Judea Pearl’s insight that, when errors are assumed independent, the Pure (aka Natural) Direct Effect (PDE) is non-parametrically identified via the Mediation Formula was “path-breaking” in more than one sense. In the same paper, Pearl described a thought-experiment as a way to motivate the PDE. Analysis of this experiment led Robins and Richardson to a novel way of conceptualizing direct effects in terms of interventions on an expanded graph in which treatment is decomposed into multiple separable components. We further develop this novel theory here, showing that it provides a self-contained framework for discussing mediation without reference to cross-world (nested) counterfactuals or interventions on the mediator. The theory preserves the dictum “no causation without manipulation” and makes questions of mediation empirically testable in future randomized controlled trials. Even so, we prove the interventionist and nested counterfactual approaches remain tightly coupled under a non-parametric structural equation model except in the presence of a “recanting witness.” In fact, our analysis also leads to a simple, sound, and complete algorithm for determining identification in the (non-interventionist) theory of path-specific counterfactuals.

38.1 Introduction

In the companion paper [Chapter 41], we described graphical counterfactual models corresponding to the finest fully randomized causally interpretable structured tree graph (FFRCISTG) models of Robins [1986]. Such models correspond to...
conditional independencies encoded by the d-separation criterion in single-world intervention graphs (SWIGs). We gave a general identification theory for treatment effects in such models with hidden variables that is a generalization of the theory described in [Tian and Pearl 2002, Shpitser and Pearl 2006a]. We also described the differences between the FFRCISTG/SWIG model and the non-parametric structural equation model with independent errors (NPSEM-IE).

Given that a treatment effect of \( A \) on \( Y \) is established, it is often desirable to try to understand the contribution to the total effect of \( A \) on \( Y \) of the other variables \( M \) that lie on causal pathways from \( A \) to \( Y \). Variables lying on the causal pathways are called mediators of the effect of \( A \) on \( Y \). This leads to consideration of direct and indirect effects. Several different types of direct effect have been considered previously. Most have asked whether “the outcome (\( Y \)) would have been different had cause (\( A \)) been different, but the level of the mediator (\( M \)) remained unchanged.” Differences arise regarding what it means to say that \( M \) remains unchanged.

In Section 38.2, we will first review these notions, the assumptions under which they are identified and the extent to which identification claims can be verified (in principle) via an experiment. These considerations will lead to a novel way of conceptualizing direct effects introduced in Robins and Richardson [2010] and generalized in Section 38.3 herein. This novel interventionist approach does not require counterfactuals defined in terms of the mediator. We first describe our interventionist approach to causal mediation analysis in the context of direct and indirect effects. This approach (i) need not assume that the mediator \( M \) has well-defined causal effects and/or counterfactuals, but (ii) instead hypothesizes that treatment variable \( A \) can be decomposed into multiple separable components each contributing to the overall effect of treatment, (iii) preserves the dictum “no causation without manipulation,” (iv) makes questions of mediation empirically testable in future randomized controlled trials, (v) may facilitate communication with subject matter experts, and (vi) when identified from data, the identifying formulae under an interventionist approach and those obtained from the other (mediator-based) approach are identical; however, the causal effects being identified differ since they refer to intervening on different variables. This theory has been extended and applied by others in the context of mediation in survival analysis [Didelez 2019, Aalen et al. 2020] and recently to competing risks [Stensrud et al. 2020a, 2020c] and interference problems [Shpitser et al. 2017]; see also Lok [2016].

Finally, in Section 38.4, we extend this approach to arbitrary (identified) path-specific effects. This allows the (generalized) ID algorithm described in Chapter 41 plus one extra step to be used as a sound and complete algorithm for determining the identification of path-specific distributions; we also provide a version
for handling conditional queries. Finally, we describe the differences between the nested counterfactual approach and the interventionist approach under non-identification.

### 38.2 Approaches to Mediation Based on Counterfactuals Defined in Terms of the Mediator: The CDE and PDE

One approach to having $M$ remain unchanged would be for $M$ to be fixed via an intervention. On this view both $A$ and $M$ are treatments, and we consider the difference between an intervention setting $A$ to $a$ and $M$ to $m$ versus an intervention setting $A$ to $a'$ and $M$ to $m$. This leads to the definition of a **controlled direct effect** (CDE):

$$CDE_{a,a'}(m) = \mathbb{E}[Y(a',m) - Y(a,m)],$$

(38.1)

where $Y(a,m)$ is the counterfactual response of $Y$ had $A$ and $M$ been set, possibly contrary to fact, to values $a$ and $m$, respectively. Note that there is a controlled direct effect for every value $m$ of $M$. Given a causal graph $\mathcal{G}$, a straightforward application of graphical causal identification theory\(^1\) determines whether the distribution $P(Y(a,m))$, and thus whether the $CDE_{a,a'}(m)$ contrast, is identified.

There are situations in which other notions of direct effect are more natural. In particular, there are contexts in which we wish to know whether $A$ taking the value of $a'$ (rather than the baseline level $a$) would lead to a change in the value of $Y$ if the effect of $A$ on $M$ were “blocked.” Specifically, if the effect on $M$ of $A$ taking $a'$ was blocked so that the mediator $M$ takes the value $M(a)$ that $M$ would take were $A$ set to the baseline value $a$.

Along these lines, in many contexts we may ask what “fraction” of the (total) effect of $A$ on $Y$ may be attributed to a particular causal pathway. For example, consider a randomized controlled trial that investigates the effect of an anti-smoking intervention ($A$) on the outcome myocardial infarction (MI) at 2 years ($Y$) among non-hypertensive smokers. For simplicity, assume everyone in the treatment arm and no one in the placebo arm quit cigarettes, that all subjects were tested for new-onset hypertension ($M$) at the end of the first year, and no subject suffered an MI in the first year. Hence $A$, $M$, and $Y$ occur in that order. Suppose the trial showed smoking cessation had a beneficial effect on both hypertension and MI. It is natural to ask: “What fraction of the total effect of smoking cessation ($A$) on MI ($Y$) is via a pathway that does not involve hypertension ($M$)?”

---

1. See the companion paper [Chapter 41] in this volume and references therein.
These ideas lead to the pure direct effect\(^2\) (PDE) defined in Robins and Greenland [1992], which in counterfactual notation can be written as follows:

\[
PDE_{a,a'} = \mathbb{E}[Y(a', M(a)) - Y(a, M(a))] = \mathbb{E}[Y(a', M(a)) - Y(a)]. \tag{38.2}
\]

This is the difference between two quantities: first, the outcome \(Y\) that would result if we set \(A\) to \(a'\), while “holding fixed” \(M\) at the value \(M(a)\) that it would have taken had \(A\) been \(a\); second, the outcome \(Y\) that would result from simply setting \(A\) to \(a\) [and thus having \(M\) again take the value \(M(a)\)]. In Robins and Greenland [1992], the first was alternately described as the result of setting \(A\) to \(a'\) on \(Y\) when the effect of \(A\) on \(M\) is blocked, thereby leaving \(M\) unchanged from its reference value \(M(a)\). Thus the PDE interprets had “\(M\) remained unchanged” to mean “had (somehow) \(M\) taken the value that it would have taken had we fixed \(A\) to \(a\).”

As another application of this idea, Pearl [2009, p. 131] cites the following legal opinion arising in a discrimination case: “The central question in any employment discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same.” [Carson vs. Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996)].

Here \(A\) corresponds to membership in a protected class, while \(M\) corresponds to criteria, such as qualifications, that are permitted to be considered in such decisions. If the PDE is non-zero, then discrimination has taken place.

A notion of the indirect effect may be defined similarly, using the nested counterfactual \(Y(a', M(a))\). On the additive scale, direct and indirect effects may be used to give a decomposition of the average causal effect [Robins and Greenland 1992]:

\[
\frac{\mathbb{E}[Y(a')] - \mathbb{E}[Y(a)]}{\text{average causal effect}} = \frac{(\mathbb{E}[Y(a')] - \mathbb{E}[Y(a', M(a))]) + (\mathbb{E}[Y(a', M(a))] - \mathbb{E}[Y(a)])}{\text{total indirect effect}} + \frac{(\mathbb{E}[Y(a')] - \mathbb{E}[Y(a, M(a'))]) + (\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a)])}{\text{pure direct effect}}.
\]

\[
\frac{\mathbb{E}[Y(a')] - \mathbb{E}[Y(a)]}{\text{average causal effect}} = \frac{(\mathbb{E}[Y(a')] - \mathbb{E}[Y(a, M(a'))]) + (\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a)])}{\text{total direct effect}} + \frac{(\mathbb{E}[Y(a')] - \mathbb{E}[Y(a, M(a'))]) + (\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a)])}{\text{pure indirect effect}}.
\tag{38.3}
\]

Notice that the PDE depends on \(Y(a', M(a))\)—a variable in which two different levels of \(a\) are nested within the counterfactual for \(Y\). Consequently, in contrast to \(CDE_{a,a'}\), this counterfactual does not correspond to any experimental intervention on \(A\) and \(M\). This is because in order to know the value \(M(a)\) for a unit, it is necessary to set \(A\) to \(a\), but this then precludes setting \(A\) to \(a'\). This is a manifestation of the fundamental problem of causal inference: it is not possible for a single unit to

---

2. Also called the natural direct effect in Pearl [2001].
receive two different levels of the same treatment at the same time. For this reason, the counterfactual \( Y(a', M(a)) \) is referred to as a cross-world counterfactual.\(^3\)

The differences in the assumptions made by the FFRCISTG and NPSEM-IE models lead to quite different identification results for the PDE in the context of the simple DAG shown in Figure 38.3(a). These differences reflect important epistemological distinctions between the two frameworks that are described in Sections 38.2.3 and 38.2.4 below.

### 38.2.1 Two Hypothetical River Blindness Treatment Studies

We will use as a running example the following pair of hypothetical studies that are represented together in the single causal graph shown in Figure 38.1(a), as described below: A random sample of individuals in an impoverished medically underserved catchment area are selected to participate in a double-blind placebo controlled randomized trial \( A = 1 \) selected, \( A = 0 \) otherwise) of single dose therapy with the drug ivermectin \( (M = 1 \) received the drug, \( M = 0 \) otherwise) for the treatment of onchocerciasis (river blindness). The outcome is diminished vision 9 months later \( (Y = 1 \) if worse than 20/100 and \( Y = 0 \) otherwise). All subjects in the trial complied with their assigned therapy. The trial was motivated in part by the fact that ivermectin was already being sold by local shop owners as a cure for river blindness without evidence for effectiveness or safety in the local population. After the trial finished, a non-governmental organization carried out a retrospective observational cohort study on a random subset of non-selected subjects \( (A = 0) \), collecting data on \( M \) and \( Y \).

In a subset of patients ivermectin can actually decrease visual acuity. This occurs when the larvae in the eye killed by ivermectin induce an overvigorous immune response. Because of this side effect, a clinic available to all trial participants was established to screen for the above side-effect and treat with immunosuppressive drugs if required \( (R = 1 \) if treated with immunosuppressants, \( R = 0 \) otherwise) to prevent further eye damage. The patients not selected for the study had neither access to the clinic nor to immunosuppressive therapy. Thus letting \( S = 1 (S = 0) \) denote access (no access) to a clinic, we have \( S = 1 \) iff \( A = 1 \) and \( R = 0 \) if \( A = 0 \). Finally we let \( U \) denote an unmeasured variable with \( U = 1 \) and \( U = 0 \) denoting individuals with greater versus lesser propensity to take medical treatments if offered. In the unselected subjects \( (A = 0) \), \( U \) is thus positively correlated with \( M \). In contrast, owing to random assignment of \( M, U \), and \( M \) are independent in selected subjects \( (A = 1) \). On the other hand, in selected subjects, conditional

---

3. Formally, we will say that a counterfactual expression is cross-world if it involves assigning more than one value to a single index such as \( a \) and \( a' \) in \( Y(a', M(a)) \).
Figure 38.1 (a) A DAG $\mathcal{G}$ representing the underlying structure in the combined observational and randomized trial of treatments for river blindness. Shaded variables are not observed. Here $A$ indicates a randomized trial ($a = 1$) or an observational study ($a = 0$); $M$ is whether the patient received ivermectin; $Y$ is the patient’s vision; $S$ indicates access to a clinic; $R$ is treatment with immunosuppressants; $U$ indicates the inclination of the patient to avail themselves of medical care that is offered. (b) The SWIG $\mathcal{G}(a, m)$ resulting from $\mathcal{G}$; (c) and (d) show SWIGs $\mathcal{G}(a = 0, m)$ and $\mathcal{G}(a = 1, m)$ that incorporate additional context-specific causal information. ($a^*$), ($b^*$), ($c^*$), ($d^*$) show the corresponding latent projections.

On $M$, $U$ is positively correlated with $R$ and thus with $Y$; in unselected subjects $U$ and $Y$ are independent given $M$ because treatment with $R$ is not available. (Here
we assume $U$ has an effect on $Y$ only through the treatments actually received.) Finally, records recording which of the selected subjects received immunosuppressive therapy in the trial were destroyed in a fire so the data available for analysis were solely $(A, M, Y)$, the same variables available in the observational study.

### 38.2.2 The PDE and CDE in the River Blindness Studies

We next explain the meaning of the CDE and PDE in the context of the ivermectin studies. There we suppose the true data-generating process is as described by the DAG in Figure 38.2(a). Recall that the corresponding latent projection is given in Figure 38.1(a∗) because in Figure 38.1(a) there is a causal path from $A$ to $Y$, and $U$ is a common cause of $M$ and $Y$. In this context $M(a = 0)$ is the ivermectin treatment that the patient would select in the observational study where $A = 0$. Likewise $Y(a = 1, M(a = 0))$ is the patient’s outcome if they received the ivermectin treatment as in the observational study (where $A = 0$) but, as in the randomized study (where $A = 1$), a clinic ($S = 1$) was made available to them. To see why, note that $Y(a = 1, M(a = 0))$ corresponds to the effect of setting $a = 1$ on $Y$ through all causal pathways ($S \rightarrow R \rightarrow Y$) not passing through ivermectin ($M$) when $M$ remains at its self-selected value $M(a = 0)$ in the observational study. Finally, $Y(a = 0) \equiv Y(a = 0, M(a = 0))$ is the patient’s outcome if they were assigned to the observational study and did not have access to the clinic. The PDE is the mean of the difference $Y(a = 1, M(a = 0)) - Y(a = 0, M(a = 0))$.

The $CDE(m = 0)$ is the mean effect on $Y$ of having ($s = 1$) versus not having ($s = 0$) a clinic available had no one received ivermectin ($m = 0$). Likewise,

\[
U \quad \xrightarrow{a = 0} \quad A \quad \xrightarrow{s = 1} \quad S(a_0) \quad \xrightarrow{m} \quad R(s, m) \quad \xrightarrow{Y(s, m)} Y(s, m)
\]

**Figure 38.2** The SWIG $G(a = 0, s = 1, m)$ associated with the DAG shown in Figure 38.1(a); here $a_0$ and $s_1$ are short for $a = 0$ and $s = 1$.

---

4. See Section 41.4.1 in the companion paper [Chapter 41] in this volume for this definition and other standard graphical definitions.

5. If patients know that a clinic is available then this may influence their decision to take ivermectin; thus, in order for the patient’s decision to be as in the observational study, they would need to make decisions relating to ivermectin treatment before being told that a clinic is available.

6. This is because $Y(a, m) = Y(S(a), m) = Y(s, m)$ since $S(a) = a$. 

\( CDE(m = 1) \) is the clinic effect when all subjects received ivermectin \((m = 1)\). In contrast, as noted above, the PDE is the mean effect of having versus not having a clinic available had subjects chosen ivermectin treatment \((M)\) as in the observational study.

**Identification of the CDE in the River Blindness Studies**

We next consider whether the \( CDE(m = 0) \) and \( CDE(m = 1) \) are identified from the available data on \( A, M, \) and \( Y \). From examining the latent projection shown in Figure 38.1(a'), one would expect that the CDE is not identified owing to the bidirected edge \( M \leftrightarrow Y \). In particular, identification does not follow from existing methods such as the do-calculus, the ID algorithm,\(^7\) or the back-door criterion [Pearl 2009], although see Tikka et al. [2019].

However, we will now show, perhaps surprisingly, that we do have identification of the CDE. This is a consequence of context specific independencies. Although such independencies cannot be represented using standard causal DAGs (or their latent projections) an extension to context-specific SWIGs due to Dahabreh et al. [2019] and Sarvet et al. [2020] makes this possible.

Consider the SWIG \( \mathcal{G}(a, m) \) resulting from a joint intervention setting \( A \) to \( a \) and \( M \) to \( m \) as shown in Figure 38.1(b) and its latent projection shown in Figure 38.1(b'). This SWIG \( \mathcal{G}(a, m) \), shown in Figure 38.1(b), represents the conditional independence relations that hold for all four of the possible counterfactual distributions \((a, m \in \{0, 1\})\) resulting from jointly intervening on \( A \) and \( M \).

Consider the two interventions setting \( a = 1 \) and \( m \in \{0, 1\} \). These interventions correspond to performing the randomized trial in which treatment \( M \) is randomly assigned and thus \( U \perp M(a = 1) \). Consequently, the distribution of the counterfactuals may be represented by the SWIG \( \mathcal{G}(a = 1, m) \), shown in Figure 38.1(d), in which the edge \( U \rightarrow M(a = 1) \) is absent.

Now consider the remaining two interventions setting \( a = 0 \), that is, assigned to the observational study. Recall that the clinic \((S)\) and (hence) treatment with immunosuppressants \((R)\) are only available to those people who were selected for the trial \((a = 1)\). Thus, for all subjects in the observational study \((a = 0)\), \( S(0) = R(0, m) = 0 \). Consequently, \( U \perp R(a = 0, m) | A, M(a = 0) \), since the inclination \((U)\) of the patient does not affect whether they have access to immunosuppressants. Hence the distribution of the counterfactuals may be represented by the

\[7\] Note that the ID algorithm is complete [Shpitser and Pearl 2006a] with respect to models defined in terms of the independences holding in a (standard) causal DAG; these independences are not context specific.
38.2 Approaches to Mediation Based on Counterfactuals Defined in Terms of the Mediator

SWIG \( \mathcal{G}(a = 0, m) \), shown in Figure 38.1(c), in which the edge \( U \rightarrow R(a = 0, m) \) is absent.\(^8\)

Applying d-separation to the latent projections in Figure 38.1(c\(^*\)) and (d\(^*\)), we see that\(^9\)

\[
Y(a, m) \perp M(a), A \quad \text{for } a = 0, 1.
\] (38.4)

Consequently, since \( Y = Y(a, M(a) = m) \) on the event \( A = a, M(a) = m \),

\[
p(Y | A = a, M = m) = p(Y(a, m)),
\] (38.5)

is identified for both \( a = 0 \) and \( a = 1 \). Hence the CDE\((m)\) of \( A \) and \( M \) on \( Y \) is identified.

As noted above, the identification (Equation (38.5)) does not follow from the DAG in Figure 38.1(a), with the latent projection in Figure 38.1(a\(^*\)). However, the independences \( U \perp R(a = 0, m) | A, M(a) = 0 \) and \( U \perp M(a = 1) | A \) can be encoded by and read from the context-specific SWIGs shown in Figure 38.1(c) and (d), respectively.\(^10\)

Together with consistency, these independences imply, respectively, \( U \perp R | A = 0 \) and \( U \perp M | A = 1 \). Since, in addition to Equation (38.4), we also have \( M(a) \perp A \) for \( a = 0, 1 \), it follows that the distribution of the counterfactuals \( \{A, M(a), Y(a, m)\} \) for all \( a, m \) obeys the FFRCISTG model associated with the graph shown in Figure 38.3(a) in which there are no bi-directed edges. Interestingly, we show below in Section 38.2.5 that \( M(a = 0) \not\perp Y(a = 1, m) \). Consequently, the distribution of the counterfactuals does not obey the NPSEM-IE associated with Figure 38.3(a), and thus as discussed below, the PDE is not identified without additional assumptions beyond those of the NPSEM-IE model.\(^11\)

---

\(^8\) Since \( S(0) \) and \( R(0, m) \) are constants we could also leave out the edges \( S(0) \rightarrow R(0, m) \) and \( m \rightarrow R(0, m) \) on the same basis, but this would not change our conclusions.

\(^9\) Recall that when testing d-separation in SWIGs, fixed nodes such as \( a = 0 \) in Figure 38.1(c\(^*\)) and \( a = 1 \) in Figure 38.1(d\(^*\)) always block paths they occur on (when they are not end points).

\(^10\) These independences cannot be read from the SWIG \( \mathcal{G}(a, m) \) shown in Figure 38.1(b) that was constructed from \( \mathcal{G} \). Note that the graph \( \mathcal{G}(a, m) \) contains the union of edges in the two context-specific SWIGs \( \mathcal{G}(a = 0, m) \) and \( \mathcal{G}(a = 1, m) \) shown in Figure 38.1(c) and (d), respectively. \( \mathcal{G}(a, m) \) thus represents the (non-context-specific) conditional independence relations that are common to all four instantiations \( a, m \in \{0, 1\} \).

\(^11\) The distribution does obey the NPSEM-IE (hence also the FFRCISTG) associated with Figure 38.1(a\(^*\)) that includes the \( M \leftrightarrow Y \) edge; see also Footnote 23.
Figure 38.3 (a) A simple DAG $G$ representing a causal model with a treatment $A$, a mediator $M$, and a response $Y$. (b) An expanded causal model where $N$ and $O$ are a decomposition of $A$; $N$ and $O$ are, respectively, the nicotine and non-nicotine components of tobacco. Thicker red edges indicate deterministic relations. (c) An expanded version $G^{ex}$ of the DAG $G$ in (a), and an edge subgraph of the DAG in (b), where $N$ does not cause $Y$ directly, and $O$ does not cause $M$ directly. In this graph the direct and indirect effects of $A$ on $Y$ may be defined via interventions on $N$ and $O$. (d) Special case of (c) in which $A$ plays the role of $O$; (e) Special case of (c) in which $A$ plays the role of $N$.

### 38.2.3 Identification of the PDE via the Mediation Formula under the NPSEM-IE for Figure 38.3(a)

We now consider the situation in which, unlike the ivermectin example above, the NPSEM-IE associated with the graph in Figure 38.3(a) holds. In this case the PDE$_{a,a'}$ is identified via the following Mediation Formula [Pearl 2001, 2012]:

$$\text{med}_{a,a'} = \sum_{m} (\mathbb{E}[Y | m, a] - \mathbb{E}[Y | m, a']) p(m | a')$$

(38.6)

$$= \left( \sum_{m} \mathbb{E}[Y | m, a] p(m | a') \right) - \mathbb{E}[Y | a'].$$

(38.7)

The proof of this result under the NPSEM-IE is as follows:

$$p(Y(a, M(a') = m) = y)$$

$$= \sum_{m} p(Y(a, M(a') = m) = y | M(a') = m) p(M(a') = m)$$

$$= \sum_{m} p(Y(a, m) = y) p(M(a') = m)$$

$$= \sum_{m} p(Y = y | A = a, M = m) p(M = m | A = a').$$

Here the first line follows from elementary probability, the second from the cross-world NPSEM-IE independence:

$$Y(a, m) \perp\!\!\!\!\perp M(a')$$

which follows from (41.5); the third follows from the FFR CSTG independence (41.6) and thus also from (41.5). Under the NPSEM-IE associated with Figure 41.5 (a*) identification fails because this cross-world independence does not hold.
38.2.4 Partial Identification of the PDE Under the FFRCISTG for Figure 38.3(a)

In contrast, under the less restrictive FFRCISTG model associated with the graph in Figure 38.3(a), the PDE is not, in general, identified. This follows from the fact that the second equality in the previous proof relies on the cross-world independence \(Y(a, m) \perp \perp M(a')\); but the FFRCISTG model does not assume any cross-world independencies. However, under the FFRCISTG the observed data implies bounds on the PDE. For example, in the case in which \(M\) and \(Y\) are binary we have the following sharp bounds [Robins and Richardson 2010]:

\[
\begin{align*}
\max\{0, p(M = 0|A = a') + p(Y = 1|A = a, M = 0) - 1\} + \\
\max\{0, p(M = 1|A = a') + p(Y = 1|A = a, M = 1) - 1\} - p(Y = 1|A = a') \\
\leq \text{PDE}_{a,a'} \leq \\
\min\{p(M = 0|A = a'), p(Y = 1|A = a, M = 0)\} + \\
\min\{p(M = 1|A = a'), p(Y = 1|A = a, M = 1)\} - p(Y = 1|A = a').
\end{align*}
\]

A proof is given in the Appendix.

38.2.5 An Example in Which an FFRCISTG Model Holds, but an NPSEM-IE Does Not

The differing results on identifiability for the PDE in the previous two sections raise the question as to whether it is most appropriate to adopt the NPSEM-IE (41.5) or FFRCISTG (41.6) assumptions in practice. As shown above, this choice matters since if one assumes the NPSEM-IE associated with the simple graph in Figure 38.3(a) then one will believe the PDE is point identified, while if one assumes the FFRCISTG associated with Figure 38.3(a) only bounds may be obtained.

It has been argued that it is hard to conceive of a realistic data-generating process under which the FFRCISTG model holds, but the NPSEM-IE model does not. However, we now show that the river blindness studies described above provides a counterexample.

Recall that due to records being destroyed in a fire, only the three variables \((A, M, Y)\) are observed. The first \((A)\) is randomly assigned and the analyst assumes (possibly incorrectly) that there is no other variable (either measured or unmeasured) that is a cause of both the mediator \(M\) and the final response \(Y\). This hypothesis would imply the causal structure depicted in Figure 38.3(a). The graph in Figure 38.3(a) would be the latent projection of Figure 38.1(a) if it were assumed incorrectly that there is no unmeasured confounder \(U\). Both the FFRCISTG and NPSEM-IE models associated with Figure 38.3(a) imply:

\[
Y(a, m) \perp \perp M(a), \quad \text{for } a = 0, 1. \tag{38.8}
\]
The NPSEM-IE also implies the cross-world independence:

\[ Y(a = 1, m) \perp \!\!\! \perp M(a = 0). \quad (38.9) \]

We have already shown that the underlying data-generating process in the ivermectin example implies Equation (38.8); see Equation (38.4) above. We now show that Equation (38.9) fails to hold for almost all laws corresponding to the NPSEM-IE associated with Figure 38.1(a). This remains true even if we impose, in addition, the context-specific counterfactual independences needed to identify the CDE that are encoded in Figure 38.1(c) and (d).

To see this, first consider the SWIG \( G(a = 0, s = 1, m) \) shown in Figure 38.2 that is constructed from the DAG in Figure 38.1(a). There is a d-connecting path \( Y(s = 1, m) \) to \( M(a = 0) \), namely \( M(a = 0) \leftarrow U \rightarrow R(s = 1, m) \rightarrow Y(s = 1, m) \). Consequently, for almost all distributions in the FFRCISTG model, it holds that \( M(a = 0) \not\!\!\! \perp \!\!\! \perp Y(s = 1, m) \). Further, we have:

\[ Y(s = 1, m) = Y(S(a = 1), m) = Y(a = 1, m), \]

where in the first equality we use that, due to determinism, \( S(a = 1) = 1 \) for all individuals; the second follows from recursive substitution.\(^{13}\) Hence Equation (38.9) does not hold\(^{14}\) from which it follows that the PDE is not identified; see Appendix 38.A.2. Consequently, the distribution of the counterfactuals \( \{A, M(a), Y(a, m) \} \) for all \( a, m \) is not in the NPSEM-IE model associated with Figure 38.3(a). Interestingly we see that by considering a SWIG with interventions on three variables \( A, S, \) and \( M \) we have shown that the FFRCISTG model plus recursive substitution and some determinism can be used to prove that a cross-world independence fails to hold; see Footnote 23 and Appendix 38.A.3.

Finally, we note that in the ivermectin example the Acyclic Directed Mixed Graph (ADMG)\(^{15}\) with fewest edges (over \( A, M, Y \)) that represents the distribution of the counterfactuals \( \{A, M(a), Y(a, m) \} \) for all \( a, m \) is Figure 38.3(a) under the FFRCISTG. In contrast, the minimal ADMG that represents this distribution under the NPSEM-IE is the graph with an additional bi-directed confounding arc shown in Figure 38.1(a*). This has the following interesting consequence: typically, people

---

12. This is also true if we assume the NPSEM-IE associated with the graph in Figure 38.1(a).
13. See Equation (41.1) in Chapter 41 in this volume.
14. Intuitively, this should not be surprising since \( U \) is a “common cause” of \( M(a = 0) \) and \( Y(a = 1, m) \) in that \( U \rightarrow M(a = 0) \) in Figure 38.1(c) while there is a path \( U \rightarrow R(a = 1, m) \rightarrow Y(a = 1, m) \) in Figure 38.1(d).
15. see Section 41.2.1 in Chapter 41 in this volume.
may make a statement such as “Figure 38.3(a) is the true causal graph.” However, we now see that this statement does not have a truth-value without clarifying whether we are referring to the NPSEM-IE or the FFRCISTG as the true underlying counterfactual model.

On the other hand, one might prefer to replace “Figure 38.3(a)” with “Figure 38.1(a*)” in the statement above, since then the modified statement holds for both counterfactual models; furthermore, Figure 38.1(a*) accurately indicates that there is confounding for the PDE. However, the disadvantage of this choice is that the inclusion of the bi-directed edge $M \leftrightarrow Y$ does not reveal that the CDE is identified via $E[Y(a, m) | A = a, M = m] = E[Y | A = a, M = m]$. 

### 38.2.6 Testable Versus Untestable Assumptions and Identifiability

Given a graph such as Figure 38.3(a), in principle, there is an empirical test of the FFRCISTG model. However, there is no additional empirical test (on the variables in the graph) for the extra assumptions made by the corresponding NPSEM-IE, which are required to identify the PDE. Consequently, there is a qualitative distinction in the testability of the identification assumptions for these two contrasts.

In more detail, identification of the $CDE(m)$, $m = 0,1$ is, in principle, subject to direct empirical test: one conducts a four-armed randomized experiment on subjects drawn from the same population, in which both $A$ and $M$ are randomly assigned to their four possible joint values. If for any $(a, m)$ the distribution in the four-arm $(A, M)$ randomized trial $p(Y(a, m) = y)$ differs from the conditional distribution in the two-armed trial $p(Y(a) = y | M(a) = m) = p(Y = y | M = m, A = a)$ in which only $A$ was randomized, then we may infer that an unmeasured common cause was present between $M$ and $Y$ and hence it was incorrect to postulate the causal DAG in Figure 38.3(a), regardless of whether we are considering the FFRCISTG or NPSEM-IE models associated with this graph.

---

16. Since the distribution of $\{A, M(a), Y(a, m)\}$ for all $a, m$ obeys the FFRCISTG corresponding to Figure 38.3(a), which is a subgraph of Figure 38.1(a*), the distribution also obeys the FFRCISTG corresponding to this latter graph.

17. Formally, this requires that one can observe the “natural” value of a variable prior to intervention.

18. Here and throughout this section, we will use $p(A, M, Y)$ to denote the observed distribution in which only $A$ is randomized; we will use $p(Y(a, m))$ for $a, m \in \{0,1\}$ to indicate the four distributions that would be observed if both $A$ and $M$ were to be randomized in a four-arm trial. This is because in such a trial in the arm in which people are assigned to $A = a, M = m$ we would observe the counterfactual $Y(a, m)$. 

In contrast, whether the PDE equals the mediation formula (38.7) cannot be empirically tested using data on \((A, M, Y)\). Even if we can directly manipulate \(M\) (in addition to \(A\)), there is no experiment involving \(A\) and \(M\) such that the resulting contrast corresponds to the PDE. This is for the following reason: to observe, for a given subject, the cross-world counterfactual \(Y(a = 1, M(a = 0))\) that occurs in the PDE, one would need to “first” assign them to \(a = 0\) and record \(M(a = 0)\), and “then” perform a “second” experiment (on the same subject) in which they are assigned to \(a = 1\) and the recorded value \(M(a = 0)\) from the “first” experiment. However, this is usually not possible for the simple reason that having assigned the patient to \(a = 0\) in the “first” experiment precludes “subsequently” assigning them to \(a = 1\), except in the rare circumstances where a valid cross-over trial is feasible.

These considerations are particularly relevant in a setting such as the ivermectin example, where, as shown above, the distribution over the counterfactual variables \(\{A, M(a), Y(a, m)\}\) for all \(a, m\) obeys the FFRCISTG but not the NPSEM-IE model associated with the causal DAG in Figure 38.3(a). In particular, an analyst who was unaware of the variables \(U, R, S\) in Figure 38.1(a) and posited the model in Figure 38.3(a) would find no evidence of confounding between \(M\) and \(Y\) even if they were to subsequently perform a four-arm \((A, M)\) randomized trial.

In summary, the PDE identification via the mediation formula (38.7) requires not only that there be no detectable single-world confounding between \(M\) and \(Y\) (as assumed by the FFRCISTG), but, in addition, that undetectable cross-world confounding also be absent.\(^{19}\) Consequently, as with the ivermectin example, it is possible for the mediation formula to give an inconsistent estimate of the PDE, yet for this to be undetectable given any randomized experiment that could be performed using the variables \(A, M, Y\) on the graph in Figure 38.3(a).

### 38.3 Interventionist Theory of Mediation

The above considerations motivate a theory of mediation based on interventions on sub-components of treatment, rather than on the mediator.

#### 38.3.1 Intervenational Interpretation of the PDE Under an Expanded Graph

As described above, the counterfactual \(E[Y(a = 1, M(a = 0))\)] and thus the PDE cannot be empirically tested by any intervention on the variables on the graph.

\(^{19}\) If, after carrying out an experiment in which \(A\) and \(M\) are randomly assigned, it is observed that \(p(Y(a, m))\) and \(p(Y | A = a, M = m)\) are statistically indistinguishable, then this would almost certainly increase the probability that a Bayesian would assign to cross-world independence holding. However, the ivermectin example shows the importance of the investigator—Bayesian or not—thinking carefully about the underlying data-generating mechanism.
However, curiously, Pearl has often argued that the PDE is a causal contrast of substantive and public-health importance by offering examples along the following lines.

**Example: Nicotine-Free Cigarette**
Consider the example discussed in Section 38.2 where we have data from a randomized smoking-cessation trial. We have data available on smoking status $A$, hypertensive status $M$ 6 months after randomization, and myocardial infarction (MI) status $Y$ at 1 year.

Following a similar argument given in Pearl [2001] to motivate the PDE, suppose that nicotine-free cigarettes will be newly available starting a year from now. The substantive goal is to use the already collected data from the smoking cessation trial to estimate the difference two years from now in the incidence of MI if all smokers were to change to nicotine-free cigarettes when they become available (in a year) compared to the incidence if all smokers were to stop smoking altogether (in a year).

Further suppose it is believed that the entire effect of nicotine on MI is through its effect on hypertensive status, while the non-nicotine toxins in cigarettes have no effect on hypertension and that there do not exist unmeasured confounders for the effect of hypertension on MI.

In this context, a researcher following the approach that has been advocated by Pearl may postulate that the smoking cessation trial is represented by the NPSEM-IE model associated with Figure 38.3(a). Under these assumptions, the MI incidence in smokers of cigarettes free of nicotine would be $E[Y(a = 1, M(a = 0))]|A = 1]$ since the hypertensive status of smokers of nicotine-free cigarettes will equal their hypertensive status under non-exposure to cigarettes. Thus $E[Y(a = 1, M(a = 0))]|A = 1]$ is precisely the incidence of MI in smokers two years from now were all smokers to change to nicotine-free cigarettes a year from now, and thus the PDE:

$$PDE = E[Y(a = 1, M(a = 0))] - E[Y(a = 0, M(a = 0))]$$

is the causal contrast of interest.

Given the assumption of an NPSEM-IE, it follows that $E[Y(a = 1, M(a = 0))]|A = 1] = \sum_m E[Y|A = 1, M = m)p(m|A = 0)$, and therefore the PDE is identified from the mediation formula applied to the data from the smoking cessation trial.

---

20. Pearl [2001, section 2] considers a similar example, but where $A$ is a drug, $M$ is aspirin taken to mitigate side-effects, and $Y$ is the final outcome.
What is interesting about Pearl’s motivation is that to argue for the substantive importance of the parameter $E[Y(a = 1, M(a = 0))]$, he tells a story about the effect of a manipulation—a manipulation that makes no reference to $M$ at all. Rather, in this context, the manipulation is to intervene to eliminate the nicotine component of cigarettes.

The most direct representation of this story is provided by the expanded DAG $\mathcal{G}^\text{ex}$ in Figure 38.3(c), where $N$ is a binary variable representing nicotine exposure and $O$ is a binary variable representing exposure to the other non-nicotine components of a cigarette. The bolded arrows from $A$ to $N$ and $O$ indicate deterministic relationships: $N(a) = O(a) = a$. This is because in the factual data (with probability one) either one smokes normal cigarettes so $A = N = O = 1$ or one is a non-smoker (i.e., ex-smoker) and $A = N = O = 0$.

This expanded graph now provides a simple interventional interpretation of the PDE. The researcher’s assumption of the NPSEM-IE associated with Figure 38.3(a) together with the existence of the variables $N$ and $O$ and associated counterfactuals imply that the nested counterfactual $Y(a = 1, M(a = 0))$ is equal to the simple counterfactual $Y(n = 0, o = 1)$, the outcome had we intervened to expose all subjects to the non-nicotine components, but not to the nicotine components.\(^{21}\) It follows that

$$PDE = E[Y(n = 0, o = 1)] - E[Y(n = 0, o = 0)]. \quad (38.11)$$

Furthermore, these assumptions imply that the graph in Figure 38.3(c) is an FFRCISTG. It then follows from proposition 41.3 in Chapter 41 in this volume that

$$E[Y(n = 0, o = 1)] = \sum_m E[Y \mid O = 1, M = m]p(m \mid N = 0), \quad (38.12)$$

where the right-hand side (RHS) is the g-formula. Thus $E[Y(n = 0, o = 1)]$ is identified provided that the terms in the g-formula are functions of the distribution of the facts $p(A, N, O, M, Y)$. Since in the factual data now available there is no subject with $N = 0$ and $O = 1$ positivity fails and one might suppose that $E[Y(n = 0, o = 1)]$ is not identified, but in fact it is. To see this, note that the event $\{O = 1, M = m\}$ is equal to the event $\{A = 1, M = m\}$ and similarly the event $\{N = 0, M = m\}$ is equal to the event $\{A = 0, M = m\}$ owing to determinism. Thus,

---

21. Notice that here we show that the “cross-world” counterfactual $Y(a = 1, M(a = 0))$ defined in the DAG in Figure 38.3(a) is equal (as a random variable) to the “non-cross cross-world” counterfactual $Y(n = 0, o = 1)$ associated with the DAG in Figure 38.3(c); see Section 38.3.6 for further discussion.
by substituting these events we conclude that

$$E[Y(n = 0, o = 1)] = \sum_m E[Y|A = 1, M = m]p(m|A = 0),$$

(38.13)

which coincides with (one of the terms in) the mediation formula.22

For a researcher following Pearl [2001], having at the outset assumed an NPSEM-IE associated with the DAG in Figure 38.3(a), the story involving N and O does not contribute to identification; rather, it served only to show that the PDE encodes a substantively important parameter.

However, from the FFRCISTG point of view, the story not only provides an interventional interpretation of the PDE but in addition makes the PDE identifiable with the mediation formula being the identifying formula. Furthermore, the story makes refutable the claim that the PDE is identified by the mediation formula. Specifically, when nicotine-free cigarettes become available, Pearl’s claim can be tested by an intervention that forces a random sample of the population to smoke nicotine-free cigarettes; if the mean of Y under this intervention differs from the RHS of Equation (38.12), Pearl’s claim is falsified. As this refers to an actual intervention, the variables (N, O) are not simply formal constructions. Without knowledge of the substantive meaning of N and O, the trial in which N is set to 0 and O is set to 1 is not possible, even in principle. See Robins and Richardson [2010] and Stensrud et al. [2020c] for discussion of substantive considerations regarding

22. Implications of Determinism: The FFRCISTG models associated with Figures 38.3(b) and (c) both imply the factuals (A,N,O,M,Y) factor with respect to the corresponding graph. Now in Figure 38.3(c) note (i) N is d-separated from Y given {M, O} and (ii) O is d-separated from M given N which imply N \perp Y | O, M and O \perp M | N, respectively. In contrast on Figure 38.3(b) neither of the above d-separations hold. Yet, since by the determinism A = N = O as random variables, both independencies also hold for the DAG in Figure 38.3(b). Furthermore, \(E[Y(n = 0, o = 1)]\) would be identified by the g-formula:

$$\sum_m E[Y|N = 0, O = 1, M = m]p(m|N = 0, O = 1)$$

were the formula a function of the factual distribution. However, it is not; the event \(\{N = 0, O = 1\}\) has probability zero due to determinism. Notwithstanding this, a naive application of the above independences might lead one to conclude that this g-formula is equal to the RHS of (Equation 38.12) and thus is identified by (Equation 38.13). The error in this argument is that \(O \perp M | N\) does not imply \(p(M = m|N = 0)\) equals \(p(M = m|N = 0, O = 1)\) when the event \(\{O = 1, N = 0\}\) has probability zero, since the latter is not well-defined.

Note that, in the presence of determinism, we have two DAGs with different adjacencies that represent the same set of factual distributions. The counterfactuals corresponding to the DAGs in Figures 38.3(b) and (c) do not represent the same set of counterfactual distributions; see Footnote 27.
whether variables \( N \) and \( O \) exist that satisfy the no direct effect assumptions of Figure 38.3(c).

**Remark 38.1**

It should be noted that, in the following sense, it is sufficient to find one of the variables \( N \) or \( O \): Specifically, if we have a well-defined intervention \( N \), satisfying:

- (n1) \( N(a) = a \) so that \( N = A \) in the observed data;
- (n2) \( N \) has no direct effect on \( Y \) relative to \( A \) and \( M \), so \( Y(a, n, m) = Y(a, m) \);
- (n3) \( A \) has no direct effect on \( M \) relative to \( N \) so \( M(a, n) = M(n) \),

then \( A \) will satisfy the conditions for \( O \); see Figure 38.3(d). Conversely, if there is a well-defined intervention \( O \) such that:

- (o1) \( O(a) = a \) so that \( O = A \) in the observed data;
- (o2) \( A \) has no direct effect on \( Y \) relative to \( O \) and \( M \), so \( Y(a, o, m) = Y(o, m) \);
- (o3) \( O \) has no direct effect on \( M \) relative to \( A \) so \( M(a, o) = M(a) \),

then \( A \) will satisfy the conditions for \( N \); see Figure 38.3(e).

We use all three variables \( (A, N, O) \) in our subsequent development since this choice is symmetric in \( N \) and \( O \), covers both cases and more closely aligns with the original motivating nicotine intervention.

Lastly, note that, as in the ivermectin example, the existence of the interventions \( N, O \) satisfying the no direct effect conditions do not imply that the mediation formula identifies the PDE because confounding between \( M \) and \( Y \) may still be present.

---

23. One can trivially construct artificial variables \( N^* \equiv A \) and \( O^* \equiv A \) such that the (degenerate) joint distribution of the factuals \( p(A, N^*, O^*, M; Y) \) will factorize according to the DAG in Figure 38.3(c) and thus satisfy \( M \perp O^*; A \mid N^* \) and \( Y \perp N^* \mid A, O^*, M \). However, these latter independencies are tautologies owing to determinism and thus do not establish, for example, \( Y(o, m) \perp M(n) \) as required by the FFRCISTG associated with Figure 38.3(c).

24. An attentive reader might wonder how it is that the first expression in the mediation formula fails to identify \( R_Y(n, o) \) in the ivermectin example (with \( A \) as “\( N \)” and \( S \) as “\( O \)”); see Remark 38.1 above). As noted earlier, the counterfactual variables \( A, M(a), Y(a, m) \) follow the conditional independences implied by the FFRCISTG model associated with the DAG in Figure 38.3(a) (in which there is no bi-directed arc between \( M \) and \( Y \)); see the end of Section 38.2.5. Under this FFRCISTG model, the absence of the \( M \leftrightarrow Y \) edge implies that there is no confounding between \( M \) and \( Y \) that is detectable via interventions on \( A \) and \( M \), since \( p(Y(a, m)) = p(Y \mid A = a, M = m) \).

However, in the ivermectin example, the expanded set of counterfactual variables \( A, O, M(a), Y(a, o) \) do not follow the FFRCISTG model corresponding to the DAG in Figure 38.3(e). To see this, consider performing, in addition to the randomized trial where \( a = 1 \) and \( s = 1 \), an intervention setting \( A \) to 0 and \( O \equiv S \) to 1, which corresponds to an observational study but with clinics and immunosuppressants. Notice that the confounding variable \( U \) in Figure 38.1(a) now becomes detectable in the following sense: If \( U \) were not present in Figure 38.13(a) then
38.3.2 Direct and Indirect Effects via the Expanded Graph

In this section we formally introduce the interventionist theory of mediation first introduced in Robins and Richardson [2010] and greatly generalized herein. We do so by continuing with the Nicotine Example. Recall that \( N \) and \( O \) were substantively meaningful variables, and the goal was to use data from a smoking cessation trial to estimate \( E[Y(n = 0, o = 1)] \), the incidence of MI through year 2 if all smokers were to change to nicotine-free cigarettes at one year. As noted above, this policy intervention was used by Pearl to motivate consideration of the PDE. However, given the public health importance of the policy question, one could instead focus directly on estimating the effect of the proposed substantive intervention given data on \((A, M, Y)\) without regard to whether it is equal to the PDE.

In fact, there are many situations where mediation analysis is applied, in which interventions on the putative mediator \( M \) are not well-defined. Consequently, substantive researchers may not wish to make reference to the corresponding counterfactuals,\(^{25}\) regardless of whether they may be formally constructed. For such researchers, the PDE parameter may not be substantively meaningful. Fortunately, the interventionist theory described herein, in contrast to Pearl’s approach, does not require reference to counterfactuals indexed by \( m \), such as \( Y(a, m) \).\(^{26}\)

As noted, the interventionist theory only requires that \( N, O, \) and interventions on them are substantively meaningful. This is a major advantage since it makes it straightforward to discuss with subject matter experts, for example, physicians, experiments that would shed light on causal pathways; this is a property not shared by the PDE.

Up to this point we have motivated our interventionist theory, based on the expanded graph, as providing an empiricist foundation for the existing mediation theory that is based on cross-world (nested) counterfactuals. However, the interventionist theory can be viewed as autonomous,\(^{27}\) providing a self-contained framework for discussing mediation without reference to cross-world (i.e., nested)

\[ p(Y(a = 0, s = 1) | M(a = 0)) = p(Y(a = 1, s = 1) | M(a = 1)). \]

This follows from Rule 3 of the po­"calculus [Malinsky et al. 2019, chapter 41] applied to the SWIG \( (a, s) \) derived from Figure 38.5(e) after replacing “\( O \)” with \( S \). However, we show in Appendix 38.A.3 that in this example if \( U \) is present then this equality does not hold in general. Consequently, the latent projection over the variables \( A, O \equiv S, M, Y \) includes an \( M \leftrightarrow Y \) edge; see also the last paragraph of Section 38.2.6.

25. We regard the existence of interventions as necessary for counterfactuals to be well-defined, but Pearl [2018, 2019] and others may take a different view.

26. The \( CDE(m) \) and \( PDE \) are defined in terms of such counterfactuals.

27. Following Wittgenstein [1922, section 6.54], we may view the theory based on nested counterfactuals as a “ladder” that we climbed to reach our interventionist theory, and now having done so, we may choose to “kick it away.”
counterfactuals. In this section, we adopt this viewpoint. However, we will see that we can prove the two theories are tightly coupled in certain settings; see Section 38.3.6.28

Concretely, consider the DAG associated with Figure 38.3(b). Unlike the model associated with Figure 38.3(a), which involves counterfactuals \( M(a) \) and \( Y(a) \), the expanded FFRCISTG model associated with Figure 38.3(b) involves counterfactuals \( M(n, o) \) and \( Y(n, o) \). Taking the interventionist view as primitive, in what follows we will discuss counterfactuals such as \( Y(n = 0, o = 1) \) that, without further assumptions, are defined solely within this larger expanded model.

**Identification of Four Arms from Two**

For the purposes of our development, consider the following three datasets all derived from the same distribution \( p \) over the one-step-ahead counterfactuals in the FFRCISTG model associated with the graph in Figure 38.3(b):

(i) The original observed data from the trial in which \( A \) was randomized, namely \( A, M, Y \);

(ii) Data from a putative four-arm \( (N, O) \) randomized trial; the data in each arm \( (n, o) \in \{0, 1\}^2 \) corresponds to \( M(n, o) \), \( Y(n, o) \);

(iii) A dataset obtained from the four-arm \( (N, O) \) trial (ii) by restricting to the two arms in which \( n = o \).

Note that in dataset (i) among people with \( A = a \) we observe \( N = O = a \), owing to determinism; hence, we observe \( M(n = a, o = a) \) and \( Y(n = a, o = a) \) on this

---

28. In related work, Lok [2016] has developed an interventional approach to mediation that also does not require interventions on the mediator in order to be well-defined. Lok introduces a notion of an ‘organic intervention’ \( (I = 1) \) that is required to satisfy certain conditions. Lok then defines notions of direct and indirect effects in terms of such organic interventions. Our interventional definitions introduced here are similar in spirit to Lok’s conditions.

However, in order to capture certain aspects of the concept of direct and indirect effects, we, unlike Lok, also require that the variables \((N \text{ and } O)\) defining our additional interventions (e.g., on \( N \)) be equal to \( A \) in the observed data. Among other things, this ensures that \( N(a) = O(a) = a \) and hence \( M(a) = M(n = a, o = a) \) and \( Y(a) = Y(n = a, o = a) \).

In contrast, an organic intervention could change the mechanism by which the mediator is produced so that the relevant counterfactual random variables for the mediator under the organic intervention \( (M(a = 0, i = 1)) \) do not correspond to those in the absence of the organic intervention \( M(a = 1) \), although they have the same distribution. See Robins [2003], section 3 for additional discussion in terms of blocking paths.
event.\textsuperscript{29} By randomization of $A$, it follows that for $a \in \{0, 1\}$,

$$p(M = m, Y = y | A = a) = p(M(n = a, o = a) = m, Y(n = a, o = a) = y | A = a)$$

$$= p(M(n = a, o = a) = m, Y(n = a, o = a) = y)$$

which is the distribution of individuals in dataset (iii). We conclude that the distribution of the data in (iii) is identified from the observed data (i). Thus our goal becomes the identification of $E[Y(n, o)]$ for $n \neq o$ from data on $M(n, o)$ and $Y(n, o)$ in the two arms with $n = o$. Motivated by the nomenclature of Stensrud et al. [2020c], when this identification is possible we will say that the effects of $N$ and $O$ on $M$ and $Y$ are separable. The following proposition provides sufficient conditions.

**Proposition 38.1** If for some $x \in \{0, 1\}$ the following two conditions hold:

$$p(M(n = x, o = 0) = m) = p(M(n = x, o = 1) = m), \quad (38.14)$$

$$p(Y(n = 1, o = x^*) = y | M(n = 1, o = x^*) = m)$$

$$= p(Y(n = 0, o = x^*) = y | M(n = 0, o = x^*) = m), \quad (38.15)$$

where $x^* = 1 - x$, then:

$$p(M(n = x, o = x^*) = m, Y(n = x, o = x^*) = y)$$

$$= p(Y(n = x^*, o = x^*) = y | M(n = x^*, o = x^*) = m)p(M(n = x, o = x) = m). \quad (38.16)$$

Note that by consistency and randomization of $A$ in dataset (i), the RHS of Equation (38.16) when summed over $m$ is simply the first expression in the mediation formula (38.7). Hence under Equations (38.15) and (38.14) $E[Y(n, o)]$ is identified from data set (iii) by this expression in the mediation formula.

**Proof.**

$$p(M(n = x, o = x^*), Y(n = x, o = x^*))$$

$$= p(Y(n = x, o = x^*) | M(n = x, o = x^*))p(M(n = x, o = x^*))$$

$$= p(Y(n = x^*, o = x^*) | M(n = x^*, o = x^*))p(M(n = x, o = x)).$$

\textsuperscript{29} Note that the assumption that $p(M(a)) = p(M(n = a, o = a))$ and $p(Y(a) | M(a)) = p(Y(n = a, o = a) | M(n = a))$. This assumption is subject to empirical test by examining whether the distribution from (i) and (iii) are the same. This corresponds to the six-arm trial described by Stensrud et al. [2020b]. The distribution of (i) and (iii) could differ when, for example, the treatment $A$ contains additional sub-components that are present in neither $N$ nor $O$. 

The constraints in Equations (38.15) and (38.14) are implied by the SWIG given in Figure 38.4(b) with treatments \( n \) and \( o \) over the random variables \( N, O, M(n, o) \) and \( Y(n, o) \).\(^{30}\) Note that \( Y(n, o) \) is d-separated from the fixed node \( n \) given \( M(n, o) \), which implies under the SWIG global Markov property that the distribution \( p(Y(n, o) | M(n, o)) \) does not depend on \( n \), which is equivalent to Equation (38.15).

Similarly, since \( M(n, o) \) is d-separated from the fixed node \( o \), \( p(M(n, o)) \) does not depend on \( o \). Thus, we see that the constraints (38.14) are implied by the FFRCISTG model.\(^{31}\)

Thus, the identifiability result in Proposition 38.1 follows from the fact that in the SWIG in Figure 38.4(b), there is no variable whose conditional distributional distribution, given its (random) parents, depends on both \( n \) and \( o \).

Proposition 38.1 above, and indeed all the results in the remainder of this subsection, holds when counterfactuals indexed by the mediator \( m \) are not well-defined. Recall that Robins [1986] and Robins and Richardson [2010] develop an FFRCISTG model in which only interventions on a subset of variables are considered well-defined.

Remarks:

1. The reader may wonder why in this SWIG we have labeled \( M \) with \((n, o)\), rather than simply with \((n)\). This is to emphasize that in this subsection our results do not require the assumption that missing arrows on a SWIG imply the absence of the associated direct effect for all individuals. Rather in this sub-section we only impose the weaker assumption that any SWIG

![Figure 38.4](image-url)  
(a, b) SWIGs derived from the corresponding expanded DAGs \( G^e \) shown in Figures 38.3(b) and (c), respectively. In the SWIG shown in (b), the mediator is labeled \( M(n, o) \) to indicate that this is a population FFRCISTG for \( G^e \), which does not assume the absence of individual-level direct effects.

---

\(^{30}\) Thus, although, as noted previously in Footnote 21, under determinism, the DAGs in Figure 38.3(b) and (c) imply the same conditional independence relations on \( p(A, M, Y, N, O) \), they lead to different counterfactual models since the constraints (Equation 38.15) and (Equation 38.14) are implied by the SWIG in Figure 38.4(b), but not the SWIG in Figure 38.4(a).

\(^{31}\) As noted in remark 38.1, in some settings \( A \) can play the role of \( N \) or \( O \).
is a “population causal graph.” A population causal graph [Richardson and Robins 2013, section 7] assumes the distribution of the variables on the graph factor according to the graph, but does not impose the assumption that a missing arrow implies no individual level effects. Thus the variable $M(n, o)$ need not equal the variable $M(n) = M(n, O)$ and thus $M(n, o)$ cannot be labeled as $M(n)$. That is, in the underlying FFRCISTG model associated with the SWIG, the one-step-ahead counterfactuals $M(n, o)$ depend on both $n$ and $o$. In other words, this population FFRCISTG contains the counterfactual variables present in the SWIG shown in Figure 38.4(a) that results from splitting $N$ and $O$ in the graph shown in Figure 38.3(b). It does not correspond to an NPSEM associated with the graph in Figure 38.3(c) since that NPSEM assumes well-defined counterfactuals intervening on $M$ and also assumes that $M(n, o) = M(n)$; see also Footnotes 17 and 22. In particular, under the SWIG derived from the population graph, the constraints (Equation 38.15) and (Equation 38.14) correspond to the absence of the edges $n \to Y(n, o)$ and $o \to M(n, o)$, respectively.

2. Since Equations (38.15) and (38.14) are restrictions on the distribution of the counterfactuals in Figure 38.4(b), there exist consistent tests of the conditions (Equation 38.15) and (Equation 38.14) given the data from (ii). These conditions cannot be tested given only the data (iii) [or equivalently (i)].

3. We have seen that, under the FFRCISTG associated with Figure 38.4(b), the distribution of $Y(n, o)$ for all four arms is identified from the two arms in which $n = o$; thus the structure of this SWIG is sufficient for this identification. However, this structure is also “necessary” in that it is the only population SWIG over $M(n, o)$ and $Y(n, o)$ where this identification is possible. To see this, first note that if $Y(n, o)$ depends on both $n$ and $o$ then $n$ and $o$ are both ancestors of $Y$. By Proposition 38.2 below, if $n$ and $o$ are both parents of $Y$, then identifiability fails to hold. Given that we only have one other measured variable then this implies that we must have one fixed node that is a parent of $M$ (and $M$ in turn a parent of $Y$); the other fixed node is then a parent of $Y$. If any other edges are present between $\{n, o\}$ and $\{M(n, o), Y(n, o)\}$, then again by Proposition 38.2 the conditional distributions will not be identifiable. Likewise, if there is an unmeasured confounder between $M$ and $Y$, then $Y(n, o)$ will not be d-separated from $n$ given $M(n, o)$.

---

32. We note that consistent tests of individual level no direct effect conditions such as $Y(n, o) = Y(o)$ do not exist since it is possible that $E[Y(n, o)] = E[Y(n, o')]$ and yet $Y(n, o) \neq Y(n, o')$ (as random variables); but see also footnote 29.

33. Here we are ignoring the structure relating the random variables $N$ and $O$. 

We have the following more general result.

**Proposition 38.2**

Assume the distribution of the variables on an unexpanded DAG $\mathcal{G}$ is positive. Under an FFRCISTG corresponding to the population SWIG $\mathcal{G}^{ex}(n,o)$ there is no vertex that has both $n$ and $o$ as parents if and only if the joint distributions $p(V(n = x, o = x^*))$ for $x \neq x^*$ are identified from the counterfactual distributions $p(V(n = x, o = x))$. Further, since $P(V(n = x, o = x)) = P(V(a = x))$, also from the distribution of the variables in $\mathcal{G}$.\textsuperscript{34}

**Proof.** Consider the $g$-formula of proposition 3 in Chapter 41 applied to the graph $\mathcal{G}^{ex}$ under an intervention on $n$ and $o$. This formula is a function of the joint distribution of the observables if and only if none of the terms in the $g$-formula have both $N$ and $O$ in the conditioning event. Note that each term in the $g$-formula is the conditional distribution of a variable given its parents on the population SWIG $\mathcal{G}^{ex}(n,o)$. The requirement here that there is no vertex that is a child of both $n$ and $o$ is directly analogous to the “no recanting witness” condition in the theory developed by Avin et al. [2005].

Consider the following examples from Robins and Richardson [2010]. Suppose our original causal graph $\mathcal{G}$ in Figure 38.3(a) for the cigarette cessation trial was incorrect and the correct causal graph is shown in Figure 38.5(a). There exist three possible $(N,O)$ elaborations of this graph, which are shown in Figure 38.6. These represent different causal theories about the causal effect of the treatment variables $N, O$ on $L, M$, and $Y$. Figure 38.7 shows the corresponding population SWIGs. Under the SWIGs in Figure 38.7(a) and (b) the distributions of the four arms $Y(n,o)$ are identified given the distributions $Y(n = x, o = x)$:

$$p(Y(x, x^*)) = \sum_{m,l} p(Y(x^*, x^*) | M(x^*, x^*), L(x^*, x^*)) p(M(x, x) | L(x, x)) p(L(x, x))$$

$$= \sum_{m,l} p(Y | m, l, a = x^*) p(m | l, a = x) p(l | a = x)$$

(38.17)

and

$$p(Y(x, x^*)) = \sum_{m,l} p(Y(x^*, x^*) | M(x^*, x^*), L(x^*, x^*)) p(M(x, x) | L(x, x)) p(L(x^*, x^*))$$

$$= \sum_{m,l} p(Y | m, l, a = x^*) p(m | l, a = x) p(l | a = x^*)$$

(38.18)

respectively, where here we are using $Y(i,j)$ to denote $Y(n = i, o = j)$. Note that the identifying formulae are different. See Stensrud et al. [2020a] for generalizations of these results.

\textsuperscript{34} Note that the result here holds because the DAG $\mathcal{G}$ here does not contain hidden variables.
38.3 Interventionist Theory of Mediation

Figure 38.5 (a) DAG containing an observed common cause $L$ of the mediator $M$ and outcome $Y$ that is also caused by $A$; (b) DAG containing an unobserved common cause $H$ of the mediator $M$ and outcome $Y$.

Figure 38.6 Elaborations of the graph in Figure 38.5(a), with additional edges. These represent different causal theories about the causal effect of the treatment variables $N$, $O$ on $L$, $M$, $Y$. As before, the thicker red edges indicate deterministic relations.

In contrast, under the SWIG in Figure 38.7(c), $p(Y(n,o))$ for $n \neq o$ is not identified from the data on the two arms with $n = o$ (equivalently the observed data) because in $\mathcal{G}^{ts}$ $L$ has both $N$ and $O$ as parents; hence the term $p(l | n = x, o = x')$ in the g-formula for $p(Y(x, x'))$ is not a function of the observed data.

Given that $N$ and $O$ are real interventions, at most one of the expanded causal graphs shown in Figure 38.6 can represent the true causal structure. If, in the future, we obtain data from a four-arm ($N$, $O$) trial we can test between the three competing theories associated with these expanded graphs.\textsuperscript{35}

\textsuperscript{35} Note however, that if the results from the four-arm trial do not correspond to the identifying formulae obtained from either Figure 38.7(a) or (b), then although it is possible that the DAG in
Figure 38.7 Three different population SWIGs associated with the three expanded DAGs shown in Figure 38.6. Under the SWIGs (a), (b) the effects of $N$ and $O$ on $Y$ are separable so that the distribution of $Y(n, o)$ in the four arms are identified given the distribution of $Y(n = x, o = x)$ although the identification formulae differ; (c) A SWIG under which the four arms $Y(n, o)$ are not identified from the two arms $Y(n = x, o = x)$.

38.3.3 Expanded Graphs for a Single Treatment

We formally define expanded graphs as follows:

Given a DAG $\mathcal{G}$ with a single treatment variable $A$, an expanded graph $\mathcal{G}^{\text{ex}}$ for $A$ is a DAG constructed by first adding a set of new variables $\{A^{(1)}, \ldots, A^{(p)}\}$ corresponding to a decomposition of the treatment $A$ into $p$ separate components (proposed by the investigator); every variable $A^{(j)}$ is a child of $A$ with the same state space and $A^{(j)}(a) = a$, but $A$ has no other children in $\mathcal{G}^{\text{ex}}$; each child $C_j$ of $A$ in $\mathcal{G}$ has in $\mathcal{G}^{\text{ex}}$ a subset of $\{A^{(1)}, \ldots, A^{(p)}\}$ as its set of parents.

Lemma 38.1 Assume the distribution of the variables on an unexpanded DAG $\mathcal{G}$ is positive. Under an FFRCISTG corresponding to an expanded (population) graph $\mathcal{G}^{\text{ex}}$ for treatment $A$, the intervention distribution $p(V(A^{(i)} = x^{(i)}, \ldots, A^{(p)} = x^{(p)})$ is identified by the g-formula applied to $\mathcal{G}^{\text{ex}}$ from the data on $\mathcal{G}$ if for every child $C_j$ of $A$ in $\mathcal{G}$ the set of parents of $C_j$ in $\mathcal{G}^{\text{ex}}$ that are components of $A$ take the same value.$^{36}$

---

36. More formally, we require that for all $C_j \in \text{ch}_{\mathcal{G}}(A)$, if $A^{(i)}, A^{(j)} \in \text{pa}_{\mathcal{G}^{\text{ex}}}(C_j) \cap \{A^{(1)}, \ldots, A^{(p)}\}$ then $x^{(i)} = x^{(j)}$. 

$^{36}$ Figure 38.7(c) holds, it is also possible that the true graph could correspond to Figure 38.7(a) or (b) with an added unmeasured confounder between any pair of the variables $L$, $M$ and $Y$. 
Corollary 38.3.4

38.1

A

sponding

standing

\( \mathcal{G} \)

38.

formula

which

separable

Under

37.

x

arise

in

then

for

See

refer

given,

the

distribution

…

A

in

the

corresponding

to

the

context

of

\( x \)

= a

by

the

edge

expanded graph for \( A \), which we often denote as \( \mathcal{G}^{\text{edge}} \). In this case \( p = \text{ch}_G(A) \) and in this special case \( \mathcal{G}^{\text{ex}} \) corresponds to the graph formed from \( \mathcal{G} \) by replacing each edge \( A \rightarrow C_j \) with \( A \rightarrow A^{(j)} \rightarrow C_j \). Note that \( \mathcal{G}^{\text{edge}} \) is unique.\(^{38}\) See Figure 38.9 for an example.

\textbf{Corollary 38.1} Under the assumptions of Lemma 38.1, if \( \mathcal{G}^{\text{ex}} \) is the edge expanded graph \( \mathcal{G}^{\text{edge}} \) for \( A \) then for all assignments \( x^{(1)}, \ldots, x^{(p)} \), \( \Pr(V(\alpha^{(1)} = x^{(1)}, \ldots, \alpha^{(p)} = x^{(p)})) \) is identified from the data on \( \mathcal{G} \).

When the conditions of this corollary hold, we will say, following the nomenclature in Stensrud et al. [2020c], that the treatment components \( \{A^{(1)}, \ldots, A^{(p)}\} \) have separable effects.

\textbf{38.3.4 On the Substantive Relationship between Different \( \mathcal{G}^{\text{ex}} \) Graphs and \( \mathcal{G}^{\text{edge}} \)}

In the context of the smoking cessation trial recall that the two expanded graphs in Figure 38.6(a) and (b) led to different identifying formulae for \( \Pr(Y(n = x, o = x')) \) given, respectively, by Equations (38.17) and (38.18). The identifying formulae also arise in the context of the graph \( \mathcal{G}^{\text{edge}} \) shown in Figure 38.9. Specifically, \( \Pr(Y(\alpha^{(1)} = x, \alpha^{(2)} = x, \alpha^{(3)} = x')) \) and \( \Pr(Y(\alpha^{(1)} = x', \alpha^{(2)} = x, \alpha^{(3)} = x')) \) are identified by Equations (38.17) and (38.18), respectively. The FFRCISTG models associated with the expanded graphs shown in Figure 38.6(a) and (b) correspond to distinct mutually exclusive causal structures, which lead to different identifying formulae for the distribution of the counterfactual \( Y(n = x, o = x') \) in an arm of the four arm \((N, O)\) trial in which one intervenes to set \( n = x \), and \( o = x' \). However, given the FFRCISTG model corresponding to the graph in Figure 38.9, we are able to interpret the identifying expressions (38.17) and (38.18) as identifying two \underline{different} interventions on \( A^{(1)}, A^{(2)}, \) and \( A^{(3)} \) on a single graph \( \mathcal{G}^{\text{edge}} \)

---

37. More formally, we have that for all children \( C_j \) of \( A \), \( \text{pa}_{\mathcal{G}^{\text{ex}}}(C_j) \cap \{A^{(1)}, \ldots, A^{(p)}\} = \{A^{(j)}\} \).

38. Since the expanded graph \( \mathcal{G}^{\text{edge}} \) for \( A \) postulates a separate component of treatment corresponding to each child of \( A \), such a graph will be unlikely to represent the substantive understanding of an investigator when \( p \) is large.
Robins and Richardson [2010] note that the above may seem to, but do not, contradict one another. Recall that $N$ and $O$ represent the substantive variables recording the presence or absence of nicotine and all other cigarette components.

Suppose we further divide the other cigarette components ($O$) into Tar ($T$) and cigarette components other than tar and nicotine ($O^*$). Thus, substantively setting $O$ to a value corresponds to setting both $O^*$ and $T$ to that value. Furthermore, the graph $G^{edge}$ in Figure 38.9 being an FFRCISTG implies that the graph $G^{CE}$ in Figure 38.6(b) formed by (re)combining $O^*$ and $T$ is an FFRCISTG. Thus an intervention setting $(O = x, N = x^*)$ on Figure 38.6(b) substantively corresponds to the intervention on the graph $G^{edge}$ in Figure 38.9 with $O^* = A^{(3)} = x$, $T = A^{(1)} = x$, $N = A^{(2)} = x^*$. Thus these interventions in this $G^{CE}$ and in $G^{edge}$ give the same identifying formula (38.18).

Given we have locked in the substantive interpretation of the $A^{(j)}$ in Figure 38.9, the intervention $O^* = A^{(3)} = x$, $T = A^{(1)} = x^*$, $N = A^{(2)} = x^*$ on Figure 38.9 corresponds to the intervention in which $N$ and $T$ are set to the same value and thus does not represent a joint intervention on the substantive variables $N$ and $O$ (since $O = O^* = T$ as random variables in the observed data). However, the identifying formula for this intervention if Figure 38.9 were the causal graph happens to have the same identifying formula (38.17) as the intervention setting $O = x$, $N = x^*$ if Figure 38.6(a) were the causal graph. This might seem surprising since, under the substantive meanings of the components ($A^{(1)}, A^{(2)}, A^{(3)}$), specifically, $A^{(2)} \equiv N$, Figure 38.9 is compatible with Figure 38.6(b) and not Figure 38.6(a). However, it is not surprising from a formal perspective, if we notice that by considering a DAG with the same structure as Figure 38.6(a), but in which $N$ is replaced by a variable $N^1 \equiv N \times T$ indicating the presence of both Tar and Nicotine, then we may represent the intervention that sets $O^* = x$, $N = T = x^*$ via an intervention on $N^1$ and $O^*$.39

If, instead of dividing $O$, one divides Nicotine, $N$, into sub-components corresponding to two different isotopes and re-defines the variables in Figure 38.9 as $A^{(1)} \equiv$ Nicotine Isotope 1, $A^{(2)} \equiv$ Nicotine Isotope 2, $A^{(3)} \equiv O$, then the mirror image of the above holds. Specifically, Figure 38.9 is compatible with Figure 38.6(a),40 and not Figure 38.6(b).

Note, however, that there is no way to re-define $A^{(1)}$, $A^{(2)}$, and $A^{(3)}$ such that Figure 38.9 is compatible with Figure 38.6(c). Specifically, an intervention setting $N = x$ and $O = x^* \neq x$ cannot be represented via an intervention on $(A^{(1)}, A^{(2)}, A^{(3)})$, since in Figure 38.6(c) $L$ has two parents $N$ and $O$ while in Figure 38.6(c) $L$ has.

---

39. Though the graph constructed in this way has the same topology as Figure 38.6(a), it represents a different substantive hypothesis since the component $T$ of $N^1$ is a parent of $L$ and not $N$ itself.

40. Figure 38.6(a) with the variables $N$ and $O$ (not $N^1$ and $O$).
only one. Of course, this must be the case because the intervention on \(N\) and \(O\) in Figure 38.6(c) is not identified from the observed data. In contrast, by Corollary 38.1 any intervention on \(A^{(1)}, A^{(2)}, A^{(3)}\) in the graph shown in Figure 38.9 is identified.

Lastly, note that \(G^{\text{edge}}\) assumes that \(A^{(1)}, A^{(2)}, \text{and } A^{(3)}\) each directly affect only \(L, M, \text{and } Y\), respectively; this hypothesis could, in principle, be tested if one were to perform an eight-arm \((A^{(1)}, A^{(2)}, A^{(3)})\) trial.

In general, if \(G^{\text{ex}}\) is an FFRCISTG with separable (hence, identified) effects and further \(G^{\text{edge}}\) is an FFRCISTG, then the counterfactual variables in \(G^{\text{ex}}\) may be obtained from those in \(G^{\text{edge}}\) by imposing the equality \(d^{(0)} = d^{(0)}\) whenever the corresponding children \(C_i\) and \(C_j\) of \(A\) in \(G\) share a common parent in \(G^{\text{ex}}\). Note that this is directly analogous to the way in which the counterfactual variables \(V_i(a)\) in \(G(a)\) are equal to the counterfactuals \(V_i(n = a, o = a)\) present in \(G(n, o)\).

### 38.3.5 Generalizations

The foregoing development may be further generalized in several ways:

(a) Rather than having data from a single treatment variable \(A\), we may consider a study in which there were multiple treatment variables \(\{A_1, \ldots, A_k\}\). In this setting it may be of interest to attempt to identify the distribution \(V(n_1, o_1, \ldots, n_k, o_k)\) of a hypothetical future study where \(N_i\) and \(O_i\) are components of \(A_i\) such that in the original study \(A_i = O_i = N_i\). See Figure 38.10. More generally each \(A_i\) may have \(p_i\) components. The natural generalization of Lemma 38.1 holds. 41

(b) We may consider a setting in which some variables \((H)\) in the underlying causal DAG \(G(V \cup H)\) are not observed; though variables we intervene on are observed, so \(A \subseteq V\).

Here, we proceed in two steps. In the first step, we check identification of a standard interventional distribution. Specifically, we construct an ADMG \(G^{\text{ex}}\) containing the variables \(\{N_1, O_1, \ldots, N_k, O_k\}\), such that \(N_i\) and \(O_i\) have only \(A_i\) as a parent; the only edges with an arrowhead at \(N_i\) and \(O_i\) are of the form \(A_i \rightarrow \) while the only edges out are of the form \(\rightarrow C\) where \(C\) is a child of \(A\) in \(G\). We then apply the extended ID algorithm described in Chapter 41 of this volume to first determine whether \(p(V(n_1, o_1, \ldots, n_k, o_k))\) would be identified given a positive distribution \(p(V \cup \{N_1, O_1, \ldots, N_k, O_k\})\) over the observed variables and the treatment components. 42

---

41. That is, we have identification if, for each \(i\) and each child \(C\) of \(A_i\) in \(G\), the subset of the \(p_i\) components of \(A_i\) that are parents of \(C\) in the expanded graph \(G^{\text{ex}}\) take the same value.

42. This would correspond to an observational study where the variables \(V \cup \{N_1, O_1, \ldots, N_k, O_k\}\) are observed in a population in which \(N_i\) and \(O_i\) are no longer deterministic functions of \(A_i\), but for which all other one-step-ahead counterfactuals remain the same; variables in \(H\) are not observed.
In the second step, we check whether the identification would hold under the weaker conditions where we only have access to a positive distribution on \( p(V) \). This corresponds to (i) making sure that identification of every term in the identifying formula given by the ID Algorithm via the inductive application of Proposition 41.5 from Chapter 41 in this volume ensures that the splitting operation is applied to any \( A_i \in A \) before any \( N_i \) or \( O_i \) (this ensures that the positivity requirement for Proposition 41.5 is met), and (ii) confirming that for every \( (N_i, O_i) \), in the identifying formula given by the ID Algorithm\(^{43}\) there is no district \( D \) that has both \( N_i \) and \( O_i \) as parents; Shpitser [2013] terms such districts, which violate this condition, *recanting districts*.\(^{44}\)

A simple example of such a structure arises when there is an unobserved common cause of \( M \) and \( Y \), as shown in Figure 38.5(b): though the four distributions \( E[Y(n, o)] \) are identified given a four-arm \( (N, O) \) trial, the distributions for which \( n \neq o \) are not identified solely given data on \( (A, M, Y) \). This is because \( M \leftarrow H \rightarrow Y \) forms a district and both \( N \) and \( O \) are parents of this district, hence it is recanting.

If the two-step procedure yields identification, the resulting functional structurally resembles the functional obtained from the ID algorithm, except each term in the functional that depends on treatments is evaluated at its own treatment value, corresponding to either \( n_i \) or \( o_i \) (but never both at once).

See Section 38.4 below for a general method of addressing all complications above simultaneously.

### 38.3.6 Identification of Cross-world Nested Counterfactuals of DAG \( \mathcal{G} \) under an FFRCISTG Model for its Expanded Graph \( \mathcal{G}^{ex} \)

From Section 38.3.2 to this point, we only studied the distribution of counterfactuals associated with interventions on \( (N_i, O_i) \) or, more generally, \( (A^{(i)}, \ldots, A^{(p)}) \); no other counterfactuals associated with an expanded graph \( \mathcal{G}^{ex} \) were mentioned. As argued in Section 38.3.2, for most purposes these counterfactuals constitute an adequate basis for formulating contrasts relating to the mediation of effects. However, since the prior literature on mediation has been formulated in terms of

\(^{43}\) See Equation (41.21) and subsequent discussion in Chapter 41 in this volume.

\(^{44}\) Note that in the special case where \( D = \{ V_i \} \) is a singleton, then \( D \) will be a recanting district in \( \mathcal{G} \) if and only if \( N_i \) and \( O_i \) have a common child in \( \mathcal{G}^{ex} \). In this case \( V_i \) is a “recanting witness” as defined by Avin et al. [2005]; see also Section 38.4. Thus, when no hidden variables exist, the first step in (b) always succeeds; hence, as implied by Lemma 38.1, identification fails if and only if there exists an \( N_i \) and \( O_i \) that have a common child.
cross-world nested counterfactuals associated with the original (unexpanded) DAG \( \mathcal{G} \), we return to our earlier discussion.

**The PDE**

Recall that in Section 38.1 we related the PDE from the NPSEM associated with Figure 38.3(a) to a four-arm \((N, O)\) trial under the FFRCISTG associated with the SWIG \( \mathcal{G}^{ex}(n, o) \) shown in Figure 38.4(b). This may be broken down into two steps:

1. Show that \( \mathbb{E}[Y(n = 0, o = 1)] \) was identified from data on \( p(A, M, Y) \) via Equation (38.12). This step only required that \( \mathcal{G}^{ex} \) was a population FFRCISTG.\(^{45}\) That is, this identification follows from the deterministic relationship \( N(a) = O(a) = a \) together with the population level conditions (Equation 38.15) and (Equation 38.14), without requiring that counterfactuals for interventions on \( M \) be well-defined.

2. Next show that \( Y(a = 1, M(a = 0)) = Y(n = 0, o = 1) \) holds under the individual level no direct effect assumptions encoded in the NPSEM associated with \( \mathcal{G}^{ex} \) in Figure 38.8. Note that this step does not require that Figure 38.3(c) is an FFRCISTG, only that it be an NPSEM associated with \( \mathcal{G}^{ex} \).

In more detail, the NPSEM associated with \( \mathcal{G}^{ex} \) implies the following:

(i) \( M(n) = M(a, n, o) \);

(ii) \( Y(o, m) = Y(a, n, o, m) \).

Under conditions (i) and (ii) we have that:

\[
\begin{align*}
M(a = 0) &= M(N(a = 0)) = M(n = 0), \\
Y(a = 1, m) &= Y(O(a = 1), m) = Y(o = 1, m), \\
Y(a = 1, M(a = 0)) &= Y(o = 1, M(n = 0)) = Y(n = 0, o = 1).
\end{align*}
\]  

\[ \text{(38.19)} \]

**Figure 38.8** A SWIG derived from the expanded DAG \( \mathcal{G}^{ex} \) in Figure 38.3(c), under the assumption that the counterfactual variables obey the NPSEM associated with \( \mathcal{G}^{ex} \). Consequently, in contrast to Figure 38.4(b), the graph contains \( M(n) \) rather than \( M(n, o) \).

\(^{45}\) Thus, this step does not require that the counterfactual variables follow an NPSEM associated with \( \mathcal{G}^{ex} \).
Consequently, under the assumptions (i) and (ii) it follows that:

\[ P(Y(a = 1, M(a = 0))) = P(Y(n = 0, o = 1)). \] (38.20)

Remarks:

1. If \( N(a) = O(a) = a \) and conditions (i) and (ii) hold, then \( PDE = E[Y(n = 0, o = 1)] - E[Y(n = 0, o = 0)] \) even if conditions (38.15) and (38.14) fail, for example due to \((M, Y)\) confounding. This is the situation discussed in Section 38.1 where data from the four-arm \((N, O)\) trial makes it possible to estimate the PDE and hence determine whether it equals the mediation formula.46

2. The conditions (38.15) and (38.14) alone, that is, without (i) and (ii) above, are not sufficient to identify the PDE,47

3.(i) If condition (i) fails, so that \( O \) has a (population) direct effect on \( M \) (relative to \( A, N \)), then the counterfactual \( M(n, o) \) is not a function of the original one-step-ahead counterfactuals \( Y(a, m) \) and \( M(a) \). This can be seen from the fact that whereas \( M \) has a single parent \( A \) in Figure 38.3(a), under the elaboration that includes \( N \) and \( O \) it now has two: \( \{N, O\} \).

3.(ii) Similarly, if condition (ii) fails so that \( N \) has a direct effect on \( Y \) (relative to \( A, M, O \)), then the counterfactual \( Y(n, o) \) is not a function of the original one-step-ahead counterfactuals \( Y(a, m) \) and \( M(a) \). \( Y \) has two parents \( \{A,M\} \) in Figure 38.3(a), under the elaboration that includes \( N \) and \( O \) it would have three \( \{N, O, M\} \).

Example: The River Blindness Studies

Returning to the river blindness study, note that intervening to set \( s = 1 \) and \( a = 0 \) gives \( Y(s = 1, a = 0) = Y(a = 1, M(a = 0)) \) as random variables. Hence the PDE is identified from data in a four-arm \((A, S)\) trial.48 It is not identified from data on

46. Note, however, that if \( P(Y(n = 0, o = 1)) \) does not equal the first expression in the mediation formula it is possible that this is solely because \((N, O)\) do not satisfy (i) and (ii) but that there are other subcomponents of \( A \), say \((N^{*}, O^{*})\) that do satisfy (i) and (ii).

47. This is because

\[
E[Y(a = 1, M(a = 0))] = E[Y(n = 1, o = 1, M(n = 0, o = 0))]
= \sum_{m} E[Y(n = 1, o = 1, m) | M(n = 0, o = 0) = m]p(M(n = 0, o = 0) = m),
\] (38.21)

\[
= E[Y(n = 0, o = 1)] - E[Y(n = 0, o = 0)].
\] (38.22)

but this latter conditional expectation term is cross-world in terms of the counterfactuals in \( \mathcal{G}^{\text{CR}} \) although, as noted, they do identify \( E[Y(n = 0, o = 1)] - E[Y(n = 0, o = 0)] \).

48. Recall that in \( Y(s = 1, a = 0), A \) is serving as “\( N\)” and \( S \) is “\( O\)” ; see Figure 38.3(e) and Footnote 23.
A, M, Y because M ↔ Y forms a “recanting district,” as defined in Shpitser [2013]. Note that had identification of the PDE failed due to a recanting witness, that is, A and S having a common child, then additional interventions on the variables in the graph would not have led to identification. Finally, we note that, owing to the context-specific conditional independences in this example, the distribution
\[ p(Y(s = 0, a = 1)) = Y(a = 0, M(a = 1)) \]
which occurs in the Total Direct Effect is identified given data from the two arms \( p(Y(s = x, a = x)), x \in \{0, 1\} \) by the formula given in (38.A), and hence also from \( p(A, M, Y) \); see Appendix 38.A.2.

Counterfactuals Related to the DAG in Figure 38.5(a)
Recall that \( E \{ Y(n = 0, o = 1) \} \) is identified under the FFRCISTG models associated with the graphs in Figures 38.6(a) and (b), but not (c).

Let \( Y(a, l, m), M(a, l), \) and \( L(a) \) denote the one-step-ahead counterfactuals associated with the graph in Figure 38.5(a). It follows from the deterministic counterfactual relation \( N(a) = O(a) = a \) and the NPSEM associated with Figure 38.5(a), and its associated expanded graph \( \mathcal{G}^{ex} \) in Figure 38.6(a) that the random variable
\[ Y(n = 0, o = 1) = Y(o = 1, L(n = 0), M(n = 0, L(n = 0))) \]

associated with the NPSEM in Figure 38.6(a) can be written in terms of the counterfactuals associated with the graph in Figure 38.5(a) as the cross-world counterfactual
\[ Y(a = 1, L(a = 0), M(a = 0)) = Y(a = 1, L(a = 0), M(a = 0, L(a = 0))). \quad (38.23) \]

Similarly, if we assume that \( \mathcal{G}^{edge} \) of Figure 38.9 is an NPSEM, the counterfactual (38.23) also equals, as a random variable, the counterfactual \( Y(A^{(1)} = 0, A^{(2)} = 0, A^{(3)} = 1) \). Thus if either Figure 38.6(a) or Figure 38.9 represented the FFRCISTG generating the data, the distribution of Equation (38.23) is identified by the same formula (38.17).

![Figure 38.9](image-url) The edge expanded graph \( \mathcal{G}^{edge} \) associated with the DAG shown in Figure 38.5(a).
Likewise,

\[ Y(n = 0, o = 1) = Y(o = 1, L(o = 1), M(n = 0, L(o = 1))) \]

associated with the graph in Figure 38.6(b) equals (as a random variable) the cross-world counterfactual

\[ Y(a = 1, L(a = 1), M(a = 0, L(a = 1))). \]  

(38.24)

associated with the graph in Figure 38.5(a). Again, if we assume that \( G^{\text{edge}} \) of Figure 38.9 is an NPSEM, the counterfactual (38.24) also equals \( Y(A^{(3)} = 1, A^{(2)} = 0, A^{(1)} = 1) \) (as a random variable). Thus, in this case if either Figure 38.6(b) or Figure 38.9 represented the FFRCISTG generating the data, the distribution of Equation (38.24) is identified by the same formula (38.18).

In contrast, \( E[Y(n = 0, o = 1)] \) associated with the graph in Figure 38.6(c) is not the mean of any counterfactual defined from \( Y(a, l, m), M(a, l) \), and \( L(a) \) under the graph in Figure 38.5(a) since \( L \), after intervening to set \( n = 0, o = 1 \), is neither \( L(a = 1) \) nor \( L(a = 0) \) as both imply a counterfactual for \( L \) under which \( n = o \). Furthermore, the parameter occurring in the PDE

\[ E[Y(a = 1, M(a = 0))] = E[Y(a = 1, L(a = 1), M(a = 0, L(a = 0)))] \]

and associated with the graph in Figure 38.5(a) is not identified under the FFRCISTGs associated with any of the three graphs in Figures 38.6(a), (b), and (c). This is because all three of these graphs are compatible with the NPSEM in Figure 38.5(a) under which, by recursive substitution, we have the following equality \( Y(a_1, M(a_0)) = Y(a_1, L(a_1), M(a_0, L(a_0))) \), as random variables. But the counterfactual \( Y(a_1, L(a_1), M(a_0, L(a_0))) \), since it involves \( L(a_0) \) and \( L(a_1) \), does not correspond to an intervention on \( n \) and \( o \) under any of the expanded graphs in Figures 38.6(a), (b), or (c). Similarly, \( Y(a_0, M(a_1)) \) is not identified under any of these graphs. These results follow from the fact that \( L \) is a recanting witness for \( A \) in Figure 38.5(a).

In the following section, we define path-specific effects associated with the NPSEM model. We show that the two identified nested cross-world counterfactuals are identified path-specific effects under the NPSEM-IE for the graph in Figure 38.5(a). In particular, \( E[Y(a = 1, L(a = 0), M(a = 0))] \) is the path-specific effect associated with the path \( A \rightarrow Y \) and \( E[Y(a = 1, L(a = 1), M(a = 0, L(a = 1)))] \) is the path-specific effect associated with the paths \( A \rightarrow Y, A \rightarrow L \rightarrow Y \), and \( A \rightarrow L \rightarrow M \rightarrow Y \). Thus, for Pearl, identification of these nested cross-world counterfactuals associated with Figure 38.5(a) follows from the assumption that the
38.4 Path-Specific Counterfactuals

In Section 38.1 we considered different notions of direct effect, which led to the notion of nested counterfactuals and the PDE, which is identified under the NPSEM-IE associated with the graph in Figure 38.3(a) via the associated Mediation Formula. In Section 38.3.2, following Robins and Richardson [2010], we discuss the notion of separability of effects in the sense of Stensrud et al. [2020c] via an expanded \((N, O)\) graph. We showed that the counterfactuals defining the PDE were equal to ordinary (non-cross-world) interventional counterfactuals in the NPSEM given by the expanded graph (with determinism). We also showed that under the corresponding FFRCISTG these effects were identified by the mediation formula. In this section, we now consider path-specific effects which generalize the notion of direct and indirect effects.

In the simplest setting, the intuition behind an indirect effect is to consider all paths from \(A\) to \(Y\) other than the edge \(A \rightarrow Y\). This can be generalized to settings where the effect along a particular set of causal paths from \(A\) to \(Y\) is of interest. In what follows we will show that each such path-specific effect will correspond to a cross-world counterfactual contrast associated with the (original) graph \(\mathcal{G}\). We will see that in the absence of recanting witnesses these cross-world counterfactuals are equal (as random variables) to interventional counterfactuals in the NPSEM associated with \(\mathcal{G}^{\text{edge}}\), the edge expanded graph associated with the set of treatment variables \(A\). Consequently, these path-specific counterfactuals will be identified if and only if the corresponding intervention is identified under the FFRCISTG associated with \(\mathcal{G}^{\text{edge}}\). Thus all identifying formulae for path-specific cross-world counterfactuals on \(\mathcal{G}\) may be derived from \(\mathcal{G}^{\text{edge}}\). Further, it follows that all identifying formulae for path-specific cross-world counterfactuals may also be obtained under the assumption that \(\mathcal{G}\) is an NPSEM-IE.

---

49. As noted earlier, the counterfactuals associated with \(\mathcal{G}^{\text{edge}}\) may not have clear substantive meaning. For \(\mathcal{G}^{\text{edge}}\) to be substantive it is necessary for each treatment variable \(A\) to be decomposable into components each of which could, in principle, be intervened on (separately) and affect one and only one of its children; see Footnote 34, Section 38.3.4 for further discussion. The graph \(\mathcal{G}^{\text{edge}}\) may still be useful as a formal construction.
The general theory developed by Shpitser [2013] associates a random variable with each subset of causal paths between a treatment $A$ and an outcome $Y$. The intuition is that this subset of proper causal paths from $A$ to $Y$ denoted $\pi$ remain active, while all other causal paths, denoted $\overline{\pi}$, from $A$ to $Y$ are blocked. Next, pick a pair of value sets $a$ and $a'$ for elements in $A$; $a$ will be associated with active paths, $a'$ with those that are blocked.

For any $V_i \in V$, define the potential outcome $V_i(\pi, a, a')$ by setting $A$ to $a$ for the purposes of paths in $\pi$ that end in $V_i$, and setting $A$ to $a'$ for the purposes of proper causal paths from $A$ to $V_i$ not in $\pi$. Formally, the definition is as follows, for any $V_i \in V$:

$$
V_i(\pi, a, a') \equiv a \text{ if } V_i \in A,
$$

$$
V_i(\pi, a, a') \equiv V_i(\{ V_j(\pi, a, a') | V_j \in pa^\pi_i \}, \{ V_j(\pi, a, a') | V_j \in pa^{\overline{\pi}}_i \}).
$$

where $V_j(\pi, a, a') \equiv a'$ if $V_j \in A$, and given by recursive substitution otherwise, $pa^\pi_i$ is the set of parents of $V_i$ along an edge which is a part of a path in $\pi$, and $pa^{\overline{\pi}}_i$ is the set of all other parents of $V_i$.

A counterfactual $V_j(\pi, a, a')$ is said to be edge inconsistent if, for some edge $A_k \rightarrow V_j$ in $G$, counterfactuals of the form $V_j(a_k, \ldots)$ and $V_j(a'_k, \ldots)$ occur in $V_j(\pi, a, a')$, otherwise it is said to be edge consistent. In the former case $V_j$ is said to be a recanting witness (for $\pi$). It is simple to verify using Equation (38.25) that edge consistent counterfactuals are precisely those where no paths in $\pi$ and $\overline{\pi}$ share the initial edge. Shpitser [2013] and Shpitser and Tchetgen Tchetgen [2016] have shown that a joint distribution $p(V(\pi, a, a'))$ containing an edge-inconsistent counterfactual $V_j(\pi, a, a')$ is not identified in the NPSEM-IE (nor weaker causal models) in the presence of a recanting witness.

As an example, consider the graph shown in Figure 38.5(a) and the counterfactual given in Equation (38.23) that corresponds to the path $\pi_1 = \{ A \rightarrow Y \}$:

$$
Y(\pi_1, a = 1, a = 0) \equiv Y(a = 1, L(a = 0), M(a = 0, L(a = 0))).
$$

50. A proper causal path intersects the set $A$ only once at the source node.
51. It is important to understand that the predicates “active” and “blocked” are applied to paths. In particular, it is possible for every edge and vertex on a blocked path to also be present on some active path, and vice-versa.
52. Note that it follows from the definition of proper causal path that each path in $\pi$ is associated with a unique vertex in $A$; similarly for each path in $\overline{\pi}$. (Two paths may be associated with the same vertex in $A$.)
The counterfactual associated with the paths \( \pi_2 = \{A \rightarrow Y, A \rightarrow L \rightarrow Y\} \) is given by:

\[
Y(\pi_2, a = 1, a = 0) \equiv Y(a = 1, L(a = 1), M(a = 0, L(a = 0))).
\]

Note that \( Y(\pi_1, a = 1, a = 0) \) is edge consistent while \( Y(\pi_2, 1, 0) \) is edge-inconsistent due to the presence of \( L(a = 0) \) and \( L(a = 1) \).

This result is proved in [Shpitser and Tchetgen Tchetgen 2016]:

**Theorem 38.1** If \( V(\pi, a, a') \) is edge consistent, then under the NPSEM-IE for the DAG \( \mathcal{G} \),

\[
p(V(\pi, a, a')) = \prod_{i=1}^{K} p(V_i \mid a \cap \text{pa}_i^+, a' \cap \text{pa}_i^-, \text{pa}_i^0 \setminus A).
\]  

(38.26)

As an example of such an identification consider the distribution \( p(Y(\pi, a, a')) \) of the edge consistent counterfactual in Figure 38.5(a). It follows from Theorem 38.1 that

\[
p(Y(\pi_1, a = 1, a = 0)) = p(Y(a = 1, L(a = 0), M(a = 0, L(a = 0))))
= \sum_{m,l} p(Y \mid m, l, a = 1) p(m \mid l, a = 0) p(l \mid a = 0),
\]

a marginal distribution derived from Equation (38.26).

In the following, we exploit an equivalence between edge consistent counterfactuals \( V(\pi, a, a') \) and standard potential outcomes based on edge expanded graphs \( \mathcal{G}^{\text{edge}} \), already defined in the case of a single treatment variable in Section 38.3.3. In this section, we will abbreviate \( \mathcal{G}^{\text{edge}} \) as \( \mathcal{G}^e \) for conciseness. The edge expanded graph both simplifies complex nested potential outcome expressions and enables us to leverage the prior result in Shpitser and Pearl [2006b] to identify conditional path-specific effects.

We now extend the definition of expanded graph to sets of treatments \( |A| > 1 \): Given an ADMG \( \mathcal{G}(V) \), define for each \( A_i \subseteq A \) a new set of variables, \( A_i^\text{Ch} \equiv \{A_i' \mid V_j \in \text{Ch}_i\} \) with state spaces \( \mathcal{X}_{A_i'} = \mathcal{X}_{A_i} \); thus for each directed edge \( A_i \rightarrow V_j \) in \( \mathcal{G}(V) \) from a treatment variable to its child, we have created a new variable \( A_i' \) with the same state space as \( A_i \). Denote the full set of new variables as \( A^\text{Ch} = \bigcup_{A_i \in A} A_i^\text{Ch} \).

We define the edge expanded graph of \( \mathcal{G}(V) \), written \( \mathcal{G'}(V \cup A^\text{Ch}) \), as the graph with

53. The above development may be generalized to \( k \) different assignments rather than two, by partitioning the set of paths into \( k \). These and other generalizations are termed *path interventions* by Shpitser and Tchetgen Tchetgen [2016].

54. A similar construction was called the “extended graph” in Malinsky et al. [2019].
the vertex set $V \cup A^{Ch}$; the edge expanded graph contains all the edges in $G$ except for the edges $A_i \rightarrow V_j$ that join a treatment variable $A_i$ to its child $V_j$, in addition we add the edges $A_i \rightarrow A_i' \rightarrow V_j$ (which thus “replace” the removed edge $A_i \rightarrow V_j$).

As an example, the edge expanded graph for the DAG in Figure 38.3(a), with $A_1 = V$, is shown in Figure 38.10. For conciseness, we will generally drop explicit references to vertices $V \cup A^{Ch}$, and denote edge expanded graph of $G(V)$ by $G^e$.

More generally, we associate a causal model with $G^e$ as follows. Let $\mathbb{V}$ be the set of one-step-ahead potential outcomes associated with the original graph $G$. Similarly, we let $\mathbb{V}^e$ denote the set of one-step-ahead potential outcomes associated with $G^e$, constructed as follows: For every $V_i(pa_i) \in \mathbb{V}$, we let $V_i(pa'_i)$ be in $\mathbb{V}^e$. Note that this is well-defined, since vertices $V_i$ in $G$ and $G^e$ share the number of parents, and the parent sets for every $V_i$ share state spaces. In addition, for every $A'_i \in A^{Ch}$, we let $A'_i(a_i)$ for $a_i \in X_i$ be in $\mathbb{V}^e$.

The edges $A_i \rightarrow A'_i$ in $G^e$ are understood to represent deterministic equality relationships, such that $A'_i = A_i$. More precisely, every $A'_i \in A^{Ch}$ has a single parent $A_i$, and we let $A'_i(a_i) = a_i$ so that $A'_i(a_i)$ is a degenerate (constant) random variable corresponding to a point-mass at $a_i$.

The FFRCISTG model associated with the edge expanded graph $G^e$ includes these deterministic relationships. Note that it follows that given a distribution $p(\mathbb{V})$ over the counterfactuals given by the NPSEM $G$ there is a unique distribution $p'(\mathbb{V}^e)$ over the counterfactuals given by $G^e$ that satisfies these deterministic relationships.

We now show the following two results. First, we show that an edge-consistent $V(\pi, a, a')$ may be represented without loss of generality by a counterfactual response to an intervention on a subset of $A^{Ch}$ in $G^e$ with the causal model defined above. Second, we show that this response is identified by the same functional (38.26).

![Figure 38.10](image)

**Figure 38.10** An edge expanded graph that considers components of the treatment variables $A_1$ and $A_2$. 

---

**Chapter 38 An Interventionist Approach to Mediation Analysis**
Given an edge consistent \( V(\pi, a, a') \), define \( G^e \) for \( A \subseteq V \). We define \( a^e \) that assigns \( a_i \) to \( A_i \in A^{Ch} \) if \( A_i \rightarrow V_j \) in \( G(V) \) is in \( \pi \), and assigns \( a'_i \) to \( A'_i \in A^{Ch} \) if \( A_i \rightarrow V_j \) in \( G(V) \) is not in \( \pi \). The resulting set of counterfactuals \( V(a^e) \) is well defined in the model for \( V^e \), and we have the following result, proved in the Appendix.

**Proposition 38.3** Fix a distribution \( p(V) \) in the NPSEM-IE for a DAG \( G(V) \), and consider the corresponding distribution \( p^e(V) \) in the FFRCI STG model associated with a DAG \( G^e(V \cup A^{Ch}) \). Then for every \( V_i \in V \), the random variable \( V_i \) in the original model associated with \( G \) is equal to the random variable \( V_j \) in the restricted model associated with \( G^e \) that includes the equalities defining the variables in \( A^{Ch} \). Moreover, for any edge consistent \( \pi, a, a' \), the random variable \( V_i(\pi, a, a') \) in the original model associated with \( G \) is equal to the random variable \( V_i(a^e) \) in the restricted model associated with \( G^e \).

**Theorem 38.2** Under the FFRCI STG model associated with the edge expanded DAG \( G^e \), for any edge consistent \( \pi, a, a' \):

\[
p^e(V(a^e)) = \prod_{i=1}^{K} p^e(V_i | a^e \cap pa_i^{G^e}, pa_i^{G^e} \setminus A).
\] (38.27)

Note that since, by definition, the distribution \( p^e(V \cup A^{Ch}) \) is a deterministic function of \( p(V) \), hence by Equation (38.27) \( p^e(V(a^e)) \) is identified by \( p(V) \). Recall that if \( G^e \) is an FFRCI STG model then this requires that all of the interventions on the variables in \( A^{Ch} \) are well-defined.

**Corollary 38.2** Under the NPSEM-IE model associated with the DAG \( G \), for any edge consistent \( \pi, a, a' \):

\[
p(V(\pi, a, a')) = p^e(V(a^e)) = \prod_{i=1}^{K} p^e(V_i | a^e \cap pa_i^{G^e}, pa_i^{G^e} \setminus A),
\]

where \( G^e \) is the edge expanded graph corresponding to \( G \).

In fact, Corollary 38.2 provides an alternative proof of Theorem 38.1 since, by definition, for any edge consistent \( \pi, a, a' \):

\[
p(V_i | a \cap pa_i^{G^e}, a' \cap pa_i^{G^e}, pa_i^{G^e} \setminus A) = p^e(V_i | a^e \cap pa_i^{G^e}, pa_i^{G^e} \setminus A).
\]

Though, as discussed previously, if the graph \( G \) is an NPSEM-IE the graph \( G^e \) may be regarded as a purely formal construction that aids in the derivation of the identifying formulae. However, if \( G^e \) is interpreted as a causal model, so that there are well-defined counterfactuals associated with intervening on each of the vertices
in $A^\text{Ch}$, then $G^\pi$ provides an interventionist interpretation of path-specific counterfactuals; in fact these interventions allow the identification results above to be checked, in principle, in a hypothetical randomized trial.

In the causal models derived from DAGs with unobserved variables (e.g., $G(V \cup H)$), identification of distributions on potential outcomes such as $p(V(a))$ or $p(V(\pi, a, a'))$ may be stated without loss of generality on the latent projection ADMG $G(V)$. A complete algorithm for identification of path-specific effects in NPSEM-IEs with hidden variable was given in Shpitser [2013] and presented in a more concise form in Shpitser and Sherman [2018].

We now show that identification theory for $p(V(\pi, a, a'))$ in latent projection ADMGs $G(V)$ may be restated, without loss of generality, in terms of identification of $p^\pi(V(a^\pi))$ in $G^\pi(V \cup A^\text{Ch})$.

**Proposition 38.4** Let $G(V \cup H)$ be a DAG and $Y \subseteq V$ be an ancestral set in $G(a)$, so $an_{G(a)}(Y(a)) = Y(a)$. Under the FFRCISTG model associated with the edge expanded DAG $G^\pi(V \cup A^\text{Ch} \cup H)$, for any edge consistent $\pi, a, a'$, it follows that $Y(\pi, a, a') = Y(a^\pi)$ and thus $p(Y(\pi, a, a'))$ is identified given $p(V)$ if and only if $p^\pi(Y(a^\pi))$ is identified from $p^\pi(V \cup A^\text{Ch})$.

This proposition is a generalization of Theorem 38.2 from DAGs to latent projection ADMGs. To determine whether $p^\pi(Y(a^\pi))$ is identified we examine the ADMG $G^\pi(V, A^\text{Ch})$. Since this graph is a standard latent projection ADMG (albeit with deterministic relationships relating $A$ and $A^\text{Ch}$), the extended ID algorithm decomposes the distribution $p^\pi(Y(a^\pi))$ into a set of factors as in (41.21). However, in order for $p^\pi(Y(a^\pi))$ to be identified given data $p(V)$, an additional requirement must be placed on the terms of this decomposition. Specifically, for each term $p^\pi(V^D(a^\pi), v_{pa^\pi_G(Y(a^\pi))}) = v_D$ in (41.21), it must be the case that $a^\pi$ assigns consistent values to each element of $A$ that is in $pa^\pi_D(Y(a^\pi)) \equiv pa^\pi_D(Y(a^\pi)) \setminus (D \cup a^\pi)$, the random parents of district $D$ that are not in $D$. This requirement corresponds to the *recanting district criterion* that was introduced and shown to be complete in Shpitser [2013]. Aside from this requirement, each term must be identified by the extended ID algorithm described in the companion paper [Chapter 41].

---

55. Note that the vertex set for $G^\pi(V, A^\text{Ch})$ comprises $V \cup A^\text{Ch}$. Further, by construction, the distribution over $p^\pi(V \cup A^\text{Ch})$ is degenerate since the variables in $A^\text{Ch}$ are deterministic functions of those in $A$.

56. See Point (b) in Section 38.3.5.
38.4.1 Conditional Path-specific Distributions

Having established that we can identify path-specific effects by working with potential outcomes derived from the $G'$ model, we turn to the identification of conditional path-specific effects using the po-calculus. In Shpitser and Pearl [2006b], the authors present the conditional identification (IDC) algorithm for identifying quantities of the form $p(Y(x) \mid W(x))$ (in our notation), given an ADMG. Since conditional path-specific effects correspond to exactly such quantities defined on the model associated with the edge expanded graph $G'$, we can leverage their scheme for our purposes. The idea is to reduce the conditional problem, identification of $p^c(Y(a^x) \mid W(a^z))$, to an unconditional (joint) identification problem for which a complete identification algorithm already exists.

The algorithm has three steps: first, exhaustively apply Rule 2 of the po-calculus to reduce the conditioning set as much as possible; second, identify the relevant joint distribution using Proposition 38.4 and the complete algorithm of Shpitser and Sherman [2018]; third, divide that joint by the marginal distribution of the remaining conditioning set to yield the conditional path-specific potential outcome distribution.

Note that we make use of SWIGs defined from edge expanded graphs of the form $G'(a^x, z)$. Beginning with $G'$ the SWIG $G'(a^x, z)$ is constructed by the usual node-splitting operation: split nodes $Z$ and $A_i$ into random and fixed halves, where $A_i'$ has a fixed copy $a_i$ if $A_i \rightarrow V_j$ in $G(V)$ is in $\pi$, and $a_i'$ if $A_i \rightarrow V_j$ in $G(V)$ is not in $\pi$. Relabeling of random nodes proceeds as previously described. This procedure is in fact complete, as shown by the following result with a proof found in Malinsky et al. [2019].

Theorem 38.3

Let $p(Y(\pi, a, a') \mid W(\pi, a, a'))$ be a conditional path-specific distribution in the NPSEM-IE model for $G'$, and let $p^c(Y(a^x) \mid W(a^z))$ be the corresponding distribution under the FFRCISTG associated with the edge expanded graph $G'(V \cup A^c)$. Let $Z$ be the maximal subset of $W$ such that $p^c(Y(a^x) \mid W(a^z)) = p^c(Y(a^x, z) \mid W(a^x, z), Z(a^x, z)).$ Then $p^c(Y(a^x) \mid W(a^z))$ is identifiable in $G'$ if and only if $p^c(Y(a^x, z), W(a^x, z), Z(a^x, z))$ is identifiable in $G'$.

As an example, $p(Y(a, M(a'))) is identified from $p(C, A, M, Y)$ in the NPSEM-IE model for the graph in Figure 38.1(a) via

$$p(Y(a, M(a'))) = \sum_m \left( \frac{\sum_c p(Y, m \mid a, c)p(c)}{\sum_c p(m \mid a, c)p(c)} \right) \left( \sum_c p(m \mid a', c')p(c') \right).$$

57. Malinsky et al. [2019] assumed an NPSEM-IE but the proof only uses the rules of the po-calculus, hence also applies under the less restrictive FFRCISTG assumptions for $G'$. 
38.5 Conclusion

We have shown here that graphical insights derived from Pearl’s thought experiment may be used to place mediation analysis on a firm interventionist footing, and yield analyses of direct, indirect, and path-specific effects that are amenable to falsification, and explainable to practitioners.

Acknowledgments

The first author was supported by the grants ONR N000141912446 and National Institutes of Health (NIH) awards R01 AG057869, R01 AI127271, R37 AI102634, U01 CA261277-01, and R01 CA222147-02.

The second author was supported by the grant ONR N00014-19-1-2446.

The third author was supported by the grants ONR N00014-18-1-2760, NSF CAREER 1942239, NSF 1939675, and R01 AI127271-01A1.

The authors would like to thank F. Richard Guo for his insightful comments that improved this manuscript.

38.A Appendix

38.A.1 Proof of PDE Bounds under the FFRCISTG Model

We here prove the bounds on the PDE implied by the FFRCISTG model associated with the graph in Figure 38.3(a).
Proof. It follows from the definition that: \( PDE_{a,a'} = P(Y(a, M(a')) = 1) - P(Y = 1 | A = a') \). Note that

\[
p(Y(a, M(a')) = 1) = p(Y(a, m = 0) = 1 | M(a') = 0)p(M(a') = 0 | A = a') \\
+ p(Y(a, m = 1) = 1 | M(a') = 1)p(M(a') = 1 | A = a') \\
= p(Y(a, m = 0) = 1 | M(a') = 0)p(M = 0 | A = a') \\
+ p(Y(a, m = 1) = 1 | M(a') = 1)p(M = 1 | A = a').
\]

The quantities \( p(Y(a, m = 0) = 1 | M(a') = 0) \) and \( p(Y(a, m = 1) = 1 | M(a') = 1) \) are constrained by the law for the observed data via:

\[
p(Y = 1 | A = a, M = 0) = p(Y(a, m = 0) = 1) \\
= p(Y(a, m = 0) = 1 | M(a') = 0)p(M(a') = 0) \\
+ p(Y(a, m = 0) = 1 | M(a') = 1)p(M(a') = 1) \\
= p(Y(a, m = 0) = 1 | M(a') = 0)p(M = 0 | A = a') \\
+ p(Y(a, m = 0) = 1 | M(a') = 1)p(M = 1 | A = a'),
\]

and

\[
p(Y = 1 | A = a, M = 1) = p(Y(a, m = 1) = 1) \\
= p(Y(a, m = 1) = 1 | M(a') = 0)p(M(a') = 0) \\
+ p(Y(a, m = 1) = 1 | M(a') = 1)p(M(a') = 1) \\
= p(Y(a, m = 1) = 1 | M(a') = 0)p(M = 0 | A = a') \\
+ p(Y(a, m = 1) = 1 | M(a') = 1)p(M = 1 | A = a').
\]

It then follows from the analysis in Section 2.2 in Richardson and Robins [2010] that the set of possible values for the pair

\[(\alpha_0, \alpha_1) \equiv (p(Y(a, m = 0) = 1 | M(a') = 0), p(Y(a, m = 1) = 1 | M(a') = 1))\]

compatible with the observed joint distribution \( p(m, y | a) \) is given by:

\[(\alpha_0, \alpha_1) \in [l_0, u_0] \times [l_1, u_1]\]

where,

\[
l_0 = \max\{0, 1 + (p(Y = 1 | A = a, M = 0) - 1)/p(M = 0 | A = a')\}, \\
u_0 = \min\{p(Y = 1 | A = a, M = 0)/p(M = 0 | A = a'), 1\}, \\
l_1 = \max\{0, 1 + (p(Y = 1 | A = a, M = 1) - 1)/p(M = 1 | A = a')\}, \\
u_1 = \min\{p(Y = 1 | A = a, M = 1)/p(M = 1 | A = a'), 1\}.
\]
38.2 Proof that the PDE is Not Identified in the River Blindness Study

Consider the NPSEM-IE corresponding to the DAG shown in Figure 38.12, which may be obtained by marginalizing R from the causal DAG representing the river blindness studies shown in Figure 38.1(a).

The one-step-ahead counterfactuals defining the model are: \( U, A, S(a), M(a, u), Y(m, s, u) \). We will assume that all variables, including \( U \), are binary. Though not shown explicitly on the graph, we also have the following constraints: (i) \( Y(m, s, u) = Y(m, s_0, u_0) \equiv Y(m, s_0) \); (ii) \( Y(m_0, s_1, u) = Y(m_0, s_1, u_1) \equiv Y(m_0, s_1) \); (iii) \( M(a_1, u_0) = M(a_1, u_1) \equiv M(a_1) \), and (iv) \( S(a) = a \), where here we use \( x_i \) as a shorthand for \( x = i \). These arise from, respectively, (i) the fact that if immunosuppressants are not available \( (s = 0) \) then the patient’s outcome \( Y \) is not influenced by their predisposition \( U \) to use medicine; (ii) the availability of immunosuppressants, and hence the patient’s predisposition \( U \) to use them, is not relevant to patients who do not receive ivermectin;\(^\text{58} \) (iii) the fact that in a randomized trial \( (a = 1) \) the treatment the patient receives is not influenced by \( U \); and (iv) the fact that the clinic is available \( (s = 1) \) if and only if a patient is in the randomized trial \( (a = 1) \). The NPSEM-IE is then defined by the following 10 parameters:

\[
\begin{align*}
U & : \quad \theta_U \equiv p(U = 1); \\
A & : \quad \theta_A \equiv p(A = 1); \\
M(a, u) & : \quad \theta_M(a_0, u) \equiv p(M(a_0, u) = 1), \quad \text{for } u \in \{0, 1\}, \\
& \quad \theta_M(a_1) \equiv p(M(a_1) = 1); \\
Y(m, s, u) & : \quad \theta_Y(m_1, s_1, u) \equiv p(Y(m_1, s_1, u) = 1), \quad \text{for } u \in \{0, 1\}, \\
& \quad \theta_Y(m, s) \equiv p(Y(m, s) = 1), \quad \text{for } (m, s) \in \{(0, 0), (0, 1), (1, 0)\}.
\end{align*}
\]

**Figure 38.12** The DAG corresponding to the projection of the DAG \( \mathcal{G} \) shown in Figure 38.1(a) after marginalizing \( R \).

\(^{58}\) This is an additional restriction not present in the original account but introduced here solely to reduce the number of parameters. This corresponds to removing the \( U \rightarrow R \) edge in the SWIG \( \mathcal{G}(a_1, m_0, s_1) \) where \( m \) is set to 0; see Figure 38.2. (Since non-identifiability under a sub-model implies non-identifiability in the larger model, we can make this assumption without loss of generality.)
Let $\theta$ denote a vector containing all of these parameters. The corresponding observed distribution $p_\theta(A, M, Y)$ is given by the following equations:

$$p_\theta(A = 1) = \theta_A,$$

$$p_\theta(M = 1|A = a) = \begin{cases} 
\theta_M(a_0, u_0)(1 - \theta_U) + \theta_M(a_0, u_1)\theta_U, & \text{if } a = 0, \\
\theta_M(a_1), & \text{if } a = 1,
\end{cases}$$

$$p_\theta(Y = 1|A = a, M = m) = \begin{cases} 
\theta_Y(m, s_a), & \text{if } (a, m) \in \{(0, 0), (0, 1), (1, 0)\}, \\
\theta_Y(m_1, s_1, u_0)(1 - \theta_U) + \theta_Y(m_1, s_1, u_1)\theta_U, & \text{if } (a, m) = (1, 1).
\end{cases}$$

Given $\theta$, consider a perturbed vector $\tilde{\theta}$ defined by taking

$$\tilde{\theta}_M(a_0, u_0) = \theta_M(a_0, u_0) + \varepsilon(1 - \theta_U),$$

$$\tilde{\theta}_M(a_0, u_1) = \theta_M(a_0, u_1) - \varepsilon\theta_U,$$

for sufficiently small $\varepsilon$ and leaving the other 8 parameters unchanged from $\theta$. It is simple to see that the resulting observed distribution is unchanged, so $p_{\tilde{\theta}}(A, M, Y) = p_\theta(A, M, Y)$, since the perturbation only changes the expression for $p(M = 1|A = 0)$, but the additional terms involving $\varepsilon$ cancel.

Turning to the PDE we see that:

$$p_{\tilde{\theta}}(Y(a_1, M(a_0)) = 1)$$

$$= p_{\tilde{\theta}}(Y(M(a_0), s_1) = 1)$$

$$= \sum_u p_{\tilde{\theta}}(Y(M(a_0), s_1) = 1|U = u)p_{\tilde{\theta}}(U = u)$$

$$= \sum_{u,k} p_{\tilde{\theta}}(Y(m_k, s_1) = 1|U = u, M(a_0) = k)p_{\tilde{\theta}}(M(a_0) = k|U = u)p_{\tilde{\theta}}(U = u)$$

$$\text{consistency} = \sum_{u,k} p_{\tilde{\theta}}(Y(m_k, s_1, u) = 1|U = u, M(a_0) = k)p_{\tilde{\theta}}(M(a_0, u) = k|U = u)p_{\tilde{\theta}}(U = u)$$

$$\text{independence} = \sum_{u,k} p_{\tilde{\theta}}(Y(m_k, s_1, u) = 1)p_{\tilde{\theta}}(M(a_0, u) = k)p_{\tilde{\theta}}(U = u)$$

$$= \theta_Y(m_0, s_1)(1 - \theta_M(a_0, u_0))(1 - \theta_U) + \theta_Y(m_0, s_1)(1 - \theta_M(a_0, u_1))\theta_U$$

$$+ \theta_Y(m_1, s_1, u_0)\theta_M(a_0, u_0)(1 - \theta_U) + \theta_Y(m_1, s_1, u_1)\theta_M(a_0, u_1)\theta_U.$$
so that the PDE will take a different value under \( p_\theta \), so long as \( \theta_Y(m_1, s_1, u_0) \neq \theta_Y(m_1, s_1, u_1) \).

That the PDE is not identified in the river blindness study example should not be surprising in light of the results in Section 38.4. In particular, we know that

\[
Y(a_1, M(a_0)) = Y(s_1, a_0).
\]

In addition, we see that in the SWIG \( \mathcal{G}(a = 0, s = 1) \) shown in Figure 38.13(a), which here plays the role of the SWIG for the expanded graph \( \mathcal{G}^e \), \( Y(a_0, s_1) \) is in the same district as \( M(a_0) \), but the fixed nodes \( a = 0 \) and \( s = 1 \) are both parents of this district, but are both associated with \( A \) in the original graph. Consequently \( M(a_0) \leftrightarrow Y(a_0, s_1) \) forms a recanting district. However, formally Section 38.4 considers models defined solely via ordinary conditional independences, whereas the model in the river blindness study also incorporated context-specific independences. For this reason, since a quantity may be unidentified in a model but identified in a submodel, we provided an explicit construction of an NPSEM-IE for the DAG in Figure 38.1(a).

Similarly, consideration of the SWIG \( \mathcal{G}(a = 1, s = 0) \) shown in Figure 38.13(b) shows that \( p(Y(a_1, s_0)) \) is identified from \( p(A, M, Y) \) because there is no recanting district.

\[\begin{align*}
\text{(a)} & \quad \text{The SWIG } \mathcal{G}(a = 0, s = 1) \text{ associated with the DAG shown in Figure 38.1(a);} \\
\text{(b)} & \quad \text{The SWIG } \mathcal{G}(a = 1, s = 0) \text{ associated with the DAG shown in Figure 38.1(a). The } U \rightarrow M \text{ edge is absent because this is a randomized study } (a = 1). \text{ The } U \rightarrow R \text{ edge is not present because the clinic is not available } (s = 0), \text{ hence patients do not have the option to take immunosuppressants.}
\end{align*}\]

59. This inequality corresponds to the presence of the edge \( U \rightarrow Y \) in Figure 38.12.

60. This assumes that missing edges correspond to the absence of direct effects at the individual level, so that \( M(a_0, s_1) = M(a_0) \).

61. Since the NPSEM-IE is a submodel of the FFRCISTG, this also establishes that the PDE is not identified under the latter interpretation of the DAG in Figure 38.1(a).

62. Note that this identification argument does not use the additional constraint (ii).
\[
p(Y(a_1, s_0) = 1) = \sum_{r, m} p(y = 1 \mid r, m)p(r \mid s = 0, m)p(m \mid a = 1)
\quad \text{g-formula}
\]

\[
\begin{align*}
\text{independence} &= \sum_{r, m} p(y = 1 \mid r, m, s = 0)p(r \mid s = 0, m)p(m \mid a = 1) \\
\text{marginalization} &= \sum_{m} p(y = 1 \mid m, s = 0)p(m \mid a = 1) \\
\text{determinism} &= \sum_{m} p(y = 1 \mid m, a = 0)p(m \mid a = 1).
\end{align*}
\]

Thus, in this example, \(p(Y(a, s)) = p(Y(s, M(a)))\) is identified for \((a, s) \in \{(0, 0), (1, 1), (1, 0)\}\) but is not identified for \((a, s) = (0, 1)\). Since \(Y(a_0, M(a_1)) = Y(a_1, s_0)\), it follows that \(p(Y(a_0, M(a_1)))\), which forms part of the Total Direct Effect (38.3), is also identified.

This conclusion also follows from Proposition 38.1, here letting \(A\) be “\(N\)” and \(S\) be “\(O\)”, since equality (38.15) holds with \(x = 1\) (though not with \(x = 0\)).

### 38.A.3 Detecting Confounding via Interventions on \(A\) and \(S\)

In the river blindness study, it is not possible to detect the confounder \(U\) via interventions on \(A\) and \(M\) since the distribution \(\{A, M(a), Y(a, m)\}\) obeys the FFRCISTG model corresponding to Figure 38.3(a). This is due to the context-specific independences that hold in this example; see Section 38.2.5. However, confounding becomes detectable if we are able to intervene on \(S\) and \(A\). To see this, note that:

\[
p(Y(a_0, s_1) = 1 \mid M(a_0) = 1) \\
= \sum_u p(Y(a_0, s_1) = 1 \mid M(a_0) = 1, U = u)p(u \mid M(a_0) = 1) \\
= \sum_u p(Y(m_1, s_1, u) = 1 \mid M(a_0) = 1, U = u)p(u \mid M(a_0) = 1) \\
= \sum_u p(Y(m_1, s_1, u) = 1)p(u \mid M(a_0) = 1) \\
= \frac{\theta_Y(m_1, s_1, u_0)(1 - \theta_U)\theta_M(a_0, u_0) + \theta_Y(m_1, s_1, u_1)\theta_U\theta_M(a_0, u_1)}{(1 - \theta_U)\theta_M(a_0, u_0) + \theta_U\theta_M(a_0, u_1)}
\]

which will depend on \(\theta_M(a_0, u_0)\) and \(\theta_M(a_0, u_1)\), (if \(\theta_Y(m_1, s_1, u_0) \neq \theta_Y(m_1, s_1, u_1)\)). However, 

\[
p(Y(a_1, s_1) \mid M(a_1) = 1) = \sum_u p(Y(a_1, s_1) \mid M(a_1) = 1, U = u)p(u \mid M(a_1) = 1) \\
= \sum_u p(Y(m_1, s_1, u) \mid M(a_1) = 1, U = u)p(u \mid M(a_1) = 1)
\]
\[
\begin{align*}
&= \sum_u p(Y(m_1, s_1, u)|M(a_1) = 1) \\
&= \sum_u p(Y(m_1, s_1, u))p(u|Y(m_1, s_1, u))p(Y(m_1, s_1, u)|M(a_1) = 1) \\
&= \theta_Y(m_1, s_1, u_0)(1 - \theta_U) + \theta_Y(m_1, s_1, u_0)\theta_U.
\end{align*}
\]

Thus \( p(Y(a_0, s_1) = 1|M(a_0) = 1) \neq p(Y(a_1, s_1) = 1|M(a_1) = 1) \), while if \( U \) were absent we would have equality.\(^{63}\) That the equality holds if \( U \) is absent may be seen by removing \( U \) from the SWIG \( \mathcal{G}(a, s) \) constructed from \( \mathcal{G} \) in Figure 38.1(a) and then applying Rule 3 of the po-calculus [Shpitser et al. 2021]: the equality follows from the fact that \( a \) is d-separated from \( Y(a, s) \) given \( M(a) \). Conversely, that the equality fails to hold when \( U \) is present is not surprising since we have the d-connecting path \( a \rightarrow M(a) \leftrightarrow U \rightarrow R(m, s) \rightarrow Y(m, s) \) in Figure 38.13(a).

### 38.4.4 Proof of Proposition 38.3

**Proof.** For any \( V_i \in V, V_i(a^\pi) \) is defined via (41.1) applied to \( \mathcal{G}^{\text{edge}} \),

\[
V_i(a^\pi) \equiv V_i \left( a^\pi_{pa^\pi_i \cap M(A_i)} V_{pa^\pi_i \cup A_i}(a^\pi) \right).
\]

Similarly, \( V_i(\pi, a, a') \) is defined via Equation (38.25) applied to \( \mathcal{G} \) as

\[
V_i(\pi, a, a') \equiv a \text{ if } V_i \in A,
\]

\[
V_i(\pi, a, a') \equiv V_i(\{V_j(\pi, a, a') | V_j \in pa^\pi \}, \{V_j(a') | V_j \in pa^\pi \})
\]

where \( V_j(a') \equiv a' \) if \( V_j \in A \), and given by (41.1) otherwise, \( pa^\pi \) is the set of parents of \( V_i \) along an edge that is a part of a path in \( \pi \), and \( pa^\pi \) is the set of all other parents of \( V_i \).

By definition of \( \mathcal{G}^{\text{edge}} \), the induction on the tree structure of Equation (38.25) in \( \mathcal{G} \) matches the induction on the tree structure of (41.1) in \( \mathcal{G}^{\text{edge}} \).

Finally, since \( V_i(\pi, a, a') \) is edge consistent, any edge of the form \( A_k \rightarrow V_j \) in \( \mathcal{G} \), for any \( A_k \in A \) that starts a proper causal path that ends at \( V_j \) is assigned precisely one value (either \( a_{A_k} \) or \( a'_{A_k} \)). Moreover, this value in the base case of the induction of Equation (38.25), by definition of \( a^\pi \) matches the value \( a^\pi_k \) assigned by the corresponding base case of the induction of (41.1). This establishes our conclusion.

That the corresponding observed data variables \( V_i \) match in models represented by the original graph \( \mathcal{G} \), and the edge expanded graph \( \mathcal{G}^{\text{edge}} \) follows by the above

\(^{63}\) Note that under the additional assumption (iii), we have: \( p(Y(a_0, s_1) = 1|M(a_0) = 0) = p(Y(a_1, s_1) = 1|M(a_1) = 0) \) since \( \theta_Y(m_1, s_1, u_0) = \theta_Y(m_1, s_1, u_0) \). Without (iii) the equality will not hold.
argument, using the fact that all elements in $A^{Ch}$ are deterministically related to the appropriate elements in $A$.

Malinsky et al. [2019] prove a weaker result establishing equality in distribution.

References

O. O. Aalen, M. J. Stensrud, V. Didelez, R. Daniel, K. Røysland, and S. Strohmaier. 2020. Time-dependent mediators in survival analysis: Modeling direct and indirect effects with the additive hazards model. *Biom. J.* 62, 3, 532–549. DOI: https://doi.org/10.1002/bimj.201800263.

C. Avin, I. Shpitser, and J. Pearl. 2005. Identifiability of path-specific effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*. Vol. 19, Morgan Kaufmann, San Francisco, 357–363.

I. J. Dahabreh, J. M. Robins, S. J. Haneuse, and M. A. Hernán. 2019. Generalizing causal inferences from randomized trials: Counterfactual and graphical identification. *arXiv preprint arXiv:1906.10792.*

V. Didelez. 2019. Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Anal.* 25, 4, 593–610. DOI: https://doi.org/10.1007/s10985-018-9449-0.

J. J. Lok. 2016. Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. *Stat. Med.* 35, 22, 4008–4020. DOI: https://doi.org/10.1002/sim.6990.

D. Malinsky, I. Shpitser, and T. S. Richardson. 2019. A potential outcomes calculus for identifying conditional path-specific effects. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. arXiv:1903.03662.

J. Pearl. 2001. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI-01)*. Morgan Kaufmann, San Francisco, 411–420.

J. Pearl. 2009. *Causality: Models, Reasoning, and Inference.* (2nd. ed.). Cambridge University Press. ISBN: 978-0521895606.

J. Pearl. 2012. The causal mediation formula—A guide to the assessment of pathways and mechanisms. *Prev. Sci.* 13, 4, 426–436. DOI: https://doi.org/10.1007/s11121-011-0270-1.

J. Pearl. 2018. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *J. Causal Inference.* 6. DOI: https://doi.org/10.1515/jci-2018-2001.

J. Pearl. 2019. On the interpretation of do(x). *J. Causal Inference.* 7. DOI: https://doi.org/10.1515/jci-2019-2002.

T. S. Richardson and J. M. Robins. 2010. Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. 415–444.

T. S. Richardson and J. M. Robins. 2013. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. preprint: http://www.csss.washington.edu/Papers/wp128.pdf.
J. M. Robins. 1986. A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Math. Model.* 7, 1393–1512. DOI: https://doi.org/10.1016/0270-0255(86)90088-6.

J. M. Robins. 2003. Highly structured stochastic systems. In P. J. Green, N. L. Hjort, and S. Richardson (Eds.), *Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects.* Oxford University Press, Oxford.

J. M. Robins and S. Greenland. 1992. Identifiability and exchangeability of direct and indirect effects. *Epidemiology* 3, 143–155. DOI: https://doi.org/10.1097/00001648-199203000-00013.

J. M. Robins and T. S. Richardson. 2010. Alternative graphical causal models and the identification of direct effects. In P. E. Shrout, K. M. Keyes, and K. Ornstein (Eds.), *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures.* Oxford University Press, Oxford.

A. L. Sarvet, K. N. Wans, M. J. Stensrud, and M. A. Hernán. 2020. A graphical description of partial exchangeability. *Epidemiology* 31, 365–368. DOI: https://doi.org/10.1097/EDE.0000000000001165.

I. Shpitser. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cog. Sci.* 37, 1011–1035. DOI: https://doi.org/10.1111/cogs.12058.

I. Shpitser and J. Pearl. 2006a. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06).* AAAI Press, Palo Alto.

I. Shpitser and J. Pearl. 2006b. Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI-06).* AUAI Press, Corvallis, OR, 437–444.

I. Shpitser and E. Sherman. 2018. Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18).*

I. Shpitser and E. J. Tchetgen Tchetgen. 2016. Causal inference with a graphical hierarchy of interventions. *Ann. Stat.* 44, 6, 2433–2466. DOI: https://doi.org/10.1214/15-AOS1411.

I. Shpitser, E. J. Tchetgen Tchetgen, and R. Andrews. 2017. *Modeling Interference Via Symmetric Treatment Decomposition.* Working paper. https://arxiv.org/abs/1709.01050.

I. Shpitser, T. S. Richardson, and J. M. Robins. 2021. Multivariate counterfactual systems and causal graphical models. This Volume.

M. Stensrud, M. Hernán, E. Tchetgen Tchetgen, J. M. Robins, V. Didelez, and J. Young. 2020a. Generalized interpretation and identification of separable effects in competing risk settings. *arXiv preprint arXiv:2004.14824.*

M. J. Stensrud, J. M. Robins, A. Sarvet, E. J. Tchetgen Tchetgen, and J. G. Young. 2020b. Conditional separable effects. *arXiv:2006.1568f* [stat.ME].

M. Stensrud, J. Young, V. Didelez, J. M. Robins, and M. Hernán, 2020c. Separable effects for causal inference in the presence of competing events. *J. Am. Stat. Assoc.* DOI: https://doi.org/10.1080/01621459.2020.1765783.
J. Tian and J. Pearl. 2002. On the testable implications of causal models with hidden variables. In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02), Vol. 18. AUAI Press, Corvallis, OR, 519–527.

S. Tikka, A. Hyttinen, and J. Karvanen. 2019. Identifying causal effects via context-specific independence relations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2800–2810. http://papers.nips.cc/paper/8547-identifying-causal-effects-via-context-specific-independence-relations.pdf.

L. Wittgenstein. 1922. Tractatus Logico-Philosophicus. Routledge, London, 1981.