Statistics Surveys
Vol. 17 (2023) 1–41
ISSN: 1935-7516
https://doi.org/10.1214/22-SS140

Causal mediation analysis: From simple to more robust strategies for estimation of marginal natural (in)direct effects*

Trang Quynh Nguyen

Johns Hopkins Bloomberg School of Public Health
e-mail: trang.nguyen@jhu.edu

Elizabeth L. Ogburn

Johns Hopkins Bloomberg School of Public Health
e-mail: eogburn@jhu.edu

Ian Schmid

Johns Hopkins Bloomberg School of Public Health
e-mail: ian_schmid@jhu.edu

Elizabeth B. Sarker

Johns Hopkins Bloomberg School of Public Health
e-mail: esarker1@jhmi.edu

Noah Greifer

Harvard University Institute for Quantitative Social Science
e-mail: ngreifer@iq.harvard.edu

Ina M. Koning

Utrecht University
e-mail: i.koning@uu.nl

Elizabeth A. Stuart

Johns Hopkins Bloomberg School of Public Health
e-mail: estuart@jhu.edu

Abstract: This paper aims to provide practitioners of causal mediation analysis with a better understanding of estimation options. We take as inputs two familiar strategies (weighting and model-based prediction) and a simple way of combining them (weighted models), and show how a range of estimators can be generated, with different modeling requirements and robustness properties. The primary goal is to help build intuitive appreciation for robust estimation that is conducive to sound practice. We do this by visualizing the target estimand and the estimation strategies. A second goal is to provide a "menu" of estimators that practitioners can

arXiv: 2102.06048
*This work is supported by NIMH grants R01MH115487 and T32MH122357 (PI Stuart).

choose from for the estimation of marginal natural (in)direct effects. The estimators generated from this exercise include some that coincide or are similar to existing estimators and others that have not previously appeared in the literature. We note several different ways to estimate the weights for cross-world weighting based on three expressions of the weighting function, including one that is novel; and show how to check the resulting covariate and mediator balance. We use a random continuous weights bootstrap to obtain confidence intervals, and also derive general asymptotic variance formulas for the estimators. The estimators are illustrated using data from an adolescent alcohol use prevention study. R-code is provided.

MSC2020 subject classifications: 62D20.
Keywords and phrases: Causal mediation analysis, robust estimation, method visualization, natural (in)direct effects.

Received March 2022.

Contents

1. Introduction

Causal mediation methodology is complex. There are different types of causal contrasts: controlled direct effect, natural (in)direct effects [34, 30], interventional (in)direct [2, 52] and other interventional effects [28], etc., each with their own set of identification assumptions [27]. The literature on effect estimation is vast, with a wide variety of estimation methods based on regression [e.g., 50, 46, 25], weighting [e.g., 7, 8, 42, 12, 51], simulation [e.g., 16, 15, 53, 51], or some combination of these strategies. Further complicating the picture, some methods estimate marginal effects [15, 7, 8, 23, 51, 43] while others estimate effects conditional on covariates [50, 46, 38, 44, 42]. Most of these methods are parametric and require all the models used to be correctly specified. Some methods have built in robustness to model misspecification; these are often presented in highly technical papers [e.g., 43, 60]. It can be difficult for researchers to find their way through this literature and identify the estimation approach most appropriate for their application.

To help ease this task, this paper explicates a range of estimation options for causal mediation, focusing on options with some robustness properties. Rather than reviewing the complex and constantly growing methodological literature [see e.g., 13], we take a concrete approach of using as inputs two strategies familiar to practitioners (weighting and regression) and a simple way of combining them, and show how to generate a range of estimators with different modeling requirements and robustness properties. The primary goal is to build intuitive appreciation for robust estimation that is conducive to sound practice (without requiring prior understanding of these methods). This will benefit from the useful notion of *pseudo samples*, as each weighting procedure can be interpreted as creating a certain meaningful pseudo sample. A secondary goal is to provide a "menu" of estimators that practitioners can choose from (depending on which modeling components they are comfortable with given the specific application).

The paper focuses on *natural (in)direct effects*. These decompose the total causal effect and (when identified) provide insight about effect mechanisms. This is a direct match to researchers' common motivation for conducting mediation analysis – a wish to understand what part of a causal effect is indirect (operating

through a specific intermediate variable) and what part is direct (not through that variable). The kind of reasoning used to build estimators here is not specific to these effects, but can also be applied to other effect types in causal mediation analysis, which we will comment on at the end of the paper. (For readers who require an orientation to different effect types, we refer to [28, 27] which discuss *interventional* and natural effects, their relevance in practice, and their identification; and to [35] which proposes *separable effects*.)

We consider *marginal* natural (in)direct effects. These effects, when defined on the additive scale, correspond to the total effect being the average treatment effect – a popular effect in causal inference. Adaptation to average effects on the treated or on the controls is trivial. This paper does not address the estimation of conditional effects as functions of covariates, which entails a different set of estimation strategies that should be tackled separately.

As our construction of estimators is a bottom-up exercise, not all the estimators generated have appeared in the literature. We connect to work that employs, or is related to, the strategies and estimators discussed in this paper, and comment on the differences (some quite subtle) between some of these estimators. In addition to giving credit where credit is due, this aims to help the reader be a more informed consumer of the related literature.

To make the paper accessible to a broad audience, all proofs (about robustness properties, large-sample variance, and weight formulas) are placed in the Technical Appendix (Supplement 1). To facilitate application, R-code to implement the estimators is provided in the R-package `mediationClarity` (available at https://github.com/trangnguyen74/mediationClarity).

2. Preliminaries

2.1. Effect definitions

Consider the setting with a binary exposure A, followed in time by a mediator variable M (which may be multivariate), followed in time by an outcome Y. We define effects using the potential outcome framework [37, 6]. The target estimands in this paper are *marginal* natural (in)direct effects, which decompose the *marginal* total effect.

On the additive scale, the marginal total effect is formally $\text{TE} := \text{E}[Y_1] - \text{E}[Y_0]$, the difference between the population mean of Y_1 (potential outcome if exposed to the active treatment) and that of Y_0 (potential outcome if exposed to the comparison condition). Definition of *natural (in)direct effects* [30] additionally employs a nested potential outcome type, $Y_{aM_{a'}}$ (for a hypothetical condition with exposure set to a and mediator set to its potential value under condition a') where a and a' can be either 0 or 1. We assume that $Y_a = Y_{aM_a}$, thus $\text{TE} = \text{E}[Y_{1M_1}] - \text{E}[Y_{0M_0}]$. Using a third potential outcome with mismatched a and a', either Y_{1M_0} (exposure set to the active treatment but mediator set to its potential value under control) or Y_{0M_1} (the other way around), TE is decomposed in two ways, giving rise to two pairs of natural (in)direct effects:

$$\mathrm{TE} = \underbrace{\mathrm{E}[Y_{1M_1}] - \mathrm{E}[Y_{1M_0}]}_{\mathrm{NIE}_1} + \underbrace{\mathrm{E}[Y_{1M_0}] - \mathrm{E}[Y_{0M_0}]}_{\mathrm{NDE}_0},$$
$$\mathrm{TE} = \underbrace{\mathrm{E}[Y_{1M_1}] - \mathrm{E}[Y_{0M_1}]}_{\mathrm{NDE}_1} + \underbrace{\mathrm{E}[Y_{0M_1}] - \mathrm{E}[Y_{0M_0}]}_{\mathrm{NIE}_0}.$$

On multiplicative scales, marginal effects are ratios of marginal means or marginal odds of potential outcomes. For example, the marginal total effect is $\mathrm{E}[Y_1]/\mathrm{E}[Y_0]$ on the mean/risk ratio scale and $\frac{\mathrm{E}[Y_1]/(1-\mathrm{E}[Y_1])}{\mathrm{E}[Y_0]/(1-\mathrm{E}[Y_0])}$ on the odds ratio scale; other effects are defined accordingly. On both scales, decomposition is by product instead of sum, $\mathrm{TE} = \mathrm{NDE}_0 \times \mathrm{NIE}_1$ and $\mathrm{TE} = \mathrm{NIE}_0 \times \mathrm{NDE}_1$.

Marginal effects on the additive scale are also average effects. The marginal additive TE is equal to the mean of the causal effect on the individual, $Y_1 - Y_0$, and thus is usually known as the *average treatment effect* in the non-mediation literature. Marginal additive natural (in)direct effects can also be seen as averages of effects on individuals. This interpretation does not apply to effects defined on multiplicative scales.

Note that each TE decomposition mentioned here includes only one indirect effect. In a situation where M is a set of more than one mediator (as in our data example), this is the effect mediated by all the mediators combined. Alternatively, one may be interested in path-specific effects involving different mediators or subsets of mediators; that problem is outside the scope of this paper.

For conciseness, the rest of the paper addresses one of the two effect pairs: the NDE_0 and NIE_1 (also called the *pure direct effect* and *total indirect effect* [34]). The other effect pair mirrors this one in all content covered here.

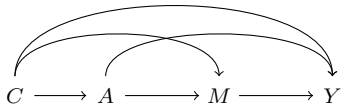
2.2. Assumptions for effect identification

As the current focus is estimation, we simply assume that the effects of interest are identified, noting that this is a matter for careful judgment in applications. By "identified" we mean that the effects, which are functions of *potential* outcomes, can be equated (under certain assumptions) to some functions of the *observed* data distribution. It is the latter that we will attempt to estimate. Below are the assumptions we make for (NDE_0, NIE_1) identification; for more detailed explication, see [49], [16], [31] or our companion paper [27].

Consistency The first assumption is that there is *consistency* between observed and potential outcomes or mediator values, and between potential outcomes of several types. Specifically,

$$\begin{aligned}
&Y = Y_a \text{ if } A = a,\\
&Y = Y_{1m} \text{ if } A = 1, M = m,\\
&M = M_0 \text{ if } A = 0,\\
&Y_a = Y_{aM_a},\\
&Y_{1M_0} = Y_{1m} \text{ if } A = 0, M = m,
\end{aligned}$$

FIG 1. *Unconfoundedness holds when no mediator-outcome confounders are influenced by exposure and a set of observed pre-exposure covariates C captures all confounding.*



for a being either 0 or 1, and m being any mediator value. Essentially we have invoked this assumption in defining the effects above.

Unconfoundedness The second assumption may be called *ignorability*, *exchangeability* or *unconfoundedness*. This assumption requires that there is a set of observed pre-exposure covariates C (where "pre-exposure" means either preceding exposure in time or simply not being influenced by exposure) that provides several conditional independence relationships. Specifically,

$$A \perp\!\!\!\perp Y_a, Y_{1m}, M_0 \mid C,$$
$$M \perp\!\!\!\perp Y_{1m} \mid C, A = 1,$$
$$M_0 \perp\!\!\!\perp Y_{1m} \mid C,$$

for a being either 0 or 1, and m being any value in the distribution of the mediator given covariates C in the unexposed. The first two of the three elements above fit with the usual notion of ignorability, where once we condition on some variables the *observed* exposure (or mediator value) does not carry any information about certain *potential* variables. The last element is different in that it involves two potential variables (M_0 and Y_{1m}) in two different worlds (thus commonly known as the cross-world independence assumption).[1]

In practice the usual way to deal with the unconfoundedness assumption is to ask (i) whether there are any mediator-outcome confounders (observed or not) that is influenced by exposure (these are often known as post-treatment confounders);[2] and if not, (ii) whether there is a set of pre-exposure covariates C (all of which observed) that captures all exposure-mediator, exposure-outcome and mediator-outcome confounders. If either the answer to (i) is yes or the answer to (ii) is no, then the unconfoundedness assumption does not hold. Note though that while we can use substantive knowledge to judge the plausibility of these assumptions, these assumptions are not testable using data.

[1]This assumption is needed to identify natural effects (which are defined based on a hypothetical situation where exposure is set to *one condition* but mediator is set to the value under *the other condition*) but is not needed to identify interventional effects – see [27].

[2]A post-treatment confounder L results in violation of the cross-world independence assumption $M_0 \perp\!\!\!\perp Y_{1m} \mid C$, due to a backdoor path connecting M_0 and Y_{1m} that is not blocked by C. This path is $M_0 \leftarrow L_0 \leftarrow U_L \rightarrow L_1 \rightarrow Y_{1m}$, where L_0, L_1 are potential values of L under exposure and nonexposure, U_L represents the unique causes of L that are not shared with A, M, Y but are shared by L_0, L_1. For more thorough treatments of the no post-treatment confounder (or cross-world independence) assumption, see [27, 49, 16, 31].

Positivity Since identification involves conditioning on covariates C, what is also required is that for all covariate levels there are positive chances of observing relevant potential mediator/outcomes. This is the third assumption, termed *positivity*. Specifically,

$$\begin{aligned}&\mathrm{P}(A = a \mid C) > 0,\\ &\mathrm{P}(M = m \mid C, A = 1) > 0,\end{aligned}$$

for a being either 0 or 1, and m being any value in the distribution of the mediator given covariates C in the unexposed. The first element implies positive chances of observing Y_1, Y_0, M_0; both combined imply positive chances of observing Y_{1m}.

In more practical terms, the positivity assumption means that (i) the covariate range is the same in both the exposed and unexposed groups; and (ii) within each subpopulation homogeneous in covariates C, the range of M in the exposed group covers the range of M in the unexposed group.

Two quick notes before we proceed. First, the unconfoundedness (and accompanying positivity) assumptions above with a single covariate set C are a simple version. Alternatively, different (yet overlapping) covariate sets could be used to deconfound the exposure-mediator, exposure-outcome and mediator-outcome relationships – see details in [27]. In that case, the estimation methods discussed here need to be adapted, which is straightforward but involves complicated expressions, and thus is not included to keep the paper manageable. Second, the assumptions above point identify the effects of interest (described shortly). There are cases where one may believe or be concerned that an assumption does not hold. For example, the no unobserved mediator-outcome confounding assumption and the cross-world independence assumption are often questioned. In these cases, one strategy is to seek bounds for the effects based on the assumptions one is willing to make (e.g., [24]), another is to conduct sensitivity analyses on the assumption that is likely violated (e.g., [9, 10, 32, 11, 16, 43]). We will return to this point at the end of the paper.

2.3. A heuristic view of identification that clarifies the estimation task

Identification of the $(\mathrm{NDE}_0, \mathrm{NIE}_1)$ pair amounts to identifying the means of the three potential outcomes Y_1, Y_0 and Y_{1M_0}. Under the assumptions above, the identification results [30, 49, 16, 31] of these three means are

$$\begin{aligned}&\mathrm{E}[Y_1] = \mathrm{E}_C\{\mathrm{E}[Y \mid C, A = 1]\},\\ &\mathrm{E}[Y_0] = \mathrm{E}_C\{\mathrm{E}[Y \mid C, A = 0]\},\\ &\mathrm{E}[Y_{1M_0}] = \mathrm{E}_C(\mathrm{E}_{M|C,A=0}\{\mathrm{E}[Y \mid C, M, A = 1]\}),\end{aligned}$$

where the right-hand sides are functions of the *observed* data distribution.

To make these results more intuitive to readers who find them unfamiliar, we offer a heuristic visualization in Figure 2. This figure has three columns. The

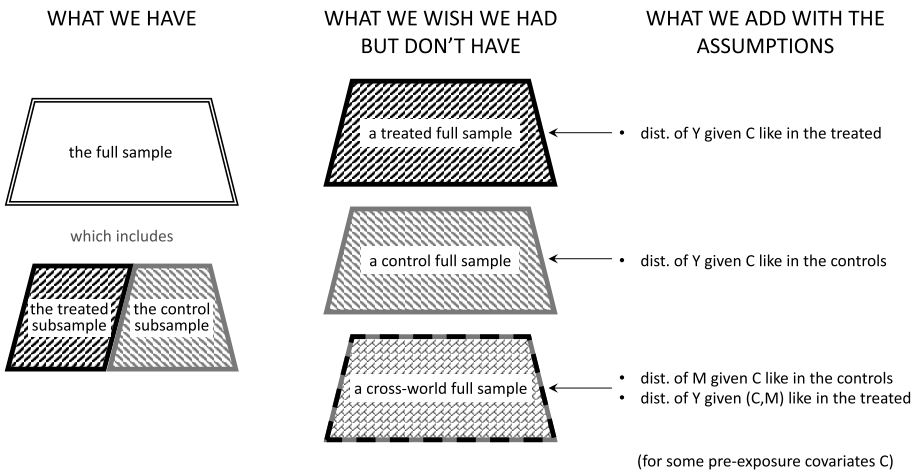


FIG 2. *Heuristic visualization of the identification result*

left column shows the data that we have: the full sample, which is comprised of the treated subsample and the control subsample. As we are interested in the NDE_0 and NIE_1 that together contrast the means of Y_1, Y_0, Y_{1M_0}, what we would ideally like to have instead is shown in the middle column: three full samples that all resemble the actual full sample pre-exposure, but are then set to three conditions: the treated (1) condition, the control (0) condition, and the cross-world condition (where exposure is set to 1 but mediator is set to M_0); this would allow us to average the outcome in the three samples to estimate the three potential outcome means. Unfortunately, we do not observe these three full samples. To remedy the situation, we invoke the assumptions above, which give us the additional information in the right column: in such a treated (control) full sample, the outcome distribution given C would be the same as that in the observed treated (control) units; and in such a cross-world full sample, the mediator distribution given C would be the same as that in the observed control units, while the outcome distribution given (C, M) would be the same as that in the observed treated units.

This sheds light on the estimation puzzle we need to solve. If we take the obvious approach of estimating the three potential outcome means, the task of estimating $\text{E}[Y_1]$ (or $\text{E}[Y_0]$) would be a puzzle of obtaining the outcome mean for a hypothetical full sample with the distribution of C from the actual full sample and the outcome distribution given C from the actual treated (control) units. The task of estimating $\text{E}[Y_{1M_0}]$ would be another puzzle of obtaining the outcome mean for a hypothetical sample with the distribution of C from the actual full sample, the mediator distribution given C from the control units, and the outcome distribution given (C, M) from the treated units. This is what is conveyed in the identification results stated above.

2.4. Preview of approaches and strategies for effect estimation

This paper makes use of two main tools: weighting and model-based prediction. We will first (in Section 3) consider weighting, which can be used as the sole puzzle-solving tool. Here we focus on conceptual clarity of the more complicated cross-world weighting component (including a new simple view based on a novel expression of the weight), methods for weight estimation, and balance checking. Then (in Section 4) we bring in the model-based prediction tool and examine pairs of estimators of potential outcome means, where each pair includes a nonrobust estimator (requiring all models to be correctly specified) and a more robust estimator (allowing some model misspecification). Here the robustness is due to strategic incorporation of weighting. Section 5 addresses the specific case of effects defined on the additive scale where there is an alternative view of the puzzle, obtaining a pair of estimators of the natural direct effect. We do our best throughout to reference work that employs or is related to the strategies mentioned. Section 6 discusses considerations in choosing an estimator.

With respect to interval estimation (see Section 8), we use a bootstrap procedure to obtain confidence intervals that applies to all the estimators discussed. We also derive general formulas for the asymptotic variance of the estimators.[3]

2.5. Illustrative example

We illustrate the estimators using a synthetic dataset generated to mimic data from The Prevention of Alcohol Use in Students (PAS) trial in the Netherlands. In the real trial, middle schools were randomized to one of four conditions: student intervention (promoting healthy attitudes and strengthening refusal skills), parent intervention (encouraging parental rule setting), student and parent combined intervention, and control condition (regular biology curriculum covering effects of alcohol). The combined intervention was effective in reducing drinking onset [22, 20] and drinking frequency [22], and [21] found that student attitudes towards alcohol, perceived self-control in situations involving alcohol, and student-reported parental rules about alcohol mediated the effect of the combined intervention on onset of weekly drinking. Our analysis of synthetic data considers the effect of the combined intervention relative to control on weekly drinking at 22 months, with the same mediators measured at six months.

The PAS sample consists of students clustered in schools, and has missing data on covariates, mediators and outcome. As our purpose is to illustrate a range of estimators, not to draw inference on the trial, we ignore the clustering for simplicity, complete the dataset with a single imputation, and use it as the basis to create a synthetic dataset. The imputation and synthesization used R-packages `mice` [47] and `synthpop` [29], and both are nonparametric (using classification and regression trees). All estimation outputs are specific to the synthetic dataset, and should not be interpreted as results of the original study.

[3]This is based on parametric specification of components of the model that need to be estimated, e.g., $\mathrm{P}(A \mid C)$, $\mathrm{P}(A \mid C, M)$, $\mathrm{E}[Y \mid C, M, A = 1]$, etc. depending on the estimator.

3. Weighting to create pseudo samples

Let us first examine one of the two tools we set out to use, that of weighting. This tool can be used by itself to estimate the effects of interest: we would weight data to create pseudo samples that stand in for the ideal treated, control and cross-world full samples we wish we had (see Figure 2), average the outcome in those pseudo samples to estimate the potential outcome means, and then contrast those means to estimate the effects. Such an estimator is consistent if the weights are consistently estimated. It tends to have large variance, and may suffer from high influence of observations with large weights. An important value of weighting, though, is that it can also be used in combination with regression-based techniques (as we shall see in Section 4) for more precise and robust estimation. It is therefore important to clarify how the weighting is done.

3.1. The pseudo treated and control samples

These pseudo samples are obtained by weighting treated units and control units to mimic the full sample covariate distribution, using the well-known inverse probability weights, $\omega_1(C) = \frac{1}{\mathrm{P}(A=1|C)}$ for treated units and $\omega_0(C) = \frac{1}{\mathrm{P}(A=0|C)}$ for control units. These weights are commonly estimated via propensity score [36] modeling. With such indirect estimation, it is common practice to check covariate balance and possibly adjust the model to achieve good balance. We will use probability models to estimate weights, and it is most familiar.

An alternative approach is to estimate the weights directly, finding weights that reduce the difference between the full and pseudo samples' covariate distributions. For example, several methods (e.g., entropy balancing [5] and covariate-balancing propensity score [17]) directly target balance on covariate moments specified by the user, and another method [14] minimizes a measure of distance between multivariate distributions called *energy distance* [41].

3.2. The pseudo cross-world sample

This pseudo sample is obtained by weighting treated units to mimic the C distribution in the full sample and the M given C distribution in control units. It stands in for the hypothetical full cross-world sample that we wish we had: in addition to these two elements, it retains its original Y given (C, M) distribution (which is that of treated units). Denote the weights that form the pseudo cross-world sample out of treated units by $\omega_{\mathrm{x}}(C, M)$. These weights have several equivalent expressions that point to several ways they may be estimated.

3.2.1. Three expressions (and views) of the cross-world weights

The first expression of $\omega_{\mathrm{x}}(C, M)$ builds on the inverse probability weights $\omega_1(C)$, which weight treated units to the full sample with respect to the covariate

distribution, in a sense doing half of the job. Such weighting does not change the mediator given covariates distribution (which is the distribution of M_1 given C). To morph this distribution to mimic the M_0 given C distribution, we use density ratio weighting (or probability ratio if the mediator is discrete) with the weighting function $\frac{\mathrm{P}(M|C,A=0)}{\mathrm{P}(M|C,A=1)}$, where the numerator and denominator are the densities (or probabilities) of the observed mediator value M conditional on C and on $A = 0$ and $A = 1$, respectively. This weighting scheme was proposed by Hong (2020) [7] [see also 8]. Thus we have

$$\omega_{\mathrm{x}}(C, M) = \frac{1}{\mathrm{P}(A = 1 \mid C)} \frac{\mathrm{P}(M \mid C, A = 0)}{\mathrm{P}(M \mid C, A = 1)}. \tag{1}$$

A second expression is due to the fact that by Bayes' rule the ratio of mediator densities above is equal to the ratio of two odds of exposure, $\frac{\mathrm{P}(A=0|C,M)}{\mathrm{P}(A=1|C,M)} \Big/ \frac{\mathrm{P}(A=0|C)}{\mathrm{P}(A=1|C)}$ (noted by Zheng et al., 2012 [60]). The resulting expression,

$$\omega_{\mathrm{x}}(C, M) = \frac{\mathrm{P}(A = 0 \mid C, M)}{\mathrm{P}(A = 1 \mid C, M)} \frac{1}{\mathrm{P}(A = 0 \mid C)}, \tag{2}$$

(which appears in an identification result in [12]) is the product of two terms: an odds weight[4] and an inverse probability weight. This formula provides another interpretation of the weighting: it could be thought of as first morphing the treated subsample to mimic the joint distribution of (C, M) in the control subsample (this is what odds weighting does), and then morphing the C distribution (which now reflects the distribution under control) to mimic that in the full sample (this is what inverse probability weighting does).

In addition, we found a novel third expression (see derivation in the Appendix). This expression is best viewed in its version for stabilized weights. The $\omega_{\mathrm{x}}(C, M)$ weights in treated units have mean equal to $\frac{1}{\mathrm{P}(A=1)}$; stabilized weights are simply $\omega_{\mathrm{x}}(C, M)$ scaled down to mean 1 by multiplying with $\mathrm{P}(A = 1)$. The third expression is

$$\omega_{\mathrm{x}}^{\mathrm{stabilized}}(C, M) = \frac{\mathrm{P}(C, M \mid A = 0) \frac{\mathrm{P}(A=0)}{\mathrm{P}(A=0|C)}}{\mathrm{P}(C, M \mid A = 1)}, \tag{3}$$

which could be seen as the ratio of two densities of (C, M): the denominator is the density in the treated subsample, and the weighted density in the numerator turns out to be the density in the pseudo control sample. That is, the weighting morphs the treated subsample such that it mimics the joint (C, M) distribution in the pseudo control sample. This makes sense, as the pseudo control sample has the C distribution of the full sample and the M given C distribution of control units – two of the three features desired for the pseudo cross-world sample.

[4]Side note: This odds weight component also appears as part of the weight formula in Tchetgen Tchetgen et al.'s inverse odds ratio weighting method [42] for estimation of the conditional natural direct effect. Both there and here, the role of weighting by this odds is to construct a weighted outcome distribution that reflects the distribution of the cross-world potential outcome given covariates.

3.2.2. Estimation of the cross-world weights

The first two of the expressions for $\omega_{\mathrm{x}}(C, M)$ above can be used directly as formulas for estimation purposes. In addition to the propensity score model, we fit either two mediator density models, $\mathrm{P}(M \mid C, A = 0)$ and $\mathrm{P}(M \mid C, A = 1)$ (if using the first formula), or a model for exposure given covariates and mediators, $\mathrm{P}(A \mid C, M)$ (if using the second formula); and plug the estimated elements into the formula. Between these two methods, the first one has the appeal that the models are variationally independent,[5] but its disadvantage is that density estimation is generally a harder problem than mean estimation (and especially so for a non-binary or multivariate mediator). The second method requires fitting fewer models (only two) and they are conditional mean models.

With the third expression of $\omega_{\mathrm{x}}(C, M)$, rather than treating it as an estimation formula (which would require estimating conditional densities of (C, M)), we can use the insight it provides – to weight the treated subsample to mimic the (C, M) distribution in the pseudo control sample – and note that this can be achieved by odds weighting. This means stacking the treated subsample with the pseudo control sample, fitting a model for A given C, M to the stacked data, and computing $\omega_{\mathrm{x}}(C, M)$ as the model-predicted odds of being in the pseudo control sample rather than the treated subsample. (This is just another instance of the connection between density ratios and odds of group membership.) This method also requires only two conditional mean models. Figure 3 visualizes these three weights estimation methods.

For readers who wish to use direct weights estimation tools such as moments balancing or distance minimizing (rather than relying on probability models), a couple of notes. First, the third expression of $\omega_{\mathrm{x}}(C, M)$ provides a simple and elegant way to use such tools: seek weights that morph the treated subsample to mimic the pseudo control sample with respect to the joint (C, M) distribution.[6] Second, while the second expression of $\omega_{\mathrm{x}}(C, M)$ suggests that direct weights estimation can be used for two-step weighting (first mimicking the control subsample's (C, M) distribution, then mimicking the full sample's C distribution), we do not recommend this, as this zigzag weighting may result in unnecessary loss of samples and suboptimal weights. [7]

[5] This weights estimation method uses models for $\mathrm{P}(A \mid C)$, $\mathrm{P}(M \mid C, A = 1)$ and $\mathrm{P}(M \mid C, A = 0)$, which correspond to a factorization of the likelihood. These model components thus do not put constraints on one another.

[6] Direct weights estimation seeks to directly mimic a target distribution, thus requires data reflecting that distribution. The pseudo control sample reflects the target (C, M) distribution.

[7] This weighting scheme is zigzag in the sense that the first step overshoots the target C distribution, as the full sample C distribution is *in between* those in the two subsamples. Therefore the first step may give very small weights to (or even drop) some observations (especially if the treated and control subsamples are dissimilar), which means those observations are essentially lost to the second step.

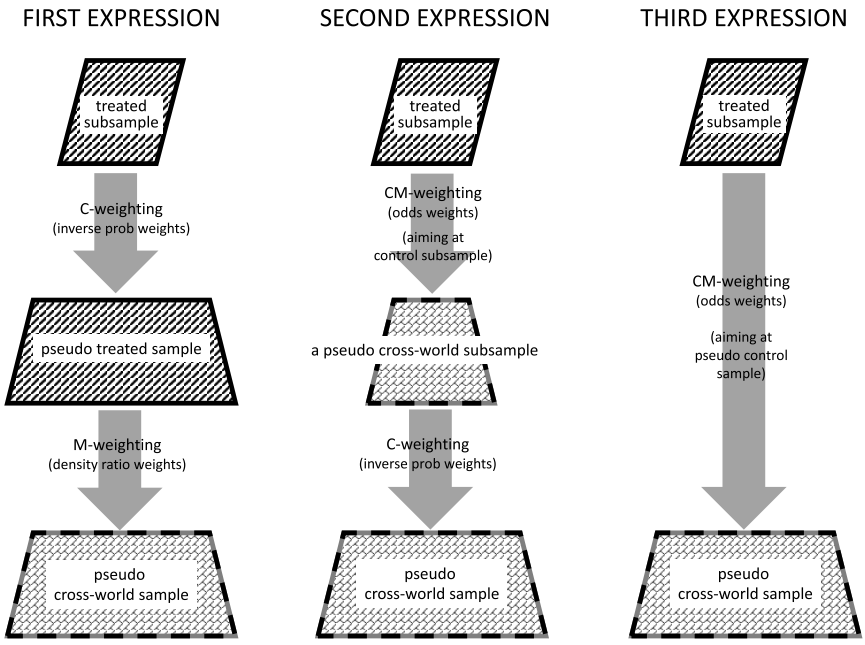


FIG 3. *Different views of cross-world weighting via alternative expressions of the weight function*

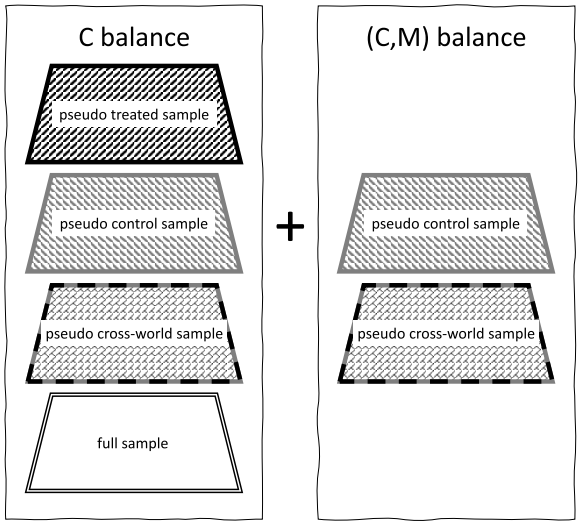


FIG 4. *Desired balance when using weighting to estimate the NDE0, NIE1 pair: covariate balance among all three pseudo samples and full sample, and covariate-and-mediator balance between pseudo cross-world and control samples.*

3.3. Balance checking

With the three pseudo samples, the desired balance includes two components (see Figure 4). The first component is *covariate balance* between the three pseudo samples and the full sample as well as among the three pseudo samples. The second component is the *covariate-and-mediator balance* between the pseudo cross-world sample and the pseudo control sample.

This full balance is important when using weighting as the pure estimation strategy, i.e., the effects are estimated by contrasting the outcome means from the pseudo samples. For some of the other estimators in the next section, certain elements of balance (which we will note) are crucial as they relate directly to the estimator's consistency, while other elements are in a sense of secondary importance as they serve mainly to induce robustness.

4. Estimating potential outcome means: pairs of nonrobust and more robust estimators

The weighting above gives us one solution to the puzzle described in Section 2.3. We can simply average the outcomes in the pseudo samples and contrast the averages to obtain estimates of the total and natural (in)direct effects. We call this the *pure weighting* estimator. This estimator is consistent only if the three weight functions are consistently estimated.

We now explore several other solutions to the puzzle, using our second tool, model-based prediction, either alone or in combination with weighting. These solutions are estimators of the means of Y_0 and Y_1 (which we refer to as *regular* potential outcomes) and of the *cross-world* potential outcome, which are to be combined to estimate marginal effects on either additive or multiplicative scale.

We present these potential outcome mean estimators in pairs. Each pair consists of a simple estimator that does the minimum needed to solve the puzzle, and a more complex estimator built on the simple one that is more robust as it provides some protection against model misspecification. Our explanations of robustness properties here strive for simple language; proofs for all estimators (here and in the next sections) are provided in the Technical Appendix.

As the paper touches on many estimators, a labeling system is needed. We use labels with two parts separated by "|", where the front part signals what is being estimated (e.g., "reg" and "crw" for regular and cross-world potential outcome means), and the back part signals the estimation method. Within each pair, the more robust estimator is distinguished from the nonrobust one by adding "MR" or "R" to the back label (the difference between these will be clear shortly). When referring to a pair, we use the base label (without MR or R).

4.1. Regular potential outcome means

We start with a single pair of estimators for the regular potential outcome means. While these may be broadly familiar, we will take time in motivating

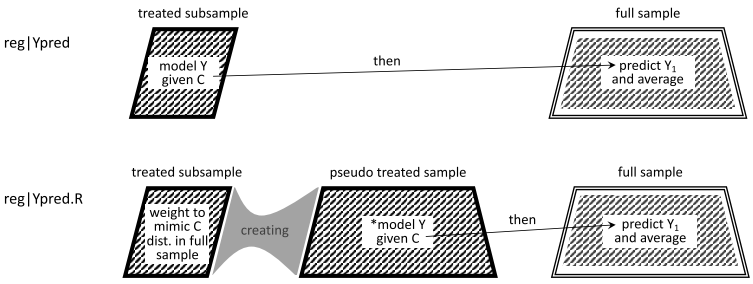


FIG 5. *A pair of regression-based estimators of regular potential outcome means, depicted as targeting* $\mathrm{E}[Y_1]$. * *indicates that the model is required to satisfy the mean recovery condition.*

them and clarifying ideas that will later apply in constructing estimators for the cross-world potential outcome mean.

4.1.1. reg/Ypred: outcome prediction given covariates

The simple version (reg/Ypred)

Recall from Figure 2 that if we were to observe a full treated/control sample, it would have the same outcome distribution given C as that in the actual treated/control units (under our assumption of unconfoundedness given C). This means we can learn this distribution from the treated/control subsample and apply it to the full sample. We thus fit a model regressing Y on C in the treated subsample, use this model to predict Y_1 for every individual in the full sample, and average the predicted values over the full sample to estimate $\mathrm{E}[Y_1]$. For $\mathrm{E}[Y_0]$, we fit the model to the control subsample and use it to predict Y_0.[8]

The two models here are models for $\mathrm{E}[Y|C, A = 1]$ and for $\mathrm{E}[Y|C, A = 0]$. Below we often refer to them collectively as $\mathrm{E}[Y|C, a]$ (keeping $a = 0, 1$ implicit); this is just an abbreviation as the key is that these are models that allow predicting two different variables, Y_1 and Y_0. Instead of fitting the models separately, we can also fit a joint model regressing Y on C, A. Separate model fitting has the advantage that it allows tailoring to the subsamples while avoiding the risk of (conscious or unconscious) fishing for a desired treatment effect estimate.

This simple estimator is nonrobust. If the outcome models are misspecified, predictions may be poor, leading to estimation bias. The problem may be exacerbated by extrapolation if the covariate distributions of the subsamples (to which the models are fit) differ substantially from that of the full sample (on which outcomes are predicted).

[8] A variant is to combine observed Y_1 and Y_0 values (in treated and control units, respectively) with predicted Y_1 and Y_0 values (for control and treated units, respectively).

The doubly robust version (reg|Ypred.R)

There is a class of estimators that are doubly robust. They combine outcome models and inverse probability weights, and are consistent if one of these two components (but not necessarily both) is correct. Many such estimators exist [see 19, 33]. We consider one estimator [33, 54] based on a strategy that readily extends to the later estimation tasks in the paper.

Like the simple estimator reg|Ypred, this robust estimator reg|Ypred.R relies on predicting Y_1 and Y_0 for the full sample and averaging predicted values (Figure 5). However, there is a key difference between the two estimators and a mild technical requirement imposed on the robust estimator. The key difference is that the outcome models used for prediction are fit to the *pseudo treated/control samples* instead of the subsamples, i.e., weighted regression models are used.[9] The technical requirement is that the outcome models satisfy a condition we label *mean recovery*: in the sample to which the model is fit (here a pseudo sample), the average model-predicted outcome equals the average observed outcome [33, equation 8]. Due to these two features combined, the estimator is doubly robust. We offer some intuition about these two points.

First, there is a simple rationale for fitting models to pseudo samples. Generally we do not know the true model that generated the data, so all models we use are just approximations of the true model. One way to improve the approximation (other than using flexible models to reduce misspecification) is to fit the model to the same covariate space on which it will be used for prediction; this is a guard against extrapolation. Compared to the treated/control subsamples, the pseudo samples have covariate distributions that are (at best) the same as or (at least) closer to that of the full sample. Fitting models to the pseudo samples is thus an improvement over the simple prediction estimator.

Second, the technical mean recovery condition serves to make sure that even if the predicted outcome values may be biased, they would be *on average unbiased* (if the weights that form the pseudo samples are correct). This condition is satisfied by generalized linear models with canonical link and an intercept (e.g., the usual linear regression, logistic regression, Poisson regression), which is the option we will use. Note that this is not the only choice. For example, an estimating equations approach may accommodate other link functions while satisfying this condition. Or if the outcome model is fit by machine learning, this condition may be achieved using targeted maximum likelihood estimation (TMLE) [48]. These topics are outside the scope of the current paper.

4.2. Cross-world potential outcome mean

Now we turn to the cross-world potential outcome. There are a range of strategies for estimating its mean. This is because the task requires combining several

[9] Related to the nonrobust variant in footnote 5, the corresponding robust variant here would predict Y_1 for control units based on an outcome model fit to a weighted treated subsample that mimics the control subsample (using odds weights $\frac{\mathrm{P}(A=0|C)}{\mathrm{P}(A=1|C)}$), and predict Y_0 for treated units based on an outcome model fit to a weighted control subsample that mimics the treated subsample (using odds weights $\frac{\mathrm{P}(A=1|C)}{\mathrm{P}(A=0|C)}$).

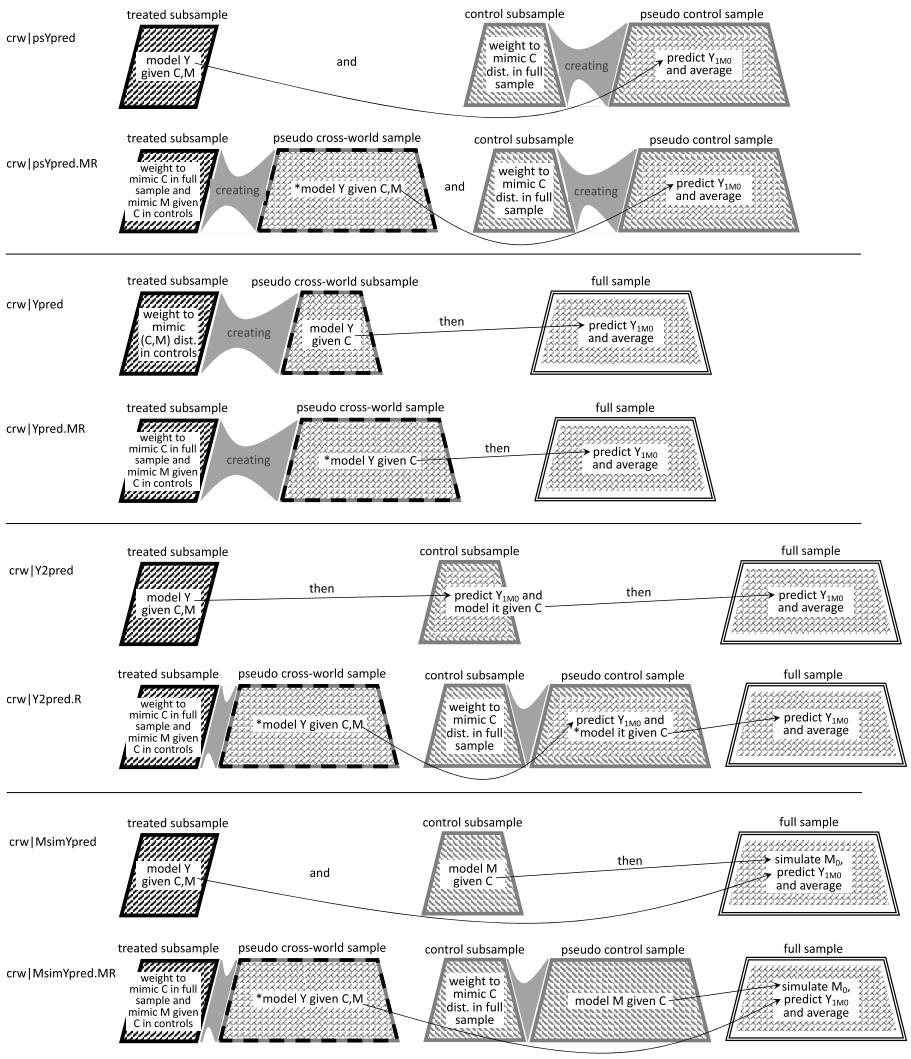


FIG 6. *Four pairs of estimators of the cross-world potential outcome mean* $\mathrm{E}[Y_{1M_0}]$. *Outcome models used for MR/R estimators (marked with *) are required to be mean-recovering.*

pieces of information from the full sample and from the treated and control conditions – recall that the cross-world sample in Figure 2 has the C distribution of the full sample, the M given C distribution of the controls, and the Y given C, M distribution of the treated – and there are different ways those pieces could be obtained and combined. We present four pairs of nonrobust and more robust estimators. The more robust estimators differ from their nonrobust counterparts in that they fit certain models to relevant pseudo samples instead of subsamples, and that outcome models used for them are required to satisfy

the mean recovery condition mentioned above. We explain the estimators' robustness properties and point out their nonrobustness if any. To aid explanation and provide a clear view, a visual representation of these four estimator pairs is provided in Figure 6. Also, Table 1 lists the steps for implementing them.

4.2.1. crw|psYpred: outcome prediction given covariates and mediators on pseudo control sample

These estimators are anchored on the pseudo control sample, where the mediator is M_0, and (if the weights are correct) the C distribution mimics that of the full sample. This gives us two of the three required pieces of information. The Y given C, M distribution, however, is off because it is that of control units. To complete the puzzle, we replace the observed outcome with predicted Y_{1M_0} given the units' (C, M) values, where the prediction is based on a model for $\mathrm{E}[Y \mid C, M, A = 1]$. We then average these predicted Y_{1M_0} values over the pseudo control sample to estimate $\mathrm{E}[Y_{1M_0}]$. (These are the only estimators in the paper that average predicted outcome *over a pseudo sample*, hence the 'ps' in the label.)

With the *nonrobust* estimator in this pair, the $\mathrm{E}[Y \mid C, M, A = 1]$ model is estimated by regressing Y on (C, M) in the treated subsample. For this estimator to be consistent, the control weights $\omega_0(C)$ have to be consistently estimated and this outcome model has to be correctly specified.

Relating to the literature, this nonrobust estimator constitutes part of VanderWeele and Vansteelandt's weighting-based estimator for the multiple mediator setting [51]. Albert [1] employs this strategy – using a model fit in one exposure condition to predict the cross-world outcome on units in the other condition and weighting those units to standardize the covariate distribution – but for a general target population (a generalization of our current purpose). Also, this strategy of predicting the cross-world outcome on a pseudo sample is one of the methods used in the R-package medflex [38] as a first step in estimating effects conditional on covariates.

The *more robust* estimator crw|psYpred.MR fits the outcome model (that satisfies the mean recovery condition) to the pseudo cross-world sample instead of the treated subsample. Like its nonrobust sibling, this estimator is not consistent if the $\omega_0(C)$ weights are not consistent. But if they are, then crw|psYpred.MR is consistent if either the outcome model is correctly specified or the cross-world weights $\omega_{\mathrm{x}}(C, M)$ are consistent. That is, crw|psYpred.MR has two chances to be correct, while crw|psYpred has only one. Here we use the MR (more robust) suffix (instead of simply R) to signal that although this estimator is more robust than its nonrobust sibling, it depends on one estimation component (here the control weights) being correct.

Because these two estimators average predicted outcome over the pseudo control sample, they depend on the covariate distribution of the pseudo sample mimicking the full sample well. Therefore, when using either of these estimators, it is crucial to obtain good covariate balance between the pseudo control sample and the full sample.

While these two estimators rely on outcome prediction, they are also weighted estimators (as predictions are averaged over weighted control units), and weighted estimators may have large variance due to the variability of the estimated weights. An alternative strategy is to find a way to predict Y_{1M_0} on the full sample instead of on the pseudo control sample. The next three pairs of estimators do this in different ways.

4.2.2. crw/Ypred: outcome prediction given covariates

In the full sample, we cannot predict Y_{1M_0} based on observed covariates and mediators (because the observed M is a mixture of both M_0 and M_1). Instead, this pair of estimators relies on Y_{1M_0} prediction based on covariates only. To do this, we need a model that informs of the mean of the cross-world potential outcome given covariates, $\mathrm{E}[Y_{1M_0} \mid C]$.

The trick employed by the simpler estimator in this pair is to weight the treated subsample to mimic the control subsample with respect to the joint distribution of (C, M) (using odds weights $o_{\mathrm{x}}(C, M) = \frac{\mathrm{P}(A=0|C,M)}{\mathrm{P}(A=1|C,M)}$, which can be estimated based on a model for exposure given covariates and mediator).[10] This weighted subsample (which we call a pseudo cross-world subsample) has two of the three desired features of the ideal cross-world full sample: the M given C distribution (like that in the controls) and the Y given (C, M) distribution (like that in the treated). That means the Y given C distribution is like that in the ideal cross-world full sample – the distribution of Y_{1M_0} given C. We thus fit to this pseudo subsample a model regressing outcome on covariates to estimate $\mathrm{E}[Y_{1M_0} \mid C]$. (To simplify language, we loosely call this model the $\mathrm{E}[Y_{1M_0} \mid C]$ model.) Based on this model, we predict Y_{1M_0} in the full sample and average the predicted values to estimate $\mathrm{E}[Y_{1M_0}]$. This estimator is *nonrobust*. For it to be consistent, the weights $o_{\mathrm{x}}(C, M)$ have to be consistently estimated and the outcome model has to be correctly specified.

The *more robust* estimator crw|Ypred.MR, on the other hand, fits an outcome given covariates model (that satisfies the mean recovery condition) to the pseudo cross-world sample (instead of the cross-world subsample above). This estimator has two chances to be correct: (1) if the weights $\omega_{\mathrm{x}}(C, M)$ are consistently estimated, crw|Ypred.MR is consistent even if the outcome model is misspecified; and (2) if the outcome model is correctly specified and only *the mediator-related part* of $\omega_{\mathrm{x}}(C, M)$ is correct, crw|Ypred.MR is consistent.

To clarify the second case, the mediator-related part of the weights is the term that controls the mediator distribution in the pseudo cross-world sample. It varies by the weights estimation method: in the first and second methods it is the $\frac{\mathrm{P}(M|C,A=0)}{\mathrm{P}(M|C,A=1)}$ and $\frac{\mathrm{P}(A=0|C,M)}{\mathrm{P}(A=1|C,M)}$ terms, respectively; in the third method it is the odds of being in the pseudo control sample rather than the treated subsample (where the pseudo control sample may be incorrectly weighted). When only

[10] An alternative is to mimic the M given C distribution using density ratio weights $\frac{\mathrm{P}(M|C,A=0)}{\mathrm{P}(M|C,A=1)}$; we do not recommend this because there is no simple way to check balance on conditional distributions.

the mediator-related part of the weights is correct, the weighting gets the C distribution wrong but gets the M given C distribution right, and it is the latter that ensures that the treated units' Y given C distribution is appropriately morphed to resemble the target Y_{1M_0} given C distribution. As the outcome regression model conditions on C, it is (if correctly specified) not affected by the incorrectly weighted C distribution.

Intuitively, both of these estimators rely completely on the weighting to obtain data that reflect the distribution of Y_{1M_0} given C (via getting the M given C distribution right). It is thus crucial to achieve good balance, specifically (C, M) balance between the pseudo cross-world subsample and the control subsample (for the nonrobust estimator) or between the pseudo cross-world sample and the pseudo control subsample[11] (for the more robust estimator).

4.2.3. crw|Y2pred: outcome prediction based on double model fit

These estimators also rely on predicting Y_{1M_0} in the full sample based on a model that estimates $\mathrm{E}[Y_{1M_0} \mid C]$. Here this model is estimated in two steps: first fitting a $\mathrm{E}[Y \mid C, M, A = 1]$ model and using it to predict Y_{1M_0} in control units, then regressing the predicted Y_{1M_0} on covariates to estimate $\mathrm{E}[Y_{1M_0} \mid C]$. We loosely refer to these two regression models with one building directly on the other as a *double outcome model fit* (or iterated regression).[12]

With the *nonrobust* estimator crw|Y2pred, these two models are fit to the treated and control subsamples, respectively. For this estimator to be consistent, both models have to be correctly specified.

The *robust* estimator crw|Y2pred.R fits the outcome models to the pseudo cross-world and pseudo control samples instead; and both models are required to satisfy the mean recovery condition. This estimator has three chances to be consistent: (1) both of the outcome models are correctly specified; (2) both the $\omega_{\mathrm{x}}(C, M)$ and $\omega_0(C)$ weights (that form pseudo cross-world and pseudo control samples) are consistently estimated; or (3) the $\omega_{\mathrm{x}}(C, M)$ weights are consistently estimated and the $\mathrm{E}[Y \mid C, A = 1, M]$ model is correctly specified. Notice that we call crw|Y2pred.R *robust* (rather than *more robust*) to signal that this estimator does not depend on any specific estimation component being correct.

A technical note: crw|Y2pred.R is a multi-step estimator based on the nonparametric influence function. As such, it has similar robustness properties to Tchetgen Tchetgen and Shpitser's estimator which solves the nonparametric influence function estimating equation [43]. The appeal of crw|Y2pred.R is that the steps are intuitively meaningful without requiring knowledge of influence function theory.

[11] Strictly speaking, the construction of the pseudo control subsample is not required to obtain a crw|Ypred estimate of $\mathrm{E}[Y_{1M_0}]$ (see Figure 6). However, this is likely not additional work because the pseudo control subsample is already constructed for estimating $\mathrm{E}[Y_0]$.

[12] This iterated regression procedure is a straightforward implementation, by regression, of the double expectation in the identification result.

TABLE 1

Implementation steps of estimators of potential outcome means in Sections 3, 4 and 5.1. nonR, MR and R stands for nonrobust, more robust and robust, respectively.

Purely weighting: nonR estimator
1. Estimate $\omega_1(C)$, $\omega_0(C)$ and $\omega_{\mathrm{x}}(C, M)$ weights
2. Average the observed outcome in the pseudo control, pseudo treated and pseudo cross-world samples

reg|Ypred pair: nonR and R estimators
1. For the R estimator, estimate $\omega_1(C)$ and $\omega_0(C)$ weights
2. Model Y given C in the treated and control subsamples (nonR) or pseudo treated and control samples* (R)
3. Based on these models, predict Y_1 and Y_0 given C in full sample
4. Average predicted Y_1 and Y_0 in full sample

crw|psYpred pair: nonR and MR estimators
1. Estimate $\omega_0(C)$ weights
2. For the MR estimator, also estimate $\omega_{\mathrm{x}}(C, M)$ weights
3. Model Y given C, M in the treated subsample (nonR) or pseudo cross-world sample* (MR)
4. Based on model, predict Y_{1M_0} given C, M in control units
5. Average predicted Y_{1M_0} in pseudo control sample

crw|Ypred pair: nonR and MR estimators
1. Estimate $o_{\mathrm{x}}(C, M)$ weights (nonR) or $\omega_{\mathrm{x}}(C, M)$ weights (MR)
2. Model Y given C in the pseudo cross-world subsample (nonR) or pseudo cross-world sample* (MR)
3. Based on model, predict Y_{1M_0} given C in full sample
4. Average predicted Y_{1M_0} in full sample

crw|MsimYpred pair: nonR and MR estimators
1. For the MR estimator, estimate $\omega_0(C)$ and $\omega_{\mathrm{x}}(C, M)$ weights
2. Model the density of M given C in the control subsample (nonR) or pseudo control sample (MR)
3. Model Y given C, M in the treated subsample (nonR) or pseudo cross-world sample* (MR)
4. Do many times in full sample:
    i. Based on first model, simulate M_0 given C
    ii. Based on second model, predict Y_{1M_0} given the combination of C and predicted M_0
5. Average all predicted Y_{1M_0} values in full sample

crw|Y2pred pair: nonR and R estimators
1. For the R estimator, estimate $\omega_0(C)$ and $\omega_{\mathrm{x}}(C, M)$ weights
2. Model Y given C, M in the treated subsample (nonR) or pseudo cross-world sample* (R)
3. Based on model, predict Y_{1M_0} given C, M in control units
4. Model predicted Y_{1M_0} given C in the control subsample (nonR) or pseudo control sample* (R)
5. Based on model, predict Y_{1M_0} given C in full sample
6. Average predicted Y_{1M_0} in full sample

NDE|YpredEpred pair: nonR and R estimators
1. For the R estimator, estimate $\omega_0(C)$ and $\omega_{\mathrm{x}}(C, M)$ weights
2. Model Y given C, M in the treated subsample (nonR) or pseudo cross-world sample* (R)
3. Based on model, predict Y_{1M_0} given C, M in control units, and compute proxy of the individual NDE_0 as predicted Y_{1M_0} minus Y
4. Model the NDE_0 proxy given C in the control subsample (nonR) or pseudo control sample* (R)
5. Based on model, predict NDE_0 given C in full sample
6. Average predicted NDE_0 in full sample

* This regression model for the MR/R estimator is required to satisfy the mean recovery condition.

4.2.4. crw|MsimYpred: mediator simulation and outcome prediction

These estimators involve fitting models for the conditional mediator density $\mathrm{P}(M \mid C, A = 0)$ and the conditional outcome mean $\mathrm{E}[Y \mid C, M, A = 1]$. Having learned these models, we put the observed mediator and outcome aside. For all units in the full sample, we simulate M_0 based on the first model, and with the simulated M_0 and observed C, predict Y_{1M_0} based on the second model. [13]

We do this multiple times, resulting in multiple sets of predicted Y_{1M_0} values, and average these predicted values to estimate $\mathrm{E}[Y_{1M_0}]$.[14]

With the *nonrobust* estimator in this pair, the mediator model is fit to the control subsample and the outcome model to the treated subsample. For this estimator to be consistent, both models have to be correctly specified. This mediator simulation strategy is used in Imai et al.'s natural (in)direct effects estimation method [15], implemented in the R package *mediation* [45]. This other estimator differs from crw|MsimYpred in that it uses this strategy for all potential outcome means (not only the cross-world one) therefore relies on more models. Also, the implementation in the package uses models for $\mathrm{P}(M \mid C, A)$ and $\mathrm{E}[Y \mid C, A, M]$ fit to the full sample rather than exposure-specific models.

The *more robust* crw|MsimYpred.MR instead fits the mediator model to the pseudo control sample and the outcome model to the pseudo cross-world sample, with the outcome model safisfying the mean recovery condition. Like its nonrobust sibling, this estimator is inconsistent if the mediator density is misspecified. Assuming correct specification of this model, this estimator has two chances to be correct: either the $\omega_{\mathrm{x}}(C, M)$ weights are consistently estimated or the outcome model is correctly specified.[15]

Relating to the existing literature, a specific version of crw|MsimYpred.MR where the cross-world weights are estimated based on mediator density models is an implementation of the estimator in section 5 in [43]. The estimator in [43] integrates the conditional outcome mean function $\mathrm{E}[Y \mid C, M, A = 1]$ over the conditional mediator density $\mathrm{P}(M \mid C, A = 0)$; the simulation-prediction-averaging procedure here is a numerical evaluation of that integral.

An interesting point: with both estimators being inconsistent if the mediator density model is misspecified, is anything gained by fitting that model to the pseudo control sample instead of the control subsample? Yes, what is gained is a partial correction in the sense that with a wrong model, the density fitted with correct weights is closer (in KL-divergence, see section B.4.2.4 of the Technical Appendix) to the true density than the density fitted without weights is.

These two estimators' dependence on correct specification of the mediator density model is an important drawback, as density estimation is a harder prob-

[13] A variant is to use the observed mediator in control units and only simulate M_0 for treated units.

[14] This is an implementation of the double expectation in the identification result where the outer expectation is evaluated by numerical integration.

[15] In the *more robust* version of the variant mentioned two foot notes ago, the mediator model is fit to a weighted control subsample that mimics the treated subsample's C distribution – using odds weights $\frac{\mathrm{P}(A=1|C)}{\mathrm{P}(A=0|C)}$.

lem than mean estimation, an issue raised in [1]. (Estimators that do not involve simulating M_0, or more generally, integrating over an estimated conditional density of M_0, avoid this problem.) For example, with a continuous variable, if the conditional mean is of interest, a common model choice is the linear model, which assumes a functional form for the mean but makes no other assumption. If the conditional density is of interest, one might still use the linear model but has to make additional distributional assumptions (e.g., the error is normally distributed or follows some other distribution) which are likely incorrect. In the special case with a single binary mediator, the distribution is fully described by the probability so the model reduces to a conditional mean model.

In a setting with a multivariate mediator (like in our data example), we need to model the joint distribution of the mediators given covariates in control units. To do this, we factor the joint into conditional densities/probabilities. Let $M = (M^a, M^b, M^c)$ where M^a, M^b, M^c are three mediators.

$$\begin{aligned}&\mathrm{P}(M \mid C, A=0) = \\ &\qquad \mathrm{P}(M^a \mid C, A=0)\,\mathrm{P}(M^b \mid C, A=0, M^a)\,\mathrm{P}(M^c \mid C, A=0, M^a, M^b).\end{aligned}$$

In the control subsample (or the pseudo control sample if using the more robust version), we fit three models for the three mediators. All three models condition on C, the second model conditions additionally on M^a, and the third model conditions additionally on both M^a and M^b. The order of variables in the factorization can be chosen for modeling convenience (see our data application for an example). Simulation follows the order of the fitted models.

As a mini recap, the four pairs of estimators of $\mathrm{E}[Y_{1M_0}]$ above represent different solutions to the puzzle of finding the outcome mean in a target condition where the C distribution is the same as that in the full sample, the M given C distribution is the same as that in the controls, and the Y given (C, M) distribution is the same as that in the treated. In each pair, the second estimator is more robust than the first as it does not require that all estimation components are correct. Among the four more robust estimators, crw|Y2pred.R is the most robust as it does not require any specific estimation component to be correctly specified/consistent; in contrast crw|psYpred.MR and crw|Ypred.MR are not robust to inconsistent weights, and crw|MsimYpred.MR is not robust to misspecification of the conditional mediator density model.

4.3. A weighting-centric view of the more robust estimators

The presentation of estimators in pairs above shows that each MR/R estimator is an improvement over a simpler regression-based estimator by incorporating weighting. Several of these estimators can also be seen as a direct improvement on the pure weighting estimator by incorporating regression-based prediction. This is easily seen from the visualization in Figures 5 and 6.

Consider reg|Ypred.R as an estimator of $\mathrm{E}[Y_1]$. As shown in Figure 5, it is a modification of the pure weighting estimator. The latter solves the puzzle by

obtaining the pseudo treated sample and stops there. reg|Ypred.R goes one step further: using regression-based prediction to correct for discrepancy in outcome mean due to the remaining difference between this pseudo treated sample and the *target covariate distribution* (contained in the full sample).

crw|Ypred.MR can also be seen as a direct improvement upon the pure weighting estimator of $\mathrm{E}[Y_{1M_0}]$. The pure weighting estimator solves the puzzle by obtaining the pseudo cross-world sample. crw|Ypred.MR takes an additional step to correct for the remaining difference between the pseudo cross-world sample and the *target covariate distribution* (but not the target conditional mediator distribution).

crw|psYpred.MR also starts with creating the pseudo cross-world sample like the pure weighting estimator. The additional regression-based prediction on the pseudo control sample adjusts for any difference between the pseudo cross-world sample and pseudo control sample. This effectively is a correction for the remaining difference between the pseudo cross-world sample and the *target conditional mediator distribution* (contained in control units).

Like the two previous estimators, crw|Y2pred.R also starts with creating the pseudo cross-world sample by weighting. Then it goes two additional steps to correct for the remaining differences from the *target conditional mediator distribution* and the *target covariate distribution*.

4.4. Combining reg| and crw| estimators to estimate the effects

The marginal natural (in)direct effects are estimated by contrasting the estimated means of the three potential outcomes, using the difference or ratio definition of choice. We combine each of the four nonrobust regression-based crw| estimators with the nonrobust reg|Ypred, and each of the more robust crw| estimators with reg|Ypred.R. We label the resulting effect estimators using simple labels that mostly reflect the crw| method, e.g., Y2pred.R is the combination of crw|Y2pred.R and reg|Ypred.R.

For two crw| strategies (crw|psYpred and crw|MsimYpred, both nonrobust and more robust versions), we also form a second combination with a modified reg| strategy. Note that crw|psYpred is anchored on the pseudo control sample. The first psYpred combination is with reg|Ypred, which is anchored on the full sample. In the second combination, however, the reg| part is also anchored on the pseudo control sample: $\mathrm{E}[Y_1]$ and $\mathrm{E}[Y_0]$ are estimated by averaging predicted Y_1 values and observed Y_0 values on the pseudo control sample. The other case is crw|MsimYpred. The first combination uses reg|Ypred for both $\mathrm{E}[Y_1]$ and $\mathrm{E}[Y_0]$; the second combination uses mediator simulation to estimate both $\mathrm{E}[Y_{1M_0}]$ and $\mathrm{E}[Y_0]$ (here seen as $\mathrm{E}[Y_{0M_0}]$), and uses reg|Ypred to estimate $\mathrm{E}[Y_1]$ only.

These estimators of natural (in)direct effects inherit the properties of the reg| and crw| estimators they combine. Table 2 summarizes the estimation components involved in, and the (non)robustness properties of, each of these effect estimators. It also covers the pure weighting estimator and a pair of estimators that will be considered in section 5.

TABLE 2. *Robustness and nonrobustness properties of estimators from Sections 3, 4 and 5*

Estimator label	**Estimator summary**	**Estimation components used to estimate $\mathrm{E}[Y_{1M_0}]$ (or NDE_0)**	**Estimation components used to estimate $\mathrm{E}[Y_1], \mathrm{E}[Y_0]$ (or TE)**	**Combination of components that need to be correct for the estimator to be consistent**	**Components not allowed to be inconsistent**
wtd	pure weighting	wts: $\omega_{\mathrm{x}}(C, M)$	wts: $\omega_1(C), \omega_0(C)$	all components correct	all
psYpred1	crw|psYpred, reg|Ypred	wts: $\omega_0(C)$ omod: $\mathrm{E}[Y \mid C, M, A = 1]$	omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	all components correct	all
psYpred2	Y_{1M_0}, Y_1|psYpred, Y_0|wtd		wts: $\omega_0(C)$ omod: $\mathrm{E}[Y \mid C, A = 1]$	all components correct	all
Ypred	crw|Ypred, reg|Ypred	wts: $o_{\mathrm{x}}(C, M)$ omod: $\mathrm{E}[Y_{1M_0} \mid C]$	omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	all components correct	all
MsimYpred1	crw|MsimYpred, reg|Ypred	mmod: $\mathrm{P}(M \mid C, A = 0)$ omod: $\mathrm{E}[Y \mid C, M, A = 1]$	omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	all components correct	all
MsimYpred2	Y_{1M_0}, Y_0|MsimYpred, Y_1|Ypred		mmod: $\mathrm{P}(M \mid C, A = 0)$ omod: $\mathrm{E}[Y \mid C, M, A = 0]$, $\mathrm{E}[Y \mid C, A = 1]$	all components correct	all
Y2pred	crw|Y2pred, reg|Ypred	omods: $\mathrm{E}[Y \mid C, M, A = 1], \mathrm{E}[Y_{1M_0} \mid C]$	omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	all components correct	all
NDEpred*	NDE|YpredEpred, TE|Ypred	omod: $\mathrm{E}[Y \mid C, M, A = 1]$ emod: $\mathrm{E}[NDE_0 \mid C]$	omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	all components correct	all
psYpred1.MR	crw|psYpred.MR, reg|Ypred.R	wts: $\omega_0(C), \omega_{\mathrm{x}}(C, M)$ omod: $\mathrm{E}[Y \mid C, M, A = 1]$	wts: $\omega_1(C), \omega_0(C)$ omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	• $\omega_0(C)$ correct, and • either $\omega_1(C)$ or $\mathrm{E}[Y \mid C, A = 1]$ correct, and • either $\omega_{\mathrm{x}}(C, M)$ or $\mathrm{E}[Y \mid C, M, A = 1]$ correct	$\omega_0(C)$
psYpred2.MR	Y_{1M_0}, Y_1|psYpred.MR, Y_0|Ypred.R		wts: $\omega_1(C), \omega_0(C)$ omod: $\mathrm{E}[Y \mid C, A = 1]$		
Ypred.MR	crw|Ypred.MR, reg|Ypred.R	wts: $\omega_{\mathrm{x}}(C, M)$ omod: $\mathrm{E}[Y_{1M_0} \mid C]$	wts: $\omega_1(C), \omega_0(C)$ omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	• either $\omega_1(C)$ or $\mathrm{E}[Y \mid C, A = 1]$ correct, and • either $\omega_0(C)$ or $\mathrm{E}[Y \mid C, A = 0]$ correct, and • either $\omega_{\mathrm{x}}(C, M)$ correct, or the M-related part of $\omega_{\mathrm{x}}(C, M)$ and $\mathrm{E}[Y_{1M_0} \mid C]$ correct	the M-related part of $\omega_{\mathrm{x}}(C, M)$
MsimYpred1.MR	crw|MsimYpred.MR, reg|Ypred.R	wts: $\omega_{\mathrm{x}}(C, M), \omega_0(C)$ mmod: $\mathrm{P}(M \mid C, A = 0)$ omod: $\mathrm{E}[Y \mid C, M, A = 1]$	wts: $\omega_1(C), \omega_0(C)$ omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	• either $\omega_1(C)$ or $\mathrm{E}[Y \mid C, A = 1]$ correct, and • either $\omega_0(C)$ or $\mathrm{E}[Y \mid C, A = 0]$ correct, and • either $\omega_{\mathrm{x}}(C, M)$ or $\mathrm{E}[Y \mid C, M, A = 1]$ correct, and • $\mathrm{P}(M \mid C, A = 0)$ correct	$\mathrm{P}(M \mid C, A = 0)$
MsimYpred2.MR	Y_{1M_0}, Y_0|MsimYpred.MR, Y_1|Ypred.R		wts: $\omega_1(C), \omega_0(C)$ mmod: $\mathrm{P}(M \mid C, A = 0)$ omods: $\mathrm{E}[Y \mid C, A = 1]$, $\mathrm{E}[Y \mid C, M, A = 0]$	• either $\omega_1(C)$ or $\mathrm{E}[Y \mid C, A = 1]$ correct, and • either $\omega_0(C)$ or $\mathrm{E}[Y \mid C, M, A = 0]$ correct, and • either $\omega_{\mathrm{x}}(C, M)$ or $\mathrm{E}[Y \mid C, M, A = 1]$ correct, and • $\mathrm{P}(M \mid C, A = 0)$ correct	
Y2pred.R	crw|Y2pred.R, reg|Ypred.R	wts: $\omega_{\mathrm{x}}(C, M), \omega_0(C)$ omods: $\mathrm{E}[Y \mid C, M, A = 1], \mathrm{E}[Y_{1M_0} \mid C]$	wts: $\omega_1(C), \omega_0(C)$ omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	• either $\omega_1(C)$ or $\mathrm{E}[Y \mid C, A = 1]$ correct, and • either $\omega_0(C)$ correct, or both $\mathrm{E}[Y \mid C, A = 0]$ and $\mathrm{E}[Y_{1M_0} \mid C]$ correct, and • either $\omega_{\mathrm{x}}(C, M)$ or $\mathrm{E}[Y \mid C, M, A = 1]$ correct	NONE
NDEpred.R*	NDE|YpredEpred.R, TE|Ypred.R	wts: $\omega_{\mathrm{x}}(C, M), \omega_0(C)$ omod: $\mathrm{E}[Y \mid C, M, A = 1]$ emod: $\mathrm{E}[NDE_0 \mid C]$	wts: $\omega_1(C), \omega_0(C)$ omods: $\mathrm{E}[Y \mid C, A = a]$ for $a = 1, 0$	• either $\omega_1(C)$ or $\mathrm{E}[Y \mid C, A = 1]$ correct, and • either $\omega_0(C)$ correct, or both $\mathrm{E}[Y \mid C, A = 0]$ and $\mathrm{E}[NDE_0 \mid C]$ correct, and • either $\omega_{\mathrm{x}}(C, M)$ or $\mathrm{E}[Y \mid C, M, A = 1]$ correct	NONE

Notes: "wts" = weights. "omod" = outcome mean model. "mmod" = mediator density model. "emod" = effect model. * = only for additive effects.

4.5. A quick note on model compatibility

This section is included for the more technically inclined readers and might not be of general interest. In response to helpful comments from the referees, we explored the topic of model compatibility or lack thereof for the estimators covered in this paper. Generally it is undesirable to use incompatible modeling components, because then at least one component in the conflict is mis-specified, regardless of what the actual distribution is. The specific concern here is whether an estimator's use of variationally dependent modeling components means the estimator has a model incompatibility issue.

Interestingly, we find that practically one needs not worry about model incompatibility for these estimators. Let us consider three relevant cases of variationally dependent models: (i) combination of two conditional outcome mean models $\mathrm{E}[Y \mid C, A = 1]$ and $\mathrm{E}[Y \mid C, M, A = 1]$; (ii) combination of two conditional exposure models $\mathrm{P}(A \mid C)$ and $\mathrm{P}(A \mid C, M)$; and (iii) combination of the last two models with the mediator density model $\mathrm{P}(M \mid C, A = 0)$. In cases (i) and (ii), although the models are variationally dependent, as long as their specification does not restrict the range of the conditional means/probabilities, they are compatible. To make this concrete, an example of restriction-induced incompatibility is that for a certain value c of C, one specify $\mathrm{P}(A \mid C = c) = .2$ but specify $\mathrm{P}(A \mid C = c, M) \in (.4, .7)$; these specifications are incompatible because there exists no density $\mathrm{P}(M \mid C = c)$ that satisfies $\mathrm{P}(A \mid C = c) = \mathrm{E}_{M|C=c}[\mathrm{P}(A \mid C = c, M]$. It is hard to think of a case where one would specify such weirdly constrained models, though, so this is not really a practical concern. We thus exclude this kind of conflicting specification from consideration.

Case (iii), the combination of the two conditional exposure models with a model for $\mathrm{P}(M \mid C, A = 0)$, is only present in a version of MsimYpred.MR that estimates the cross-world weights $\omega_{\mathrm{x}}(C, M)$ using the second formula. (The other choice for MsimYpred.MR is to estimate $\omega_{\mathrm{x}}(C, M)$ using the first formula based on mediator densities, where model components are variationally independent so there is no incompatibility.) The case of combining $\mathrm{P}(M \mid C, A = 0)$ for mediator simulation with $\mathrm{P}(A \mid C)$ and $\mathrm{P}(A \mid C, M)$ for $\omega_{\mathrm{x}}(C, M)$ estimation is an interesting case where it turns out that there is also no model incompatibility. Here the explicit specification of $\mathrm{P}(A \mid C)$ and $\mathrm{P}(A \mid C, M)$ implies an implicit specification of the ratio $\frac{\mathrm{P}(M|C,A=0)}{\mathrm{P}(M|C,A=1)}$ (as this is equal to $\frac{\mathrm{odds}(A=1|C)}{\mathrm{odds}(A=1|C,M)}$). This implicit specification combined with the explicit specification of $\mathrm{P}(M \mid C, A = 0)$ implies an implicit specification of $\mathrm{P}(M \mid C, A = 1)$. Since we do not explicitly model $\mathrm{P}(M \mid C, A = 1)$, there is no model incompatibility. This estimator essentially "escapes" model incompatibility by simulating the mediator only for the cross-world condition, thus relying on estimating the mediator density under one treatment only.

We note that the assurance of model compatibility here does not tell us whether the model is correct. Model compatibility is a quality of the estimator; correct or mis-specification is a quality of the correspondence between the model/estimator and the truth.

At the request of the Editor, we now respond to a specific point raised by a Referee in the second round review of this paper, which references model compatibility but in our opinion is more about model mis-specification. The point raised is: when modeling the two conditional outcome mean functions $\mathrm{E}[Y \mid C, A = 1]$ and $\mathrm{E}[Y \mid C, M, A = 1]$ (case (i) above), this implies an implicit specification of the mediator distribution, and the concern is that this implicit specification may be mis-specified. The Referee comments that this would not be an issue with the alternative choice of modeling $\mathrm{E}[Y \mid C, M, A = 1]$ and $\mathrm{P}(M \mid C, A = 1)$. We respond in two parts, one technical and one practical. The technical part is that the latter choice also has the same issue, as it implies an implicit specification for $\mathrm{E}[Y \mid C, A = 1]$, and one may also be concerned that this implicit specification is incorrect. In fact, since these three functions (two outcome mean and one mediator density) are tied together by the relationship $\mathrm{E}[Y \mid C, A = 1] = \mathrm{E}_{M|C,A=1}\{\mathrm{E}[Y \mid C, M, A = 1]\}$, explicit specification of any two of the three implies an implicit specification of the third. Also, all specifications, explicit or implicit, may be incorrect. The practical part of our response is that modeling choices should be guided by the specific estimation strategy. For strategies that require an estimate of the function $\mathrm{E}[Y \mid C, A = 1]$ (as part of estimating $\mathrm{E}[Y_1]$) and an estimate of the function $\mathrm{E}[Y \mid C, M, A = 1]$ (as part of estimating $\mathrm{E}[Y_{1M_0}]$), we choose to estimate both of these target functions directly, so (roughly speaking) both have equal chance of being estimated well. This also means they have equal chance of being estimated poorly. The suggested alternative means estimating $\mathrm{E}[Y \mid C, A = 1]$ indirectly by estimating the other two functions and putting them together; this would double this target function's chance of being estimated poorly because either of the component functions could be poorly estimated; and perhaps this chance is more than doubled because density estimation is harder than mean estimation.

In making the point above, the Referee also comments that using logit models for both $\mathrm{E}[Y \mid C, A = 1]$ and $\mathrm{E}[Y \mid C, M, A = 1]$ implies that $\mathrm{P}(M \mid C, A = 1)$ is a bridge distribution [57, 58]. This is not the case, though, because the models bridged by a bridge distribution are fundamentally different from our outcome models. We explain this in section D of the Technical Appendix.

5. If targeting effects on the additive scale: marginal effect as the mean of individual specific effects

If the marginal effects being targeted are defined on the additive scale, there is an alternative view of the puzzle, where what we wish we had is a single full sample in which all potential outcomes are simultaneously observed, which means for each individual the effects are observed. Then the individual TE, NDE_0, NIE_1, etc. are variables that could simply be averaged to estimate the average (which are also the marginal additive) effects. While these effect variables are not observed (this is the fundamental problem of causal inference), this view suggests we might learn an average effect if we have a good proxy for the individual effect. It turns out that this works for natural direct effects.

5.1. NDE|YpredEpred: effect prediction based on a proxy model

The key to this method is to choose a proxy for the individual effect that has the same mean given covariates as the effect itself. Consider the individual $NDE_0 = Y_{1M_0} - Y_0$. For control units, we observe Y_0 but not Y_{1M_0}. The idea is to replace the unobserved Y_{1M_0} with its predicted value based on an appropriate model.

This leads to an estimator pair that is a slight modification of crw|Y2pred. Recall that crw|Y2pred involves a double model fit where the second model regresses predicted Y_{1M_0} values (in control units) on covariates. The modification is that to estimate NDE_0, that second model instead regresses the difference between predicted Y_{1M_0} and observed Y_0 (a proxy for the individual NDE_0) on covariates. This model, which we loosely call the $\text{E}[NDE_0 \mid C]$ model, is then used to predict NDE_0 for all units in the full sample, and these predicted individual effects are averaged to estimate the average NDE_0.

Like the crw|Y2pred pair, the NDE|YpredEpred pair includes a nonrobust and a robust estimator. For the nonrobust one, the $\text{E}[Y \mid C, M, A = 1]$ and $\text{E}[NDE_0 \mid C]$ models are fit to the treated and control subsamples, respectively. For the robust estimator NDE|YpredEpred.R, these models (which now are required to satisfy the mean recovery condition) are fit to the pseudo cross-world and pseudo control samples instead. This robust estimator has three chances to be correct: (i) if both the outcome and effect models are correctly specified; or (ii) if the $\omega_0(C)$ and $\omega_{\text{x}}(C, M)$ weights that form the two pseudo samples are consistent; or (iii) if the $\omega_{\text{x}}(C, M)$ weights and the outcome model are consistent.

An aside: Instead of using the observed outcome in the construction of the proxy for NDE_0 in control units, a variant replaces it with a predicted value of this outcome based on a $\text{E}[Y \mid C, M, A = 0]$ model. The robust version of this variant (where all models are fit to relevant pseudo samples) is closely related to Zheng and van der Laan's TMLE estimator [60]. This alternative estimator also has three chances to be correct under similar conditions to those listed above, except that condition (i) additionally requires correct specification of the $\text{E}[Y \mid C, M, A = 0]$ model.

Note that the current strategy only works for direct effects, as no similar proxy for the individual $NIE_1 = Y_1 - Y_{1M_0}$ is available. To estimate the indirect effect, one can subtract an NDE|YpredEpred estimator off of a total effect estimate. We obtain the latter using reg|Ypred estimators.

6. How to choose an estimator

In addition to the primary goal of building intuition for (more) robust estimation, a secondary goal of this paper is to provide a menu of estimation options for the specific estimands considered. Table 2 summarizes the estimators of marginal natural (in)direct effects discussed so far, with nonrobust estimators in the top panel and more robust estimators in the bottom panel. Columns 3 and 4 of this table list the components involved in each estimator, under the

groupings of outcome models (omod), effect models (emod), mediator models (mmod) and weights (wts). Column 5 lists what is required of these components for the estimator to be consistent. For the (more) robust estimators, each bullet in this column is one requirement, and any bullet of the either-or form indicates a robustness property, while any bullet not in either-or form is a nonrobust component that needs to be consistent for the estimator to be consistent. The nonrobust component is also pointed out specifically in column 6.

Given this menu of estimators, which one should be used for a particular application? We take the pragmatic viewpoint that the choice of methods should partly depend on the user's level of comfort with the different types of methods, because implementation is more error prone if a method is more complex and not well understood by the user. Therefore we do not intend to propose or advocate for a single method but to lay out a range of potential choices. We offer some considerations below.

As the nonrobust estimators are simpler, one approach is to pick one of those estimators. Among the nonrobust options we do not recomend the weighting-based estimators, as they are inefficient. Otherwise, we recommend considering the set of estimation components (weights, mediator density and outcome/effect mean models) required by each estimator and deciding which set is most feasible to implement well. The disadvantage of the simpler estimators, of course, is the lack of robustness to model mis-specification. Another approach is to consider the more robust estimators, looking at components that must be correctly specified as a way to rule out estimators that depend on hard-to-model components.

Again, we note that density estimation is generally more challenging than mean estimation, and the difficulty of the task depends on the number and types of mediators. It is easier for certain types of variables (e.g., binary variables) than others (e.g., continuous variables). When there are multiple mediators, there are multiple models to fit and more chances to mis-specify them. With multiple mediators, we need to choose an order of factorization for the mediators, and therefore may prefer an order that makes models slightly easier or more convenient to specify (see our data example). We might want to consider alternatives to density estimation where possible, e.g., using the second formula of the cross-world weight rather than the first formula. When using the MsimYpred estimators, we should keep in mind that these methods depend on a correct specification of the $\mathrm{P}(M \mid C, A = 0)$ model.

Several estimators depend on certain weights being correct, including the nonrobust weighting-based estimators, as well as the more robust estimators psYpred.MR (requiring correct control weights) and Ypred.MR (requiring that the M-related component of the cross-world weight is correct). While correct specification cannot be determined, balance checking is a useful guard against severe misspecification. If certain weights do not achieve excellent balance, it is advisable not to use the estimators whose validity hangs on those weights.

Lastly, on the menu there are two fully robust estimators, Y2pred.R and NDEpred.R, which do not depend on any specific modeling component being correct. Note that for each of these, we still need enough of the estimation components to be correct.

7. A comment on a common practice

Above we have shown how simple estimation strategies can be made more robust. Here we comment on a common practice in applied research that looks similar to robust estimation to point out that it should not be seen as such, but should simply be seen as a method to improve precision.

For simplicity, first consider the non-mediation setting where the estimand is the average treatment effect, $\mathrm{E}[Y_1] - \mathrm{E}[Y_0]$. Here this common practice involves first balancing covariates to justify comparing outcomes between the two groups, and then fitting a simple model regressing outcome on exposure and covariates, usually in the form of main effects. Seen through our pseudo samples lens, when covariate balancing is done by propensity score weighting, we achieve the pseudo treated and pseudo control samples, and the regression model is fit to the combination of these two pseudo samples. With the combination of weighting and an outcome model, this type of analysis looks similar to a doubly robust method, and we have heard practitioners describe it as doubly robust. However, the regression model is likely too simple to have a chance at being correct. As the consistency of the method depends on the pseudo samples, this practice should not be seen as a robust method. It would be appropriate, though, to refer to this use of the simple regression model as *leveraging covariates to improve precision*. Leveraging covariates to improve precision is an approach for analysis of randomized trials, where the effect is identified so no covariate adjustment is needed, but the use of a *working regression model* (not assumed to be correct) helps explain outcome variance and thereby makes the effect estimate more precise [55, 40]. Intuitively, the pseudo samples mimic a randomized trial, so the simple regression is just a working model to improve precision.

In the current setting of estimating marginal natural (in)direct effects, methods that first create three pseudo samples representing the conditions being contrasted (by weighting only or weighting combined with prediction/imputation of Y_{1M_0}) and then fit a simple model regressing outcome on covariates and conditions (indicated by dummy variables) should be seen as methods to leverage covariates to improve precision, not as robust methods. For interested readers, a preprint of this paper [26] (version 3, section 6) includes a translation to the current setting of techniques for using covariates to improve precision that apply to different outcome types. It also comments on similarity with and key distinctions from several methods that estimate conditional effects [53, 38].

8. Confidence interval estimation

The previous sections cover point estimation of the effects. We now turn to interval estimation. All the estimators in this paper are M-estimators; they are solutions of generalized estimating equations. Applying the calculus of M-estimators [39], we derive general formulas for the asymptotic variance of each of the estimators when all estimation components (weights, conditional mediator density, conditional mean outcome/effect) are based on (semi)parametric models. The derivations are placed in the Technical Appendix.

As such variance estimators depend on the specific models used for the different estimation components, they are clunky to use in practice. We use bootstrapping as a generic tool to obtain confidence intervals for all the estimators. As the data example includes quite a few categorical covariate/mediator variables, a challenge when using the simple resampling bootstrap [3] is that some bootstrap samples do not cover all values (and combinations of values) of those variables, resulting in predictors being dropped from models and predictions being distorted. To avoid this problem we instead use a continuous weights bootstrap [59]. With both bootstrap procedures, the making of a bootstrap sample can be seen as weighting the observations of the original sample by a set of random weights that are identically distributed: the resampling bootstrap uses integer weights drawn from a uniform Multinomial distribution; the continuous weights bootstrap draws weights from a continuous distribution. We use the version proposed by Xu et al. [59] based on the uniform Dirichlet distribution, where the weights sum to sample size n, have mean 1 and variance $(n-1)/(n+1)$. (For comparison, resampling weights sum to n, have mean 1 and variance $(n-1)/n$.) Bootstrap samples based on continuous bootstrap weights retain all observations, thus they do not lose data patterns.

9. Data example application

In this example A is a binary variable `treat` indicating whether a student is in the treatment (i.e., combined intervention) or control condition. Y is binary variable `drink` indicating whether the student engages in weekly drinking at 22 months. M consists of three mediators measured at six months: attitudes towards alcohol consumption (binary variable `att` indicating attitudes against consumption), self-control in situations involving alcohol (continuous variable `sfc`), and parental rules regarding alcohol (binary variable `rul` indicating strict rules). Baseline covariates C include demographic variables age, sex, religion, education track (academic or vocational); baseline measures of the mediators (`att0`, `sfc0`, `rul0`); and baseline measure of the outcome (`drink0`). Table 3 summarizes the baseline covariates in the synthetic sample, showing some covariate imbalance between the intervention and control conditions. The unconfoundedness assumption means that the listed baseline covariates are sufficient to remove exposure-mediator, exposure-outcome and mediator-outcome confounding.

With this example, we target marginal effects on the additive scale. The total effect can be understood as a reduction in weekly drinking prevalence that would occur had all students received the treatment versus no students received the treatment. The natural indirect effect is roughly interpreted as the component of that prevalence reduction that is due to the intervention's impact on the mediators, and the natural direct effect is the remaining component.

9.1. Weighting

Weights estimation

The estimators have different requirements in terms of weights – see Tables 2. The pure weighting estimator and all the MR/R estimators involve the trio of

TABLE 3
Baseline covariates in the synthetic dataset based on the PAS study

	Treated (n=778)		**Control** (n=907)		**Total** (n=1685)	
Age						
11	35	(4.5%)	38	(4.2%)	73	(4.3%)
12	559	(71.9%)	682	(75.2%)	1241	(73.6%)
13	184	(23.7%)	187	(20.6%)	371	(22.0%)
Sex						
female	305	(39.2%)	448	(49.4%)	753	(44.7%)
male	473	(60.8%)	459	(50.6%)	932	(55.3%)
Religion						
Catholic	62	(8.0%)	319	(35.2%)	381	(22.6%)
Protestant/other Christian	84	(10.8%)	114	(12.6%)	198	(11.8%)
Islam	45	(5.8%)	34	(3.7%)	79	(4.7%)
not religiously socialized	552	(71.0%)	416	(45.9%)	968	(57.4%)
other	35	(4.5%)	24	(2.6%)	59	(3.5%)
Education tract						
vocational	276	(35.5%)	547	(60.3%)	823	(48.8%)
academic	502	(64.5%)	360	(39.7%)	862	(51.2%)
Baseline weekly drinking						
yes	96	(12.3%)	166	(18.3%)	262	(9.0%)
no	625	(80.3%)	647	(71.3%)	1272	(75.5%)
no answer	57	(7.3%)	94	(10.4%)	151	(9.0%)
Baseline attitude						
negative re. alcohol use	518	(66.6%)	594	(65.5%)	1112	(66.0%)
less negative.	260	(33.4%)	313	(34.5%)	573	(34.0%)
Baseline parental rule						
strict	561	(72.1%)	580	(63.9%)	1141	(67.7%)
not strict	271	(27.9%)	327	(36.1%)	544	(32.3%)
Baseline self control						
mean (SD)	3.59	(0.55)	3.57	(0.53)	3.58	(0.54)
median [min, max]	3.62	[1.69,4.85]	3.62	[2.00,4.92]	3.62	[1.69,4.92]

$\omega_0(C)$, $\omega_1(C)$ and $\omega_x(C, M)$ weights. Several nonrobust estimators involve some (but not all) weights: psYpred estimators involve $\omega_0(C)$ and Ypred involves $o_x(C, M)$ weights. The nonrobust Y2pred and MsimYpred estimators do not require weights. Here we focus on the $\omega_0(C)$, $\omega_1(C)$ and $\omega_x(C, M)$ weights.

We estimate weights via parametric models. $\omega_1(C)$ and $\omega_0(C)$ are estimated via on a propensity score model, i.e., a model for $\mathrm{P}(A \mid C)$. $\omega_x(C, M)$ is estimated by the second method, using the combination of this propensity score model and a model for $\mathrm{P}(A \mid C, M)$. (We avoid the first method which would require fitting six models for the three mediators.) For both models we use logistic regression with spline terms on continuous predictors and some interaction terms in the second model (the result of several rounds of model fitting and balance checking).[16] The model formulas used for $\mathrm{P}(A \mid C)$ and $\mathrm{P}(A \mid C, M)$ are:

```
treat ~ sex + age + edu + religion + drink0 + att0 + rul0 + ns(sfc0, 4),
treat ~ sex + age + edu + religion + drink0 + att0 * att + rul0 * rul +
        ns(sf0, 4) * ns(sfc, 4).
```

The distributions of these weights (in stabilized form, i.e., with mean 1 in each group) are shown in Figure 7. Some of the weights are large, but not extreme.

Balance checking

Balance on the means of covariates and mediators for the pseudo treated, control and cross-world samples are shown in Figure 8 (based on the prescription

[16]The function `ns(v,d)` from R-package `splines` implements cubic splines on variable `v` with `d` degrees of freedom.

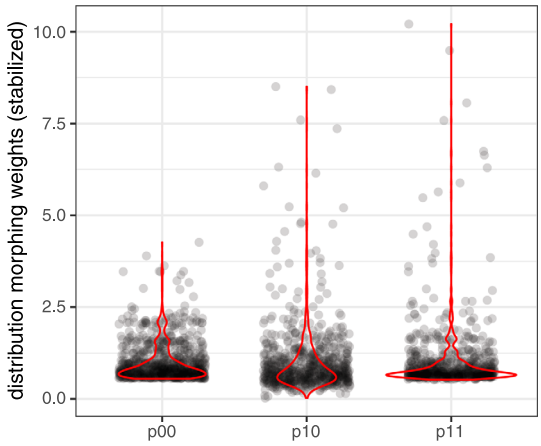


FIG 7. *Distributions of weights for the pseudo control (p00), treated (p11) and cross-world (p10) samples. For comparability, stabilized weights are shown.*

in Figure 4). Overall, balance improves after weighting; this is prominent for covariates sex, education track, religion, and the three mediators. Interestingly, balance on baseline self-control (`sfc0`) is slightly worsened, although the standardized mean difference is still modest. In addition to mean balance, distributional balance on continuous covariates and mediators should also be checked, e.g., using the R-package `cobalt` [4].

Note that the comments in the plot labels specifically address estimators that depend on all balance components, here the pure weighting estimator (wtd). For estimators with some robustness, we recommend a combination of two plots: a *full balance* plot capturing balance resulting from all the weighting involved in the estimator, and a *key balance* plot capturing the balance component that the estimator absolutely depends on. For Ypred.MR, for example, the full balance plot is the same plot in Figure 8 without the comments in the plot labels, and the key balance plot picks out the 'p10–p00' component.

For nonrobust estimators that depend on a weighting element, balance checking is specific to the weighting. For example, psYpred1 and psYpred2 require covariate balance between the pseudo control sample and the full sample; Ypred requires covariate-and-mediator balance between the pseudo cross-world subsample and the control subsample. All these variants of balance checking are implemented in the `mediationClarity` package (see details in package vignette).

9.2. Other estimation components

Outcome mean models

The various estimators call for fitting models for the outcome given covariates or given covariates and mediators to subsamples or pseudo samples. We use logistic

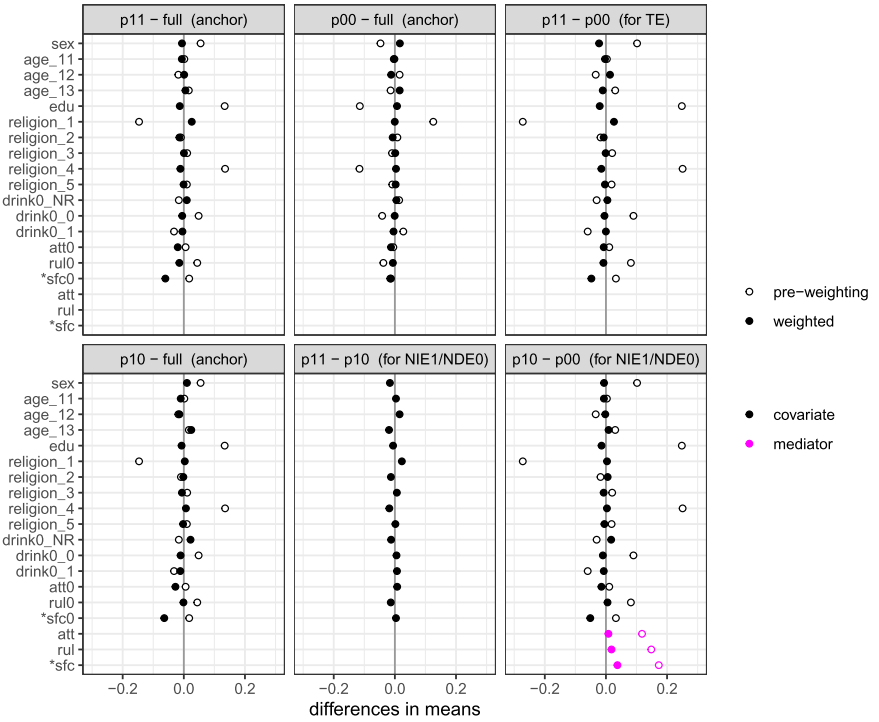


FIG 8. *Covariate and mediator balance for pseudo treated (p11), pseudo control (p00) and pseudo cross-world (p10) samples. For continuous covariate* `sfc0` *and continuous mediator* `sfc` *(marked with *), the mean differences are standardized. The parenthesized comments are specific to the wtd and wt-Cadj estimators, indicating that all balance components are important to those estimators.*

regression for the binary outcome. Models that regress outcome on covariates (estimating $\mathrm{E}[Y \mid C, A = 1]$, $\mathrm{E}[Y \mid C, A = 0]$ or $\mathrm{E}[Y_{1M_0} \mid C]$) use formula

```
drink ~ sex + age + edu + religion + drink0 + att0 + rul0 + ns(sfc0, 3).
```

Models that regress outcome on both covariates and mediators (estimating $\mathrm{E}[Y \mid C, M, A = 1]$) use formula

```
drink ~ sex + age + edu + religion + drink0 + att0 + rul0 + ns(sfc0, 3) +
        att + rul + ns(sfc, 3).
```

Mediator density model

The MsimYpred estimators require mediator density modeling in the control subsample or pseudo control sample. We fit logit models for the two binary mediators `att` (M^a) and `rul` (M^b), and a linear model for the continuous mediator `sfc` (M^c), with formulas:

```
att ~ age + sex + edu + religion + drink0 + att0 + rul0 + ns(sfc0, 3)
rul ~ age + sex + edu + religion + drink0 + att0 + rul0 + ns(sfc0, 3) + att
sfc ~ age + sex + edu + religion + drink0 + att0 + rul0 + ns(sfc0, 3) + att + rul
```

and assume the errors in the third model are normally distributed and homoscedastic. These models estimate $\mathrm{P}(M^a \mid C, A=0)$, $\mathrm{P}(M^b \mid C, M^a, A=0)$, and $\mathrm{P}(M^c \mid C, M^a, M^b, A=0)$, respectively.[17]

Model for NDE_0 given covariates

The NDEpred estimators involve regressing the proxy of the individual NDE_0 (predicted Y_{1M_0} minus observed Y_0) on covariates in the control subsample or pseudo control sample. As the difference between the two (predicted and observed) binary outcomes is bounded in the $[-1, 1]$ interval, we transform it by adding 1 then dividing by 2 to map to the $[0, 1]$ interval, and fit the regression model to the transformed difference using logit link. Predictions based on this model are back transformed by multiplying by 2 then subtracting 1.[18] The formula we use for this model is:

```
trans.diff ~ age + sex + edu + religion + drink0 + att0 + rul0 + ns(sfc0, 3).
```

9.3. Results

Effect estimates from different estimators are shown in Figure 9. To avoid clutter, for estimator types that have multiple versions, we show only one version. The estimates are quite similar. Overall, it appears that the effect of the intervention on weekly drinking at follow-up consists of a small part mediated by the mediators being considered (alcohol-related attitudes, parental rules, and self-control), and a larger direct effect.

As these are estimates from one dataset, one should be cautious not to infer characteristics of the estimators. That said, we note that within each pair of nonrobust and (more) robust estimators, the (more) robust one tends to have larger variance than the nonrobust one, with wider confidence intervals. The pure weighting estimator by theory has the largest variance, although for this dataset this is not obvious from the confidence interval widths.

10. Concluding remarks

In this paper we have shown how a range of estimators may be constructed based on two strategies that are familiar to many who are involved in statistical analyses (weighting and model-based prediction) and a simple way of combining them (weighted models); this is the paper's primary goal. The key ideas of this exercise, which are not specific to natural effects but apply generally, are (i) to use these tools flexibly to put together the different pieces of the estimation puzzle, where the puzzle is defined by the identification result of the effects of

[17] Of the three mediators, we choose to model `sfc` last for convenience. This avoids having to specify models with `sfc` as another continuous predictor, which would be more complicated.

[18] This is equivalent to using the tanh link for a response variable bounded in $[-1, 1]$ [56]; the transformation trick allows fitting the model using standard software with logit link.

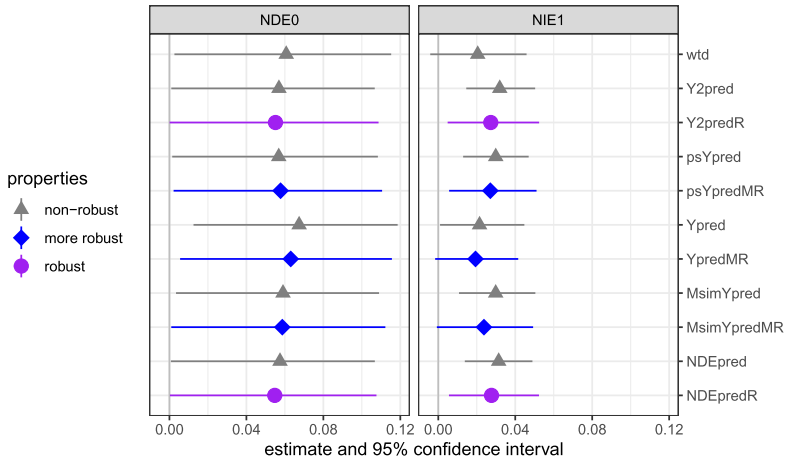


FIG 9. *Effect estimates from different estimators, shown as reduction in outcome (weekly drinking) prevalence. The psYpred and MsimYpred estimators shown here are the psYpred2 and MsimYpred1 versions in Table 2.*

interest; and (ii) to induce robustness on pieces of the puzzle by using weighted models.

Thinking more broadly, this approach to constructing estimators could be applied to other marginal estimands, including interventional effects of various kinds [28] and causal decomposition of disparities [18]. A key difference is that these other estimands involve setting the mediator (or the target variable of causal decomposition) to one of a range of interventional distributions (depending on the estimand) which may condition on or marginalize over certain pre-exposure and post-exposure covariates. This means the weighting scheme needs to be tailored and may be more complicated, and the density being mimicked may condition (or not) on different types of variables, and may be known or need to be estimated. Whatever the case, the idea of visualizing the identification result to bring clarity to where different types of information (pieces of the puzzle) come from, and a similar exercise of assembling them, will be productive in generating estimators, and importantly will make the estimator transparent to the user (as a sound solution to the puzzle).

On a technical note, the focal estimand in the current case (the identification result of the mean cross-world potential outcome) is an iterated expectation, and there are different ways an iterated expectation can be estimated. For the current case, one way involves fitting repeated conditional mean models (iterated regression), and another involves integrating an inner expectation over an estimated conditional density (here via simulation). Weights can be used to fit the models to the space of predictors where they are used for prediction/simulation, to help correct bias due to model mis-specification. One point we noted is that this provides only a partial correction for misspecified conditional density models. This point is likely relevant to some of the more complicated estimands

for which some density estimation cannot be avoided.

One important topic that was not covered in this paper is sensitivity analysis to violation of identifying assumptions. While several sensitivity analysis methods have been proposed, there is room for work that connects each of the many estimation methods that exist and may be used in practice to relevant sensitivity analyses, or at the least point out which estimation methods can (and which cannot) be appropriately paired with which sensitivity analyses. This would be very helpful to the use of methods in practice.

Acknowledgments

The authors appreciate Drs. Guanglei Hong and Fan Yang for helpful feedback on an earlier draft; Drs. Ilya Shpitser and Eric Tchetgen Tchetgen for insightful discussions on robust estimation and their seminal article [43]; participants of our summer institute mediation course from 2021 and 2022 and participants of the second term 2021 seminar on statistical methods for mental health research at Johns Hopkins Bloomberg School of Public Health for fruitful discussion; and two anonymous Referees, the Associate Editor and the Editor Dr. Richard Lockhart for their thoughtful and constructive comments. The authors thank the participants, staff and investigators of the PAS trial.

Supplementary Material

Supplement 1
(doi: 10.1214/22-SS140SUPPA; .pdf). This is the technical appendix to the paper, with three parts. Part A derives and connects alternative expressions for the cross-world weight function, including the novel third expression. Part B contains proofs of the robustness (and any nonrobustness) properties of the robust and more robust estimators. Part C derives general asymptotic variance formulas for all the estimators, where estimation components (weights, conditional mediator density functions, conditional outcome/effect mean functions) are based on parametric models. Part D explains our response in section 4.5 that two logit outcome models do not imply a bridge mediator distribution.

Supplement 2
(doi: 10.1214/22-SS140SUPPB; .pdf). Vignette of the R package `mediationClarity`, which implements the estimators.

References

[1] Albert, J. M. (2012). Distribution-free mediation analysis for nonlinear models with confounding. *Epidemiology* **23** 879–88.

[2] Didelez, V., Dawid, A. P. and Geneletti, S. (2006). Direct and Indirect Effects of Sequential Treatments. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence* 138–146. AUAI Press.

[3] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **11** 1–26.
[4] Greifer, N. (2022). cobalt: Covariate Balance Tables and Plots R package version 4.3.2.
[5] Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* **20** 25–46.
[6] Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* **81** 945.
[7] Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. In *Proceedings of the American Statistical Association, Biometrics Section* 2401–2415.
[8] Hong, G., Deutsch, J. and Hill, H. D. (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics* **40** 307–340.
[9] Hong, G., Qin, X. and Yang, F. (2018). Weighting-Based Sensitivity Analysis in Causal Mediation Studies. *Journal of Educational and Behavioral Statistics* **43** 32–56.
[10] Hong, G., Yang, F. and Qin, X. (2021). Post-Treatment Confounding in Causal Mediation Studies: A Cutting-Edge Problem and A Novel Solution via Sensitivity Analysis.
[11] Hong, G., Yang, F. and Qin, X. (2021). Did you conduct a sensitivity analysis? A new weighting-based approach for evaluations of the average treatment effect for the treated. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184** 227–254. MR4204918
[12] Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics* **29** 920–943.
[13] Huber, M. (2020). Mediation Analysis. In *Handbook of Labor, Human Resources and Population Economics* (K. F. Zimmermann, ed.) Springer.
[14] Huling, J. D. and Mak, S. (2020). Energy balancing of covariate distributions. *arXiv* 1–68.
[15] Imai, K., Keele, L. and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods* **15** 309–34.
[16] Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* **25** 51–71.
[17] Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **76** 243–263.
[18] Jackson, J. W. (2021). Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework. *Epidemiology* **32** 282–290.
[19] Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science* **22** 523–539.

[20] Koning, I. M., Van Den Eijnden, R. J., Verdurmen, J. E., Engels, R. C. and Vollebergh, W. a. (2011). Long-term effects of a parent and student intervention on alcohol use in adolescents: A cluster randomized controlled trial. *American Journal of Preventive Medicine* **40** 541–547.

[21] Koning, I. M., van den Eijnden, R. J. J. M., Engels, R. C. M. E., Verdurmen, J. E. E. and Vollebergh, W. a. M. (2010). Why target early adolescents and parents in alcohol prevention? The mediating effects of self-control, rules and attitudes about alcohol use. *Addiction* **106** 538–46.

[22] Koning, I. M., Vollebergh, W. a. M., Smit, F., Verdurmen, J. E. E., Van Den Eijnden, R. J. J. M., Ter Bogt, T. F. M., Stattin, H. and Engels, R. C. M. E. (2009). Preventing heavy alcohol use in adolescents (PAS): cluster randomized trial of a parent and student intervention offered separately and simultaneously. *Addiction* **104** 1669–78.

[23] Lange, T., Vansteelandt, S. and Bekaert, M. (2012). A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology* **176** 190–195.

[24] Miles, C., Kanki, P., Meloni, S. and Tchetgen Tchetgen, E. (2017). On Partial Identification of the Natural Indirect Effect. *Journal of Causal Inference* **5**. MR4328876

[25] Muthén, B. O. and Asparouhov, T. (2015). Causal effects in mediation modeling: An introduction with applications to latent variables. *Structural Equation Modeling* **22** 12–23.

[26] Nguyen, T. Q., Ogburn, E. L., Schmid, I., Sarker, E. B., Greifer, N., Koning, I. M. and Stuart, E. A. (2022). Causal mediation analysis: From simple to more robust strategies for estimation of marginal natural (in)direct effects. *arXiv:2102.06048*. Version 3.

[27] Nguyen, T. Q., Schmid, I., Ogburn, E. L. and Stuart, E. A. (2022). Clarifying Causal Mediation Analysis: Effect Identification via Three Assumptions and Five Potential Outcomes. *Journal of Causal Inference* **10** 246–279.

[28] Nguyen, T. Q., Schmid, I. and Stuart, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods* **26** 255–271.

[29] Nowok, B., Raab, G. M. and Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software* **74** 1–26.

[30] Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence* 411–420.

[31] Pearl, J. (2012). The causal mediation formula–a guide to the assessment of pathways and mechanisms. *Prevention Science* **13** 426–36.

[32] Qin, X. and Yang, F. (2021). Simulation-based sensitivity analysis for causal mediation studies. *Psychological Methods*.

[33] Robins, J., Sued, M., Lei-Gomez, Q. and Rotnitzky, A. (2007). Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable. *Statistical Science* **22** 544–559.

[34] Robins, J. M. and Greenland, S. (1992). Identifiability and exchange-

ability for direct and indirect effects. *Epidemiology* **3** 143–155.
[35] Robins, J. M., Richardson, T. S. and Shpitser, I. (2022). An Interventionist Approach to Mediation Analysis. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, first ed. **36** 713–764. Association for Computing Machinery, New York, NY, USA.
[36] Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70** 41.
[37] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.
[38] Steen, J., Loeys, T., Moerkerke, B. and Vansteelandt, S. (2017). Medflex: An R package for flexible mediation analysis using natural effect models. *Journal of Statistical Software* **76**.
[39] Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56** 29–38.
[40] Steingrimsson, J. A., Hanley, D. F. and Rosenblum, M. (2017). Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. *Contemporary Clinical Trials* **54** 18–24.
[41] Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* **143** 1249–1272.
[42] Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in Medicine* **32** 4567–4580. MR3118376
[43] Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics* **40** 1816–1845. MR3015045
[44] Tchetgen Tchetgen, E. J. and Shpitser, I. (2014). Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika* **101** 849–864.
[45] Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software* **59** 1–38.
[46] Valeri, L. and VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological methods* **18** 137–150.
[47] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45** 1–67.
[48] van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics*. Springer New York.

[49] VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface* **2** 457–468.

[50] VanderWeele, T. J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology* **172** 1339–1348.

[51] VanderWeele, T. J. and Vansteelandt, S. (2013). Mediation analysis with multiple mediators. *Epidemiologic Methods* **2** 95–115.

[52] VanderWeele, T. J., Vansteelandt, S. and Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* **25** 300–6.

[53] Vansteelandt, S., Bekaert, M. and Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods* **1** 7.

[54] Vansteelandt, S. and Keiding, N. (2011). Invited commentary: G-computation–Lost in translation? *American Journal of Epidemiology* **173** 739–742.

[55] Wang, B., Ogburn, E. L. and Rosenblum, M. (2019). Analysis of Covariance in Randomized Trials: More Precision, Less Conditional Bias, and Valid Confidence Intervals, Without Model Assumptions. *Biometrics* **75** 1391–1400.

[56] Wang, L. and Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **80** 531–550. MR3798877

[57] Wang, Z. and Louis, T. A. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika* **90** 765–775.

[58] Wang, Z. and Louis, T. A. (2004). Marginalized Binary Mixed-Effects Models with Covariate-Dependent Random Effects and Likelihood Inference. *Biometrics* **60** 884–891. MR2133540

[59] Xu, L., Gotwalt, C., Hong, Y., King, C. B. and Meeker, W. Q. (2020). Applications of the Fractional-Random-Weight Bootstrap. *American Statistician* **1305** 1–32. MR4168255

[60] Zheng, W. and van der Laan, M. J. (2012). Targeted maximum likelihood estimation of natural direct effects. *The International Journal of Biostatistics* **8**.