
and Nima S. Hejazi  
University of California, Berkeley, USA  

[Received January 2019. Final revision December 2019]

Summary. Mediation analysis in causal inference has traditionally focused on binary exposures and deterministic interventions, and a decomposition of the average treatment effect in terms of direct and indirect effects. We present an analogous decomposition of the population intervention effect, defined through stochastic interventions on the exposure. Population intervention effects provide a generalized framework in which a variety of interesting causal contrasts can be defined, including effects for continuous and categorical exposures. We show that identification of direct and indirect effects for the population intervention effect requires weaker assumptions than its average treatment effect counterpart, under the assumption of no mediator–outcome confounders affected by exposure. In particular, identification of direct effects is guaranteed in experiments that randomize the exposure and the mediator. We propose various estimators of the direct and indirect effects, including substitution, reweighted and efficient estimators based on flexible regression techniques, allowing for multivariate mediators. Our efficient estimator is asymptotically linear under a condition requiring $n^{1/4}$-consistency of certain regression functions. We perform a simulation study in which we assess the finite sample properties of our proposed estimators. We present the results of an illustrative study where we assess the effect of participation in a sports team on the body mass index among children, using mediators such as exercise habits, daily consumption of snacks and overweight status.

Keywords: Continuous exposures; Cross-world independence; Efficient estimation; Intermediate confounders; Non-parametric mediation analysis; Stochastic interventions

1. Introduction

Mediation analysis is a powerful analytical tool that enables scientists to unveil the mechanisms through which causal effects operate. The development of tools for mediation analysis has a long history in the statistical sciences, starting with the early work of Wright (1921, 1934) on path analysis, which provided the foundations for the later development of mediation analysis using structural equation models (Goldberger, 1972). Indeed, one of the most widely used techniques for mediation analysis is based on structural equations (Baron and Kenny, 1986). Recent decades have seen a revolution in the field of causal inference from observational and randomized studies, starting with the seminal work of Rubin (1974) on the potential outcomes framework, which is itself rooted in ideas dating back to Neyman (1923). More recently, Pearl (1995, 2000) developed a causal inference framework using non-parametric structural equation
models (NPSEMs), directed acyclic graphs and the ‘do’ calculus. Related approaches have been proposed by Robins (1986), Spirtes et al. (2000), Dawid (2000) and Richardson and Robins (2013). These frameworks enable researchers to define causal effects non-parametrically, and to assess the conditions under which causal effects can be identified from data. In particular, novel tools have uncovered important limitations of the earlier work on parametric structural equation models for mediation analysis (Pearl, 1998; Imai et al., 2010). Essentially, structural equation models impose implausible assumptions on the data-generating mechanism and are thus of limited applicability to complex phenomena in biology, health, economics and the social sciences. For example, modern causal models have revealed crucial flaws in the widely popular method of Baron and Kenny (1986) in several important cases, such as in the presence of confounders of the mediator–outcome relationship (Cole and Hernán, 2002).

Using the potential outcomes framework, Robins and Greenland (1992) introduced a non-parametric decomposition of the causal effect of a binary exposure into so-called natural indirect and direct effects. The indirect effect quantifies the effect on the outcome through the mediator and the direct effect quantifies the effect through all other mechanisms. Pearl (2001) arrived at an equivalent effect decomposition by using NPSEMs. The identification of these natural (in)direct effects relies on cross-world counterfactual independences, i.e. independences on counterfactual variables indexed by distinct hypothetical interventions. An important consequence of this definition is that the natural (in)direct effect is not identifiable in a randomized trial without assumptions, which is problematic as it implies that scientific claims that are obtained from these models are not falsifiable through experimentation (Popper, 1934; Dawid, 2000; Robins and Richardson, 2010).

In an attempt to solve these problems, several researchers have proposed methods that do away with cross-world counterfactual independences. These methods can be divided into two types: identification of bounds (Robins and Richardson, 2010; Tchetgen Tchetgen and Phiri, 2014; Miles et al., 2015) and alternative definitions of the (in)direct effect (Petersen et al., 2006; van der Laan and Petersen, 2008; Vansteelandt and VanderWeele, 2012; VanderWeele et al., 2014). Here, we take the second approach, defining the (in)direct effect in terms of a decomposition of the total effect of a stochastic intervention on the population exposure.

Most causal inference problems consider deterministic interventions that set the exposure of each unit to some fixed value that could be a function of the unit’s baseline variables. Stochastic interventions are a generalization of this framework and are loosely defined as interventions which yield an exposure that is a random variable after conditioning on baseline variables. Estimation of total effects of stochastic interventions was first considered by Stock (1989) and has been the subject of recent study (Robins et al., 2004; Didelez et al., 2006; Tian, 2008; Pearl, 2009; Taubman et al., 2009; Stitelman et al., 2010; Díaz and van der Laan, 2013; Dudík et al., 2014; Haneuse and Rotnitzky, 2013; Young et al., 2014). Particularly relevant to this work are the methods of Díaz and van der Laan (2012), Haneuse and Rotnitzky (2013), who defined total effects for modified treatment policies, and Kennedy (2019), who studied identification and estimation of the total effect of propensity score interventions that shift a binary exposure distribution. These works do not address decomposition of the effects of stochastic interventions on the exposure into direct and indirect effects, which is the central theme of the present work.

Our methods also share similarities with a family of new direct and indirect effects (Didelez et al., 2006; VanderWeele et al., 2014; Lok, 2016; Vansteelandt and Daniel, 2017; Zheng and van der Laan, 2017; Rudolph et al., 2017; Lok, 2019), which have been collectively termed interventional effects (Nguyen et al., 2019). This family of effects deals with binary exposures and deterministic interventions on the exposure and is thus not entirely related to our approach, which deals with both continuous and categorical exposures and stochastic interventions on
the exposure. Similarly to the effects on the treated of Vansteelandt and VanderWeele (2012), interventional effects share the no cross-world independence property of our methods. The interested reader is referred to Nguyen et al. (2019) for a taxonomy of the several mediation analyses that have been proposed in the causal inference literature to date.

Stochastic interventions have analytical advantages compared with their deterministic counterparts, such as allowing for the seamless definition of causal effects of continuous exposures with an interpretation that is familiar to regular users of linear regression adjustment. For example, Haneuse and Rotnitzky (2013) assessed the effect of an intervention that reduces a patient’s operating time (i.e. the time spent in surgery) on the risk of post-operative outcomes among patients undergoing surgical resection of non-small-cell lung cancer. Diaz and van der Laan (2012) studied the effect of increasing the amount of leisure time physical activity in the elderly on subsequent all-cause mortality. Diaz and van der Laan (2013) studied the effect of a (hypothetical) policy that enforces pollution levels below a certain cut-off point. Kennedy (2019) showed that stochastic interventions can also be used in longitudinal studies to define and estimate total effects without reliance on the positivity assumption.

In the present work, we propose a decomposition of the effect of a stochastic intervention into a direct and an indirect effect, with interpretation analogous to that originally proposed by Robins and Greenland (1992) and Pearl (2001). We show that the identification of (in)direct effects based on stochastic interventions does not require cross-world counterfactual independences, therefore yielding scientific results that can be tested through experimentation on both the exposure and the mediator. Of high practical relevance, our proposal also allows the definition and estimation of non-parametric mediated effects for continuous exposures: a problem for which no methods or software currently exist. Parametric mediation methods such as those discussed by Vansteelandt et al. (2012) may induce large biases due to untestable parametric assumptions on the distribution of cross-world counterfactuals.

We develop a one-step non-parametric estimator based on the efficient influence function (EIF), incorporating flexible regression tools from the machine learning literature, and we provide \( n^{1/2} \)-rate convergence and asymptotic linearity results. We propose methods to use these asymptotic distributions to construct confidence regions and to test the null hypothesis of no direct effect. Our estimator has roots in semiparametric estimation theory (e.g. Pfanzagl and Wefelmeyer (1985), Begun et al. (1983), van der Vaart (1991), Newey (1994) and Bickel et al. (1997)), and in the targeted learning framework of van der Laan and Rubin (2006) and van der Laan and Rose (2011, 2018). In particular, we use cross-fitting to obtain \( n^{1/2} \)-convergence of our estimators while avoiding entropy conditions that may be violated by the data-adaptive estimators that we recommend (Bickel, 1982; Zheng and van der Laan, 2011; Chernozhukov et al., 2018). Our estimators use a reparameterization of certain integrals as conditional expectations to accommodate multivariate mediators. Software implementing our methods is provided in the form of the open source R package \texttt{medshift} that is freely available on GitHub (https://github.com/nhejazi/medshift).

2. Mediation analysis for population intervention effects

Let \( A \) denote a continuous or categorical exposure variable, let \( Y \) denote a continuous or binary outcome, let \( Z \) denote a multivariate mediator and let \( W \) denote a vector of observed covariates. Let \( O = (W, A, Z, Y) \) represent a random variable with distribution \( \mathbb{P} \). We use \( \mathbb{P}_n \) to denote the empirical distribution of a sample of \( n \) independent and identically distributed observations \( O_1, \ldots, O_n \). We let \( \mathbb{P} f = \int f(o) d\mathbb{P}(o) \) for a given function \( f(o) \), and we use \( \mathbb{E} \) to denote expectations with respect to \( \mathbb{P} \). We assume that \( \mathbb{P} \in \mathcal{M} \), where \( \mathcal{M} \) is the non-parametric statistical model.
defined as all continuous densities on $O$ with respect to a dominating measure $\nu$. Let $p$ denote
the corresponding probability density function. We use $g(a|w)$ to denote the probability density
function or the probability mass function of $A$ conditional on $W = w$, $m(a, z, w)$ and $b(a, w)$
to denote the outcome regression functions $\mathbb{E}(Y|A = a, Z = z, W = w)$ and $\mathbb{E}(Y|A = a, W = w)$
respectively, and $e(a|z, w)$ to denote the conditional density or probability mass function of $A$
conditional on $(Z, W)$. Let $g(a|w)$ be dominated by a measure $\kappa(a)$ (e.g. the counting measure
for binary $A$ and the Lebesgue measure for continuous $A$). We use $q(z|a, w)$ and $r(z|w)$ to denote
the corresponding conditional densities of $Z$. Note that $b, g$ and $r$ can be obtained from $m, e$ and
$q$ by averaging over the corresponding distributions of $Z$ and $A$. The parameterization $e = qg/r$
will prove fundamental in the construction of our estimators, since it will enable us to avoid
estimation of multivariate conditional densities. A similar parameterization was used by Zheng
and van der Laan (2012) and Tchetgen Tchetgen (2013) to estimate mediated effects under de-
terministic interventions. We use $W, A, Z$ and $Y$ to denote the support of the corresponding
random variables.

We formalize the definition of our counterfactual variables by using the following NPSEM,
but note that equivalent methods may be developed by taking the counterfactual variables as
primitives. Assume that

$$
\begin{align*}
W &= f_W(U_W), \\
A &= f_A(W, U_A), \\
Z &= f_Z(W, A, U_M), \\
Y &= f_Y(W, A, Z, U_Y).
\end{align*}
$$

(1)

This set of equations represents a mechanistic model generating the observed data $O$; fur-
thermore, the NPSEM encodes several fundamental assumptions. Firstly, an implicit tempo-
ral ordering is assumed—i.e. $Y$ occurs after $Z, A$ and $W$, $Z$ occurs after $A$ and $W$, and $A$
occurs after $W$. Secondly, each variable (i.e. $\{W, A, Z, Y\}$) is assumed to be generated from
the corresponding deterministic function (i.e. $\{f_W, f_A, f_Z, f_Y\}$) of the observed variables that
precede it temporally, plus an exogenous variable, denoted by $U$. Each exogenous variable is
assumed to contain all unobserved causes of the corresponding observed variable. Independence
assumptions on $U = (U_W, U_A, U_Z, U_Y)$ that are necessary for identification will be clari-
fied in Section 2.3. Furthermore, we note that we have explicitly excluded confounders of the
mediator–outcome relationship affected by exposure. Mediation analysis in the presence of
such variables is notoriously difficult (Avin et al., 2005); the adaptation of our methods to this
problem is possible but requires a new set of tools that are outside the scope of the present
paper.

Causal effects are defined in terms of hypothetical interventions on the NPSEM (1). In par-
icular, consider an intervention in which the equation corresponding to $A$ is removed, and the
exposure is drawn from a user-specified distribution $g_\delta(a|w)$, which may depend on $g$ and is
indexed by a user-specified parameter $\delta$. We assume without loss of generality that $g_\delta=0 = g$. Let
$A_\delta$ denote a draw from $g_\delta(a|w)$. Alternatively, such modifications can sometimes be described in
terms of an intervention in which the equation corresponding to $A$ is removed and the exposure
is set equal to a hypothetical regime $A_\delta = d(A, W)$. Regime $d$ depends on the natural (i.e. under
no intervention) exposure level $A$ and covariates $W$. This intervention is sometimes referred to
as depending on the natural value of exposure, or as a modified treatment policy (Haneuse and
Rotnitzky, 2013). Young et al. (2014) have provided a discussion of the differences and similar-
ities in the interpretation and identification of these two interventions. Below, we discuss two
examples of stochastic interventions:
(a) modified treatment policies and
(b) exponential tilting.

2.1. Example 1 (modified treatment policy)
Let $A$ denote a continuous exposure, such as operating time in non-small-cell lung cancer (Haneuse and Rotnitzky, 2013). Assume that the distribution of $A$ conditional on $W = w$ is supported in the interval $(l(w), u(w))$, i.e. the minimum possible operating time for an individual with covariates $W = w$ is $l(w)$. Then we may define a hypothetical post-intervention exposure $A_{\delta} = d(A, W)$, where

$$d(a, w) = \begin{cases} a - \delta & \text{if } a > l(w) + \delta, \\ a & \text{if } a \leq l(w) + \delta, \end{cases}$$

(2)

where $0 < \delta < u(w)$ is an arbitrary user-given value. Interesting modifications to this regime may be obtained by allowing $\delta$ to be a function of $w$, therefore enabling the researcher to specify a different change in operating time as a function of covariates such as comorbidities and age. This intervention was first introduced by Díaz and van der Laan (2012) and has been further discussed in Díaz and van der Laan (2018) and Haneuse and Rotnitzky (2013).

2.2. Example 2 (exponential tilting)
Alternatively, a tilted intervention distribution may be defined as

$$g_{\delta}(a|w) = \frac{\exp(\delta a) g(a|w)}{\int \exp(\delta a) g(a|w) d\kappa(a)},$$

(3)

for $\delta \in \mathbb{R}$, letting the hypothetical post-intervention exposure $A_{\delta}$ be a random draw from $g_{\delta}$, conditional on the natural value of the observed covariates $W$. For binary $A$, Kennedy (2019) proposed to evaluate the total effect of a binary exposure $A$ in terms of incremental propensity score interventions that replace the propensity score $g(1|w)$ with a shifted version based on multiplying the odds of exposure by a user-given parameter $\delta'$. In particular, the post-intervention propensity score is given by

$$g_{\delta'}(1|w) = \frac{\delta' g(1|w)}{\delta' g(1|w) + 1 - g(1|w)},$$

(4)

for $0 < \delta' < \infty$. The proposal of Kennedy (2019) is thus a case of exponential tilting (3) under the parameterization $\delta' = \exp(\delta)$. This choice of parameterization is motivated by the fact that $\delta'$ can be interpreted as an odds ratio indicating how the intervention changes the odds of exposure. The extremes of $\delta' = 0$ and $\delta' = \infty$ correspond to the standard static interventions $A = 0$ and $A = 1$ considered in the definition of the average treatment effect.

We now turn our attention to defining the population intervention effect (PIE) of $A$ on $Y$. To proceed, for any values $(a, z)$, consider the counterfactual outcomes $Y_{a, z} = f_Y(W, a, z, U_Y)$ and $Y(a) = f_Y(W, a, Z(a), U_Y)$, as well as the counterfactual mediator $Z(a) = f_Z(W, a, U_Z)$. The counterfactual $Y_{a, z}$ is the outcome in a hypothetical world in which $(A, Z) = (a, z)$ is fixed externally. The PIE is defined as a contrast comparing the expectation of the outcome under no intervention with the expectation of the counterfactual outcome obtained under an intervention $A_{\delta}$:

$$\mathbb{E}\{Y(A_{\delta}) - Y\}.$$

Note that the interpretation of the PIE depends on the stochastic intervention that is considered. For example, for the modified treatment policies of example 1, the PIE describes the difference in outcomes obtained by a reduction of $\delta$ in operating time. In the case of the incremental propensity
score intervention (4), the PIE is interpreted as the difference in outcomes obtained by an intervention under which the odds of exposure is $\delta'$ times higher compared with current practice.

Since $A$ is a cause of $Z$, an intervention that changes the exposure to $A_\delta$ also induces a counterfactual mediator $Z(A_\delta)$. As a consequence of the consistency that is implied by the NPSEM, we have $Y(A, Z) = Y$. Similarly, the expression $Y\{A_\delta, Z(A_\delta)\} = Y(A_\delta)$ follows from the law of composition (Pearl, 2000). Thus, the PIE may be decomposed in terms of a population intervention direct effect and a population intervention indirect effect:

$$
\text{indirect effect} = \mathbb{E}\{Y\{A_\delta, Z(A_\delta)\} - Y(A_\delta, Z)\} + \mathbb{E}\{Y(A_\delta, Z) - Y(A, Z)\}.
$$

This decomposition of the PIE as the sum of direct and indirect effects has an interpretation that is analogous to the corresponding standard decomposition of the average treatment effect (Pearl, 2001). In particular, the direct effect represents the effect of an intervention that changes the distribution of the exposure while keeping the distribution of the mediators fixed at the value that it would have taken under no intervention. The indirect effect measures the effect of an indirect intervention on the mediators generated by intervening on the exposure while holding the intervention on the exposure constant.

The intervention in example 1 arises naturally as a modified treatment policy. In contrast, the intervention in example 2 arises directly as a stochastic intervention that modifies the distribution of the exposure variable—it is unclear yet whether this quantity may be interpreted as a modified intervention in example 2 arises directly as a stochastic intervention that modifies the distribution of the exposure while holding the distribution of the mediators generated by intervening on the exposure while keeping the distribution of the mediators fixed at the value $h_j(a, w)$ with derivative $h'_j(a, w)$.

Under this assumption, the distribution of a modified treatment policy $A_\delta = d(A, W)$ may be recovered through

$$
g_\delta(a|w) = \sum_{j=1}^{J(w)} I_{\delta, j}\{h_j(a, w), w\} g\{h_j(a, w)|w\} h'_j(a, w),
$$

where $I_{\delta, j}\{u, w\} = 1$ if $u \in \mathcal{I}_{\delta, j}(w)$ and $I_{\delta, j}\{u, w\} = 0$ otherwise. In example 1, the stochastic intervention becomes

$$
g_\delta(a|w) = g(a|w) \mathbb{I}\{l(w) \leq a \leq l(w) + \delta\} + g(a + \delta|w) \mathbb{I}\{l(w) \leq a \leq u(w) - \delta\}.
$$

Therefore, under assumption 1, a modified treatment policy may also be represented as a change by which the equation $f_A$ is removed from the NPSEM and $A$ is replaced by a draw $A_\delta$ from the distribution $g_\delta(a|w)$. As a result of these two representations, the intervention may be interpreted in two different ways:

(a) a change in the probabilistic mechanism that is used to assign exposure level and
(b) a subject-specific change in exposure from $A$ to $A_\delta = d(A, W)$,

where only interpretation (a) requires assumption 1. Note, however, that the population distribution of the exposure is the same under both interventions (Young et al., 2014); thus, both
representations lead to exactly the same marginal counterfactual outcome distributions. In what follows we denote $\psi(\delta) = \mathbb{E}\{Y(A_\delta)\}$ and $\theta(\delta) = \mathbb{E}\{Y(A_\delta, Z)\}$.

2.3. Identification

We introduce the following identification assumptions.

Assumption 2 (common support). Assume that $\text{supp}\{g_\delta(\cdot|w)\} \subseteq \text{supp}\{g(\cdot|w)\}$ for all $w \in W$.

Assumption 3 (conditional exchangeability of exposure assignment). For all $(a, w) \in A \times W$, assume that $\mathbb{E}\{Y(a) | A = a, W = w\} = \mathbb{E}\{Y(a) | W = w\}$.

Assumption 4 (conditional exchangeability of exposure and mediator assignment). For all $(a, w, z) \in A \times W \times Z$, assume that $\mathbb{E}\{Y(a, z) | A = a, W = w, Z = z\} = \mathbb{E}\{Y(a, z) | W = w, Z = z\}$.

Assumptions 2 and 3 are standard in the analysis of causal effects. Assumption 2 simply states that the $\delta$-specific intervention of interest is supported in the data. This assumption holds for all $\delta$ in the interventions that were described in examples 1 and 2 (Díaz and van der Laan, 2012; Kennedy, 2019). Assumption 3 is standard exchangeability of exposure assignment. Assumption 4 is related to the assumption that Vansteelandt and VanderWeele (2012) used for identification of mediated effects among the treated. In that proposal Vansteelandt and VanderWeele (2012) assumed that $Y(a, z) \perp\!(A, Z) | W$, which would imply the stronger assumption $\mathbb{E}\{Y(a, z) | A = a, W = w, Z = z\} = \mathbb{E}\{Y(a, z) | W = w\}$. This assumption would be satisfied for any pre-exposure variable $W$ in a randomized experiment in which the exposure and mediator are randomized. Thus, the direct effect for a population intervention corresponds to contrasts between treatment regimes of a randomized experiment via interventions on $A$ and $Z$, unlike the natural direct effect for the average treatment effect (Robins and Richardson, 2010). This claim is made rigorous in the identification result of theorem 1 that is presented below. A proof is available in the online supplementary materials, together with the assumptions on the exogenous errors $U$ of the NPSEM, which are compatible with assumption 4.

Theorem 1 (identification). Assume assumption 2.

(a) Under assumption 3, $\psi(\delta)$ is identified and is given by

$$\psi(\delta) = \int m(a, z, w)g_\delta(a|w)q(z|a, w)p(w)d\nu(a, z, w).$$

(b) Under assumption 4, $\theta(\delta)$ is identified and is given by

$$\theta(\delta) = \int m(a, z, w)g_\delta(a|w)p(z, w)d\nu(a, z, w).$$

Remark 1 (exposure–mediator confounders). Note that identification of $\theta(\delta)$ does not require identification of the effect of $A$ on $Z$ (see the web supplementary materials), and thus the indirect effect is identified in the presence of unmeasured exposure–mediator confounders. The direct effect, however, requires that all exposure–mediator confounders are measured, as this is necessary for assumption 3.

Remark 2 (mediator–outcome confounder not affected by exposure). Like the natural direct effect of Pearl (2001), we require that all confounders of the mediator–outcome relationship be measured. This assumption is implicit in assumption 4. To see why, consider the directed acyclic
I. Díaz and N. S. Hejazi

(...)

Remark 3 (mediator–outcome confounded by exposure). The methods that are presented here cannot be used if the mediator–outcome confounder $V$ is affected by exposure. This is due to the introduction of a new counterfactual variable $V(a)$. In particular, consider the directed acyclic graph in Fig. 2, where we have included only the relevant factual and counterfactual variables. In this case, conditioning on the collider $V$ would open a path $A \rightarrow V \leftarrow U_V \rightarrow V(a) \rightarrow Y(a, z)$, thus making assumption 4 invalid. However, conditioning on $V$ is necessary for assumption 4 to close the path $A \rightarrow Z \leftarrow V \rightarrow U_V \rightarrow V(a) \rightarrow Y(a, z)$, which is itself opened on conditioning on the collider $Z$. A comprehensive discussion of issues in identification of path effects, addressing this issue as a particular case of the general problem, may be found in Avin et al. (2005). VanderWeele et al. (2014) proposed a solution to this problem which involves a stochastic intervention on the mediator $Z$. We note that this is intrinsically different from the problem that is treated here, as we are interested in stochastic interventions on $A$ (not on $Z$) and thus do not address mediator–outcome confounders affected by exposure.

Several estimators of the functional $\psi(\delta)$ have previously been proposed. For the case of a continuous exposure, Díaz and van der Laan (2012) developed inverse probability weighted, outcome regression and doubly robust estimators based on the framework of targeted minimum loss estimation (van der Laan and Rose, 2011), using data-adaptive estimators of the relevant nuisance parameters. Díaz and van der Laan (2018) improved on the previous methodology by constructing a targeted minimum loss estimation algorithm with substantially lower computational complexity that preserves the desirable asymptotic properties of the original approach. Haneuse and Rotnitzky (2013) proposed estimators that rely on correctly specified parametric models. Such methods are of limited applicability since they are reliable only in situations where the nuisance parameters involve only few categorical variables or where correctly specified (i.e. saturated) parametric models can conscientiously be constructed. For the binary case with $g_\delta$ as in example 2, Kennedy (2019) proposed an estimator for $\psi(\delta)$. This estimator is efficient and asymptotically linear, while allowing incorporation of data-adaptive estimators of the nuisance parameters.

Since $E(Y)$ is trivially estimated by the empirical mean in the sample, our optimality theory and estimators focus on $\theta(\delta)$. We present two types of result: for general modified treatment policies satisfying assumption 1, and for the particular stochastic intervention of example 2. We compare the assumptions that are required for both.
3. Optimality theory for estimation of the direct effect

Thus far we have discussed the decomposition of the effect of a stochastic intervention into direct and indirect effects, and we have provided identification results under weaker assumptions in comparison with the natural direct effect. In what follows, we turn our attention to a discussion of efficiency theory for the estimation of $\theta(\delta)$ in the non-parametric model $\mathcal{M}$. The EIF is a key object in semiparametric estimation theory, as it characterizes the asymptotic behaviour of all regular and efficient estimators (Bickel et al., 1997; van der Vaart, 2002). Knowledge of the EIF has important practical implications. First, the EIF is often useful in constructing locally efficient estimators. There are three common approaches for this:

(a) using the EIF as an estimating equation (e.g. van der Laan and Robins (2003)),
(b) using the EIF in a one-step bias correction (e.g. Pfanzagl and Wefelmeyer (1985)) and
(c) targeted minimum loss estimation (van der Laan and Rubin, 2006; van der Laan and Rose, 2011, 2018).

Second, the EIF estimating equation often enjoys desirable properties such as multiple robustness, which allows for some components of the data distribution to be inconsistently estimated while preserving consistency of the estimator. Third, asymptotic analysis of estimators constructed by using the EIF often yields second-order bias terms, which require slow convergence rates ($e.g. n^{-1/4}$) for the nuisance parameters that are involved, thereby enabling the use of flexible regression techniques in estimating these quantities.

In theorem 2 we present the EIF for a general stochastic intervention. Although the component of the EIF that are associated with $Y$ and $Z, W$ are the same, the component that is associated with the model for the distribution of $A$ must be computed case by case, i.e. for each intervention of interest. Proofs for all results are available in the on-line supplementary materials.

**Theorem 2 (EIF).** Let $\eta = (g, m, e)$. The EIF for $\theta(\delta)$ in the non-parametric model $\mathcal{M}$ is $D_Y^{\eta,\delta}(o) + D_A^{\eta,\delta}(o) + D_Z^{Z,W}(o) - \theta(\delta)$, where

$$D_Y^{\eta,\delta}(o) = \frac{g_\delta(a|w)}{e(a|z,w)} \{y - m(a,z,w)\},$$

$$D_Z^{Z,W}(o) = \int m(a,z,w)g_\delta(a|w)d\kappa(a),$$

and $D_A^{\eta,\delta}(o)$ is the efficient score corresponding to the non-parametric model for $g$.

An immediate consequence of theorem 2 is that, in a randomized experiment where $g(a|w)$ is known by design, we have $D_A^{\eta,\delta}(o) = 0$. Lemmas 1 and 2 below present the $D_A^{\eta,\delta}(o)$ components for modified treatment policies satisfying assumption 1 and for the exponential tilting of example 2 respectively. To do so, we define the nuisance parameter

$$\phi(a, w) = \int m(a,z,w)r(z|w)d\nu(z) \quad (9)$$

$$= \mathbb{E}\left\{ \frac{g(A|W)}{e(A|Z,W)}m(A,Z,W)|A=a, W=w \right\}, \quad (10)$$

and augment $\eta$ as $\eta = (g, m, e, \phi)$.

**Lemma 1 (modified treatment policies).** If the modified treatment policy $d(A, W)$ satisfies assumption 1, then
Lemma 2 (exponential tilt). If the stochastic intervention is the exponential tilt (3), then

\[
D^{A}_{\eta, \delta}(o) = \delta g(a|w) \left\{ \phi(a, w) - \frac{\int \phi(d(a, w), w) g(a|w) \, d\kappa(a)}{\int g(a|w) \, d\kappa(a)} \right\}.
\]

Expression (9) shows that estimators based on the influence functions require integration with respect to the mediator \( Z \), as well as estimation of the possibly multivariate conditional density \( r(z|w) \), which may pose an estimation challenge due to the curse of dimensionality. To solve the issue, we propose an alternative parameterization (10) of the EIF based on a sequential regression \( \phi \), rather than using the density of \( Z \) conditional on \( (A, W) \) and \( W \). This choice has important consequences for estimation, as it helps to bypass estimation of the (possibly high dimensional) conditional density of the mediators, instead allowing for regression methods, which are far more commonly found in the statistics literature and software, to be used for estimation of the relevant quantity. If \( r(z|w) \) should prove challenging to estimate, estimators of \( \phi \) may be computed by first estimating \( g, m \) and \( e \), computing the pseudo-outcomes defined in the lemmas and applying regression techniques to estimate the outer conditional expectation.

For binary exposures, the EIF corresponding to the incremental propensity score intervention may be simplified as in the following corollary.

**Corollary 1 (EIF for incremental propensity score interventions).** Let \( A \) take values on \( \{0, 1\} \), and let the exponentially tilted intervention \( g_{\delta}(1|W) \) be as in expression (4). Then, the EIF of lemma 2 may be simplified as follows. Define the nuisance parameter

\[
\phi(w) = \mathbb{E}\{m(1, Z, W) - m(0, Z, W) | W = w\},
\]

and let \( \eta = (g, m, e, \phi) \). Then

\[
D^{A}_{\eta, \delta}(o) = \frac{\delta \phi(w) \{a - g(1|w)\}}{\{\delta g(1|w) + 1 - g(1|w)\}^2}.
\]

Note that, in lemmas 1 and 2, and corollary 1, we have used \( \phi \) to represent different parameters. We have allowed this abuse of notation because the nature of this auxiliary parameter is the same for all cases; thus, this economy of notation will enable us to state our estimation results in some generality. In what follows, the difference will always be clear from the context. Note also that \( g(a|w) \) could be pulled outside the expectation in the definition of \( \phi(a, w) \); however, we choose to leave it within the expectation as we conjecture that it may act as a stabilizing factor for the inverse probability weights \( \{e(a|z, w)\}^{-1} \) based on exploratory numerical experimentation.

In contrast with the EIF for the natural direct effect (Tchetgen Tchetgen and Shpitser, 2012), the contribution of the exposure process to the EIF for the mediated PIE is non-zero. This is a direct consequence of the fact that the parameter of interest depends on \( g \); moreover, this implies that, unlike the natural direct effect, the efficiency bound in observational studies differs from the efficiency bound in randomized studies. As we see in the following lemmas, this also implies that it is not generally possible to obtain estimating equations that are robust to inconsistent estimation of \( g \). However, such robustness will be possible if the stochastic intervention is also a modified treatment policy satisfying assumption 1.

**Lemma 3 (multiple robustness for modified treatment policies).** Let the modified treatment policy satisfy assumption 1, and let \( \eta_1 = (g_1, e_1, m_1, \phi_1) \) be such that one of the two following conditions hold:
(a) $g_1 = g$ and either $e_1 = e$ or $m_1 = m$;
(b) $m_1 = m$ and $\phi_1 = \phi$.

Then $\mathbb{P} D_{\eta_1, \delta} = \theta(\delta)$, with $D_{\eta_1, \delta}$ as defined in theorem 2 and lemma 1.

Lemma 3 implies that it is possible to construct consistent estimators for $\theta$ under consistent estimation of at least two of the nuisance parameters in $\eta$, in the configurations that are described in lemma 3. This lemma is a direct consequence of theorem 5, found in the on-line supplementary materials. We note, however, that condition (b) of the lemma may be uninteresting if the parameterization (10) is used to estimate $\phi$. In that case $\phi_1 = \phi$ will generally require $g_1 = g$, $e_1 = e$ and $m_1 = m$, as well as consistency of the estimator for the outer expectation. In contrast, if the parameterization (9) is used to estimate $\phi$, then the case $m_1 = m$ and $\phi_1 = \phi$ would be trivially satisfied if $m_1 = m$ and $r_1 = r$, where $r_1$ is the density that is used to compute $\phi_1$. It may surprise some readers that estimation of $\theta(\delta)$ may be robust to estimation of $g$, even when the parameter definition in equation (8) is explicitly dependent on $g$. We offer some intuition into this surprising result by noting that assumption 1 allows use of the change-of-variable formula to obtain

$$\theta(\delta) = \mathbb{E} \left[ \int m(A, W, Z, W) r(z, W) d\nu(z) \right].$$

Estimation of this parameter without relying on $g$ may be carried out by consistently estimating $m$ and $r$, and using the empirical distribution as an estimator of the outer expectation. This behaviour has been previously observed for the effect $\psi(\delta)$ under assumption 1 (Díaz and van der Laan, 2012; Haneuse and Rotnitzky, 2013).

The robustness properties of the EIF for an exponential tilt are presented in what follows.

**Lemma 4** (robustness for exponential tilting). Let $g_\delta$ be defined as in equation (3). Let $\eta_1 = (g_1, e_1, m_1, \phi_1)$ be such that $g_1 = g$ and either $e_1 = e$ or $m_1 = m$. Then $\mathbb{P} D_{\eta_1, \delta} = \theta(\delta)$, with $D_{\eta_1, \delta}$ as defined in theorem 2 and lemma 2.

Lemma 4 is a direct consequence of theorem 6 in the on-line supplementary materials. The corresponding proof reveals that the EIF for the binary distribution is not multiply robust—i.e. the intervention fails to satisfy assumption 1 and integrals over the range of $A$ cannot be computed by using the change-of-variable formula. This behaviour has been previously observed for other interventions that do not satisfy assumption 1 (Díaz and van der Laan, 2013). Even though this lemma implies that consistent estimation of $g$ is required, the bias terms are still second order, so an estimator of $g$ converging at rate $n^{1/4}$ or faster is sufficient, as we shall see in what follows.

### 4. Estimation and statistical inference

We begin by describing two simple estimators: the substitution and reweighted estimators. These estimators are motivated by the fact that $\theta(\delta)$ has the two following alternative representations:

$$\theta(\delta) = \mathbb{E} \left\{ \int m(a, Z, W) g_\delta(a | W) d\kappa(a) \right\}$$

$$= \mathbb{E} \left\{ \frac{g_\delta(A | W)}{e(A | Z, W)} \right\},$$

where we remind the reader that $e(a | Z, W)$ denotes the probability density function of $A$ conditional on $(Z, W)$. Equation (12) follows from noting that $gq/r = e$. This parameterization
has the advantage that only the univariate conditional density \( e(a|z,w) \) must be estimated, instead of the conditional densities of the possibly high dimensional mediator \( Z \). A similar result was also used by Zheng and van der Laan (2012) and Tchetgen Tchetgen (2013) to develop a targeted minimum loss estimator of natural direct effects under a binary exposure variable.

The substitution estimator is simply defined by plugging estimators of \( m \) and \( g_{\delta} \) into the identification result given in equation (11). Consistency of this estimator requires consistent estimation of the outcome regression \( m \) and the intervention distribution \( g_{\delta} \). The second estimator is a reweighting estimator based on the alternative representation of the identification result given in equation (12), which requires consistent estimation of \( g_{\delta} \) and \( e \). In the remainder of this section, we discuss an efficient estimator that combines ideas from the previous two estimators as well as the EIF that was derived in the previous section, to build an estimator that is both efficient and robust to model misspecification. We discuss an asymptotic linearity result for the doubly robust estimator that enables the construction of asymptotically correct confidence intervals and hypothesis tests.

In what follows, we assume that preliminary estimators \( \hat{m}, \hat{g}_{\delta}, \hat{\phi} \) and \( \hat{e} \) of \( m, g_{\delta}, \phi \) and \( e \) respectively are available. These estimators may be obtained from flexible regression techniques such as support vector machines, regression trees, boosting, neural networks, splines or ensembles thereof (Breiman, 1996; van der Laan et al., 2007). As previously discussed, the consistency of these estimators will determine the consistency of our estimators of the population mediation intervention mean \( \theta \).

4.1. Substitution estimator and reweighted estimators
First, we discuss a substitution estimator based on equation (11), computed as

\[
\hat{\theta}_{\text{sub}}(\delta) = \int \frac{1}{n} \sum_{i=1}^{n} \hat{m}(a, Z_i, W_i) \hat{g}_{\delta}(a|W_i) d\kappa(a),
\]

where we have substituted estimators of \( m \) and \( g_{\delta} \) in equation (11) and have estimated the expectation with respect to the joint density \( p(z,w) \) by the empirical mean. The reweighted estimator is based on equation (12) and is defined by

\[
\hat{\theta}_{\text{re}}(\delta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{g}_{\delta}(A_i|W_i)}{\hat{e}(A_i|Z_i, W_i)} Y_i.
\]

If \( \hat{m}, \hat{g}_{\delta} \) and \( \hat{e} \) are estimated within prespecified parametric models, then, by the delta method, both \( \hat{\theta}_{\text{sub}}(\delta) \) and \( \hat{\theta}_{\text{re}}(\delta) \) are asymptotically linear and \( n^{1/2} \) consistent. The bootstrap or an influence-function-based estimator may be used to construct asymptotically correct confidence intervals. However, if either the mediators or confounders are high dimensional, the required consistency of \( \hat{m}, \hat{g}_{\delta} \) and \( \hat{e} \) will hardly be achievable within parametric models. This issue may be alleviated through the use of data-adaptive estimators. Unfortunately, \( n^{1/2} \)-consistency of \( \hat{\theta}_{\text{sub}}(\delta) \) and \( \hat{\theta}_{\text{re}}(\delta) \) will generally require that \( \hat{m}, \hat{g}_{\delta} \) and \( \hat{e} \) are consistent in \( L_2(\mathbb{P}) \) norm at parametric rate, which is generally not possible when utilizing data-adaptive estimation of high dimensional regressions. Thus, the asymptotic distribution will generally be unknown, rendering the construction of confidence intervals and hypothesis tests impossible. In the following section, we use the EIF to propose an estimator that is \( n^{1/2} \)-consistent and efficient under a weaker assumption, requiring only \( n^{1/2} \)-convergence of second-order regression bias terms.
4.2. Efficient one-step estimator

We propose to use the EIF $D_{\eta,\delta}$ to construct a robust and efficient estimator, constructed as the solution to the estimating equation $P_n D_{\hat{\eta},\delta} = 0$ in $\theta$, for a preliminary estimator $\hat{\eta}$ of $\eta$. To avoid imposing entropy conditions on the initial estimators, we use sample splitting and cross-fitting in the estimation procedure (Bickel, 1982). Let $V_1, \ldots, V_J$ denote a random partition of the index set $\{1, \ldots, n\}$ into $J$ prediction sets of approximately the same size, i.e. $V_j \subset \{1, \ldots, n\}$, $\bigcup_{j=1}^J V_j = \{1, \ldots, n\}$ and $V_j \cap V_{j'} = \emptyset$. In addition, for each $j$, the associated training sample is given by $T_j = \{1, \ldots, n\} \setminus V_j$. Denote by $\hat{\eta}_j$ the estimator of $\eta = (g, m, \epsilon, \phi)$, obtained by training the corresponding prediction algorithm by using only data in the sample $T_j$. Further, let $j(i)$ denote the index of the validation set which contains observation $i$. The estimator is thus defined as

$$\hat{\theta}(\delta) = \frac{1}{n} \sum_{i=1}^n D_{\hat{\eta}_{j(i)},\delta}(O_i) = \frac{1}{n} \sum_{i=1}^n \{D_{\hat{\eta}_{j(i)},\delta}^Y(O_i) + D_{\hat{\eta}_{j(i)},\delta}^A(O_i) + D_{\hat{\eta}_{j(i)},\delta}^{Z,W}(O_i)\}. \quad (15)$$

In a randomized trial the estimator may also be computed by setting $D_{\hat{\eta}_{j(i)},\delta}^A(O_i) = 0$. M-estimation theory may be used to derive the asymptotic distribution of $\hat{\theta}(\delta)$. Asymptotic linearity and efficiency of the estimator for modified treatment policies are detailed in the following theorem.

**Theorem 3** (pointwise weak convergence for modified treatment policies). Let $\| \cdot \|$ denote the $L_2(\mathbb{P})$ norm defined as $\|f\|^2 = \int f^2 d\mathbb{P}$. Assume that

(a) $\|\hat{m} - m\| + \|\hat{g} - g\| + \|\hat{\epsilon} - e\| + \|\hat{\phi} - \phi\| = o_P(n^{-1/2})$,

(b) $\mathbb{P}\{|D_{\eta,\delta}(O)\| \leq C\} = \mathbb{P}\{|D_{\hat{\eta},\delta}(O)\| \leq C\} = 1$ for some $C < \infty$ and

(c) the modified treatment policy $d(a, w)$ is piecewise smooth invertible (assumption 1).

Then

$$\sqrt{n}\{\hat{\theta}(\delta) - \theta(\delta)\} \overset{d}{\rightarrow} N\{0, \sigma^2(\delta)\},$$

where $\sigma^2(\delta) = \text{var}\{D_{\eta,\delta}(O)\}$ is the efficiency bound.

Theorem 3 establishes the weak convergence of $\hat{\theta}(\delta)$ pointwise in $\delta$. This convergence is useful to derive confidence intervals in situations where the modified treatment policy has a suitable scientific interpretation for a given $\delta$, such as in our example 1. Under the assumptions of theorem 3, an estimator $\hat{\sigma}^2(\delta)$ of $\sigma^2(\delta)$ may be obtained as the empirical variance of $D_{\hat{\eta}_{j(i)},\delta}(O_i)$, and a Wald-type confidence interval may be constructed as $\hat{\theta}(\delta) \pm z_{1-\alpha/2} \hat{\sigma}(\delta) / \sqrt{n}$.

For the remainder of this section, we turn our attention to a discussion of uniform convergence of $\hat{\theta}(\delta)$. Such a convergence result will prove useful in the following section, where we establish a hypothesis test of no direct effect. Such a test is constructed by rejecting the hypothesis if the direct effect is non-significant (at level $\alpha$), uniformly in $\delta$. To build such a testing procedure, we focus on the intervention that is defined in terms of exponential tilting (3). Results for modified treatment policies are possible as well; however, these require smoothness assumptions on the map $\delta \mapsto g_\delta(a|w)$. Inspection of expression (6) reveals that this may in turn require smoothness assumptions on $a \mapsto g(a|w)$, which may not be justifiable in some applications. We thus focus on exponential tilting, which yields smooth maps $\delta \mapsto g_\delta(a|w)$ by construction. This discussion, together with lemmas 1 and 2, reveals a trade-off between smoothness and robustness in estimation of modified treatment policies and exponential tilting.

**Theorem 4** (uniform weak convergence for exponential tilting). Let $g_\delta$ be the exponential tilting intervention distribution (3) and let $\Delta = [\delta_l, \delta_u]$ denote an interval with $0 < \delta_l \leq \delta_u < \infty$. Define $c(w) = \left\{ \int g(\delta|a)g(a|w) \right\}^{-1}$. Assume that $\|\hat{c} - c\|^2 = o_P(n^{-1/2})$ as well as assumptions (a) and (b) stated in theorem 3. Then
\[ \sqrt{n} \{ \hat{\theta}(\delta) - \theta(\delta) \} \sim \mathcal{G}(\delta) \]

in \( L^\infty(\Delta) \), where, for any \( \delta_1, \delta_2 \in \Delta \), \( \mathbb{G}(\cdot) \) is a mean 0 Gaussian process with covariance function \( \mathbb{E}\{ \mathbb{G}(\delta_1) \mathbb{G}(\delta_2) \} = \mathbb{E}\{ D_{\eta,\delta_1}(O) D_{\eta,\delta_2}(O) \} \).

### 4.3. Uniform inference and testing the hypothesis of no direct effect

In this section, we consider estimation of the direct effect \( \beta(\delta) = \theta(\delta) - \mathbb{E}(Y) \). Define the corresponding (unc centred) influence function \( S_{\eta,\delta}(o) = D_{\eta,\delta}(o) - y \). A straightforward extension of theorem 4 shows that \( \hat{\beta}(\delta) = \hat{\theta}(\delta) - \hat{Y} \) converges weakly to a process \( \mathbb{G}(\delta) \) with covariance function \( \mathbb{E}\{ \mathbb{G}(\delta_1) \mathbb{G}(\delta_2) \} = \mathbb{E}\{ S_{\eta,\delta_1}(O) S_{\eta,\delta_2}(O) \} \).

We now present an approach to constructing uniform confidence bands on the function \( \beta(\delta) \), allowing testing of the null hypothesis of no direct effect \( H_0 : \sup_{\delta \in \Delta} \beta(\delta) = 0 \). This hypothesis test is useful for checking the existence of a direct effect even if the interpretation of the exponential tilt \( y_k \) (e.g., as the odds ratio comparing post versus preintervention odds of exposure) does not answer a particularly meaningful question in a given application. Let \( \hat{\sigma}(\delta) \) denote the empirical variance of \( S_{\hat{\eta}(\delta),\delta}(O) \). Our goal will be achieved by finding a value \( c_\alpha \) such that \( \hat{\rho}(c_\alpha) = 1 - \alpha \), where \( \hat{\rho} \) is a function satisfying

\[ \hat{\rho}(t) = \mathbb{P}\left\{ \sup_{\delta \in \Delta} \left| \frac{\hat{\beta}(\delta) - \beta(\delta)}{\hat{\sigma}(\delta) / \sqrt{n}} \right| \leq t \right\} + o(1). \]  

(16)

Confidence bands may be computed as \( \hat{\theta} \pm n^{-1/2} c_\alpha \hat{\sigma}(\delta) \), and \( p \)-values for \( H_0 \) can be computed by evaluating \( 1 - \hat{\rho}(t) \) at the observed value of the supremum test statistic. The function \( \hat{\rho}(t) \) may be obtained by approximating the distribution of \( \sup_{\delta \in \Delta} \mathbb{G}(\delta) \), where \( \mathbb{G}(\delta) \) is the Gaussian process defined above. We take the approach that was proposed by Kennedy (2019), using the multiplier bootstrap (Giné and Zinn, 1984; van der Vaart and Wellner, 1996; Chernozhukov et al., 2013; Belloni et al., 2018). We omit the relevant proofs as they are identical to those presented by Kennedy (2019). In comparison with the non-parametric bootstrap, the multiplier bootstrap has the computational advantage that the nuisance estimators \( \hat{\eta} \) need not be re-estimated. In comparison with directly sampling \( \sup_{\delta \in \Delta} \mathbb{G}(\delta) \), the procedure proposed does not require the evaluation of potentially large covariance matrices; therefore, it is far more computationally efficient and convenient.

The multiplier bootstrap approximates the distribution of \( \sup_{\delta \in \Delta} \mathbb{G}(\delta) \) with the supremum of the process

\[ \mathbb{M}(\delta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\xi_i S_{\hat{\eta}(\delta),\delta}(O_i) - \hat{\beta}(\delta)}{\hat{\sigma}(\delta)}, \]

where randomness is introduced through sampling the multipliers \( (\xi_1, \ldots, \xi_n) \), although the process is conditional on the observed data \( O_1, \ldots, O_n \). The multiplier variables are independent and identically distributed with mean 0 and unit variance, and are drawn independently from the sample. Typical choices are Rademacher \( (\mathbb{P}(\xi = -1) = \mathbb{P}(\xi = 1) = 0.5) \) or Gaussian multipliers. Under the assumptions of theorem 4, plus uniform consistency of \( \hat{\sigma}(\delta) \), it can be shown that condition (16) holds for

\[ \hat{\rho}(t) = \mathbb{P}\left\{ \sup_{\delta \in \Delta} \left| \mathbb{M}(\delta) \right| \leq t \left| O_1, \ldots, O_n \right. \right\}. \]

As a consequence, computation of the critical value, \( p \)-values and confidence intervals requires only simulation of a large number of realizations of the multipliers over a fine grid over \( \Delta \).
5. Simulation study

We now turn to comparing the three estimators of the direct effect, previously considered in Section 4. In particular, we investigate the performance of the substitution, (11) and (13), reweighted, (12) and (14), and efficient, (15), estimators in the case of an incremental propensity score intervention on a binary intervention variable of interest. The estimators are evaluated on data simulated from the following data-generating mechanism:

\[
W_1 \sim \text{Bern}(0.50); \\
W_2 \sim \text{Bern}(0.65); \\
W_3 \sim \text{Bern}(0.35); \\
A \sim \text{Bern}\left(\frac{1}{4} \sum_{j=1}^{3} W_j + 0.1\right); \\
Z_1 \sim \text{Bern}\left\{1 - \exp\left(\frac{A + W_1}{A + W_1 + 0.5}\right)\right\}; \\
Z_2 \sim \text{Bern}\left\{\exp\left(\frac{A - 1 + W_2}{W_3 + 3}\right)\right\}; \\
Z_3 \sim \text{Bern}\left\{\exp\left(\frac{A - 1 + 2W_1 - 1}{2W_1 + 0.5}\right)^2\right\}; \\
Y = Z_1 + Z_2 - Z_3 + A - 0.1 \left(\sum_{j=1}^{3} W_j\right)^2 + \epsilon,
\]

where \(\text{Bern}(p)\) is the Bernoulli distribution with parameter \(p\), \(\expit(x) = \{1 + \exp(-x)\}^{-1}\) is the cumulative density function of the logistic distribution (as implemented in the \(\text{plogis}\) function in the R programming language) and \(\epsilon \sim N(0, 0.25)\). The data that are available on a single observational unit are denoted by the random variable \(O = (W_1, W_2, W_3, A, Z_1, Z_2, Z_3, Y)\), where, in any given simulation, we consider observing \(n\) independent and identically distributed copies of \(O\) for one of seven sample sizes \(n \in \{400, 900, 1600, 2500, 3600, 4900, 6400\}\).

Under the above data-generating mechanism, we seek to estimate the direct effect under an incremental propensity score intervention \(\delta = 0.5\), for which the true value of the natural direct effect is approximately 0.137. We approximated this effect by using the alternative representation of \(\theta(\delta)\) as

\[
\theta(\delta) = \mathbb{E}\left\{\int m(a, Z, W) g_{\delta}(a|W) d\nu(a)\right\},
\]

where the inner integral is approximated by Monte Carlo integration through a large sample \(a_1, \ldots, a_m\) of uniformly distributed numbers in the range of \(A\), and the outer expectation is approximated through the law of large numbers by drawing a large sample \((W_1, Z_1), \ldots, (W_k, Z_k)\) from the joint distribution of \((W, Z)\). Each of the estimators is evaluated by contrasting regimes in which the appropriate nuisance parameters are fitted via a well-specified non-parametric regression or misspecified by fitting an intercept model. The enumerated set of estimators and regimes is summarized in Table 1 and Fig. 3. To ensure a well-specified non-parametric regression for the nuisance parameters, we rely on the highly adaptive lasso estimator, which is a recently proposed non-parametric regression function with properties guaranteeing convergence of estimated nuisance components at the \(n^{1/4}\)-rates that are required by our theorems (Benkeser and van der Laan, 2016; van der Laan, 2017; van der Laan and Benkeser, 2018).
Table 1. Mean-squared errors, scaled by $n$, of the three key estimators of the direct effect under an incremental propensity score intervention $\delta = 0.5$, across 1000 Monte Carlo simulations for each of seven sample sizes†

| Estimator                     | Results for the following values of $n$: |
|-------------------------------|------------------------------------------|
|                               | 400 | 900 | 1600 | 2500 | 3600 | 4900 | 6400 |
| Substitution                  | 0.083| 0.086| 0.084| 0.077| 0.072| 0.074| 0.075|
| Reweighted (inverse probability weighted) | 0.105| 0.120| 0.111| 0.116| 0.107| 0.112| 0.109|
| Efficient                     | 0.092| 0.086| 0.071| 0.072| 0.068| 0.067| 0.065|
| Efficient (E mis.)            | 0.060| 0.060| 0.060| 0.059| 0.054| 0.059| 0.058|
| Efficient (M mis.)            | 0.165| 0.130| 0.110| 0.107| 0.099| 0.103| 0.097|
| Efficient (G mis.)            | 0.436| 0.829| 1.255| 1.912| 2.662| 3.543| 4.519|

†Substitution and reweighted estimators are computed by using the highly adaptive lasso for $g$, $m$ and $e$. ‘E mis.’ denotes that $e$ was inconsistently estimated via an intercept-only logistic regression model; ‘M mis.’ and ‘G mis.’ denote analogous estimators.

Fig. 3. Statistics for the three key estimators (and variations thereof) of the direct effect under an incremental propensity score intervention $\delta = 0.5$, across 1000 Monte Carlo simulations for each of seven sample sizes

The specific value of $\delta$ is arbitrary; we chose this value arbitrarily because the objective of this application is purely to illustrate the use of the methods. In practice, this value would need to be agreed on with subject matter experts who can assess the importance and plausibility of an intervention that would modify the odds of exposure.
Inspection of the mean-squared error, after scaling by $\sqrt{n}$, reveals that the substitution estimator and the efficient one-step estimator both display excellent, essentially equivalent performance when nuisance components are estimated by using the highly adaptive lasso. The one-step estimator has slightly better performance, which seems to be driven by a better bias–variance trade-off. In contrast with the substitution estimator, the efficient one-step estimator carries the advantage of being double robust, allowing misspecification of either the outcome regression (denoted ‘M’ in Table 1) or the mediator-inclusive propensity score (denoted ‘E’). The robustness of the efficient estimator to the misspecification of these nuisance components—and the lack of robustness to the mediator-exclusive propensity score (denoted ‘G’)—are demonstrated in the last three rows of Table 1. Fig. 3 visualizes the performance of the estimators, and their misspecified variants, in terms of both the mean-squared error (as presented in Table 1) and its individual components, the bias and standard error. This comparison of the estimators reveals that the correctly specified one-step efficient estimator displays excellent performance in terms of both bias and variance whereas its non-robust misspecified variant displays an asymptotic bias that grows with sample size. Interestingly, the one-step estimator with $e$ inconsistently estimated displayed better performance than the fully efficient version. This is possibly an idiosyncrasy of this simulation because of misspecification through an intercept-only model generates smaller variable weights. Altogether, these numerical investigations demonstrate the utility of the proposed estimators in settings where the non-parametric estimation of nuisance components is viable; moreover, in applied data analytic settings where this procedure may be of interest, the one-step efficient estimator is clearly preferable on account of its multiple robustness. All numerical studies of the estimators were performed by using the implementations that are available in the medshift software package (Hejazi and Díaz, 2019) for the R language and environment for statistical computing (R Core Team, 2019). All of the results of our numerical study may be reproduced by using R scripts made publicly available at https://github.com/nhejazi/pub_medshift_jrssb.

6. Application

We now turn to considering a scenario in which the decomposition that was proposed in expression (5) and the proposed efficient estimator (15) may be used to estimate direct and indirect effects. To proceed, we take as example a simple data set from an observational study of the relationship between body mass index (BMI) and children’s behaviour, distributed as part of the mma R package, available via the Comprehensive R Archive Network (https://CRAN.R-project.org/package=mma). The documentation of this data set describes it as a

‘database obtained from the Louisiana State University Health Sciences Center, New Orleans, by Dr. Richard Scribner [who] explored the relationship between BMI and kids behavior through a survey at children, teachers and parents in Grenada in 2014. This data set includes 691 observations and 15 variables.’

In particular, we consider a modified version of this data set with all missing values removed, as these are irrelevant to the demonstration of the methodology proposed. In data analytic practice, we advocate for the use of the proposed methodology in tandem with a correction for any missingness in the observed data, such as imputation or reweighting by inverse probability of censoring (Carpenter et al., 2006; Vansteelandt et al., 2010; Seaman et al., 2012).

To demonstrate the assessment of the direct and indirect effect with this observational data set, we consider the effect of participation in a sports team on the BMI of children, taking as mediators snacking behaviour, exercising and overweight status. All other individual level and family level covariates—which included sex, age, race, number of individuals in the child’s
immediate family, the method of transport to school, the number of cars owned by the family, number of hours of television watched per week, number of hours spent using a computer per week, number of hours spent playing on a cell phone per week and number of hours spent in sweating activities per week—were treated as baseline confounders. All these baseline variables are potential confounders of the three relationships of interest (treatment–mediator, treatment–outcome and mediator–outcome) and are thus included in the $W$-vector. We note that, because these variables are measured at baseline, the mediator–outcome confounders could not possibly be affected by treatment. However, unmeasured variables such as the post-exposure number of hours spent by using a computer may be problematic in this example, as they are potential confounders of the mediator–outcome that are affected by treatment. We shall proceed with the illustration of our method, warning the reader that this is an important caveat of the results that we present.

As the intervention variable is binary, we frame our proposal in terms of an incremental propensity score intervention (Kennedy, 2019), wherein the odds of participating in a sports team is increased by a factor of $\delta = 2$ for each individual. Such a stochastic exposure regime may be interpreted as the introduction of a school programme or policy that motivates children to opt into participating in a sports team, doubling the odds of such voluntary participation.

6.1. Estimation strategy
As noted in expression (5), the PIE admits a decomposition in terms of components that enable estimation of the direct and indirect effects. We compute each of the components of the direct and indirect effects by using appropriate estimators as follows:

(a) for $E\{Y(A, Z)\} = E(Y)$, the natural value of the outcome under no intervention, the empirical mean in the sample serves as an efficient estimator;
(b) for $E\{Y(A_\delta, Z)\} = \theta(\delta)$, the mean outcome under an intervention altering the exposure mechanism but not the mediation mechanism, a one-step efficient estimator, denoted $\hat{\theta}(\delta)$, is proposed as equation (15) and made available via the medshift R package (Hejazi and Diaz, 2019);
(c) for $E\{Y(A_\delta)\} = \psi(\delta)$, the mean outcome under an intervention altering both the exposure and the mediation mechanisms, a one-step efficient estimator, denoted $\hat{\psi}(\delta)$ in what follows, is easily estimable by using the npcausal R package (Kennedy, 2018).

In the construction of estimators for $\theta(\delta)$ and $\psi(\delta)$, data-adaptive non-parametric regression procedures are incorporated to enable the relevant nuisance parameters of each estimator to be computed in a flexible manner using various R packages. The npcausal package enables the estimator $\hat{\psi}(\delta)$ to be constructed by using the ranger algorithm (Wright and Ziegler, 2015), which is an efficient and fast implementation of random forests (Breiman, 2001). In constructing $\hat{\psi}(\delta)$, the medshift package provides facilities for estimating nuisance parameters data adaptively via the Super Learner algorithm (van der Laan et al., 2007) for constructing ensemble learners through cross-validation, using its implementation in the sl3 package (Coyle et al., 2019). In particular, the Super Learner procedure was used to create a weighted ensemble of algorithms from a library including extreme gradient boosting via the xgboost package (Chen and Guestrin, 2016), with variants including 50, 100 and 300 boosting iterations, variants of random forests using 50, 100 and 500 trees, the $L_1$-penalized lasso and $L_2$-penalized ridge generalized linear models via the glmnet package (Friedman et al., 2010), an elastic net generalized linear model with equally weighted $L_1$- and $L_2$-penalization terms (also via glmnet), a main terms generalized linear model, an intercept model and the highly adaptive lasso (Benkeser
Table 2. Point estimates and 95% confidence intervals for the direct effect and indirect effect for an incremental propensity score intervention of $\delta = 2$ applied to the data set from the mma R package

| Parameter | Lower 95% confidence interval | Estimate | Upper 95% confidence interval |
|-----------|-------------------------------|----------|-------------------------------|
| Direct effect | -0.458                       | 0.011    | 0.479                         |
| Indirect effect | -0.672                       | -0.157   | 0.357                         |

and van der Laan, 2016), with fivefold cross-validation and up to either three-way or five-way interaction terms, using the hal9001 package (Coyle and Hejazi, 2018).

6.2. Estimating the direct and indirect effects
From the decomposition given in equation (5), the direct effect may be denoted $\beta(\delta) = \theta(\delta) - \mathbb{E}(Y)$. An estimator of the direct effect $\hat{\beta}(\delta)$ may be expressed as a composition of estimators of its constituent parameters:

$$\hat{\beta}(\delta) = \hat{\theta}(\delta) - \frac{1}{n} \sum_{i=1}^{n} Y_i.$$ 

Using the estimation strategies that were previously outlined, we may construct an estimate of the direct effect through a straightforward application of the delta method, yielding both a point estimate and associated standard errors under our proposed stochastic intervention policy. Similarly, the indirect effect $\psi(\delta) - \theta(\delta)$ may be estimated as $\hat{\psi}(\delta) - \hat{\theta}(\delta)$. We provide both point estimates and associated inference under our proposed stochastic intervention policy in Table 2.

From the estimates in Table 2, the conclusion may be easily drawn that there is little total effect of doubling the odds of participation in a sports team on the BMI of children, based on the data collected in the observational study made available in the mma R package. For reference, the marginal odds of participating in a sports team in the observed data are 0.69, whereas the odds under the intervention considered are 1.38. On the basis of the 95% confidence intervals around the point estimates, we cannot conclude that the proposed incremental propensity score intervention is sufficiently efficacious to decrease children’s BMI. However, the magnitude of the effects seem to be in the correct direction, with increased participation in a sports team causing a reduction in BMI of 0.157 through changes in behaviours such as snacking and exercise. Using an approach similar to that demonstrated with this data set, the direct and indirect effects that are attributable to interventions with higher odds of participating in a sports team are easily estimable. Results of this data analytic investigation may be reproduced by using the R code made publicly available at https://github.com/nhejazi/pub_medshift_jrssb.

7. Discussion
We have proposed a novel mediation analysis based on the decomposition of the causal effect of a stochastic intervention on the population, focusing on two types of intervention: modified treatment policies and exponential tilting. Unlike the natural direct effect of Pearl (2001), identification of the (in)direct effect that is proposed here does not require cross-world counter-
factual independences and is therefore achievable in an experimental setting randomizing both the exposure and the mediator.

We present results for stochastic interventions defined as a modified treatment policy and explicitly defined in terms of the post-intervention distribution (exponential tilting). In addition to the considerations about robustness and smoothness that are discussed in this paper, the choice between these two options may also be guided by the fact that modified treatment policies are more useful in practical settings as they can be used to inform feasible interventions.

Note that our effect decomposition and estimators allow for multivariate mediators. The interpretation of the (joint) indirect effect in this case is entirely context dependent. For example, the multivariate mediators may represent an innately multivariate construct (e.g. a psychological construct such as personality or behaviour); in this case the indirect effect could be interpreted as an effect through the construct (e.g. personality). Nonetheless, our approach does not require that the multivariate mediators are part of a single construct; the interpretation in these cases requires more care.

We assume that there is no mediator–outcome confounder affected by exposure. Point identification of natural (in)direct effects in the presence of such variables is not generally possible, and its partial identification is an area of active research (Robins and Richardson, 2010; Tchetgen Tchetgen and Phiri, 2014; Miles et al., 2015).

Lastly, for simplicity we focus on estimators constructed as solutions to the EIF estimating equation; moreover, we have made implementations of each of the proposed estimators available in the free and open source medshift software package (Hejazi and Diaz, 2019) for the R language and environment for statistical computing (R Core Team, 2019). Alternative estimation strategies, such as targeted minimum loss-based estimation (van der Laan and Rubin, 2006; van der Laan and Rose, 2011, 2018), may have better performance than our proposed estimators in finite samples. The development of such estimators will be the subject of future research and software development.

Acknowledgement

Iván Díaz gratefully acknowledges support by National Institute of Mental Health grant P50 MH113838 for this work.

References

Avin, C., Shpitser, I., and Pearl, J. (2005) Identifiability of path-specific effects. In Proc. 19th Int. Jt Conf. Artificial Intelligence, Edinburgh, July 30th–Aug. 5th, pp. 357–363.

Baron, R. M. and Kenny, D. A. (1986) The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Personlity Socl Psychol., 51, 1173–1182.

Begun, J. M., Hall, W. J., Huang, W.-M. and Wellner, J. A. (1983) Information and asymptotic efficiency in parametric-nonparametric models. Ann. Statist., 11, 432–452.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Wei, Y. (2018) Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. Ann. Statist., 46, no. 6B, 3643–3675.

Benkeser, D. and van der Laan, M. (2016) The highly adaptive lasso estimator. In Proc. Int. Conf. Data Science and Advanced Analytics, pp. 689–696. New York: Institute of Electrical and Electronics Engineers.

Bickel, P. J. (1982) On adaptive estimation. Ann. Statist., 10, 647–671.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1997) Efficient and Adaptive Estimation for Semiparametric Models. New York: Springer.

Breiman, L. (1996) Stacked regressions. Mach. Learn., 24, 49–64.

Breiman, L. (2001) Random forests. Mach. Learn., 45, 5–32.

Carpenter, J. R., Kenward, M. G. and Vansteelandt, S. (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. J. R. Statist. Soc. A, 169, 571–584.
Hejazi, N. S. and Díaz, I. (2019) medshift: causal mediation analysis for stochastic interventions. 

Haneuse, S. and Rotnitzky, A. (2013) Estimation of the effect of interventions that modify the received treatment.

Goldberger, A. S. (1972) Structural equation methods in the social sciences.

Cole, S. R. and Hernán, M. A. (2002) Fallibility in estimating direct effects.

Díaz, I. and van der Laan, M. J. (2018) Stochastic treatment regimes. In Díaz, I. and van der Laan M. J. (2013) Assessing the causal effect of policies: an example using stochastic interventions. Biometrics, 68, 541–549.

Díaz, I. and van der Laan M. J. (2013) Assessing the causal effect of policies: an example using stochastic interventions. Int. J. Biostatist., 9, 161–174.

Díaz, I. and van der Laan, M. J. (2018) Stochastic treatment regimes. In Targeted Learning in Data Science, pp. 219–232. New York: Springer.

Didelez, V., Dawid, A. P. and Geneletti, S. (2006) Direct and indirect effects of sequential treatments. In Proc. 22nd Conf. Uncertainty in Artificial Intelligence, pp. 138–146. Association for Uncertainty in Artificial Intelligence Press.

Dudík, M., Erhan, D., Langford, J. and Li, L. (2014) Doubly robust policy evaluation and optimization. Statist. Sci., 29, 485–511.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. J. Statist. Softw., 33, 1–22.

Giné, E. and Zinn, J. (1984) Some limit theorems for empirical processes. Ann. Probab., 12, 929–989.

Goldberger, A. S. (1972) Structural equation methods in the social sciences. Econometrica, 40, 979–1001.

Haneuse, S. and Rotnitzky, A. (2013) Estimation of the effect of interventions that modify the received treatment. Statist. Med., 32, 5260–5277.

Hejazi, N. S. and Díaz, I. (2019) medshift: causal mediation analysis for stochastic interventions. R Package Version 0.1.1. (Available from https://github.com/rhejazi/medshift.)

Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation analysis. Psychol. Meth., 15, 309–334.

Kennedy, E. H. (2018) npcausal: nonparametric causal inference methods. R Package Version 0.1.0. (Available from https://github.com/ehkennedy/npcausal.)

Kennedy, E. H. (2019) Nonparametric causal effects based on incremental propensity score interventions. J. Am. Statist. Ass., 114, 645–656.

van der Laan, M. J. (2017) A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. Int. J. Biostatist., 13, no. 2.

van der Laan, M. J. and Benkeser, D. (2018) Highly adaptive lasso (HAL). In Targeted Learning in Data Science (eds M. J. van der Laan and D. Benseker), pp. 77–94. New York: Springer.

van der Laan, M. J. and Petersen, M. L. (2008) Direct effect models. Int. J. Biostatist., 4, no. 1, article 23.

van der Laan, M. J., Polley, E. and Hubbard, A. (2007) Super learner. Statist. Appl. Genet. Molec. Biol., 6, article 25.

van der Laan, M. J. and Robins, J. M. (2003) Unified Methods for Censored Longitudinal Data and Causality. New York: Springer.

van der Laan, M. J. and Rose, S. (2011) Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer.

van der Laan, M. J. and Rose, S. (2018) Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies. New York: Springer.

van der Laan, M. J. and Rubin, D. (2006) Targeted maximum likelihood learning. Int. J. Biostatist., 2, no. 1.

Lok, J. J. (2016) Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. Statist. Med., 35, 4008–4020.

Lok, J. J. (2019) Causal organic direct and indirect effects: closer to Baron and Kenny. Preprint arXiv:1903.04697.

Miles, C. H., Kanki, P., Meloni, S. and Tchetgen Tchetgen, E. J. (2015) On partial identification of the pure direct effect. Preprint arXiv:1509.01652.

Newey, W. K. (1994) The asymptotic variance of semiparametric estimators. Econometrica, 62, 1349–1382.

Neyman, J. (1923) Sur les applications de la theóre des probabilités aux experiences agricaires: essay des principes. (Engl. transl. D. M. Dabrowska and T. P. Speed). Statist. Sci., 5 (1990), 463–472.

Nguyen, T. Q., Schmid, I. and Stuart, E. A. (2019) Clarifying causal mediation analysis for the applied researcher: defining effects based on what we want to learn. Preprint arXiv:1904.08515.

Pearl, J. (1995) Causal diagrams for empirical research. Biometrika, 82, 669–688.

Pearl, J. (1998) Graphs, causality, and structural equation models. Sociol. Meth. Res., 27, 226–284.
Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearl, J. (2001) Direct & indirect effects. In *Proc. 17th Conf. Uncertainty in Artificial Intelligence*, pp. 411–420. San Francisco: Morgan Kaufmann.

Pearl, J. (2009) Myth, confusion, and science in causal analysis. *Technical Report R-348*. Cognitive Systems Laboratory, Computer Science Department, University of California at Los Angeles, Los Angeles.

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

Pfanzagl, J. and Wefelmeyer, W. (1985) Contributions to a general asymptotic statistical theory. *Statist. Risk Modling*, 3, 379–388.

Popper, K. R. (1934) *The Logic of Scientific Discovery*. London: Hutchinson.

R Core Team (2019). *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Richardson, T. S. and Robins, J. M. (2013) Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Working Paper 128*. Center for Statistics and the Social Sciences, University of Washington, Seattle.

Robins, J. M. (1986) A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math. Modelling*, 7, 1393–1512.

Robins, J. M. and Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3, 143–155.

Robins, J. M., Hernán, M. A. and Siebert, U. (2004) Effects of multiple interventions. In *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, vol. 1, pp. 2191–2230.

Robins, J. M. and Richardson, T. S. (2010) Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*, pp. 103–158.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 688–701.

Rudolph, K. E., Sofrygin, O., Zheng, W. and van der Laan, M. J. (2017) Robust and flexible estimation of stochastic mediation effects: a proposed method and example in a randomized trial setting. *Epidem. Meth.*, 7, no. 1.

Seaman, S. R., White, I. R., Copas, A. J. and Li, L. (2012) Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68, 129–137.

Spirites, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. and Richardson, T. (2000) *Causation, Prediction, and Search*. Cambridge: MIT Press.

Stitelman, O. M., Hubbard, A. E. and Jewell, N. P. (2012) Combining multiple imputation and inverse-probability weighting. *Epidemiology*, 7, no. 1.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 688–701.

Rudolph, K. E., Sofrygin, O., Zheng, W. and van der Laan, M. J. (2017) Robust and flexible estimation of stochastic mediation effects: a proposed method and example in a randomized trial setting. *Epidem. Meth.*, 7, no. 1.

Seaman, S. R., White, I. R., Copas, A. J. and Li, L. (2012) Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68, 129–137.

Spirites, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. and Richardson, T. (2000) *Causation, Prediction, and Search*. Cambridge: MIT Press.

Stitelman, O. M., Hubbard, A. E. and Jewell, N. P. (2012) Combining multiple imputation and inverse-probability weighting. *Epidemiology*, 7, no. 1.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 688–701.

Rudolph, K. E., Sofrygin, O., Zheng, W. and van der Laan, M. J. (2017) Robust and flexible estimation of stochastic mediation effects: a proposed method and example in a randomized trial setting. *Epidem. Meth.*, 7, no. 1.

Seaman, S. R., White, I. R., Copas, A. J. and Li, L. (2012) Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68, 129–137.

Spirites, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. and Richardson, T. (2000) *Causation, Prediction, and Search*. Cambridge: MIT Press.

Stitelman, O. M., Hubbard, A. E. and Jewell, N. P. (2012) Combining multiple imputation and inverse-probability weighting. *Epidemiology*, 7, no. 1.
Young, J. G., Hernán, M. A. and Robins, J. M. (2014) Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidem. Meth.*, 3, 1–19.

Zheng, W. and van der Laan, M. J. (2011) Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pp. 459–474. New York: Springer.

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

Zheng, W. and van der Laan, M. (2017) Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *J. Causal Inf.*, 5, no. 2, 1–24.

**Supporting information**

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary materials for Causal mediation analysis for stochastic interventions’.