Journal of Machine Learning Research 25 (2024) 1-54 Submitted 8/23; Revised 6/24; Published 11/24

Causal effects of intervening variables in settings with unmeasured confounding

Lan Wen LAN.WEN@UWATERLOO.CA
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, ON N2L 3G1, Canada

Aaron L. Sarvet ASARVET@UMASS.EDU
Department of Biostatistics and Epidemiology
University of Massachusetts Amherst
Amherst, MA 01003, United States

Mats J. Stensrud MATS.STENSRUD@EPFL.CH
Department of Mathematics
Ecole Polytechnique Fédérale de Lausanne
Lausanne, 1015, Switzerland

Editor: Ilya Shpitser

Abstract

We present new results on average causal effects in settings with unmeasured exposure-outcome confounding. Our results are motivated by a class of estimands, e.g., frequently of interest in medicine and public health, that are currently not targeted by standard approaches for average causal effects. We recognize these estimands as queries about the average causal effect of an *intervening* variable. We anchor our introduction of these estimands in an investigation of the role of chronic pain and opioid prescription patterns, and illustrate how conventional approaches will lead to non-replicable estimates with ambiguous policy implications. We argue that our alternative effects are replicable and have clear policy implications, and furthermore are non-parametrically identified by the classical frontdoor formula. As an independent contribution, we derive a new semiparametric efficient estimator of the frontdoor formula with a uniform sample boundedness guarantee. This property is unique among previously-described estimators in its class, and we demonstrate superior performance in finite-sample settings. The theoretical results are applied to data from the National Health and Nutrition Examination Survey.

Keywords: Causal inference; Double robustness; Estimands; Frontdoor formula; Intervening variable; Separable effect

1. Introduction

Unmeasured confounding and ill-defined interventions are major contributing factors to the replication crisis for policy-relevant parameters, e.g., in medical research. When there is unmeasured confounding, standard covariate-adjustment approaches will lead to biases that would likely differ across studies. Further, when an analysis is based on variables that do not correspond to well-defined interventions, it is nearly impossible for future analyses and experiments to ensure that these variables are operationalized identically with the original study. In each case, replication is infeasible.

To counteract these challenges, there exists a diverse set of strategies to confront unmeasured confounding between exposure and outcome, which leverage the measurement of auxiliary variables, including instruments and other proxies (Angrist et al., 1996; Lipsitch et al., 2010; Tchetgen Tchetgen et al., 2020), as well as mediators (Pearl, 2009; Fulcher et al., 2020). At the same time, there is a long history of calls for an "interventionist" approach to causal analyses in statistics (Holland, 1986; Dawid, 2000; Richardson and Robins, 2013; Robins et al., 2022) wherein investigators focus on the effects of manipulable, or *intervenable*, variables. Use of such variables ensures that the targets of inference are clearly defined by interventions that can be implemented in principle (see, e.g., Hernán, 2005; Hernán and VanderWeele, 2011; Galea, 2013). Furthermore, the seemingly disparate challenges of ill-defined interventions and unmeasured exposure-outcome confounding are often considered separately, but systematically co-occur in practice: when an exposure variable does not correspond to a well-defined intervention, investigators will often face exceptional challenges in sufficiently describing and measuring the causes of that exposure. In such cases, investigators will often also have little confidence in the assumption of no unmeasured confounding (Hernán and Taubman, 2008).

In this article we consider jointly the twin challenges of ill-defined interventions and unmeasured confounding. Building on new results for a generalized theory of separable effects (Robins and Richardson, 2010; Shpitser and Sherman, 2018; Robins et al., 2022; Stensrud et al., 2021), our contributions concern effects of an *intervening* variable: a manipulable descendant of a (possibly ill-defined) exposure (or treatment) that precedes the outcome. We argue that such average causal effects, rather than those of (possibly ill-defined) exposures, are frequently of interest in practice. We give results on their interpretation, identification and estimation. In doing so, we develop a novel semiparametric efficient estimator for the canonical front-door functional with superior finite sample performance properties compared with existing strategies.

1.1 Related approaches

Our results are related to previous work on identification of causal effects in the presence of unmeasured confounding. As we expound in Section 3, the causal effect of an *intervening* variable may be identified by the frontdoor formula, which coincidentally also allows non-parametric identification of the average causal effect (ACE) of the exposure on the

outcome, even in the presence of unmeasured exposure-outcome confounding (Pearl, 1993, 2009). However, meaningful applications of the frontdoor criterion have been scarce. One problem is that conventional application of the frontdoor criterion requires a strict exclusion restriction, which is often infeasible: an investigator must measure at least one mediator intersecting each causal pathway from the exposure to the outcome. In contrast to conventional approaches, however, the specification of *intervening* variables often renders such exclusion restrictions plausible.

Our results are also related to the work by Fulcher et al. (2020), who gave conditions under which the frontdoor formula identifies the so-called Population Intervention Indirect Effect (PIIE). Fulcher et al. (2020) interpreted the PIIE as a "contrast between the observed outcome mean for the population and the population outcome mean if contrary to fact the mediator had taken the value that it would have in the absence of exposure [page 200]." Thus, unlike the conventional identification result for the ACE, the frontdoor formula can be used to identify the PIIE even in the presence of a direct effect of the exposure on the outcome not mediated by an intermediate variable (or intermediate variables). We distinguish the assumptions for the two aforementioned estimands in Section 4. To fix ideas about the relation to previous work on the frontdoor formula, we introduce the following running example.

Example 1 (Chronic pain and opioid use, Inoue et al., 2022) Chronic pain is associated with use and abuse of opioids, which subsequently can lead to death. Moreover, chronic pain can affect mortality outside of its effect on opioid use (Dowell et al., 2016, 2022), e.g., by causing long-term stress and undesirable lifestyle changes. Inoue et al. (2022) studied the effect of chronic pain (exposure to pain versus no pain) on mortality (outcome) mediated by opioid use. Chronic pain is notoriously treatment-resistant, and unmeasured confounders between chronic pain and mortality may include social, physiologic, and psychological factors (Inoue et al., 2022). Using data from the National Health and Nutrition Examination Survey (NHANES) from 1999–2004 with linkage to mortality databases through 2015, the investigators studied a causal effect related to the PIIE. Specifically, Inoue et al. (2022) considered a "path-specific frontdoor effect" that they interpreted as "the change in potential outcomes that follows a change in the mediator (opioid) which was caused by changing the exposure."

As indicated by Inoue et al. (2022) there likely exist unmeasured common causes of the exposure and the outcome in Example 1. Relatedly, it is unclear how one could intervene on chronic pain, or even whether any intervention on chronic pain could be well-defined. Therefore, the interpretation of the PIIE in our example has dubious public health implications (Holland, 1986). In contrast, we suggest effects of *intervening* variables, which do correspond to interventions that are feasible to implement in practice. To concretely motivate our intervention, consider a doctor who determines if a patient should receive opioids to relieve symptoms. As suggested in Inoue et al. (2022), the doctor's decision to prescribe opioids could be determined by their patient's chronic pain status. Our intervention is motivated by current guidelines of the Centers for Disease Control and Prevention (CDC)

(Dowell et al., 2016, 2022; Inoue et al., 2022), which suggest that chronic pain status should no longer be used to determine opioid prescription. Thus, our interest is in evaluating the efficacy of a modified prescription policy (the policy-relevant intervention of interest) such that the doctor regards their patient as not having chronic pain in their decisions on opioid prescriptions.

A realistic and implementable example of such a modified prescription policy is one where doctors are mandated to participate in courses, e.g. as part of their Continuing Medical Education (CME), where they were trained to base their prescription decisions on what they would have done if they did not consider, or *perceive*, their patient as having chronic pain and supposed they had none. This modified policy does not require us to conceptualize interventions on a patient's chronic pain – an intervention that would have been hard, or impossible, to specify. While a doctor's beliefs about their patient's chronic pain status is not usually directly recorded in observed data, an analyst might reasonably assume a deterministic relation between a patient's chronic pain status and the *doctor's perception* of that chronic pain status. The doctor's perception of the patient's chronic pain status during prescription decision-making process is considered to be an *intervening* variable that in turn determines the doctor's prescription in the observed data.

The remainder of the article is organized as follows. In Section 2 we describe the observed and counterfactual data structure. In Section 3 we precisely define our interventionist estimand and derive identification results. In Section 4 we relate our identification results to non-parametric conditions that allow identification of relevant estimands that have been proposed in the past, and discuss the plausibility of these conditions. In Section 5 we present a new sample-bounded semiparametric estimator of the frontdoor formula. In Section 6 we give simulation results suggesting superior finite sample performance compared with existing estimators of the frontdoor formula (Fulcher et al., 2020). In Section 7, we apply our new results to study the effect of opioid prescription policies on chronic pain using data from NHANES, and discuss the practical implications in Section 8.

2. Observed and counterfactual data structure

Consider a study of n i.i.d. individuals randomly sampled from a large superpopulation. Let A denote the observed binary exposure taking values $a^\dagger$ or a° (e.g., chronic pain). Consider also M (e.g., opioid usage), a mediator variable and Y, an outcome of interest (e.g., survival at three years or five years), both of which can be binary, categorical or continuous. Furthermore, consider L, a vector of pre-exposure covariates measured at baseline, which can include common causes of A, M and Y. Without loss of generality, we suppose covariates are absolutely continuous with respect to a counting measure. However, our arguments extend to settings with continuous covariates and the Lebesgue measure. We indicate counterfactuals in superscripts; for example M^a and Y^a denote the counterfactual mediator and outcome variables if, possibly contrary to fact, the exposure had taken a value $a \in \{a^\dagger, a^\circ\}$. Extensions to discrete exposure variables with more than two levels are discussed in Appendix H.

3. Effects of the intervening variable

In the analysis of the chronic pain example, Inoue et al. (2022) concluded that their findings "highlight the importance of careful guideline-based chronic pain management to prevent death from possibly inappropriate opioid prescriptions driven by chronic pain." However, this argument does not translate to an intervention on chronic pain A; rather, we interpret their policy concern as one that directly involves a modifiable *intervening* variable: a care provider's *perception* of the patient's chronic pain in standard-of-care pain management decisions, A_M. If we define A_M as binary (taking values $a^\dagger$ or a°), and assume that a care provider's natural consideration corresponds exactly with a patient's chronic pain experience, then in this setting, $A{=}A_M$ with probability one in the observed data. Despite this feature of the observed data, we could nevertheless conceive an intervention that modifies this *intervening* variable A_M without changing the non-modifiable exposure A, and thus appease any policy concerns at the causal estimand-formulation stage of the analysis. Our consideration of the chronic pain example motivates estimands under interventions on the modifiable *intervening* variable A_M, not A. As we subsequently review sufficient conditions for identification include that (i) A_M is deterministically equal to A in the observed data, and (ii) A_M captures the effects of A on Y through M. Analogous conditions have been considered for the identification of separable effects (Robins and Richardson, 2010; Robins et al., 2022; Stensrud et al., 2021, 2022a).

Definition 1. (Average causal effect of an intervening variable) The *intervening* variable estimand is the expected counterfactual outcome $E(Y^{a_M})$, where the intervening variable a_M can take values in the sample space of A. We consider the average causal effects of interventions on an *intervening* variable A_M on an outcome Y as a contrast between $E(Y^{a_M=a^\dagger})$ and another causal estimand such as $E(Y^{a_M=a^\circ})$. More specifically, an average causal effect of interest could be $E(Y^{a_M=a^\dagger}){-}E(Y^{a_M=a^\circ})$; another average causal effect of interest could be $E(Y){-}E(Y^{a_M=a^\dagger})$.

As in the separable effects literature, our chronic pain example is amenable to graphical representation. Consider an extended causal directed acyclic graph (DAG) (Robins and Richardson, 2010), which not only includes A but also the *intervening* variable A_M. Figure 1a shows this extended DAG with node set $V{=}(U,L,A,A_M,M,Y)$. The bold arrow from A to A_M in Figure 1a represents the assumption of a deterministic relationship between A and A_M in the observed data: with probability one under $f(v)$, either $A{=}A_M{=}a^\dagger$ or $A{=}A_M{=}a^\circ$. These graphs will be used to illustrate our subsequent results, where we give formal conditions under which the effects of an intervention that sets A_M to a_M, as represented in the Single World Interventions Graph (SWIG; Richardson and Robins, 2013) Figure 1b, can be identified. We adopt the convention that the absence of an arrow in a SWIG encodes the absence of individual level effects in that context, as described in Richardson and Robins (2013). Causal effects of intervening variables are of interest beyond our running chronic pain example. In Appendix A, we consider an example on racial disparity and an obstetrics example from Fulcher et al. (2020).

Like Robins and Richardson (2010), our formalized interventions on the intervening variable is conceptually related to edge interventions described by Shpitser and Sherman (2018). To formally state our identifiability conditions of an intervening variable estimand, we first invoke the assumption of a deterministic relationship between the exposure and the intervening variable in the observed data.

Assumption 1 (Intervening variable determinism) $A{=}A_M$ *w.p.1.*

Assumption 1 implies that A will almost always equal A_M. In the chronic pain example, this will be violated if there is a large number of chronic pain diagnoses that incorrectly capture the true underlying pain status of the patients. Nevertheless, beyond patient-reported pain, doctors can utilize many tools (e.g., electromyography, nerve conduction studies, reflex and balance tests) to diagnose a patient's pain status, and thus in our context, we believe that Assumption 1 is reasonable. We also invoke a positivity condition which only involves observable laws and requires that for all joint values of L, there is a positive probability of observing $A{=}a$ and $M{=}m$, $\forall a,m$.

Assumption 2 (Positivity) $f(m,a|l){>}0$, $\forall(m,\ a,\ l){\in}\text{supp}(M,\ A,\ L)$.

Following Stensrud et al. (2021), let "(G)" refer to a future trial where A_M is randomly assigned, and consider the following dismissible component conditions.

Assumption 3 (Dismissible component conditions)

$$Y(G) \perp\!\!\!\perp A_M(G) | A(G), L(G), M(G), \tag{1}$$
$$M(G) \perp\!\!\!\perp A(G) | A_M(G), L(G). \tag{2}$$

Assumption 3 would fail if A_M exerts effects on Y not intersected by M, or when A exerts effects on M not intersected by A_M.[1] Furthermore, Assumption 3 requires that L should capture all common causes of M and Y and all common causes of A and M.[2] It is necessary to adjust for all common causes of (M,Y) and (A,M) for all estimands described herein (see Section 4 for other relevant estimands).

In the context of our chronic pain example, it is reasonable to assume that an opioid prescription depends on a doctor's perceived chronic pain status of a patient, A_M. Moreover, it is also reasonable to assume that A_M (doctor's perception of chronic pain) affects Y (mortality) only through M (opioid prescription), because a doctor's perception of a patient's chronic pain status can ultimately lead to overdosing events and thus death *only* via their decision to prescribe opioids.[3]

1. We discuss additional scenarios involving a measured recanting witness as mentioned in Remark 1 below.
2. The dismissible component conditions are related to so-called partial isolation conditions, see, e.g., (Stensrud et al., 2021) for more details.
3. A_M may also affect usage of non-opioid pain-killers (e.g., acetaminophen), but these are generally safe and unlikely to result in death.

Assumption 3 may fail if there exists a covariate, say anti-depressant medication prescription, that potentially affects opioid prescription and mortality but is omitted in the data analysis. Thus, it is important to measure all sufficient common causes of opioid prescription and mortality (and similarly, common causes of chronic pain and opioid prescription) in any data analysis of chronic pain studies.

Our results rely on distributional consistencies between the conditional distributions of Y^{a_M} and M^{a_M}, and the conditional distributions of $Y(G)$ and $M(G)$ corresponding to a future trial (G), where A_M is randomly assigned.[4] Furthermore, we use the following consistency assumption stating that the interventions on the intervening variable A_M is well-defined:

Assumption 4 (Consistency) *If* $A_M{=}a_M$*, then* $M^{a_M}{=}M,\ \ Y^{a_M}{=}Y, \forall\ a_M{\in}\text{supp}(A)$.

Theorem 1 *The average counterfactual outcome under an intervention on* A_M *is identified by the frontdoor formula* (3) *under Assumptions 1-4. That is,*

$$E(Y^{a_M=a^\dagger}){=}\sum_{m,l} f(m|a^\dagger,l)f(l)\sum_a E(Y|L{=}l,A{=}a,M{=}m)f(a|l). \tag{3}$$

Furthermore, we let $\Psi{:=}\sum_{m,l} f(m|a^\dagger,l)f(l)\sum_a E(Y|L{=}l,A{=}a,M{=}m)f(a|l)$. As we formally show in Appendix D, the causally manipulable estimand $E(Y^{a_M=a^\dagger})$ is identified by (3) even when A is a direct cause of Y, not mediated by M. See also Theorem 2 and Proposition 4 of Robins et al. (2022) for an alternative derivation of identification results of a related estimand via the extended ID algorithm of Shpitser et al. (2022),[5] in addition to identification of other interventionist estimands related to separable effects.

Henceforth, we refer to the Ψ as the generalized frontdoor formula as it allows for some baseline covariates L. Note that Fulcher et al. (2020) used the term generalized frontdoor criterion (not formula) to describe the identification conditions for the PIIE, which we discuss in Section 4. They argued that their "identification criterion generalizes Judea Pearl's front door criterion as it does not require no direct effect of exposure not mediated by the intermediate variable". However, their strategy requires an untestable cross-world exchangeability assumption (see Assumption 9, subsequently reviewed), which is strictly not necessary to identify the ACE. Thus, their approach should not be viewed as a generalization.

4. Specifically, because A_M is randomly assigned in G, we have that:

$$E(Y(G)|A(G){=}a,A_M(G){=}a,M(G){=}m,L(G){=}l){=}E(Y^{a_M=a}|A{=}a,M^{a_M=a}{=}m,L{=}l),$$
$$P(M(G){=}m|A(G){=}a,A_M(G){=}a,L(G){=}l){=}P(M^{a_M=a}{=}m|A{=}a,L{=}l).$$

One sufficient way to ensure this is to presume that participation in the trial can only influence M and Y through A_M. That is, we define a trial (G) as one that that precludes any participation effects on M and Y not mediated through A_M (see Appendix D).

5. See Shpitser (2013) and Shpitser and Sherman (2018) who also provide a complete algorithm for identifying path-specific effects with hidden variables.

To motivate our novel estimation results in Section 5, we emphasize that, in the absence of covariates L, the frontdoor formula can be expressed as

$$\begin{aligned}\Psi &:= \sum_m f(m|a^\dagger) \sum_a E(Y|A{=}a, M{=}m) f(a) \\ &= P(A{=}a^\dagger) E(Y|A{=}a^\dagger) + P(A{=}a^\circ) \sum_m E(Y|A{=}a^\circ, M{=}m) f(m|a^\dagger).\end{aligned} \tag{4}$$

Thus, $E(Y^{a_M=a^\dagger})$ is a weighted average of a conditional mean, $E(Y|A{=}a^\dagger)$, and a term that appears in a known identification formula for separable effects (i.e., $E(Y^{a_M=a^\dagger, a_Y=a^\circ}$)[6]) of A on Y (Robins and Richardson, 2010; Robins et al., 2022; Stensrud et al., 2021, 2022a). This decomposition is instrumental in formulating new semiparametric estimators, see Appendix B for further details. In Appendix B, we provide further intuition for this re-expression, and in Appendix C, we give identification and estimation results for our new causally manipulable estimand in absence of L.

4. Other estimands that can be identified using the frontdoor formula

To give context to our identification results, we review assumptions proposed by other authors sufficient for the identification of the ACE (in the absence of a direct effect; Pearl, 2009) and the PIIE (Fulcher et al., 2020), defined respectively as

$$E(Y^{a^\dagger}) \text{ vs. } E(Y^{a^\circ}), \quad \text{and} \quad E(Y) \text{ vs. } E(Y^{M^{a^\dagger}, A}).$$

4.1 Identification of the ACE in the absence of direct effects

The ACE is identified by the generalized frontdoor formula (3) under Assumptions 2 and 5-8, detailed below.

Assumption 5 (Consistency) *If $A{=}a$ and $M{=}m$, then*

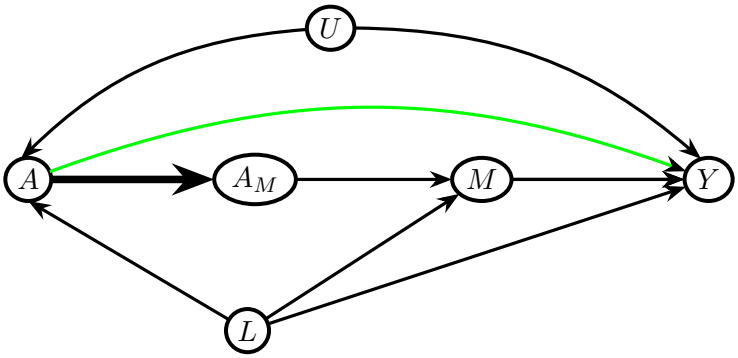
$$M^a{=}M, \quad Y^a{=}Y, \quad Y^{a,m}{=}Y, \; \forall \; (a,\, m) \in \text{supp}(A,\, M).$$

Assumption 6 (Exposure – Mediator Exchangeability)

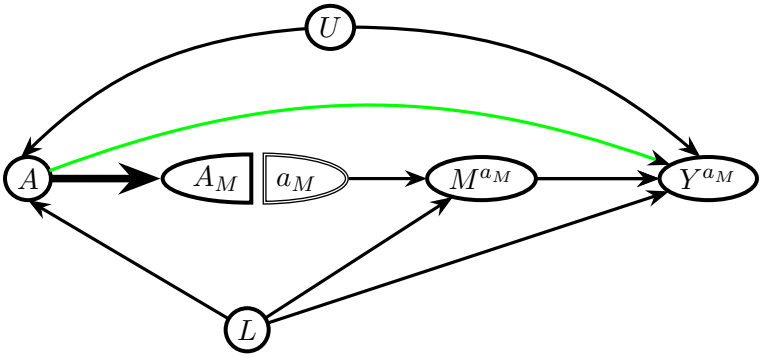
$$M^a \perp\!\!\!\perp A|L, \forall \; a \in \text{supp}(A).$$

Consistency Assumption 5 requires well-defined interventions on both A and M, which are arguably implausible in Example 1. Exchangeability Assumption 6 ensures that the

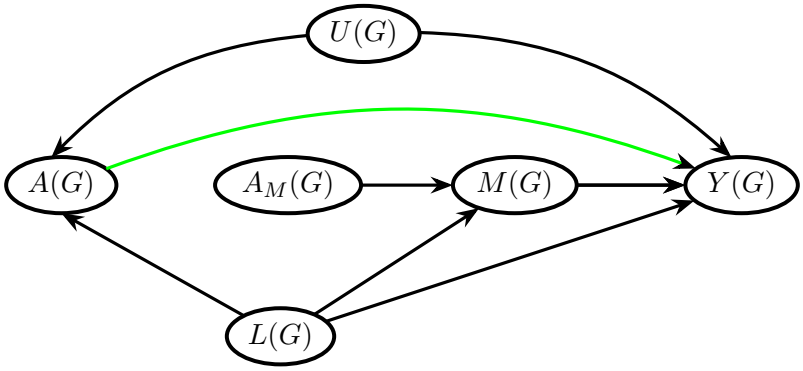
6. This estimand can be identified from the observed data under assumptions that are described in an extended DAG, or a SWIG, which results from splitting the treatment node A into two sub-components, namely A_M and A_Y.



(a) Extended DAG.



(b) SWIG transformation of the extended graph in Figure 1a, corresponding to an intervention on A_M .



(c) DAG for reading off Assumption 3.

Figure 1

exposure-mediator association is unconfounded, which holds, for instance, in the SWIG in Figure 2c. Both assumptions are used in the identification of the ACE and the PIIE in Fulcher et al. (2020), but are not strictly necessary for frontdoor identification of the ACE (see Didelez, 2018 for assumptions without intervention on M).

Assumption 7 (No Direct Effect) $Y^{a,m}=Y^m$.

Assumption 8 (Mediator – Outcome Exchangeability) $Y^{a,m} \perp\!\!\!\perp M^a | A,L$.

Assumption 7 ensures that A only affects Y through M which holds, for instance, if the green arrow is removed from the graphs in Figure 2a–2c. Assumption 8 ensures that there is no unmeasured mediator-outcome confounding, which holds, for instance, in the SWIG in Figure 2c. Together, Assumptions 2 and 5–8 allow unmeasured confounders between exposure and outcome given measured covariates such that $Y^a \not\perp\!\!\!\perp A|L$ as illustrated in the SWIG in Figure 2b (see Appendix B for details).

4.2 Identification of the Population Intervention Indirect Effect

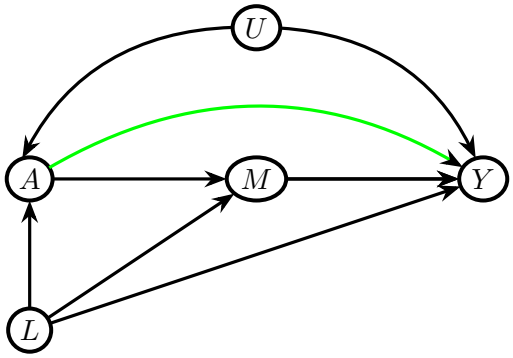
The PIIE (Fulcher et al., 2020) is a contrast between the observed mean outcome and the counterfactual outcome if, possibly contrary to fact, the mediator had taken the value that it would have when exposure equals $a^\dagger$. An example of such a contrast is $E(Y)-E(Y^{M^{a^\dagger},A})$. In the context of the chronic pain example, $E(Y^{M^{a^\dagger=0},A})=E(Y^{M^{a^\dagger=0}})$ is then interpreted as the counterfactual cumulative risk had there been no intervention on a patient's chronic pain status, but had opioid prescription been set to the value it would have taken had chronic pain been *eliminated* in all patients. This is thus considered a cross-world counterfactual estimand. Inoue et al. (2022) studied a closely related estimand, $E(Y^{M^{a^\dagger},A})-E(Y^{M^{a^\circ},A})$, which they called the path specific frontdoor effect.

Instead of imposing Assumptions 7–8, Fulcher et al. (2020) only relied on Assumptions 2, 5–6 and the following cross-world exchangeability assumption to identify the PIIE:

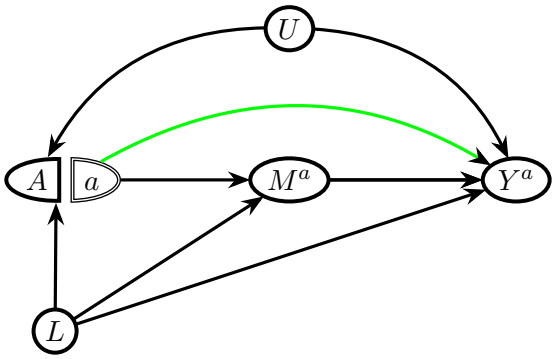
Assumption 9 (Cross-world exchangeability) $Y^{a,m} \perp\!\!\!\perp M^{a^\dagger} | A,L$ *for all values of* $a,a^\dagger,m$.

Assumption 6 is a statement about the absence of discernible single-world confounding between M and A given L, while Assumption 9 is a statement about the absence of indiscernible cross-world confounding between Y and M conditional on A and L.

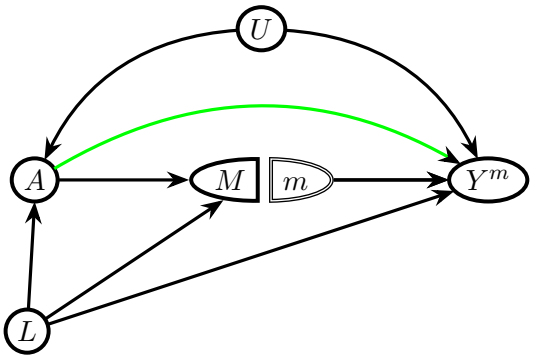
While the identification strategy proposed by Fulcher et al. (2020) permits the presence of unmeasured common causes of exposure and the outcome, and a direct effect of the exposure on the outcome, their remaining assumptions are not innocuous. First, consistency (Assumption 5) is likely to be violated in our running example; for instance, it is not clear



(a) DAG of the observed random variables.



(b) SWIG of the counterfactual variables corresponding to Figure 2a with intervention on treatment or exposure A.



(c) SWIG of the counterfactual variables corresponding to Figure 2a with intervention on mediator variable M.

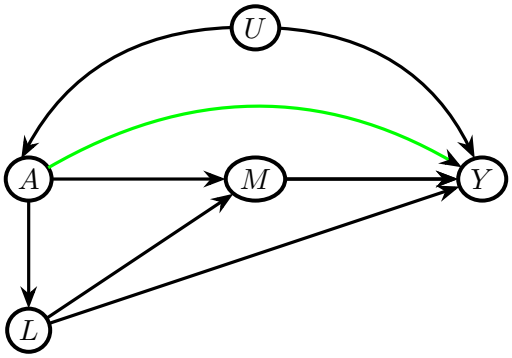
Figure 2: DAG and SWIGs of random variables under which average counterfactual outcomes can be identified.

how to specify a well-defined intervention on chronic pain. Second, cross-world independence assumptions like Assumption 9 are untestable even in principle (Robins and Richardson, 2010; Dawid, 2000): the PIIE is a cross-world estimand that cannot be realized in a real-world (Richardson and Robins, 2013). Because the PIIE can never be directly observed, it is not clear how this estimand can be used as a justification for real-world decisions. This is reflected in our difficulty in finding any correspondence between the PIIE and the substantive questions that were of direct interest to the investigators in our example. More generally, in settings where cross-world estimands often are advocated, other estimands seem to better correspond to questions of practical interest. For example, Robins and Richardson (2010) illustrated that the practical relevance of natural (pure) effects tend to be justified by interventionist estimands – defined by interventions on modified treatments – using an example on smoking and nicotine. Several other examples have since been given in different settings where other causal effects, such as principal stratum estimands, have been advocated in the past (Didelez, 2018; Stensrud et al., 2022a). Analogous to Robins and Richardson (2010), we believe that the ACE of an intervening variable is arguably the interventionist estimand of actual interest whenever the PIIE is advocated.

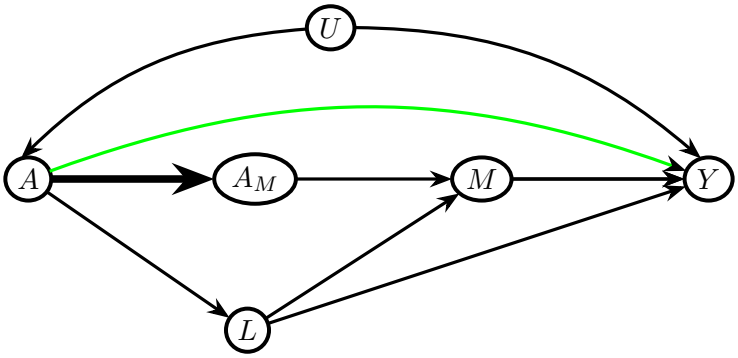
Instead of assuming well-defined interventions on an unmodifiable exposure (Assumption 5) for a parameter only identified under empirically unverifiable cross-world assumptions (Assumption 9), we consider an intervention on a modifiable exposure A_M that is identifiable under conditions that, in principle, are empirically testable (Robins and Richardson, 2010; Robins et al., 2022; Stensrud et al., 2022a).

Remark 1 As the identifying formula for the PIIE and our proposed estimand coincide under the intersection of the model in Theorem 1 and that proposed by (Fulcher et al., 2020), there is clearly a close connection between identification results for our proposed estimands and the PIIE, specifically when there is no so-called *recanting witness* (Avin et al., 2005). However, when there exists a recanting witness, the PIIE is no longer identified,[7] yet contrasts defined by $E(Y^{a_M=a^\dagger})$ can still be identified and meaningfully interpreted.[8] Such a scenario is given by the DAGs in Figure 3. These distinctions in identifiability at laws in DAG models with a recanting witness have practical implications, as we will illustrate with our chronic pain example. Specifically, the use of *Selective Serotonin Reuptake Inhibitors (SSRIs) antidepressants for clinically diagnosed depression* (L) could be a plausible recanting witness: suppose that the example is described by the DAG in Figure 3, where a patient's chronic pain status can affect SSRI use through diagnosed clinical depression, but SSRI use is unaffected by the doctor's perception of chronic pain (A_M). Then, SSRI is a recanting witness and thus the PIIE is not identified, but $E(Y^{a_M=a^\dagger})$ is still identified under our set of assumptions. In particular, under the intervention defining our estimand, a doctor's decision-making will proceed under the perception that the patient's chronic pain is absent while factoring in the patient's (factual) SSRI usage. More broadly, this feature

7. The reason being that the PIIE involves $E(Y^{A,M^{a^\dagger}})$=$E(Y^{A,L^A,M^{a^\dagger},L^{a^\dagger}})$, which involves both L and $L^{a^\dagger}$, and neither can be eliminated in the identifying formula (see also Avin et al., 2005; Robins and Richardson, 2010; Robins et al., 2022).
8. This is because Assumption 4 (Dismissible component conditions), still holds in this setting.



(a) DAG of the observed random variables with a recanting witness.



(b) Modified extended DAG with a recanting witness.

Figure 3: DAG and extended DAG with a recanting witness L.

illustrates the *interventionist* interpretation of $E(Y^{a_M=a^\dagger})$, which is analogous to motivations for separable effects (Robins and Richardson, 2010; Didelez, 2019; Robins et al., 2022; Stensrud et al., 2022b,a). These estimands can have an unambiguous interventionist interpretation, regardless of the existence of, e.g., mediators and recanting witnesses, and are meaningful without any engagement with conventional mediation analysis. As the example of antidepressant use illustrates, the existence of a recanting witness is not a problem for the interpretation of the intervention on A_M. In this case, it would not even hinder identification.

5. New estimators based on new representations of the efficient influence function of the frontdoor formula

Because our proposed estimand $E(Y^{a_M})$ is identified by the generalized frontdoor formula, we can apply existing estimators, such as the Augmented Inverse Probability Weighted (AIPW) semiparametric estimator in Fulcher et al. (2020). However, existing approaches do not guarantee that estimates are bounded by the parameter space. Such estimators have been shown to result in poor performance in finite samples (Kang and Schafer, 2007; Robins et al., 2007). This is a concern in Example 1, where the outcome is a binary indicator of

mortality. Hence, here we develop new semiparametric estimators that are guaranteed to be bounded by the parameter space.

Suppose that the observed data $\mathcal{O}=(L,A,M,Y)$ follow a law P which is known to belong to a model $\mathcal{M}=\{P_\theta:\theta\in\Theta\}$, where Θ is the parameter space. The efficient influence function $\varphi^{\text{eff}}(\mathcal{O})$ for a causal parameter $\Psi\equiv\Psi(\theta)$ in a non-parametric model $\mathcal{M}_{\text{np}}$ that imposes no restrictions on the law of $\mathcal{O}$ other than positivity is given by $d\Psi(\theta_t)/dt|_{t=0}=E\{\varphi^{\text{eff}}(\mathcal{O})S(\mathcal{O})\}$, where $d\Psi(\theta_t)/dt|_{t=0}$ is known as the pathwise derivative of the parameter Ψ along any parametric submodel of the observed data distribution indexed by t, and $S(\mathcal{O})$ is the score function of the parametric submodel evaluated at $t=0$ (Newey, 1994; Van Der Vaart, 2000).

We will first re-express our causal estimand – identified by the generalized frontdoor formula – as a weighted average of two terms given by ψ_2 and ψ_3:

$$\Psi:=P(A=a^\dagger)\underbrace{E(Y|A=a^\dagger)}_{\psi_2}+P(A=a^\circ)\underbrace{\sum_{m,l}E(Y|M=m,L=l,A=a^\circ)f(m|a^\dagger,l)f(l|a^\circ)}_{\psi_3}.$$

Thus, using differentiation rules (Van Der Vaart, 2000; Ichimura and Newey, 2022; Kennedy, 2022), the efficient influence function of Ψ can be realized by finding the efficient influence function of the following parameters: $\psi_1:=P(A=a)$; $\psi_2:=E(Y|A=a^\dagger)$; and $\psi_3:=\sum_{m,l}E(Y|M=m,L=l,A=a^\circ)f(m|a^\dagger,l)f(l|a^\circ)$. We can view ψ_3 as the identifying formula for $E(Y^{a_M=a^\dagger,a_Y=a^\circ}|A=a^\circ)$, a conditional mean that would appear in estimands for separable effects in the treated, which is identified in an extended DAG in the absence of U (see Robins and Richardson, 2010; Robins et al., 2022; Stensrud et al., 2021, 2022a). The following Theorem motivates semiparametric estimators based on this particular weighted representation of Ψ.

Theorem 2 *The efficient influence function $\varphi^{eff}(\mathcal{O})$ of the generalized front-door formula in $\mathcal{M}_{np}$ for $\mathcal{O}=(L,A,M,Y)$ is given by*

$$\varphi^{eff}(\mathcal{O})=I(A=a^\dagger)Y+I(A=a^\circ)\psi_3+\frac{I(A=a^\circ)P(A=a^\dagger|M,L)P(A=a^\circ|L)}{P(A=a^\circ|M,L)P(A=a^\dagger|L)}\{Y-b_0(M,L)\}+$$
$$\frac{I(A=a^\dagger)P(A=a^\circ|L)}{P(A=a^\dagger|L)}\{b_0(M,L)-h_\dagger(L)\}+I(A=a^\circ)\{h_\dagger(L)-\psi_3\}-\Psi, \quad (5)$$

where in the equation, $b_0(M,L):=E(Y|M,L,A=a^\circ)$, $h_\dagger(L):=E(b_0(M,L)|A=a^\dagger,L)$ and the expression in blue is proportional to the efficient influence function for ψ_3. The efficient influence function (5) can also be re-expressed as

$$\varphi^{eff}(\mathcal{O}) := I(A=a^\dagger)Y + I(A=a^\circ)\psi_3 + \frac{I(A=a^\circ)f(M|A=a^\dagger,L)}{f(M|A=a^\circ,L)}\{Y - b_0(M,L)\} + \frac{I(A=a^\dagger)P(A=a^\circ|L)}{P(A=a^\dagger|L)}\{b_0(M,L) - h_\dagger(L)\} + I(A=a^\circ)\{h_\dagger(L) - \psi_3\} - \Psi. \quad (6)$$

A proof can be found in Appendix E. Algebraic manipulation of (6) obtains the efficient influence function representation given in Equation (5) in Theorem 1 of Fulcher et al. (2020). As such, for any regular and asymptotically linear estimator $\hat{\Psi}$ of Ψ in $\mathcal{M}_{\text{np}}$, it must be that $\sqrt{n}(\hat{\Psi} - \Psi) = n^{-1/2}\sum_{i=1}^n \varphi^{\text{eff}}(\mathcal{O}_i) + o_p(1)$. Furthermore, all regular and asymptotically linear estimators in $\mathcal{M}_{\text{np}}$ with efficient influence function equaling to $\varphi^{\text{eff}}(\mathcal{O})$ are asymptotically equivalent and attain the semiparametric efficiency bound (Bickel et al., 1998).

5.1 Semiparametric efficient estimators for the generalized frontdoor formula

Writing the efficient influence function for the generalized frontdoor formula (Ψ) given in Expressions (5) or (6) motivates estimators that guarantee sample-boundedness. A weighted iterative conditional expectation (Weighted ICE) estimator with this property is presented in Algorithm 1. In what follows, we let $\mathbb{P}_n(X) := n^{-1}\sum_{i=1}^n X_i$ and let g^{-1} denote a known inverse link function satisfying $\inf(\mathbf{Y}) \leq g^{-1}(u) \leq \sup(\mathbf{Y})$, for all u, where $\mathbf{Y}$ is the sample space of Y (e.g., a logit link for dichotomous Y).[9]

Algorithm 1 Algorithm for Weighted ICE (generalized frontdoor formula)

1: Non-parametrically compute $P(A=a^\circ)$ and $P(A=a^\dagger)$.

2: Compute the maximum likelihood estimate (MLE) $\hat{\kappa}$ of κ from the observed data for the exposure or treatment model $P(A=a|L;\kappa)$. In addition, compute the MLE $\hat{\alpha}$ of α from the observed data for the exposure or treatment model $P(A=a|M,L;\alpha)$, or compute the MLE $\hat{\gamma}$ of γ from the observed data for the mediator model $P(M=m|A,L;\gamma)$.

3: In the individuals whose $A=a^\circ$, fit a regression model $Q(M,L;\theta) := g^{-1}\{\theta^T\phi(M,L)\}$ for $b_0(M,L)$ where the score function for each observation is weighted by $\hat{W}_1$. Here, $\hat{W}_1$ equals

$$\frac{P(A=a^\circ|L;\hat{\kappa})P(A=a^\dagger|M,L;\hat{\alpha})}{P(A=a^\dagger|L;\hat{\kappa})P(A=a^\circ|M,L;\hat{\alpha})}$$

if $\hat{\alpha}$ was estimated in the previous step, or $\hat{W}_1$ equals

$$\frac{f(M|A=a^\dagger,L;\hat{\gamma})}{f(M|A=a^\circ,L;\hat{\gamma})}$$

9. For bounded continuous outcome, obtain a transformed outcome $Y^* = \frac{Y-b}{c-b}$, where b and c denote the minimum and maximum of $\mathbf{Y}$, respectively. The algorithm proceeds with Y^* in place of Y, and the estimate obtained from Step 7 of the algorithm is transformed back to the original scale by multiplying it by $(c-b)$ and adding b; see also Gruber and van der Laan, 2010.

if $\hat{\gamma}$ was estimated in the previous step, and $\phi(M,L)$ is a known function of M and L. More specifically, we solve for θ in the following estimating equations:

$$\mathbb{P}_n\Big[I(A=a^\circ)\phi(M,L)\hat{W}_1\{Y-Q(M,L;\theta)\}\Big]=0.$$

4: In those whose $A=a^\dagger$, fit a model $R(L;\eta)\coloneqq g^{-1}\{\eta^T\Gamma(L)\}$ for $h_\dagger(L)$ where the score function for each observation is weighted by

$$\hat{W}_2\coloneqq\frac{P(A=a^\circ|L;\hat{\kappa})}{P(A=a^\dagger|L;\hat{\kappa})}.$$

Here, $\Gamma(L)$ is a known function of L. More specifically, we solve for η in the following estimating equations:

$$\mathbb{P}_n\Big[I(A=a^\dagger)\Gamma(L)\hat{W}_2\Big\{Q(M,L;\hat{\theta})-R(L;\eta)\Big\}\Big]=0.$$

5: In those whose $A=a^\circ$, fit an intercept-only model $T(\beta)\coloneqq g^{-1}(\beta)$ for ψ_3. More specifically, we solve for β in the following estimating equations:

$$\mathbb{P}_n[I(A=a^\circ)\{R(L;\hat{\eta})-T(\beta)\}]=0.$$

6: Compute $\hat{T}\coloneqq T(\hat{\beta})$ for all observations.
7: Estimate $\hat{\Psi}_{WICE}\coloneqq\mathbb{P}_n\{I(A=a^\dagger)Y+I(A=a^\circ)\hat{T}\}$.

In Algorithm 1, steps 3, 4 and 5 ensure that the estimates for ψ_3 are sample bounded. Moreover, in Step 6 it is clear that $\hat{\Psi}_{WICE}$ is a convex combination of Y and estimates for ψ_3, both of which are bounded by the range of the outcome Y. Thus, $\hat{\Psi}_{WICE}$ will also be sample-bounded. In Appendix F, we prove that the proposed estimator, which is based on the efficient influence function given by (5) and (6), is robust against 3 classes of model misspecification scenarios.

Theorem 3 *Under standard regularity conditions, the weighted ICE estimator $\hat{\Psi}_{WICE}$ where a model for $P(A=a|M,L)$ is specified will be consistent and asymptotically normal under the union model $\mathcal{M}_{union}=\mathcal{M}_1\cup\mathcal{M}_2\cup\mathcal{M}_3$ where we define:*

1. *Model $\mathcal{M}_1$: working models for $P(A=a|M,L)$ and $P(A=a|L)$ are correctly specified.*

2. *Model $\mathcal{M}_2$: working models for $b_0(M,L)$ and $h_\dagger(L)$ are correctly specified.*

3. *Model $\mathcal{M}_3$: working models for $b_0(M,L)$ and $P(A=a|L)$ are correctly specified.*

Moreover, $\hat{\Psi}_{WICE}$ is locally efficient in the sense that it achieves the semiparametric efficiency bound for Ψ in $\mathcal{M}_{np}$, i.e., $E\big[\varphi^{eff}(\mathcal{O})^2\big]$, at the intersection model given by $\mathcal{M}_{intersection}=\mathcal{M}_1\cap\mathcal{M}_2\cap\mathcal{M}_3$.

Alternatively, the weighted ICE estimator $\hat{\Psi}_{WICE}$ where a model for $P(M{=}m|A,L)$ is specified will be consistent and asymptotically normal under the union model $\mathcal{M}_{union}{=}\mathcal{M}_1{\cup}\mathcal{M}_2{\cup}\mathcal{M}_3$ where we define (1) Model $\mathcal{M}_1$: working models for $P(M{=}m|A,L)$[10] and $P(A{=}a|L)$ are correctly specified; (2) Model $\mathcal{M}_2$: working models for $b_0(M,L)$ and $h_\dagger(L)$ are correctly specified; and (3) Model $\mathcal{M}_3$: working models for $b_0(M,L)$ and $P(A{=}a|L)$ are correctly specified. Furthermore, this $\hat{\Psi}_{WICE}$ is also locally efficient in the sense that, when all working models are correctly specified, it achieves the semiparametric efficiency bound for estimating Ψ in $\mathcal{M}_{np}$.

Remark 2 It is possible that the weighted ICE estimator is consistent when the models for $b_0(M,L)$ and $E(M|A,L)$ are correctly specified. For instance, this would be the case when Y and M are continuous, in which case the model $R(L;\eta)$ for $h_\dagger(L)$ can also be correctly specified in this model specification scenario.

Unlike the estimator proposed by Fulcher et al. (2020), the estimator proposed here requires specification of four models instead of three, and thus our proposed estimator offers a different robustness property against model misspecification compared to that of the AIPW estimator in Fulcher et al. (2020). Consequently, the AIPW estimator in Fulcher et al. (2020), which requires specification of only three (vs. four) working models, is robust against two (vs. three) classes of model misspecification scenarios. Specifically, it will be consistent when at least (1) the models for $b_0(M,L)$ and $P(A{=}a|L)$ are correctly specified, *or* (2) the model for $P(M{=}m|A,L)$ is correctly specified. In Appendix G, we describe an iterative algorithm that preserves the double robustness property of Fulcher et al. (2020) for binary mediators. However, we do not pursue this iterative algorithm in our main manuscript as it is more computationally intensive and can run into convergence issues in finite samples (see van der Laan and Gruber, 2009; Van der Laan and Rose, 2018).

We believe that our proposed estimator will be particularly useful in cases where (i) M is a continuous variable, as in the 'Safer deliveries' application in Fulcher et al. (2020), and/or (ii) there are multiple mediator variables (M_1, M_2, $M_3\ldots$). In both of these cases, the model(s) for the mediator variable(s) will be difficult to correctly specify. In Appendix F, we prove that in the absence of L our weighted ICE estimator based on the efficient influence function for the (non-generalized) frontdoor formula is doubly robust in the sense that it will be consistent as long as the model for $P(A{=}a|M)$ – or $P(M{=}m|A)$, depending on the representation – or the model for $E(Y|A{=}a^\circ,M)$ is correctly specified. As such, for the (non-generalized) frontdoor formula in the absence of L, the double robustness property of our estimator is the same as that of Fulcher et al. (2020).

We also describe a targeted maximum likelihood estimator (TMLE) for Ψ, which is a variation of the weighted ICE estimator given above. The TMLE can accommodate complex machine learning algorithms for estimating all nuisance functions (Van der Laan and Rose, 2011; Chernozhukov et al., 2018; Wen et al., 2023).

10. or model for the density ratio of $P(M{=}m|A{=}a^\dagger,L)/P(M{=}m|A{=}a^\circ,L)$

Algorithm 2 Algorithm for Targeted maximum likelihood (generalized frontdoor formula)

1: Non-parametrically compute $P(A{=}a^{\circ})$ and $P(A{=}a^{\dagger})$.
2: Obtain estimates $\hat{P}(A{=}a|L)$, $\hat{P}(A{=}a|M,L)$ (or $\hat{P}(M{=}m|A,L)$) of $P(A{=}a|L)$, $P(A{=}a|M,L)$ (or $P(M{=}m|A,L)$), respectively, possibly using machine learning methods.
3: A. In the individuals whose $A{=}a^{\circ}$, compute $\hat{Q}(M,L)$ by regressing Y on (M,L). Here, $\hat{Q}(M,L)$ is possibly estimated using machine learning methods, and it denotes an initial estimate for $b_0(M,L)$.

B. Update the previous regression. Specifically, in the individuals whose $A{=}a^{\circ}$, fit a intercept-only regression model $Q^*(M,L;\delta){:=}g^{-1}[g\{\hat{Q}(M,L)\}{+}\delta]$ where the score function for each observation is weighted by $\hat{W}_1$ (defined previously). More specifically, we solve for δ in the following estimating equations:

$$\mathbb{P}_n\Big[I(A{=}a^{\circ})\hat{W}_1\{Y{-}Q^*(M,L;\delta)\}\Big]{=}0.$$

4: A. In those whose $A{=}a^{\dagger}$, compute $\hat{R}(L)$ by regressing $Q^*(M,L;\hat{\delta})$ (as outcome, obtained from last step) on L. Here, $\hat{R}(L)$ is possibly estimated using machine learning methods, and it denote an initial estimate of $h_{\dagger}(L)$.

B. Update the previous regression. Specifically, in those whose $A{=}a^{\dagger}$, fit an intercept-only regression model $R^*(L;\nu){:=}g^{-1[}\{g\{\hat{R}(L)\}{+}\nu]$ where the score function for each observation is weighted by $\hat{W}_2$ (defined previously). More specifically, we solve for ν in the following estimating equations:

$$\mathbb{P}_n\Big[I(A{=}a^{\dagger})\hat{W}_2\Big\{Q^*(M,L;\hat{\delta}){-}R^*(L;\nu)\Big\}\Big]{=}0.$$

5: In those whose $A{=}a^{\circ}$, fit another regression model $T(\beta){:=}g^{-1}(\beta)$ for ψ_3 with just an intercept. More specifically, we solve for β in the following estimating equations:

$$\mathbb{P}_n[I(A{=}a^{\circ})\{R^*(L;\hat{\nu}){-}T(\beta)\}]{=}0.$$

6: Compute $\hat{T}{:=}T(\hat{\beta})$ for all observations.
7: Estimate $\hat{\Psi}_{TMLE}{:=}\mathbb{P}_n\{I(A{=}a^{\dagger})Y{+}I(A{=}a^{\circ})\hat{T}\}$.[11]

Theorem 4 (Weak convergence of TMLE) *Suppose that the conditions given in Appendix J hold, and further suppose that the following condition also holds:*

$$\|\hat{h}_{\dagger}(L){-}h_{\dagger}(L)\|\ \|\hat{f}(A|L){-}f(A|L)\|{+}\|\hat{b}_0(M,L){-}b_0(M,L)\|\ \|\hat{f}(A|L){-}f(A|L)\|{+}$$
$$\|\hat{b}_0(M,L){-}b_0(M,L)\|\ \|\hat{f}(A|M,L){-}f(A|M,L)\|{=}o_p(n^{-1/2}),$$

where $\|f(x)\|_2{=}\big\{\int|f(x)|^2dP(x)\big\}^{1/2}$, *i.e. the* $L_2(P)$ *norm. Then,*

$$\sqrt{n}\Big(\widehat{\Psi}_{TMLE}{-}\Psi\Big)\rightsquigarrow N(0,\sigma^2),\quad \textit{where } \sigma^2{=}Var(\varphi^{\textit{eff}}(\mathcal{O})).$$

11. See Footnote 8 on straightforward extensions to bounded continuous outcomes.

When the nuisance functions are estimated with machine learning algorithms, the variance of the TMLE can be estimated empirically with $\mathbb{P}_n\left[\left(\hat{\varphi}^{\text{eff}}(\mathcal{O})\right)^2\right]$, where all nuisance functions are replaced with their estimates. When the nuisance functions are estimated using parametric models, this variance estimator remains valid as long as all nuisance functions are correctly specified. However, in practice, we recommend using the non-parametric bootstrap to estimate the variance, as parametric models are more prone to misspecification.

The advantage of using TMLE (or the AIPW estimator of Fulcher et al., 2020) with machine learning algorithms for the nuisance functions is that the estimator is still consistent and asymptotically normal as long as the nuisance functions converge to the truth at rates faster than $n^{-1/4}$ (Robins et al., 2008; Chernozhukov et al., 2018) (see Appendix J). Nevertheless, in real world applications there is no guarantee that such rates of convergence can be attained when more flexible algorithms such as neural network or random forest are used. Moreover, there is no guarantee that these machine learning methods will exhibit more or less bias compared with parametric models (Liu et al., 2020).

6. Simulation study

We conducted a simulation study to demonstrate that (1) our estimand, like the PIIE, is robust to unmeasured confounding between exposure and outcome, (2) our proposed estimator is more robust to model misspecification compared with estimators such as the Inverse Probability Weighted (IPW) estimator and the ICE estimator (described in Appendix G), and (3) unlike the AIPW estimator of Fulcher et al. (2020) our estimator is sample bounded in any finite sample size setting.

The simulation study was based on 1000 simulated data sets of sample sizes n=100, 250 and 500. We compared the bias, standardized bias and efficiency of IPW, ICE, AIPW and weighted ICE estimators. Standardized bias is 100×[(Average Estimate−True Parameter)/empirical standard deviation of the parameter estimates]. Larger standardized bias will have a bigger impact on efficiency, coverage, and error rates. It has been suggested that for n=500, anything greater than an absolute standardized bias of approximately 40% will have a 'noticeable adverse impact on efficiency, coverage, and error rates' (Schafer and Kang, 2008; Collins et al., 2001). The true value of Ψ=0.0144 was calculated by generating a Monte Carlo sample of size 10^7. We intentionally chose a rare outcome to compare the performance of the estimators where the AIPW estimator had a non-trivial chance of falling outside of the parameter space.

The data-generating mechanism for our simulations and model specifications are provided in Table 1. We consider four scenarios to illustrate the robustness of our proposed estimator to model misspecification: (1) all models are approximately correctly specified, (2) only the models for $b_0(M,L)$ and $h_\dagger(L)$ are approximately correctly specified, (3) only the models for $b_0(M,L)$ and $P(A{=}a|L)$ are approximately correctly specified, and (4) only the model for $P(M{=}m|A,L)$ is correctly specified *and* the model for $P(A{=}a|L)$ is approximately correctly specified. The correct mediator model in the specification scenarios is the one used

in the data generation process, and the exposure and outcome models are approximately correctly specified by including pairwise interactions between all the variables to ensure model flexibility.

Table 2 shows the results from the simulation study. Consistent with our theoretical derivations, when all of the working models are (approximately) correctly specified, all of the estimators become nearly unbiased as the sample size increases. For n=500, the AIPW estimator and our proposed weighted ICE estimator are also nearly unbiased in the three model misspecification settings whereas the IPW and ICE estimators are not all unbiased. In Appendix I, we show an additional simulation study where the variables are all binary and thus the correct exposure and outcome models are saturated models that cannot be misspecified. The results further show the robustness of our estimator in model misspecification scenarios.

Finally, ICE, AIPW and weighted ICE estimators have comparable standard errors when all of the models are correctly specified, but all three had lower standard errors compared with IPW. However, our results also show that AIPW estimator has poorer finite sample performance when n=100 compared with the weighted ICE estimator. Moreover, for n= 100, there were 90, 59 and 88 simulated data sets where estimates from the AIPW estimator fell below 0 in scenarios 1, 3 and 4, respectively.

Data generating mechanism	
$U\sim$	Ber(0.5)
$L_1\sim$	Normal(0,1)
$L_2\sim$	Ber$\{$expit$(1+2L_1)\}$
$A\sim$	Ber$\{$expit$(-1-3L_1+L_2+5L_1L_2+2U)\}$
$M\sim$	Ber$\{$expit$(1-A-2L_1+2L_2+3L_1L_2)\}$
$Y\sim$	Ber$\{$expit$(-4+2A+M-2AM+2L_1-2L_2-5L_1L_2-U)\}$
Model misspecification	
Scenario 2	$P(M=m\|L,A;\gamma)$=expit$(\gamma_0+\gamma_1A+\gamma_2L_1+\gamma_3L_2)$ $P(A=a\|L;\kappa)$=expit$(\kappa_0+\kappa_1L_1+\kappa_2L_1^2)$
Scenario 3	$P(M=m\|L,A;\gamma)$=expit$(\gamma_0+\gamma_1A+\gamma_2L_2)$ $R(L;\theta)$=expit$(\eta_0+\eta_1L_2)$
Scenario 4	$Q(M,L;\theta)$=expit$(\theta_0+\theta_1M+\theta_2L_1+\theta_3L_2)$ $R(L;\theta)$=expit$(\eta_0+\eta_1L_2(1-L_1))$

Table 1: Data generating mechanism and model misspecifications for scenarios in simulation study.

7. Application

Motivated by Example 1, we applied our results to study the effect of modified prescription policies for opioids on mortality in patients with chronic pain. The intervention of interest is

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Bias	SE	Bias_s	Bias	SE	Bias_s	Bias	SE	Bias_s	Bias	SE	Bias_s
n=100												
IPW	0.21	2.19	9.52	-0.23	1.38	-16.67	0.21	2.19	9.52	3.13	2.12	147.8
ICE	0.15	1.74	8.42	0.15	1.74	8.42	-0.28	1.36	-20.7	0.55	1.93	28.53
AIPW	0.49	1.94	25.03	-	-	-	0.61	2.19	27.71	0.31	1.72	18.31
WICE	0.20	1.81	10.82	0.15	1.75	8.57	0.17	1.96	8.551	0.23	1.72	13.52
n=250												
IPW	0.12	1.33	9.05	-0.33	0.86	-38.04	0.12	1.33	9.05	3.45	1.33	259.62
ICE	0.09	1.05	8.97	0.09	1.05	8.97	-0.36	0.92	-38.93	0.39	1.13	34.17
AIPW	0.16	1.07	14.89	-	-	-	0.16	1.18	13.16	0.07	0.92	7.26
WICE	0.09	1.02	9.06	0.08	1.03	7.92	0.09	1.22	7.38	0.01	0.88	1.26
n=500												
IPW	0.07	0.94	7.75	-0.40	0.54	-74.8	0.07	0.94	7.75	3.60	0.98	365.37
ICE	0.04	0.68	6.34	0.04	0.68	6.34	-0.45	0.54	-83.93	0.36	0.72	49.71
AIPW	0.04	0.68	5.73	-	-	-	0.05	0.88	5.72	0.03	0.57	4.89
WICE	0.05	0.68	6.96	0.04	0.67	6.51	0.04	0.83	5.33	0.01	0.55	1.02

Table 2: Simulation results: Bias, standard error (SE), and standardized bias (Bias_s) are multiplied by 100, and true value was Ψ=0.0144. For n=100, 90, 59 and 88 simulated data sets had estimates from the AIPW estimator that fell below 0 in scenarios 1, 3 and 4, respectively. For n=250, 16, 38 and 14 simulated data sets had estimates that fell below 0 in scenarios 1, 3 and 4, respectively. For n=500, 5, 31, and 1 simulated data set(s) had estimates that fell below 0 in scenarios 1, 3 and 4, respectively.

a new prescription policy, in which the doctor is instructed to consider each patient's chronic pain as absent for the purpose of opioid prescription decisions, and otherwise use measured covariates as usual (according to standard treatment guidelines). Thus, this policy is an intervention on a *doctor's perception of chronic pain*, A_M. This intervention is consistent with a recently proposed policy by the CDC that instructs practicing physicians to no longer consider chronic pain as an indication for opioid therapy. While the doctors' perceptions are not explicitly measured in the observed data, we may reasonably assume that, in the absence of an intervention, such perceptions perfectly correspond with the medical reality of the patient, that is, A=A_M almost surely.

We used the dataset from Inoue et al. (2022), which includes observations from the NHANES study linked to a national mortality database (National Death Index). The NHANES study consists of a series of in-depth in-person interviews, medical and physical examinations, and laboratory tests aimed at understanding various emerging needs in public health and nutrition. Sample data are from 1999–2004 and include information on individuals' chronic pain status (A), opioid prescriptions (M), mortality (Y), and covariates (L) including age, sex assigned at birth (male and female), race (non-Hispanic White, non-Hispanic Black, Mexican-American, or others), education levels (less than high school, high school or General Education Degree, or more than high school), poverty-income ratio (the ratio of household income to the poverty threshold), health insurance coverage, marital status, smoking, alcohol intake, and anti-depressant medication prescription. Let A_M denote the *intervening* variable: the doctor's perception of the patient's chronic pain status.

Our sample included 12037 individuals. Following Inoue et al. (2022), an individual is considered to have chronic pain if they reported pain for at least three months by the International Association for the Study of Pain criteria (Merskey and Bogduk, 1994). Moreover, data on prescription medications for pain relief used in the past 30 days were collected in the in-person interview. Opioids identified through the process include codeine, fentanyl, oxycodone, pentazocine and morphine. About 16% of the individuals in the sample experienced chronic pain and approximately 5% of the individuals in the sample reported using opioids. Detailed data description can be found in Inoue et al. (2022).

We estimated the cumulative incidence $E(Y^{a_M=0})$ and the causal contrast $E(Y)-E(Y^{a_M=0})$ using ICE, IPW and our weighted ICE estimator by specifying logistic regression models for the outcome, mediator and exposure. We used the same logistic regression models as those of Inoue et al. (2022) by adjusting for all the previously-listed measured covariates as potential confounders. All 95% confidence intervals were based on the 2.5 and 97.5 percentiles of a non-parametric bootstrap procedure with 1000 bootstrapped samples.

The ICE procedure estimated the $E(Y^{a_M=0})$ to be 4.76% (95% CI=(4.36, 5.16)) in three years and 8.55% (95% CI=(8.04, 9.09)) in five years. The IPW procedure estimated this cumulative incidence to be 4.95% (95% CI=(4.57, 5.35)) in three years and 8.82% (95% CI=(8.30, 9.35)) in five years. The weighted ICE procedure estimated this cumulative incidence to be 4.76% (95% CI=(4.36, 5.16)) in three years and 8.55% (95% CI=(8.04, 9.09)) in five years.

Moreover, the ICE procedure estimated $E(Y)-E(Y^{a_M=0})$ to be 0.22% (95% CI= $(-0.32,\ 0.82)$) in three years and 0.25% (95% CI=(0.09, 0.40)) in five years. The IPW procedure estimated this causal contrast to be 0.02% (95% CI=$(-0.38,\ 0.40)$) in three years and -0.03% (95% CI=$(-0.55,\ 0.49)$) in five years. Finally, the weighted ICE procedure estimated this causal contrast to be 0.22% (95% CI=$(-0.33,\ 0.82)$) in three years and 0.25% (95% CI=(0.09,0.40)) in five years.

We also applied the TMLE estimator described herein where the nuisance functions were estimated using the Super Learner ensemble (library of candidates including generalized additive models and multivariate adaptive regression Splines) and found similar results. For instance, the TMLE procedure estimated $E(Y)-E(Y^{a_M=0})$ to be 0.24% (95% CI= (0.08,0.39)) in five years.

The results from the data analysis illustrate that our causal estimand $E(Y)-E(Y^{a_M=0})$ has a clear interpretation and is practically relevant. We interpret $E(Y^{a_M=0})$ as a counterfactual cumulative incidence resulting from a policy where doctors are trained to consider their patients as not having chronic pain and fully adhere to this training. Our analysis suggests that under such an intervention, the cumulative incidence of death is almost identical to the cumulative incidence in the observed data after three years, but decreases very slightly after five years.

8. Discussion

We have derived identification results that justify the use of the frontdoor formula in new settings, reflecting questions of practical interest, where unmeasured confounding is a serious concern. Our identification results do not rely on ill-defined interventions or cross-world assumptions (see also Shpitser and Sherman, 2018; Robins et al., 2022). Specifically, we proposed an estimand defined by an intervention on a modifiable descendant of an exposure or treatment, which we call an *intervening* variable. Like the previously proposed PIIE, our proposed estimand is identified by the frontdoor formula even in the presence of a direct effect of the exposure on the outcome not mediated by an intermediate variable. But unlike the PIIE, our estimand is identifiable under conditions that, in principle, are empirically testable. In addition, we presented an example in which our proposed estimand is practically relevant. In this example, the exposure variable – chronic pain – was difficult, if not impossible, to intervene on. However, we argued that interventions on the intervening variable, rather than the exposure, are often of practical interest in settings where interventions on exposures are ill-defined.

Existing estimators of the frontdoor formula, including the AIPW estimator of Fulcher et al. (2020), can also be used to estimate estimands identified by the frontdoor formula, including the one proposed in this manuscript. When our proposed estimand is identified by the frontdoor formula in the absence of L, our proposed estimator and the existing AIPW estimator of Fulcher et al. (2020) are both doubly robust in the sense that they are consistent as long as (1) the model for $P(A{=}a|M)$ (or $P(M{=}m|A)$) is correctly specified or (2) the model for $b_0(M)$ is correctly specified. For the generalized frontdoor formula, the AIPW estimator is doubly robust when (1) the models for $b_0(M,L)$ and $P(A{=}a|L)$ are correctly specified, *or* (2) the model for $P(M{=}m|A,L)$ is correctly specified. Compared with existing AIPW estimator of Fulcher et al. (2020), the one disadvantage of our estimator for the generalized g-formula is that it requires specification of four models instead of three. Nevertheless, our proposed estimator then offers three chances for consistent estimation and is therefore triply robust. Specifically, it is consistent when (1) the models for $P(A{=}a|M,L)$ (or $P(M{=}m|A,L)$) and $P(A{=}a|L)$ are correctly specified, (2) the models for $b_0(M,L)$ and $h_\dagger(L)$ are correctly specified, or (3) the models for $b_0(M,L)$ and $P(A{=}a|L)$ are correctly specified. A key advantage of our semiparametric estimator is that estimates of Ψ are ensured to be bounded by the parameter space, regardless of the sample size and variability of inverse probability weights.

In practice it is advised to define richly parameterized models for $P(A{=}a|L)$ and $h_\dagger(L)$ to ameliorate model incompatibility issues between $P(A{=}a|L)$ and $P(A{=}a|M,L)$ and between $b_0(M,L)$ and $h_\dagger(L)$ (Vansteelandt et al., 2007; Tchetgen Tchetgen and Shpitser, 2014). However, similar to the AIPW estimator in Fulcher et al. (2020), our proposed TMLE estimator can also accommodate machine learning algorithms, which in principle can achieve $\sqrt{n}$-consistency as long as the nuisance functions converge at sufficiently fast rates. Moreover, when the mediator variable is continuous, the AIPW estimator of Fulcher et al. (2020) involves a preliminary estimator of a density ratio. Direct estimation of the density ratio is often cumbersome as correct specification of the probability density of M

is difficult. This is particularly challenging task when data-adaptive estimators are used for estimating high-dimensional nuisance parameters. Our proposal allows for an alternative estimation procedure to accommodate continuous mediator variables by modeling the propensity score of exposure/treatment instead. Finally, our semiparametric estimator can be applied whenever the frontdoor formula identifies the parameter of interest, which e.g., could be the ACE, PIIE or our interventionist estimand. Our results also motivate future methodological work. In particular, we aim to generalize our results to longitudinal settings, involving time-varying treatments.

Acknowledgments

The authors thank Dr. Kosuke Inoue for the access to the NHANES dataset used in the data application. Lan Wen is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant [RGPIN-2023-03641, DGECR-2023-00455]. Mats J. Stensrud and Aaron L. Sarvet are supported by the Swiss National Science Foundation, grant 200021_207436.

Appendix A. Additional examples of interventions on intervening variables

We illustrate another application in the following example considering race, an exposure often under-theorized in statistical analyses, and job interview discrimination.

Example 2 (Race and job interview discrimination) Consider the extended graph in Figure 1a, where A denotes the race of an individual, A_M denotes the race that an individual indicates on a job application for a company, M denotes the indicator that the individual receives an interview, and Y denotes whether an individual is hired for the job (1 if the individual is hired and 0 otherwise). Further, the unmeasured variable U includes complex historical processes that ultimately affect both a person's perceived or declared race on the job market and may influence whether or not they are hired by the company. Several studies have documented that resumé 'whitening' (e.g., deleting any references or connotations to a non-White race) can increase an applicant's chance of receiving an interviews (Kang et al., 2016; Gerdeman, 2017). However, it is unclear if this strategy leads to a higher chance of actually being hired, as e.g., unconscious bias can also affect whether or not an individual is hired, even if the candidate's background and qualifications are the same as other candidates. Thus, in a group of individuals with the same qualifications, the difference between $E(Y)$ and $E(Y^{a_M=\text{white}})$ could indicate the effect among non-White candidates of resumé whitening in the screening process on the probability of being hired.

Next, we consider an obstetrics example from Fulcher et al. (2020).

Example 3 (The safer deliveries program, (Fulcher et al., 2020)) The 'Safer deliveries' program was designed to reduce the relatively high rates of maternal and neonatal mortality in Zanzibar, Tanzania. The program provided counselling to pregnant women preparing for delivery. Women deemed to be in a high pregnancy "risk category", based on a mobile device algorithm, were instructed to deliver at a referral hospital, a specialty healthcare resource that generally incurred higher expenses for the women's family. Then, given the recommended delivery location, the mobile device algorithm also calculated an amount that they recommended the family should save in anticipation of future obstetric care costs (i.e. a "tailored savings recommendation"). At a later point in the study, the amount that the families actually saved for this purpose was recorded in the data (i.e. the "actual savings").

Using data from the 'Safer deliveries' program, Fulcher et al. (2020) aimed to evaluate "the effectiveness of this tailored savings recommendation by risk category on actual savings". They reported estimates of the PIIE of delivery risk (high risk versus low/medium risk exposure) on actual savings at delivery (outcome), mediated by a recommended savings amount calculated by the mobile device algorithm. As noted in Fulcher et al. (2020), there may be unmeasured confounding between a participant's recommended risk category and actual savings at delivery, for example by socioeconomic factors and individual's health-seeking behaviour.

Fulcher et al. (2020) argued that the PIIEs in Example A.3 was an appropriate estimand to study "the effectiveness of this tailored savings recommendation" for pregnant women. However, it is not clear that the plain English justification translates to a PIIE defined by interventions on a woman's recommended risk category, an exposure that is non-interveneable or of limited scientific interest. To interpret the results of their analysis we either have to consider:

1. obstetric risk category to be *defined* as a composite of various embodied socio-demographic and clinical features, in which case intervention on obstetric risk category can only be defined as interventions on the constituent components used to characterize the exposure. Such an intervention would be difficult to imagine as all of these embodied socio-demographic and clinical features may be hard to identify; or

2. obstetric risk category to be *defined* as a conceptually distinct feature from the measured socio-demographic and clinical features used to compute it, and thus possibly manipulable separately from these features. In this case the "risk category" variable would simply be no more or less than the computed "risk category" that appears on the screens of mobile devices, and these risk categories could be manipulated simply by intervening on the software run on these devices. However, such interventions would be of a substantively different nature with profound differences in interpretation and will have different implications for policy-makers. Moreover, the exposure "risk category" would not be susceptible to unmeasured confounding.

Considering effects of intervening variables ameliorates this ambiguity and also clarifies assumptions. Specifically, an intervention that avoids these difficulties would be to fix the *output* value from the algorithm, so that it recommends a delivery location as usual, but the patient's recommended savings amount is based on the delivery location the original algorithm would recommend if that patient had been deemed to be at low risk for obstetric complications. We can explicitly define the algorithm's computed risk category as A_M (a modifiable *intervening* variable) that is distinct from the patient's non-modifiable embodied risk category (A). In the observed data, $A{=}A_M$ with probability 1. However, we could conceive an intervention that modifies this intervening variable A_M without changing the exposure A.

Appendix B. Proofs of identification of the Frontdoor Formula

The proof for the frontdoor identifying formula (4) is given as follows:

$$\begin{aligned}
E(Y^{a^\dagger}) &= \sum_m E(Y^{a^\dagger}|M^{a^\dagger}=m)P(M^{a^\dagger}=m) \\
&= \sum_m E(Y^{a^\dagger}|M^{a^\dagger}=m)P(M=m|A=a^\dagger) \quad \text{(By A5, A6)} \\
&= \sum_m E(Y^{a^\dagger,m}|M^{a^\dagger}=m)p(m|a^\dagger) \quad \text{(By A5)} \\
&= \sum_{a,m} E(Y^{a^\dagger,m}|A=a,M^{a^\dagger}=m)p(a)p(m|a^\dagger) \\
&= \sum_m p(m|a^\dagger)\sum_a E(Y|A=a,M=m)p(a) \quad \text{(By A5, A7, A8).}
\end{aligned}$$

Alternatively, we also provide a slightly different proof:

$$\begin{aligned}
E(Y^{a^\dagger}) &= \sum_u E(Y^{a^\dagger}|U=u)f(u) \\
&= \sum_u E(Y^{a^\dagger}|A=a^\dagger,U=u)f(u) \\
&= \sum_u E(Y|A=a^\dagger,U=u)f(u) \\
&= \sum_{u,m} E(Y|A=a^\dagger,U=u,M=m)f(m|a^\dagger,u)f(u) \\
&= \sum_m f(m|a^\dagger)\sum_u E(Y|A=a^\dagger,U=u,M=m)f(u) \\
&= \sum_m f(m|a^\dagger)\underbrace{\sum_u E(Y|U=u,M=m)f(u)}_{(*)} \\
&= \sum_m f(m|a^\dagger)\sum_a E(Y|A=a,M=m)f(a).
\end{aligned}$$

Aside from probability laws, we note the following conditions that are used in the proof above: line 2 follows by conditional exchangeability of $Y^{a^\dagger}$ and A conditional on U seen in the SWIG in Figure 2b and follows from Assumption 6[12]; line 3 follows by consistency that is implied from recursive substitution of underlying NPSEMs; line 5 follows from Assumption 6[13]; line 6 follows from Assumptions 7 and 8 and can be seen from the conditional indepen-

12. Since there is no unmeasured common causes of exposure-mediator by Assumption 6, the only path from A to $Y^{a^\dagger}$ is a backdoor path via U.
13. If U has a direct arrow to M not through A, then this will violate Assumption 6.

dence of Y and A given U and M as seen in DAG 2a[14]; and line 7 follows from the SWIG in Figure 2c where it can be seen that $E(Y^m)=\sum_u E(Y|U=u,M=m)f(u)=\sum_a E(Y|A=a,M=m)f(a)$ as U or A blocks the backdoor back from M to Y^m. Alternatively, line 7 holds by algebraically by realizing the following:

$$\begin{aligned}\sum_a E(Y|A=a,M=m)f(a) &= \sum_{a,u} E(Y|A=a,M=m,U=u)f(u|a,m)f(a)\\ &= \sum_u E(Y|U=u,M=m)\sum_a f(u|a)f(a)\\ &= \sum_u E(Y|U=u,M=m)f(u).\end{aligned}$$

Remark 3 (The front door formula is a weighted average) *Consider a binary treatment A taking values $a^\dagger$ and a°. It can be trivially shown that the frontdoor formula is a weighted average of $E(Y|A=a^\dagger)$ and a separable estimand of treatment on Y denoted by $E(Y^{a_M=a^\dagger,a_Y=a^\circ})$ that can be identified from the observed data via an extended DAG. This extended DAG results from splitting the treatment node A into two sub-components, namely A_M and A_Y. The bolded arrows from A to A_M and A_Y indicate a deterministic relationship.[15] More specifically, the aforementioned two estimands are weighted by the probability of receiving treatment $a^\dagger$ and a° such that the frontdoor formula equals*

$$\begin{aligned}&\sum_m P(M=m|A=a^\dagger)\sum_a E(Y|A=a,M=m)P(A=a)=\\ &\quad P(A=a^\dagger)E(Y|A=a^\dagger)+P(A=a^\circ)\underbrace{\sum_m E(Y|A=a^\circ,M=m)f(m|a^\dagger)}_{E(Y^{a_M=a^\dagger,a_Y=a^\circ})}.\end{aligned}$$

We utilize this decomposition in deriving efficient influence functions for the frontdoor formula. We note that both estimands $E(Y|A=a^\dagger)$ and $E(Y^{a_M=a^\dagger,a_Y=a^\circ})$ allow for a direct path from A to Y not mediated through M. Moreover, $E(Y|A=a^\dagger)$ is identified even if there are unmeasured common causes of A and Y. This decomposition implies that when all individuals in the observed data take treatment $a^\dagger$, then the frontdoor formula equals $E(Y|A=a^\dagger)$; when all individuals in the observed data take treatment a°, then the frontdoor formula equals $E(Y^{a_M=a^\dagger,a_Y=a^\circ})$.[16] Clearly, when the frontdoor formula equals $E(Y^{a_M=a^\dagger,a_Y=a^\circ})$, then it must be that $A=A_Y=a^\circ$ for all observations. But this is precisely a scenario where intervening on A_Y (and creating a mediated path from $A\rightarrow A_Y\rightarrow Y$) is unnecessary, since $A=A_Y$ necessarily. Thus, this decomposition serves as a motivation for the construction of our proposed estimand.

14. Assumption 7 ensures that there is no direct path from A to Y not mediated by M. In addition, since there is no unmeasured common causes of mediator-outcome by Assumption 8, the only path from A to $Y^{a^\dagger}$ is a backdoor path via U and a frontdoor path via M.
15. Specifically in the observed data, $A=A_M=A_Y=1$ with probability 1, and $A=A_M=A_Y=0$ with probability 1.
16. although this would not be identified from observed data unless $f(m|a^\dagger)$ is known a priori for all m

Appendix C. Identification and estimation of new causally manipulable estimand in absence of L

In this section, we will assume that L is the empty set.

Theorem 5 *The average counterfactual outcome under an intervention on A_M is identified by the frontdoor formula* (4) *under Assumptions 1-4, that is,*

$$E(Y^{a_M=a^\dagger})=\sum_m P(M=m|A=a^\dagger)\underbrace{\sum_a E(Y|A=a,M=m)P(A=a)}_{(**)}.$$

Suppose that the observed data $\mathcal{O}=(A,M,Y)$ follow a law P which is known to belong to $\mathcal{M}=\{P_\theta:\theta\in\Theta\}$, where Θ is the parameter space. The efficient influence function $\varphi^{\text{eff}}(\mathcal{O})$ for a causal parameter $\Psi\equiv\Psi(\theta)$ in a non-parametric model $\mathcal{M}_{\text{np}}$ that imposes no restrictions on the law of $\mathcal{O}$ other than positivity is given by $d\Psi(\theta_t)/dt|_{t=0}=E\{\varphi^{\text{eff}}(\mathcal{O})S(\mathcal{O})\}$, where $d\Psi(\theta_t)/dt|_{t=0}$ is known as the pathwise derivative of the parameter Ψ along any parametric submodel of the observed data distribution indexed by t, and $S(\mathcal{O})$ is the score function of the parametric submodel evaluated at $t=0$ (Newey, 1994; Van Der Vaart, 2000).

The frontdoor formula can be re-expressed as a weighted average,

$$\Psi:=P(A=a^\dagger)E(Y|A=a^\dagger)+P(A=a^\circ)\sum_m E(Y|A=a^\circ,M=m)f(m|a^\dagger), \tag{C.7}$$

and thus the efficient influence function can be broken into two components. Using the chain rule, the efficient influence function of $\Psi=\sum_m f(m|a^\dagger)\sum_a E(Y|A=a,M=m)f(a)$ can be derived by finding the efficient influence function of (1) $\psi_1=P(A=a)$, (2) $\psi_2=E(Y|A=a^\dagger)$ and (3) $\psi_3=\sum_m E(Y|A=a^\circ,M=m)f(m|a^\dagger)$. We will use the fact that ψ_3 is an established identifying formula for $E(Y^{A_Y=a^\dagger,A_D=a^\circ})$, a term in the identification formula for separable effects. This is equal to the same functional of the observed data law $p(o)$ as $E(Y^{a^\circ}|A=a^\dagger)$ in the average treatment effect on the treated (ATT) when $a^\circ=0$ and $a^\dagger=1$ (Tchetgen and Shpitser, 2012).

Theorem 6 *The efficient influence function $\varphi^{\text{eff}}(\mathcal{O})$ of the frontdoor formula in $\mathcal{M}_{np}$ is given by*

$$\begin{aligned}\varphi^{\text{eff}}(\mathcal{O})=&\left[I(A=a^\dagger)-P(A=a^\dagger)\right]\psi_2+P(A=a^\dagger)\frac{I(A=a^\dagger)}{P(A=a^\dagger)}(Y-\psi_2)+\\&\frac{P(A=a^\circ)}{P(A=a^\dagger)}\left[I(A=a^\circ)\frac{P(A=a^\dagger|M)}{P(A=a^\circ|M)}\{Y-b_0(M)\}+I(A=a^\dagger)\{b_0(M)-\psi_3\}\right]+\\&[I(A=a^\circ)-P(A=a^\circ)]\psi_3,\end{aligned}$$

where the terms in red are the efficient influence function for $P(A{=}a^\dagger)\psi_2$*, the terms in blue is the efficient influence function for* $P(A{=}a^\circ)\psi_3$*, and* $b_0(M){=}E(Y|A{=}a^\circ,M)$*. The efficient influence function can be reduced to the following,*

$$\varphi^{\text{eff}}(\mathcal{O}){=}I(A{=}a^\dagger)Y{+}I(A{=}a^\circ)\psi_3{+} \tag{C.8}$$
$$\frac{P(A{=}a^\circ)}{P(A{=}a^\dagger)}\left[I(A{=}a^\circ)\frac{P(A{=}a^\dagger|M)}{P(A{=}a^\circ|M)}\{Y{-}b_0(M)\}{+}I(A{=}a^\dagger)\{b_0(M){-}\psi_3\}\right]{-}\Psi,$$

which can be re-expressed as:

$$\varphi^{\text{eff}}(\mathcal{O}){=}I(A{=}a^\dagger)Y{+}I(A{=}a^\circ)\psi_3{+} \tag{C.9}$$
$$\left[I(A{=}a^\circ)\frac{f(M|a^\dagger)}{f(M|a^\circ)}\{Y{-}b_0(M)\}{+}\frac{I(A{=}a^\dagger)P(A{=}a^\circ)}{P(A{=}a^\dagger)}\{b_0(M){-}\psi_3\}\right]{-}\Psi,$$

by realizing that $I(A{=}a^\circ)P(A{=}a^\circ)P(A{=}a^\dagger|M)\{P(A{=}a^\dagger)P(A{=}a^\circ|M)\}^{-1}{=}I(A{=}a^\circ)f(M|a^\dagger)\{f(M|a^\circ)\}^{-1}$*.*

After some algebra, it can be shown that (C.9) can be written as the form of the efficient influence function for Ψ given by Equation (5) in Theorem 1 in Fulcher et al. (2020) with $C{=}\emptyset$. Following Theorem 1 in Fulcher et al. (2020), the semiparametric efficiency bound for Ψ in $\mathcal{M}_{\text{np}}$ is given by $\text{var}\big(\varphi^{\text{eff}}\big)$.

C.0.1 On semiparametric estimators for the frontdoor formula

The efficient influence function for the frontdoor formula (Ψ) given by in Expressions (C.8) or (C.9) allows us to construct estimators that guarantee sample-boundedness. A weighted iterative conditional expectation (Weighted ICE) estimator that guarantee sample-boundedness is given in the following algorihtm. In what follows, we let $\mathbb{P}_n(X){=}n^{-1}\sum_{i=1}^n X_i$ and let g^{-1} denote a known inverse link function[17].

Algorithm 3 Algorithm for Weighted ICE (frontdoor formula)

1: Non-parametrically compute $P(A{=}a^\circ)$ and $P(A{=}a^\dagger)$.
2: Compute the MLEs $\hat{\alpha}$ of α from the observed data for the treatment model $P(A{=}a|M;\alpha)$, or compute the MLEs $\hat{\gamma}$ of γ from the observed data for the mediator model $P(M{=}m|A;\gamma)$.
3: In the individuals whose $A{=}a^\circ$, fit a regression model $Q(M;\theta){=}g^{-1}\{\theta^T\phi(M)\}$ for $b_0(M){=}E(Y|M,A{=}a^\circ)$ where the score function for each observation is weighted by $\hat{W}_1$ where $\hat{W}_1$ equals

$$\frac{P(A{=}a^\circ)P(A{=}a^\dagger|M;\hat{\alpha})}{P(A{=}a^\dagger)P(A{=}a^\circ|M;\hat{\alpha})}$$

17. For instance, if Y is dichotomous, the g is the logit link function

if $\hat{\alpha}$ was estimated in the previous step, or $\hat{W}$ equals

$$\frac{f(M|A{=}a^{\dagger};\hat{\gamma})}{f(M|A{=}a^{\circ};\hat{\gamma})}$$

if $\hat{\gamma}$ was estimated in the previous step. Moreover, $\phi(M)$ is a known function of M. More specifically, we solve for θ in the following estimating equations:

$$\mathbb{P}_n\Big[I(A{=}a^{\circ})\phi(M)\hat{W}_1\{Y-Q(M;\theta)\}\Big]{=}0.$$

4: In those whose $A{=}a^{\dagger}$, fit an intercept-only model $T(\beta){=}g^{-1}(\beta)$ for $\psi_3{=}E\{b_0(M)|A{=}a^{\dagger}\}$, where the score function for each observation is weighted by

$$\hat{W}_2{=}\frac{P(A{=}a^{\circ})}{P(A{=}a^{\dagger})}.$$

More specifically, we solve for β in the following estimating equations:

$$\mathbb{P}_n\Big[I(A{=}a^{\dagger})\hat{W}_2\Big\{Q(M;\hat{\theta})-T(\beta)\Big\}\Big]{=}0.$$

5: Compute $\hat{T}{\equiv}T(\hat{\beta})$ for all observations.
6: Estimate $\hat{\Psi}_{WICE}{=}\mathbb{P}_n\{I(A{=}a^{\dagger})Y+I(A{=}a^{\circ})\hat{T}\}$.

Steps 3 and 4 ensure that the estimates for $\psi_3{=}E\{b_0(M)|A{=}a^{\dagger}\}$ are sample bounded. Step 6 confirms that $\hat{\Psi}_{WICE}$ is a convex combination of Y and estimates for ψ_3, both of which are bounded by the range of the outcome Y. Thus, $\hat{\Psi}_{WICE}$ will also be sample-bounded. For instance if the outcome is binary, then $\hat{\Psi}_{WICE}$ will always be bounded between 0 and 1. Note that estimators based on Expression (C.8) are more convenient to construct than estimators based on Expression (C.9) when (1) M is continuous, and/or (2) there are multiple mediator variables (M_1, M_2, $M_3\ldots$).

In Appendix F, we prove that an estimator based on the efficient influence function given by (C.8) is doubly robust in the sense that it will be consistent as long as the model for $P(A{=}a|M)$ or the model for $b_0(M){=}E(Y|A{=}a^{\circ},M)$ is correctly specified, and an estimator based on the efficient influence function given by (C.9) is doubly robust in the sense that it will be consistent as long as the model for $P(M{=}m|A)$ or the model for $b_0(M)$ is correctly specified.

Appendix D. Interventionist identification with the frontdoor formula

Consider an extended causal DAG, which includes A and also the variable A_M, where the bold arrow from A to A_M indicates a deterministic relationship. That is, Figure 1a is the extended DAG of such a split with $V=(U,L,A,A_M,M,Y)$, and in the observed data, with probability one under $f(v)$, either $A=A_M=1$ or $A=A_M=0$.

Here and henceforth we use "(G)" to indicate that the variables refer to those realized in the hypothetical trial where A_M is randomly assigned (possibly dependent on $L(G)$ and $A(G)$), as illustrated in 1c. In particular, conditioning sets that include $A(G)=a, A_M(G)=a^\dagger$ refer to the hypothetical trial G in which A_M is randomly assigned (or where random assignment is dependent on L and/or A). Then, by our presumed definition of G that satisfies the distributional consistencies described in the main text, we have that $E(Y^{a_M=a^\dagger})=\sum_{l,a}E(Y(G)|A_M(G)=a^\dagger,L(G)=l,A(G)=a)P(A(G)=a,L(G)=l)$, where the right-hand side refers to data from the hypothetical experiment. More specifically:

$$
\begin{aligned}
E(Y^{a_M=a^\dagger}) &= \sum_{l,a} E(Y(G)|A_M(G)=a^\dagger,L(G)=l,A(G)=a)P(A(G)=a|L(G)=l)P(L(G)=l) \\
&= \sum_{m,l,a} E(Y(G)|M(G)=m,A_M(G)=a^\dagger,L(G)=l,A(G)=a) \\
&\qquad P(M(G)=m|A_M(G)=a^\dagger,A(G)=a,L(G)=l)P(A(G)=a|L(G)=l)P(L(G)=l) \\
&= \sum_{m,l} P(M(G)=m|A_M(G)=a^\dagger,A(G)=a^\dagger,L(G)=l)P(L(G)=l) \\
&\qquad \sum_{a} E(Y(G)|M(G)=m,L(G)=l,A(G)=a,A_M(G)=a)P(A(G)=a|L(G)=l) \\
&= \sum_{m,l} P(M=m|A=a^\dagger,L=l)P(L=l)\sum_{a} E(Y|M=m,L=l,A=a)P(A=a|L=l).
\end{aligned}
$$

Aside from probability laws, we note the following conditions that are used in the proof above: equality 1 follows by definition of G; equality 3 holds by the dismissible component conditions such that conditional independence $M(G)\perp\!\!\!\perp A(G)|A_M(G),L(G)$ and $Y(G)\perp\!\!\!\perp A_M(G)|A(G),M(G),L(G)$ hold; and equality 4 holds by definition of G, consistency and determinism in the observed data such that the event $\{A=a,A_M=a\}$ is the same as the event $\{A=a\}$.

Appendix E. Derivation of efficient influence function in Section 5

Proof For binary treatment variables, the generalized frontdoor formula $\Psi := \sum_m f(m|a^\dagger)\sum_a E(Y|A=a,M=m)f(a)$ is equivalent to the following:

$$\Psi := P(A=a^\dagger)E(Y|A=a^\dagger)+P(A=a^\circ)\sum_{m,l} E(Y|M=m,L=l,A=a^\circ)f(m|a^\dagger,l)f(l|a^\circ).$$

The efficient influence function in the non-parametric model $\mathcal{M}_{NP}$ is defined as the unique mean zero, finite variance random variable $\varphi^{\text{eff}}(\mathcal{O})$ such that

$$\frac{d\Psi(\theta_t)}{dt}\Big|_{t=0} = E\{\varphi^{\text{eff}}(\mathcal{O})S(\mathcal{O})\}$$

where $\mathcal{O}=(L,A,M,Y)$, $d\Psi(\theta_t)/dt|_{t=0}$ is known as the pathwise derivative of parameter Ψ along a parametric submodel indexed by t, and $S(\mathcal{O})$ is the score function of the parametric submodel evaluated at $t=0$.

The efficient influence function of Ψ can be realized by finding the efficient influence function of (1) $\psi_1 := P(A=a)$, (2) $\psi_2 := E(Y|A=a^\dagger)$ and (3) $\psi_3 := \sum_{m,l} E(Y|M=m,L=l,A=a^\circ)f(m|a^\dagger,l)f(l|a^\circ)$. In particular, using differentiation rules, the efficient influence function is given by:

$$\varphi^{\text{eff}}(\mathcal{O}) = \underbrace{[I(A=a^\dagger)-P(A=a^\dagger)]}_{(*)}\psi_2+\psi_2^{\text{eff}}P(A=a^\dagger)+\psi_3^{\text{eff}}P(A=a^\circ)+\underbrace{[I(A=a^\circ)-P(A=a^\circ)]}_{(**)}\psi_3$$

where expression $(*)$ is the efficient influence function of $P(A=a^\dagger)$ and expression $(**)$ is the efficient influence function of $P(A=a^\circ)$. Moreover,

$$\begin{aligned}\frac{d\psi_2(\theta_t)}{dt}\Big|_{t=0} &= E\{YS(Y|A=a^\dagger)|A=a^\dagger\}\\ &= E\left[\left\{Y-E(Y|A=a^\dagger)\right\}S(Y|A=a^\dagger)|A=a^\dagger\right]\\ &= E\left[\frac{I(A=a^\dagger)(Y-\psi_2)}{P(A=a^\dagger)}S(\mathcal{O})\right]\end{aligned}$$

and

$$\begin{aligned}\frac{d\psi_3(\theta_t)}{dt}\Big|_{t=0} = &\underbrace{E\left(E\left[E\{YS(Y|A=a^\circ,M,L)|A=a^\circ,M,L\}|A=a^\dagger,L\right]|A=a^\circ\right)}_{\text{Ⓐ}}+\\ &\underbrace{E\left[E\left\{E(Y|A=a^\circ,M,L)S(M|A=a^\dagger,L)|A=a^\dagger,L\right\}|A=a^\circ\right]}_{\text{Ⓑ}}+\\ &\underbrace{E\left[E\left\{E(Y|A=a^\circ,M,L)|A=a^\dagger,L\right\}S(L|A=a^\circ)|A=a^\circ\right]}_{\text{Ⓒ}}.\end{aligned}$$

We look at each expression separately. First, we consider Expression Ⓐ:

$$
\begin{aligned}
Ⓐ &= E\Big(E\Big[E\{YS(Y|A{=}a^\circ,M,L)|A{=}a^\circ,M,L\}|A{=}a^\dagger,L\Big]|A{=}a^\circ\Big)\\
&= E\left(E\left[E\left\{Y\frac{I(A{=}a^\circ)}{P(A{=}a^\circ|L,M)}S(Y|A,M,L)|M,L\right\}|A{=}a^\dagger,L\right]|A{=}a^\circ\right)\\
&= E\left(E\left[E\left\{Y\frac{I(A{=}a^\circ)}{P(A{=}a^\circ|L,M)}S(Y|A,M,L)|M,L\right\}\frac{I(A{=}a^\dagger)}{P(A{=}a^\dagger|L)}|L\right]\frac{I(A{=}a^\circ)}{P(A{=}a^\circ)}\right)\\
&= E\left(E\left[E\left\{Y\frac{I(A{=}a^\circ)}{P(A{=}a^\circ|L,M)}S(Y|A,M,L)|M,L\right\}\frac{P(A{=}a^\dagger|L,M)}{P(A{=}a^\dagger|L)}|L\right]\frac{P(A{=}a^\circ|L)}{P(A{=}a^\circ)}\right)\\
&= E\left(E\left[E\left\{\Big(Y-b_0(M,L)\Big)\frac{I(A{=}a^\circ)}{P(A{=}a^\circ|L,M)}S(Y|A,M,L)|M,L\right\}\frac{P(A{=}a^\dagger|L,M)}{P(A{=}a^\dagger|L)}|L\right]\frac{P(A{=}a^\circ|L)}{P(A{=}a^\circ)}\right)\\
&= E\left[\Big\{Y-b_0(M,L)\Big\}\frac{I(A{=}a^\circ)P(A{=}a^\dagger|L,M)P(A{=}a^\circ|L)}{P(A{=}a^\circ|L,M)P(A{=}a^\dagger|L)P(A{=}a^\circ)}S(\mathcal{O})\right]
\end{aligned}
$$

which also equals

$$
E\left[\Big\{Y-b_0(M,L)\Big\}\frac{I(A{=}a^\circ)f(M|A{=}a^\dagger,L)}{P(A{=}a^\circ)f(M|A{=}a^\circ,L)}S(\mathcal{O})\right].
$$

Next, we consider Ⓑ:

$$
\begin{aligned}
Ⓑ &= E\left[E\left\{\underbrace{E(Y|A{=}a^\circ,M,L)}_{b_0(M,L)}S(M|A{=}a^\dagger,L)|A{=}a^\dagger,L\right\}|A{=}a^\circ\right]\\
&= E\left[E\left\{b_0(M,L)\frac{I(A{=}a^\dagger)}{P(A{=}a^\dagger|L)}S(M|A,L)|L\right\}\frac{P(A{=}a^\circ|L)}{P(A{=}a^\circ)}\right]\\
&= E\left[\Big\{b_0(M,L)-E\Big(b_0(M,L)|L,A\Big)\Big\}\frac{I(A{=}a^\dagger)P(A{=}a^\circ|L)}{P(A{=}a^\dagger|L)P(A{=}a^\circ)}S(\mathcal{O})\right].
\end{aligned}
$$

Finally, we consider Ⓒ:

$$
\begin{aligned}
Ⓒ &= E\left[\underbrace{E\Big\{E(Y|A{=}a^\circ,M,L)|A{=}a^\dagger,L\Big\}}_{h_\dagger(L)}S(L|A{=}a^\circ)|A{=}a^\circ\right]\\
&= E\left\{h_\dagger(L)\frac{I(A{=}a^\circ)}{P(A{=}a^\circ)}S(L|A)\right\}\\
&= E\left\{\Big(h_\dagger(L)-\psi_3\Big)\frac{I(A{=}a^\circ)}{P(A{=}a^\circ)}S(\mathcal{O})\right\}.
\end{aligned}
$$

Thus, putting everything together and after some further algebraic simplification, we can see that the efficient influence function is indeed given by:

$$\varphi^{\text{eff}}(\mathcal{O}) = I(A=a^{\dagger})Y + I(A=a^{\circ})\psi_3 + \frac{I(A=a^{\circ})P(A=a^{\dagger}|M,L)P(A=a^{\circ}|L)}{P(A=a^{\circ}|M,L)P(A=a^{\dagger}|L)}\{Y - b_0(M,L)\} + \frac{I(A=a^{\dagger})P(A=a^{\circ}|L)}{P(A=a^{\dagger}|L)}\{b_0(M,L) - h_{\dagger}(L)\} + I(A=a^{\circ})\{h_{\dagger}(L) - \psi_3\} - \Psi.$$

It is trivial to realize that this Expression of the efficient influence function can also be re-expressed as the following:

$$\varphi^{\text{eff}}(\mathcal{O}) := I(A=a^{\dagger})Y + I(A=a^{\circ})\psi_3 + \frac{I(A=a^{\circ})f(M|A=a^{\dagger},L)}{f(M|A=a^{\circ},L)}\{Y - b_0(M,L)\} + \frac{I(A=a^{\dagger})P(A=a^{\circ}|L)}{P(A=a^{\dagger}|L)}\{b_0(M,L) - h_{\dagger}(L)\} + I(A=a^{\circ})\{h_{\dagger}(L) - \psi_3\} - \Psi.$$

■

Appendix F. Robustness against model misspecification

F.1 Non-generalized front door formula

We show that an estimator based on Equation (C.8) is doubly robust in the sense that it will be consistent as long as

1. the model for $P(A{=}a|M)$ is correctly specified, or
2. the model for $E(Y|A{=}a^\circ,M)$ is correctly specified.

and that an estimator based on Equation (C.9) is doubly robust in that it will be consistent as long as

1. the model for $P(M{=}m|A)$ is correctly specified, or
2. the model for $E(Y|A{=}a^\circ,M)$ is correctly specified.

We consider an estimator based on Equation (C.8). Suppose that α^*, θ^* and β^* are probability limits of α, θ and β, respectively. Furthermore, let $b_0^*(M){=}Q(M;\theta^*)$ where as before $Q(M;\theta)$ is a model for $b_0(M){=}E(Y|M,A{=}a^\circ)$, and let $\psi_3^*{=}T(\beta^*)$ where $T(\beta)$ is a nonparametric model for $\psi_3{=}E\{b_0(M)|A{=}a^\dagger\}$. Under Equation (C.8) suffices to show that

$$E\Bigg(I(A{=}a^\dagger)Y+I(A{=}a^\circ)\psi_3^*+ \\ \frac{P(A{=}a^\circ)}{P(A{=}a^\dagger)}\left[I(A{=}a^\circ)\frac{P(A{=}a^\dagger|M;\alpha^*)}{P(A{=}a^\circ|M;\alpha^*)}\{Y-b_0^*(M)\}+I(A{=}a^\dagger)\{b_0^*(M)-\psi_3^*\}\right]-\Psi\Bigg){=}0$$

under scenario **(1)** where $\alpha^*{=}\alpha$ and thus $P(A{=}a^\dagger|M;\alpha^*){=}P(A{=}a^\dagger|M)$, **or** under scenario **(2)** where $\theta^*{=}\theta$ and thus $b_0^*(M){=}b_0(M)$ and $\psi_3^*{=}\psi_3$.

Proof Suppose first that only the model for $P(A{=}a|M)$ is correctly specified. Then,

$$E\Bigg(I(A{=}a^\dagger)Y+I(A{=}a^\circ)\psi_3^*+ \\ \frac{P(A{=}a^\circ)}{P(A{=}a^\dagger)}\left[I(A{=}a^\circ)\frac{P(A{=}a^\dagger|M;\alpha^*)}{P(A{=}a^\circ|M;\alpha^*)}\{Y-b_0^*(M)\}+I(A{=}a^\dagger)\{b_0^*(M)-\psi_3^*\}\right]-\Psi\Bigg)$$

$$={E}\Bigg(P(A{=}a^\dagger|M)E(Y|A{=}a^\dagger,M)+P(A{=}a^\circ|M)\psi_3^*+ \\ \frac{P(A{=}a^\circ)}{P(A{=}a^\dagger)}\left[\cancel{P(A{=}a^\circ|M)}\cancel{\frac{P(A{=}a^\dagger|M;\alpha^*)}{P(A{=}a^\circ|M;\alpha^*)}}\{b_0(M)-b_0^*(M)\}+P(A{=}a^\dagger|M)\{b_0^*(M)-\psi_3^*\}\right]-\Psi\Bigg)$$

$$
\begin{aligned}
=&\sum_m \Big(P(A{=}a^\dagger|M{=}m)E(Y|A{=}a^\dagger,M{=}m)+P(A{=}a^\circ|M{=}m)\psi_3^*+ \\
&\frac{P(A{=}a^\circ)}{P(A{=}a^\dagger)}\Big[P(A{=}a^\dagger|M{=}m;\alpha^*)b_0(m)-P(A{=}a^\dagger|M{=}m)\psi_3^*\}\Big]\Big)P(M{=}m)-\Psi \\
=&E(Y|A{=}a^\dagger)P(A{=}a^\dagger)+\sum_m P(A{=}a^\circ)P(M{=}m|A{=}a^\dagger)b_0(M)-\Psi \\
=&0.
\end{aligned}
$$

Next, suppose that only the model for $E(Y|A{=}a^\circ,M)$ is correctly specified. Then,

$$
\begin{aligned}
&E\Bigg(I(A{=}a^\dagger)Y+I(A{=}a^\circ)\psi_3^*+ \\
&\frac{P(A{=}a^\circ)}{P(A{=}a^\dagger)}\left[I(A{=}a^\circ)\frac{P(A{=}a^\dagger|M;\alpha^*)}{P(A{=}a^\circ|M;\alpha^*)}\{Y-b_0^*(M)\}+I(A{=}a^\dagger)\{b_0^*(M)-\psi_3^*\}\right]-\Psi\Bigg) \\
=&E\Bigg(P(A{=}a^\dagger|M)E(Y|A{=}a^\dagger,M)+P(A{=}a^\circ|M)\psi_3^*+ \\
&\frac{P(A{=}a^\circ)}{P(A{=}a^\dagger)}\left[P(A{=}a^\circ|M)\frac{P(A{=}a^\dagger|M;\alpha^*)}{P(A{=}a^\circ|M;\alpha^*)}\{\underbrace{b_0(M)-b_0^*(M)}_{=0}\}+P(A{=}a^\dagger|M)\{b_0^*(M)-\psi_3^*\}\right]-\Psi\Bigg) \\
=&\sum_m \Bigg(P(A{=}a^\dagger|M{=}m)E(Y|A{=}a^\dagger,M{=}m)+P(A{=}a^\circ|M{=}m)\psi_3^*+ \\
&\frac{P(A{=}a^\circ)}{P(A{=}a^\dagger)}\Big[P(A{=}a^\dagger|M{=}m)\{b_0^*(M)-\psi_3^*\}\Big]\Bigg)P(M{=}m)-\Psi \\
=&E(Y|A{=}a^\dagger)P(A{=}a^\dagger)+\sum_m P(A{=}a^\circ)P(M{=}m|A{=}a^\dagger)b_0(m)-\Psi \\
=&0.
\end{aligned}
$$

■

The proof of double robustness for estimators based on Equation (C.9) follows analogously as the proof shown above.

F.2 Generalized front door formula

Our proposed estimator based on the efficient influence function given by (5) and (6) is robust against 3 classes of model misspecification scenarios. Specifically, the weighted ICE estimator where a model for $P(A{=}a|M,L)$ is specified will be consistent when at least one of the following holds:

1. the models for $P(A{=}a|M,L)$ *and* $P(A{=}a|L)$ are correctly specified, or

2. the models for $b_0(M,L)$ and $h_\dagger(L)$ are correctly specified, or

3. the models for $b_0(M,L)$ and $P(A{=}a|L)$ are correctly specified.

The weighted ICE estimator where a model for $P(M{=}m|A,L)$ is specified will be consistent when at least one of the following holds

1. the models for $P(M{=}m|A,L)$ *and* $P(A{=}a|L)$ are correctly specified, or

2. the models for $b_0(M,L)$ and $h_\dagger(L)$ are correctly specified, or

3. the models for $b_0(M,L)$ and $P(A{=}a|L)$ are correctly specified.

We will prove robustness for an estimator based on (6) where a model for $P(M{=}m|A,L)$ is specified. Proof of robustness for estimator based on (5) where a model for $P(A{=}a|M,L)$ is specified follows analogously.

Suppose that γ^*, κ^*, θ^*, η^* and β^* are probability limits of γ,κ, θ η and β, respectively. Furthermore, let $b_0^*(M,L){=}Q(M,L;\theta^*)$ where as before $Q(M,L;\theta)$ is a model for $b_0(M){=}E(Y|M,L,A{=}a^\circ)$, let $h_\dagger^*(L){=}R(L;\eta^*)$ where $R(L;\eta)$ is a model for $h_\dagger(L)$, and let $\psi_3^*{=}T(\beta^*)$ where $T(\beta)$ is a non-parametric model for $\psi_3{=}E\{h_\dagger(L)|A{=}a^\circ\}$.

Under Equation (5) suffices to show that

$$E\Bigg(I(A{=}a^\dagger)Y+I(A{=}a^\circ)\psi_3^*+\frac{I(A{=}a^\circ)f(M|A{=}a^\dagger,L;\gamma^*)}{f(M|A{=}a^\circ,L;\gamma^*)}\{Y-b_0^*(M,L)\}+$$
$$\frac{I(A{=}a^\dagger)P(A{=}a^\circ|L;\kappa^*)}{P(A{=}a^\dagger|L;\kappa^*)}\{b_0^*(M,L)-h_\dagger^*(L)\}+I(A{=}a^\circ)\{h_\dagger^*(L)-\psi_3^*\}-\Psi\Bigg){=}0$$

under scenario **(1)** where $\gamma^*{=}\gamma$ and $\kappa^*{=}\kappa$ and thus $P(M{=}m|A,L;\gamma^*){=}P(M{=}m|A,L)$ and $P(A{=}a^\dagger|L;\kappa^*){=}P(A{=}a^\dagger|L)$, **or** under scenario **(2)** where $\theta^*{=}\theta$ and $\eta^*{=}\eta$ and thus $b_0^*(M,L){=}b_0(M,L)$, $h_\dagger^*(L){=}h_\dagger(L)$ and $\psi_3^*{=}\psi_3$, **or** under scenario **(3)** where $\theta^*{=}\theta$ and $\kappa^*{=}\kappa$ and thus $b_0^*(M,L){=}b_0(M,L)$ and $P(A{=}a^\dagger|L;\kappa^*){=}P(A{=}a^\dagger|L)$.

Proof Suppose first that only the models for $P(M{=}m|A,L)$ *and* $P(A{=}a|L)$ are correctly specified. Then,

$$E\Bigg(I(A{=}a^\dagger)Y+I(A{=}a^\circ)\psi_3^*+\frac{I(A{=}a^\circ)f(M|A{=}a^\dagger,L;\gamma^*)}{f(M|A{=}a^\circ,L;\gamma^*)}\{Y-b_0^*(M,L)\}+$$
$$\frac{I(A{=}a^\dagger)P(A{=}a^\circ|L;\kappa^*)}{P(A{=}a^\dagger|L;\kappa^*)}\{b_0^*(M,L)-h_\dagger^*(L)\}+I(A{=}a^\circ)\{h_\dagger^*(L)-\psi_3^*\}-\Psi\Bigg)$$
$$={E}\Bigg(\sum_m P(A{=}a^\dagger|L)E(Y|A{=}a^\dagger,M{=}m,L)f(m|a^\dagger,L)+P(A{=}a^\circ|L)\psi_3^*+$$

$$\sum_m \frac{P(A=a^\circ|L)f(m|A=a^\dagger,L;\gamma^*)}{\cancel{f(m|A=a^\circ,L;\gamma^*)}}\{b_0(m,L)-b_0^*(m,L)\}\cancel{f(m|a^\circ,L)}+$$
$$\frac{\cancel{P(A=a^\dagger|L)}P(A=a^\circ|L;\kappa^*)}{\cancel{P(A=a^\dagger|L;\kappa^*)}}\left\{\sum_m b_0^*(m,L)f(m|a^\dagger,L)-h_\dagger^*(L)\right\}+$$
$$P(A=a^\circ|L)\{h_\dagger^*(L)-\psi_3^*\}-\Psi\Bigg)$$
$$=E\left(\sum_m P(A=a^\dagger|L)E(Y|A=a^\dagger,M=m,L)f(m|a^\dagger,L)+\sum_m P(A=a^\circ|L)f(m|A=a^\dagger,L;\gamma^*)b_0(m,L)\right)$$
$$=P(A=a^\dagger)E(Y|A=a^\dagger)+P(A=a^\circ)\sum_{m,l} b_0(m,l)f(m|a^\dagger,l)f(l|a^\circ)$$
$$=0.$$

Next, suppose that only the models for $b_0(M,L)$ and $h_\dagger(L)$ are correctly specified. Then,

$$E\Bigg(I(A=a^\dagger)Y+I(A=a^\circ)\psi_3^*+\frac{I(A=a^\circ)f(M|A=a^\dagger,L;\gamma^*)}{f(M|A=a^\circ,L;\gamma^*)}\{Y-b_0^*(M,L)\}+$$
$$\frac{I(A=a^\dagger)P(A=a^\circ|L;\kappa^*)}{P(A=a^\dagger|L;\kappa^*)}\{b_0^*(M,L)-h_\dagger^*(L)\}+I(A=a^\circ)\{h_\dagger^*(L)-\psi_3^*\}-\Psi\Bigg)$$
$$=E\Bigg(\sum_m P(A=a^\dagger|L)E(Y|A=a^\dagger,M=m,L)f(m|a^\dagger,L)+\cancel{P(A=a^\circ|L)\psi_3^*}+$$
$$\sum_m \frac{P(A=a^\circ|L)f(m|A=a^\dagger,L;\gamma^*)}{f(m|A=a^\circ,L;\gamma^*)}\{\underbrace{b_0(m,L)-b_0^*(m,L)}_{=0}\}f(m|a^\circ,L)+$$
$$\frac{P(A=a^\dagger|L)P(A=a^\circ|L;\kappa^*)}{P(A=a^\dagger|L;\kappa^*)}\left\{\underbrace{\sum_m b_0^*(m,L)f(m|a^\dagger,L)-h_\dagger^*(L)}_{=0}\right\}+$$
$$P(A=a^\circ|L)\{h_\dagger^*(L)-\cancel{\psi_3^*}\}-\Psi\Bigg)$$
$$=E\left(\sum_m P(A=a^\dagger|L)E(Y|A=a^\dagger,M=m,L)f(m|a^\dagger,L)+\sum_m P(A=a^\circ|L)f(m|A=a^\dagger,L;\gamma^*)b_0(m,L)\right)$$
$$=P(A=a^\dagger)E(Y|A=a^\dagger)+P(A=a^\circ)\sum_{m,l} b_0(m,l)f(m|a^\dagger,l)f(l|a^\circ)$$
$$=0.$$

Finally, suppose that only the models for $b_0(M,L)$ and $P(A=a|L)$ are correctly specified. Then,

$$
\begin{aligned}
&E\Bigg(I(A=a^\dagger)Y + I(A=a^\circ)\psi_3^* + \frac{I(A=a^\circ)f(M|A=a^\dagger,L;\gamma^*)}{f(M|A=a^\circ,L;\gamma^*)}\{Y-b_0^*(M,L)\}+ \\
&\quad \frac{I(A=a^\dagger)P(A=a^\circ|L;\kappa^*)}{P(A=a^\dagger|L;\kappa^*)}\{b_0^*(M,L)-h_\dagger^*(L)\}+I(A=a^\circ)\{h_\dagger^*(L)-\psi_3^*\}-\Psi\Bigg) \\
&=E\Bigg(\sum_m P(A=a^\dagger|L)E(Y|A=a^\dagger,M=m,L)f(m|a^\dagger,L)+\cancel{P(A=a^\circ|L)\psi_3^*}+ \\
&\quad \sum_m \frac{P(A=a^\circ|L)f(m|A=a^\dagger,L;\gamma^*)}{f(m|A=a^\circ,L;\gamma^*)}\underbrace{\{b_0(m,L)-b_0^*(m,L)\}}_{=0}f(m|a^\circ,L)+ \\
&\quad \frac{\cancel{P(A=a^\dagger|L)}P(A=a^\circ|L;\kappa^*)}{\cancel{P(A=a^\dagger|L;\kappa^*)}}\left\{\sum_m b_0^*(m,L)f(m|a^\dagger,L)-h_\dagger^*(L)\right\}+ \\
&\quad P(A=a^\circ|L)\{h_\dagger^*(L)-\cancel{\psi_3^*}\}-\Psi\Bigg) \\
&=E\Bigg(\sum_m P(A=a^\dagger|L)E(Y|A=a^\dagger,M=m,L)f(m|a^\dagger,L)+ \\
&\quad P(A=a^\circ|L;\kappa^*)\left\{\sum_m b_0^*(m,L)f(m|a^\dagger,L)-\cancel{h_\dagger^*(L)}\right\}+\cancel{P(A=a^\circ|L)h_\dagger^*(L)}-\Psi\Bigg) \\
&=P(A=a^\dagger)E(Y|A=a^\dagger)+P(A=a^\circ)\sum_{m,l} b_0(m,l)f(m|a^\dagger,l)f(l|a^\circ) \\
&=0.
\end{aligned}
$$

■

Appendix G. Other relevant estimators

G.1 Inverse probability weighted estimator

We describe one class of inverse probability weighted estimator that was used in the simulation and data analysis (see Fulcher et al., 2020 for other inverse probability weighted estimators). Specifically, we can solve for Ψ_{IPW} in the following IPW estimator to estimate Ψ:

$$\mathbb{P}_n\left[\frac{I(A=a^\dagger)}{f(A|L;\hat{\kappa})}\left\{\sum_a E(Y|A=a,M,L;\hat{\theta})f(a|L;\hat{\kappa})-\Psi_{IPW}\right\}\right]=0$$

where $\mathbb{P}_n(X)=n^{-1}\sum_{i=1}^n X_i$ and $E(Y|A,M,L;\hat{\theta})$ is an estimate of $E(Y|A,M,L)$ such that $E(Y|A=a^\circ,M,L)=b_0(M,L)$.

G.2 Iterative conditional expectation estimator

The ICE estimator that was used in the simulation and data analysis follows from the weighted ICE procedure, whereby we set $\hat{W}=1$ for all regression steps.

G.3 Iterative TMLE

Herein, we describe an iterative TMLE algorithm that is doubly robust in the sense of Fulcher et al. (2020) for binary mediators.

Algorithm 4 Algorithm for iterative Targeted maximum likelihood (generalized frontdoor formula)

1: Non-parametrically compute $P(A=a^\circ)$ and $P(A=a^\dagger)$.
2: Obtain estimates $\hat{P}(A=a|L)$ and $\hat{R}^{(0)}(A,L):=\hat{P}(M=m|A,L)$ of $P(A=a|L)$ and $P(M=m|A,L)$, respectively, possibly using machine learning methods.
3: In the individuals whose $A=a^\circ$, compute $\hat{Q}^{(0)}(M,L)$ by regressing Y on (M,L). Here, $\hat{Q}^{(0)}(M,L)$ is possibly estimated using machine learning methods, and it denotes an initial estimate for $b_0(M,L)=E(Y|M,L,A=a^\circ)$.
4: Define $t=0$. Iteratively update the following until convergence (i.e., until parameters $\delta^{(K)}\approx 0$ and $\nu^{(K)}\approx 0$):

   A. Update the previous regression. Specifically, in the individuals whose $A=a^\circ$, fit a intercept-only regression model $Q^{(t+1)}(M,L;\delta^{(t+1)})=g^{-1}[g\{\hat{Q}^{(t)}(M,L)\}+\delta^{(t+1)}]$ where the score function for each observation is weighted by $\hat{W}_1^{(t)}=\frac{\hat{f}^{(t)}(M|A=a^\dagger,L)}{\hat{f}(M|A=a^\circ,L)}$. More specifically, we solve for δ in the following estimating equations:

$$\mathbb{P}_n\left[I(A=a^\circ)\hat{W}_1^{(t)}\left\{Y-Q^{(t+1)}(M,L;\delta^{(t+1)})\right\}\right]=0.$$

B. Define $\hat{Q}^{(t+1)}_{\text{diff}}=Q^{(t+1)}(m=1,L;\hat{\delta}^{(t+1)})-Q^{(t+1)}(m=0,L;\hat{\delta}^{(t+1)})$. In those whose $A=a^\dagger$, fit a single covariate regression model (with no intercept) for the conditional distribution of M given by

$$R^{(t+1)}(a^\dagger,L;\nu^{(t+1)})=g^{-1[}\{g\{\hat{R}^{(t)}(a^\dagger,L)\}+\nu^{(t+1)}\hat{Q}^{(t+1)}_{\text{diff}}]$$

with observational weights given by $\hat{W}_2=\frac{\hat{P}(A=a^\circ|L)}{\hat{P}(A=a^\dagger|L)}$. More specifically, we solve for $\nu^{(t+1)}$ in the following estimating equations:

$$\mathbb{P}_n\left[I(A=a^\dagger)\hat{W}_2\hat{Q}^{(t+1)}_{\text{diff}}\left\{M-R^{(t+1)}(a^\dagger,L;\nu^{(t+1)})\right\}\right]=0.$$

C. Update $\hat{Q}^{(t+1)}(M,L)\coloneqq Q^{(t+1)}(M,L;\hat{\delta}^{(t+1)})$ and $\hat{R}^{(t+1)}(a^\dagger,L)\coloneqq R^{(t+1)}(a^\dagger,L;\hat{\nu}^{(t+1)})$.

D. t+=1

5: Upon convergence at iteration K, define $\hat{b}^{(K)}_0(m,L)=\hat{Q}^{(K)}(m,L)$ for $m=0,1$ and $\hat{f}^{(K)}(M|A=a^\dagger,L)=\hat{R}^{(K)}(a^\dagger,L)$. In those whose $A=a^\circ$, fit another regression model $T(\beta)=g^{-1}(\beta)$ for $\psi_3=E\{h_\dagger(L)|A=a^\circ\}$ with just an intercept. More specifically, we solve for β in the following estimating equations:

$$\mathbb{P}_n\left[I(A=a^\circ)\left\{\sum_{m=0}^{1}\hat{b}^{(K)}_0(m,L)\hat{f}^{(K)}(m|a^\dagger,L)-T(\beta)\right\}\right]=0.$$

6: Compute $\hat{T}\coloneqq T(\hat{\beta})$ for all observations.

7: Estimate $\hat{\Psi}_{iTMLE}=\mathbb{P}_n\{I(A=a^\dagger)Y+I(A=a^\circ)\hat{T}\}$

Appendix H. Extensions to discrete exposure variables with more than two levels

Extensions to discrete exposure variables with more than two levels is straightforward. To see this, we can show that our estimand can be written as follows:

$$\Psi := P(A=a^\dagger)E(Y|A=a^\dagger) + \sum_{\forall a^\circ \neq a^\dagger} P(A=a^\circ) \sum_{m,l} E(Y|M=m,L=l,A=a^\circ) f(m|a^\dagger,l) f(l|a^\circ).$$

It then follows that in this extension, the efficient influence function for Ψ is given by: The efficient influence function $\varphi^{\text{eff}}(\mathcal{O})$ for $A_Y=(L,A,M,Y)$ is given by

$$\begin{aligned}\varphi^{\text{eff}}(\mathcal{O}) = I(A=a^\dagger)Y + \sum_{\forall a^\circ \neq a^\dagger} I(A=a^\circ)\psi_3 + \\ \sum_{\forall a^\circ \neq a^\dagger} \Bigg[\frac{I(A=a^\circ)P(A=a^\dagger|M,L)P(A=a^\circ|L)}{P(A=a^\circ|M,L)P(A=a^\dagger|L)} \{Y - b_0(M,L)\} + \\ \frac{I(A=a^\dagger)P(A=a^\circ|L)}{P(A=a^\dagger|L)} \{b_0(M,L) - h_\dagger(L)\} + \\ I(A=a^\circ)\{h_\dagger(L) - \psi_3\} \Bigg] - \Psi,\end{aligned}$$

which can also be re-expressed as

$$\begin{aligned}\varphi^{\text{eff}}(\mathcal{O}) = I(A=a^\dagger)Y + \sum_{\forall a^\circ \neq a^\dagger} I(A=a^\circ)\psi_3 + \\ \sum_{\forall a^\circ \neq a^\dagger} \Bigg[\frac{I(A=a^\circ)f(M|A=a^\dagger,L)}{f(M|A=a^\circ,L)} \{Y - b_0(M,L)\} + \\ \frac{I(A=a^\dagger)P(A=a^\circ|L)}{P(A=a^\dagger|L)} \{b_0(M,L) - h_\dagger(L)\} + \\ I(A=a^\circ)\{h_\dagger(L) - \psi_3\} \Bigg] - \Psi.\end{aligned}$$

The weighted estimator will still be sample-bounded, but will need to be slightly modified in the following way:

Algorithm 5 Algorithm for Weighted ICE (generalized frontdoor formula for discrete exposure with more than two levels)

1: Non-parametrically compute $P(A=a)$ for all values of $a \in \mathcal{A}$.
2: Compute the MLEs $\hat{\kappa}$ of κ from the observed data for the treatment model $P(A=a|L;\kappa)$. In addition, compute the MLEs $\hat{\alpha}$ of α from the observed data for the treatment model

$P(A{=}a|M,L;\alpha)$, or compute the MLEs $\hat{\gamma}$ of γ from the observed data for the mediator model $P(M{=}m|A,L;\gamma)$

3: For all levels of $a^\circ \in \mathcal{A}$ that is not equal to $a^\dagger$, do the following:

A. In the individuals whose $A{=}a^\circ$, fit a regression model $Q_{a^\circ}(M,L;\theta_{a^\circ}){=}g^{-1}\{\theta_{a^\circ}^T\phi_{a^\circ}(M,L)\}$ for $b_{0,a^\circ}(M,L){=}E(Y|M,L,A{=}a^\circ)$ where the score function for each observation is weighted by $\hat{W}_{1,a^\circ}$ where $\hat{W}_{1,a^\circ}1$ equals

$$\frac{P(A{=}a^\circ|L;\hat{\kappa})P(A{=}a^\dagger|M,L;\hat{\alpha})}{P(A{=}a^\dagger|L;\hat{\kappa})P(A{=}a^\circ|M,L;\hat{\alpha})}$$

if $\hat{\alpha}$ was estimated in the previous step, or $\hat{W}_{1,a^\circ}$ equals

$$\frac{f(M|A{=}a^\dagger,L;\hat{\gamma})}{f(M|A{=}a^\circ,L;\hat{\gamma})}$$

if $\hat{\gamma}$ was estimated in the previous step. Moreover, $\phi_{a^\circ}(M,L)$ is a known function of M and L. More specifically, we solve for θ in the following estimating equations:

$$\mathbb{P}_n\Big[I(A{=}a^\circ)\phi_{a^\circ}(M,L)\hat{W}_{1,a^\circ}\{Y-Q_{a^\circ}(M,L;\theta_{a^\circ})\}\Big]{=}0.$$

B. In those whose $A{=}a^\dagger$, fit a regression model $R_{a^\circ}(L;\eta_{a^\circ}){=}g^{-1}\{\eta_{a^\circ}^T\Gamma_{a^\circ}(L)\}$ for $h_\dagger(L){=}E(b_{a^\circ}(M,L)|L,A{=}a^\dagger)$ where the score function for each observation is weighted by

$$\hat{W}_{2,a^\circ}{=}\frac{P(A{=}a^\circ|L;\hat{\kappa})}{P(A{=}a^\dagger|L;\hat{\kappa})}.$$

Here, $\gamma_{a^\circ}(L)$ is a known function of L. More specifically, we solve for η in the following estimating equations:

$$\mathbb{P}_n\Big[I(A{=}a^\dagger)\Gamma_{a^\circ}(L)\hat{W}_{2,a^\circ}\Big\{Q_{a^\circ}(M,L;\hat{\theta}_{a^\circ})-R_{a^\circ}(L;\eta_{a^\circ})\Big\}\Big]{=}0.$$

C. In those whose $A{=}a^\circ$, fit another regression model $T(\beta_{a^\circ}){=}g^{-1}(\beta_{a^\circ})$ for $\psi_3{=}E\{h_{\dagger,a^\circ}(L)|A{=}a^\circ\}$ with just an intercept. More specifically, we solve for β_{a° in the following estimating equations:

$$\mathbb{P}_n[I(A{=}a^\circ)\{R_{a^\circ}(L;\hat{\eta}_{a^\circ})-T(\beta_{a^\circ})\}]{=}0.$$

D. Compute $\hat{T}_{a^\circ}{\equiv}T(\hat{\beta}_{a^\circ})$ for all observations.

4: Estimate $\hat{\Psi}_{WICE}{=}\mathbb{P}_n\Big\{I(A{=}a^\dagger)Y+\sum_{\forall a^\circ\neq a^\dagger}I(A{=}a^\circ)\hat{T}_{a^\circ}\Big\}$

Appendix I. Additional simulation study

The data-generating mechanism for our second simulation study and model specifications are provided in Table 3. We consider four scenarios to illustrate the robustness of our proposed estimator to model misspecification. We consider four model specification scenarios: (1) all models are correctly specified, (2) only the models for $b_0(M,L)$ and $h_\dagger(L)$ are correctly specified, (3) only the models for $b_0(M,L)$ and $P(A{=}a|L)$ are correctly specified, and (4) only the models for $P(M{=}m|A,L)$ *and* $P(A{=}a|L)$ are correctly specified. The correct mediator model in the specification scenarios is the one used in the data generation process, and the exposure and outcome models are approximately correctly specified by including pairwise interactions between all the variables to ensure flexibility.

Table 4 shows the results from the simulation study. As expected by our theoretical derivations, when all of the working models are correctly specified, all of the estimators are nearly unbiased. The AIPW estimator and our proposed weighted ICE estimator are also nearly unbiased in the three model misspecification settings whereas the IPW and ICE estimators are not all unbiased.

Data generating mechanism	
$U\sim$	$\text{Ber}\{0.3\}$
$L_1\sim$	$\text{Ber}\{0.6\}$
$L_2\sim$	$\text{Ber}\{\text{expit}(1{+}4L_1)\}$
$A\sim$	$\text{Ber}\{\text{expit}(-1{+}L_1{+}2L_2{+}5L_1L_2{+}U)\}$
$M\sim$	$\text{Ber}\{\text{expit}(1{+}A{-}3L_1{+}2L_2{-}5L_1L_2)\}$
$Y\sim$	$\text{Ber}\{\text{expit}(-2.25{+}2A{-}5M{-}2AM{+}2L_1{-}2L_2{-}5L_1L_2{+}U)\}$
Model misspecification	
Scenario 2	$P(M{=}m\|L,A;\gamma){=}\text{expit}(\gamma_0{+}\gamma_1A{+}\gamma_2L_2)$ $P(A{=}a\|L;\kappa){=}\text{expit}(\kappa_0{+}\kappa_1L_2)$
Scenario 3	$P(M{=}m\|L,A;\gamma){=}\text{expit}(\gamma_0{+}\gamma_1A{+}\gamma_2L_2)$ $R(L;\theta){=}\text{expit}(\eta_0{+}\eta_1L_2)$
Scenario 4	$Q(M,L;\theta){=}\text{expit}(\theta_0{+}\theta_1M{+}\theta_2L_1{+}\theta_3L_2)$ $R(L;\theta){=}\text{expit}(\eta_0{+}\eta_1L_2(1{-}L_1))$

Table 3: Data generating mechanism and model misspecifications for scenarios in simulation study.

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Bias	SE	Bias_s	Bias	SE	Bias_s	Bias	SE	Bias_s	Bias	SE	Bias_s
n=100												
IPW	0.05	1.16	4.24	0.80	2.13	*37.41*	0.05	1.16	4.24	3.40	1.64	**208.00**
ICE	0.05	1.10	4.70	0.05	1.10	4.70	0.64	2.02	*31.51*	0.82	2.45	*33.33*
AIPW	0.11	1.20	9.56	-	-	-	0.12	1.23	9.64	0.11	1.22	9.16
WICE	0.06	1.11	5.06	0.05	1.10	4.76	0.00	1.08	0.00	0.01	1.13	1.18
n=250												
IPW	-0.03	0.77	-4.24	0.92	1.98	**46.55**	-0.03	0.77	-4.24	3.43	1.01	**338.39**
ICE	-0.01	0.70	-1.28	-0.01	0.70	-1.28	0.81	1.93	**42.10**	1.09	2.61	**41.87**
AIPW	0.00	0.70	0.31	-	-	-	0.00	0.73	0.64	0.00	0.70	-0.11
WICE	-0.01	0.70	-1.32	-0.01	0.70	-1.07	-0.04	0.70	-6.09	-0.04	0.69	-6.40
n=500												
IPW	0.01	0.93	1.17	1.16	2.19	**52.69**	0.01	0.93	1.17	3.45	0.73	**471.59**
ICE	-0.01	0.48	-1.46	-0.01	0.48	-1.46	0.95	1.91	**49.67**	1.21	2.65	**45.61**
AIPW	0.00	0.49	0.77	-	-	-	0.02	0.57	4.17	0.00	0.48	-0.66
WICE	-0.01	0.48	-1.21	-0.01	0.48	-1.20	-0.02	0.49	-4.29	-0.02	0.49	-4.80

Table 4: Simulation results: Bias, standard error (SE), and standardized bias (Bias_s) are multiplied by 100.

Appendix J. Asymptotic properties

In observational studies, model misspecification in the estimation of nuisance functions can induce biased estimates of the ACE. In recent years, there has been an explosion in developing flexible data-adaptive methods (e.g. kernel smoothing, generalized additive models, ensemble learners, random forest) combined with doubly robust estimators that can reduce the risk of model misspecification and provide valid causal inference. These machine learning techniques offer more protection against model misspecficiation than the parametric models.

From first order expansion of a singly-robust plug-in estimator (IPW and ICE estimators), it can be shown that we require the nuisance parameter estimators to converge to the truth at rate $n^{-1/2}$. However, this is not possible for non-parametric conditional mean functions as this rate is not attainable for these types of functions. However when doubly robust estimators are used with data-adaptive methods this issue largely disappears as doubly robust estimators enjoy the small bias property (Newey et al., 2004).

In this section we will examine the Remainder or Bias term from the following decomposition. For notational brevity, we suppress $\mathcal{O}$ in the equations below. For generality, suppose that $\Psi(\hat{P})$ is an estimator that solves the estimating equations based on the efficient influence function. We have that

$$\sqrt{n}(\Psi(\hat{P})-\Psi(P))=\sqrt{n}\left[\mathbb{P}_n(\varphi^{eff}(\hat{P}))-P(\varphi^{eff}(\hat{P}))\right]+\sqrt{n}\left[\Psi(\hat{P})+P(\varphi^{eff}(\hat{P}))-\Psi(P)\right]$$
$$=\mathbb{G}_n(\varphi^{eff}(P))+\mathbb{G}_n[\varphi^{eff}(\hat{P})-\varphi^{eff}(P)]+$$

$$\begin{aligned}&\sqrt{n}\Big[\Psi(\hat{P})+P(\varphi^{eff}(\hat{P}))-\Psi(P)\Big]\\=&\underbrace{\mathbb{G}_n(\varphi(P))}_{T_1}+\underbrace{\mathbb{G}_n[\varphi(\hat{P})-\varphi(P)]}_{T_2}+\\&\sqrt{n}\Bigg[\underbrace{\Psi(\hat{P})+P(\varphi^{eff}(\hat{P}))-\Psi(P)}_{R}\Bigg]\end{aligned}$$

where $\mathbb{G}_n[X]=\sqrt{n}(\mathbb{P}_n-P)(X)$ for any X and we define $\varphi(\mathcal{O};\tilde{P})=\varphi^{eff}(\mathcal{O};\tilde{P})+\Psi(\tilde{P})$ for any $\tilde{P}$. The first term given by T_1 is a centered sample average which converges to a mean zero Normal distribution by the central limit theorem. The second term is known as an empirical process term, which can be shown to be $o_p(1)$ if we assume that nuisance functions and their corresponding estimators are not too complex and belong to Donsker class. Alternatively, one can use sample splitting and cross fitting to overcome issues with overfitting (Chernozhukov et al., 2018).

Formally, we assume the following conditions for the first two terms:

C1. $E\left[\varphi(\mathcal{O})^2\right]<\infty$

C2. $\varphi(\mathcal{O})$ and $\hat{\varphi}(\mathcal{O})$ belong to a Donsker family.

C3. $\|\hat{\varphi}(\mathcal{O})-\varphi(\mathcal{O})\|_2^2 \xrightarrow{p} 0$

where $\hat{\varphi}(\mathcal{O})$ is analogously as $\varphi(\mathcal{O})$ but evaluated at $\hat{P}$, and where all nuisance functions estimators are exactly the same as those in the TMLE estimator. Similarly, $\hat{\varphi}^{eff}(\mathcal{O})$ can be defined analogously. It is not hard to show that $\mathbb{P}_n\{\hat{\varphi}^{\text{eff}}(\mathcal{O})\}=0$ by construction.

The last term is known as the remainder or bias term. We will need to show that $R=o_p(1)$ under some conditions about the convergence rates of the nuisance functions.

$$\Psi(\hat{P})+P(\varphi^{eff}(\hat{P}))-\Psi(P)=$$
$$E_P\Bigg[\underbrace{\frac{I(A=a^\circ)\hat{f}(M|a^\dagger,L)}{\hat{f}(M|a^\circ,L)}(Y-\hat{b}_0(M,L))}_{(A)}+\underbrace{\frac{I(A=a^\dagger)\hat{f}(a^\circ|L)}{\hat{f}(a^\dagger|L)}(\hat{b}_0(M,L)-\hat{h}_\dagger(L))+I(A=a^\circ)\hat{h}_\dagger(L)-P(A=a^\circ)\psi_3}_{(B)}\Bigg].$$

We examine the terms in blue (henceforth denoted as (A)) and term in red (henceforth denoted as (B)) in detail. Starting with the term in (B):

$$
\begin{aligned}
(B) &= E_P\left[I(A{=}a^\circ)(\hat{h}_\dagger(L){-}h_\dagger(L)){+}\frac{I(A{=}a^\dagger)\hat{f}(a^\circ|L)}{\hat{f}(a^\dagger|L)}(\hat{b}_0(M,L){-}\hat{h}_\dagger(L))\right]\\
&= E_P\left[I(A{=}a^\circ)(\hat{h}_\dagger(L){-}h_\dagger(L)){+}\frac{I(A{=}a^\dagger)\hat{f}(a^\circ|L)}{\hat{f}(a^\dagger|L)}\Big\{E_P(\hat{b}_0(M,L)|A{=}a^\dagger,L){-}\hat{h}_\dagger(L)\Big\}\right]\\
&= E_P\Bigg[f(a^\circ|L)(\hat{h}_\dagger(L){-}h_\dagger(L)){+}\frac{f(a^\dagger|L)\hat{f}(a^\circ|L)}{\hat{f}(a^\dagger|L)}\Big\{E_P(\hat{b}_0(M,L)|A{=}a^\dagger,L){-}\hat{h}_\dagger(L)\Big\}+\\
&\qquad f(a^\circ|L)\Big\{E_P(\hat{b}_0(M,L)|A{=}a^\dagger,L){-}E_P(\hat{b}_0(M,L)|A{=}a^\dagger,L)\Big\}\Bigg]\\
&= E_P\left[\Big\{E_P(\hat{b}_0(M,L)|A{=}a^\dagger,L){-}\hat{h}_\dagger(L)\Big\}\left\{\frac{f(a^\dagger|L)\hat{f}(a^\circ|L)}{\hat{f}(a^\dagger|L)}{-}f(a^\circ|L)\right\}\right]+\\
&\qquad E_P\Big\{E_P\Big(\hat{b}_0(M,L){-}b_0(M,L)|A{=}a^\dagger,L\Big)I(A{=}a^\circ)\Big\}\\
&= E_p\left[\Big\{E_P(\hat{b}_0(M,L)|A{=}a^\dagger,L){-}\hat{h}_\dagger(L){+}h_\dagger(L){-}h_\dagger(L)\Big\}\Big\{\hat{f}(a^\circ|L){-}f(a^\circ|L)\Big\}\frac{1}{\hat{f}(a^\dagger|L)}\right]\\
&\qquad E_P\Big\{E_P\Big(\hat{b}_0(M,L){-}b_0(M,L)|A{=}a^\dagger,L\Big)I(A{=}a^\circ)\Big\}\\
&= E_P\left[\{h_\dagger(L){-}\hat{h}_\dagger(L)\}\Big\{\hat{f}(a^\circ|L){-}f(a^\circ|L)\Big\}\frac{1}{\hat{f}(a^\dagger|L)}\right]+\\
&\qquad E_P\left[E_P(\hat{b}_0(M,L){-}b_0(M,L)|A{=}a^\dagger,L)\Big\{\hat{f}(a^\circ|L){-}f(a^\circ|L)\Big\}\frac{1}{\hat{f}(a^\dagger|L)}\right]+\\
&\qquad \underbrace{E_P\Big\{E_P\Big(\hat{b}_0(M,L){-}b_0(M,L)|A{=}a^\dagger,L\Big)I(A{=}a^\circ)\Big\}}_{(B.2)}.
\end{aligned}
$$

We keep in mind the term in purple (term (B.2)) as we expand upon term (A):

$$
\begin{aligned}
(A) &= E_P\left[\frac{I(A{=}a^\circ)\hat{f}(M|a^\dagger,L)}{\hat{f}(M|a^\circ,L)}(Y{-}\hat{b}_0(M,L))\right]\\
&= E_P\left[\frac{I(A{=}a^\circ)\hat{f}(M|a^\dagger,L)}{\hat{f}(M|a^\circ,L)}\Big(b_0(M,L){-}\hat{b}_0(M,L)\Big)\right]\\
&= E_P\left(I(A{=}a^\circ)E_P\left[\frac{\hat{f}(M|a^\dagger,L)f(M|a^\circ,L)}{\hat{f}(M|a^\circ,L)f(M|a^\dagger,L)}\Big\{b_0(M,L){-}\hat{b}_0(M,L)\Big\}|A{=}a^\dagger,L\right]\right).
\end{aligned}
$$

Now, adding (A) and (B.2) (term in purple) we get the following:

$$
E_P\left(I(A{=}a^\circ)E_P\left[\left\{\frac{\hat{f}(M|a^\dagger,L)f(M|a^\circ,L)}{\hat{f}(M|a^\circ,L)f(M|a^\dagger,L)}{-}1\right\}\Big\{b_0(M,L){-}\hat{b}_0(M,L)\Big\}|A{=}a^\dagger,L\right]\right)
$$

$$=E_P\left(I(A=a^\circ)E_P\left[\left\{\frac{f(M|a^\circ,L)}{f(M|a^\dagger,L)}\frac{1}{\hat{f}(M|a^\circ,L)}\right\}\left\{\hat{f}(M|a^\dagger,L)-f(M|a^\dagger,L)\right\}\left\{b_0(M,L)-\hat{b}_0(M,L)\right\}|A=a^\dagger,L\right]\right)+$$
$$E_P\left(I(A=a^\circ)E_P\left[\left\{\frac{1}{\hat{f}(M|a^\circ,L)}\right\}\left\{f(M|a^\circ,L)-\hat{f}(M|a^\circ,L)\right\}\left\{b_0(M,L)-\hat{b}_0(M,L)\right\}|A=a^\dagger,L\right]\right).$$

Thus, together we have (A)+(B) equals:

$$E_P\left[\{h_\dagger(L)-\hat{h}_\dagger(L)\}\left\{\hat{f}(a^\circ|L)-f(a^\circ|L)\right\}\frac{1}{\hat{f}(a^\dagger|L)}\right]+$$
$$E_P\left[E_P(\hat{b}_0(M,L)-b_0(M,L)|A=a^\dagger,L)\left\{\hat{f}(a^\circ|L)-f(a^\circ|L)\right\}\frac{1}{\hat{f}(a^\dagger|L)}\right]+$$
$$E_P\left(I(A=a^\circ)E_P\left[\left\{\frac{f(M|a^\circ,L)}{f(M|a^\dagger,L)}\frac{1}{\hat{f}(M|a^\circ,L)}\right\}\left\{\hat{f}(M|a^\dagger,L)-f(M|a^\dagger,L)\right\}\left\{b_0(M,L)-\hat{b}_0(M,L)\right\}|A=a^\dagger,L\right]\right)+$$
$$E_P\left(I(A=a^\circ)E_P\left[\left\{\frac{1}{\hat{f}(M|a^\circ,L)}\right\}\left\{f(M|a^\circ,L)-\hat{f}(M|a^\circ,L)\right\}\left\{b_0(M,L)-\hat{b}_0(M,L)\right\}|A=a^\dagger,L\right]\right).$$

By an application of Cauchy-Schwartz, we can show that as long as:

1. $\|\hat{h}_\dagger(L)-h_\dagger(L)\|\ \|\hat{f}(a^\circ|L)-f(a^\circ|L)\|=O_p(n^{-\nu})$, and
2. $\|\hat{b}_0(M,L)-b_0(M,L)\|\ \|\hat{f}(a^\circ|L)-f(a^\circ|L)\|=O_p(n^{-\nu})$, and
3. $\|\hat{b}_0(M,L)-b_0(M,L)\|\ \|\hat{f}(M|a,L)-f(M|a,L)\|=O_p(n^{-\nu}),\ \ \forall a$

for $\nu>1/2$ and where $\|f(x)\|=\left\{\int|f(x)|^2dP(x)\right\}^{1/2}$, i.e. the $L_2(P)$ norm. Then, $\sqrt{n}\Big\{\Psi(\hat{P})+P(\varphi^{eff}(\hat{P}))-\Psi(P)\Big\}=o_p(1)$. This can be accomplished, for example, if the nuisance functions are each consistently estimated at a rate faster than $n^{-1/4}$.

Note that $h_\dagger(L)=\sum_m b_0(m,L)f(m|a^\dagger,L)$. In our estimators, we propose estimating $h_\dagger(L)$ by regressing $b_0(M,L)$ on L in those whose $A=a^\dagger$ to ensure sample-boundedness. However, if we estimate $h_\dagger(L)$ by calculating $\sum_m \hat{b}_0(m,L)\hat{f}(m|a^\dagger,L)$ explicitly (e.g., as in Fulcher et al., 2020), under the AIPW estimator of Fulcher et al. (2020) (and the iterative TMLE), the remainder term reduces to the following:

$$E_P\left[\sum_m \hat{b}_0(m,L)\left\{f(m|A=a^\dagger,L)-\hat{f}(m|A=a^\dagger,L)\right\}\left\{\hat{f}(a^\circ|L)-f(a^\circ|L)\right\}\frac{1}{\hat{f}(a^\dagger|L)}\right]+$$
$$E_P\left(I(A=a^\circ)E_P\left[\left\{\frac{f(M|a^\circ,L)}{f(M|a^\dagger,L)}\frac{1}{\hat{f}(M|a^\circ,L)}\right\}\left\{\hat{f}(M|a^\dagger,L)-f(M|a^\dagger,L)\right\}\left\{b_0(M,L)-\hat{b}_0(M,L)\right\}|A=a^\dagger,L\right]\right)+$$

$$E_P\left(I(A{=}a^\circ)E_P\left[\left\{\frac{1}{\hat{f}(M|a^\circ,L)}\right\}\left\{f(M|a^\circ,L){-}\hat{f}(M|a^\circ,L)\right\}\left\{b_0(M,L){-}\hat{b}_0(M,L)\right\}|A{=}a^\dagger,L\right]\right).$$

Then we can show that if the model for $f(M|A,L)$ is correctly specified, then the remainder term reduces to the following asymptotically:

$$E_P\left(I(A{=}a^\circ)E_P\left[\left\{\frac{f(M|a^\circ,L)}{f(M|a^\dagger,L)}\frac{1}{f^*(M|a^\circ,L)}\right\}\left\{f^*(M|a^\dagger,L){-}f(M|a^\dagger,L)\right\}\left\{b_0(M,L){-}b_0^*(M,L)|A{=}a^\dagger,L\right\}\right]\right)+$$
$$E_P\left(I(A{=}a^\circ)E_P\left[\left\{\frac{1}{f^*(M|a^\circ,L)}\right\}\{f(M|a^\circ,L){-}f^*(M|a^\circ,L)\}\left\{b_0(M,L){-}b_0^*(M,L)|A{=}a^\dagger,L\right\}\right]\right)+o_p(1)$$

where $f^*(M|A,L)$ and $b_0^*(M,L)$ denote the limiting values of $\hat{f}(M|A,L)$ and $\hat{b}_0(M,L)$. This gives intuition to why the augmented inverse probability weighted estimator proposed in Fulcher et al. (2020) is consistent when models for $b_0(M,L)$ and $P(A{=}a|L)$ are correctly specified, or when the model for $P(M{=}m|A,L)$ is correctly specified.

Similarly, it can also be shown that as long as:

1. $\|\hat{h}_\dagger(L){-}h_\dagger(L)\|\ \|\hat{f}(a^\circ|L){-}f(a^\circ|L)\|{=}O_p(n^{-\nu})$, and
2. $\|\hat{b}_0(M,L){-}b_0(M,L)\|\ \|\hat{f}(a|L){-}f(a|L)\|{=}O_p(n^{-\nu}),\ \ \forall a$, and
3. $\|\hat{b}_0(M,L){-}b_0(M,L)\|\ \|\hat{f}(a|M,L){-}f(a|M,L)\|{=}O_p(n^{-\nu}),\ \ \forall a$

for $\nu{>}1/2$ and where $\|f(x)\|{=}\left\{\int|f(x)|^2dP(x)\right\}^{1/2}$, i.e. the $L_2(P)$ norm, then, $\sqrt{n}\Big\{\Psi(\hat{P}){+}P(\varphi^{eff}(\hat{P})){-}\Psi(P)\Big\}{=}o_p(1)$. This can be accomplished, for example, if the nuisance functions are each consistently estimated at a rate faster than $n^{-1/4}$.

References

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 357–363, 2005.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 2. New York: Springer, 1998.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

L.M. Collins, J.L. Schafer, and C.M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001.

A Philip Dawid. Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450):407–424, 2000.

Vanessa Didelez. Causal concepts and graphical models. In *Handbook of graphical models*, pages 353–380. CRC Press, 2018.

Vanessa Didelez. Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime data analysis*, 25:593–610, 2019.

Deborah Dowell, Tamara M Haegerich, and Roger Chou. Cdc guideline for prescribing opioids for chronic pain—united states, 2016. *Jama*, 315(15):1624–1645, 2016.

Deborah Dowell, Kathleen R Ragan, Christopher M Jones, Grant T Baldwin, and Roger Chou. Cdc clinical practice guideline for prescribing opioids for pain—united states, 2022. *MMWR Recommendations and Reports*, 71(3):1, 2022.

I. R. Fulcher, I. Shpitser, S. Marealle, and E. J. Tchetgen Tchetgen. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):199–214, 2020.

Sandro Galea. An argument for a consequentialist epidemiology. *American journal of epidemiology*, 178(8):1185–1191, 2013.

D. Gerdeman. Minorities who 'whiten' job resumes get more interviews. *Harvard Business School: Working Knowledge*, 2017.

Susan Gruber and Mark J van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 2010.

Miguel A Hernán. Invited commentary: hypothetical interventions to define causal effects—afterthought or prerequisite? *American journal of epidemiology*, 162(7):618–620, 2005.

Miguel A Hernán and Sarah L Taubman. Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *International journal of obesity*, 32 (3):S8–S14, 2008.

Miguel A Hernán and Tyler J VanderWeele. Compound treatments and transportability of causal inference. *Epidemiology (Cambridge, Mass.)*, 22(3):368, 2011.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

Kosuke Inoue, Beate Ritz, and Onyebuchi A Arah. Causal effect of chronic pain on mortality through opioid prescriptions: Application of the front-door formula. *Epidemiology*, 33(4): 572–580, 2022.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.

S. K. Kang, K. A. DeCelles, A. Tilcsik, and S. Jun. Whitened résumés: Race and self-presentation in the labor market. *Administrative Science Quarterly*, 61(3):469–502, 2016.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383, 2010.

L. Liu, R. Mukherjee, and J. M. Robins. On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3):518–539, 2020.

H. Merskey and N. Bogduk. Classification of chronic pain, iasp task force on taxonomy, 1994.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 62(6):1349–1382, 1994.

Whitney K Newey, Fushing Hsieh, and James M Robins. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72(3):947–962, 2004.

Judea Pearl. Mediating instrumental variables. *Technical report*, 1993.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *CSSS, University of Washington Series. Working Paper*, 128(30):2013, 2013.

J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.

J. Robins, L. Li, E. Tchetgen, and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and statistics: essays in honor of David A. Freedman*, 2:335–421, 2008.

J. M. Robins and T. S. Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, 84:103–158, 2010.

James M Robins, Thomas S Richardson, and Ilya Shpitser. An interventionist approach to mediation analysis. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 713–764. 2022.

J. Schafer and J. Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, 13(4):279, 2008.

Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035, 2013.

Ilya Shpitser and Eli Sherman. Identification of personalized effects associated with causal pathways. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2018. NIH Public Access, 2018.

Ilya Shpitser, Thomas S Richardson, and James M Robins. Multivariate counterfactual systems and causal graphical models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 813–852. 2022.

M. J. Stensrud, M. A. Hernán, E. J. Tchetgen Tchetgen, J. M. Robins, V. Didelez, and J. G. Young. A generalized theory of separable effects in competing event settings. *Lifetime data analysis*, 27:588–631, 2021.

Mats J Stensrud, James M Robins, Aaron Sarvet, Eric J Tchetgen Tchetgen, and Jessica G Young. Conditional separable effects. *Journal of the American Statistical Association*, 118(544):1–29, 2022a.

Mats J Stensrud, Jessica G Young, Vanessa Didelez, James M Robins, and Miguel A Hernán. Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, 117(537):175–183, 2022b.

Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816, 2012.

E. J. Tchetgen Tchetgen and I. Shpitser. Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101(4):849–864, 2014.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*, volume 10. Springer, 2011.

M. J. Van der Laan and S. Rose. *Targeted learning in data science*. Springer, 2018.

Mark van der Laan and Susan Gruber. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *International Journal of Biostatistics*, 12(1):351–378, 2009.

A. W. Van Der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

S. Vansteelandt, A. Rotnitzky, and J.M. Robins. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94(4):841–860, 2007.

L. Wen, J. Marcus, and J. Young. Intervention treatment distributions that depend on the observed treatment process and model double robustness in causal survival analysis. *Statistical Methods in Medical Research*, 32(3):509–523, 2023.