

Research Article

Trang Quynh Nguyen*, Ian Schmid, Elizabeth L. Ogburn, and Elizabeth A. Stuart

Clarifying causal mediation analysis: Effect identification via three assumptions and five potential outcomes

<https://doi.org/10.1515/jci-2021-0049>

received September 21, 2021; accepted August 08, 2022

Abstract: Causal mediation analysis is complicated with multiple effect definitions that require different sets of assumptions for identification. This article provides a systematic explanation of such assumptions. We define five potential outcome types whose means are involved in various effect definitions. We tackle their mean/distribution's identification, starting with the one that requires the weakest assumptions and gradually building up to the one that requires the strongest assumptions. This presentation shows clearly why an assumption is required for one estimand and not another, and provides a succinct table from which an applied researcher could pick out the assumptions required for identifying the causal effects they target. Using a running example, the article illustrates the assembling and consideration of identifying assumptions for a range of causal contrasts. For several that are commonly encountered in the literature, this exercise clarifies that identification requires weaker assumptions than those often stated in the literature. This attention to the details also draws attention to the differences in the positivity assumption for different estimands, with practical implications. Clarity on the identifying assumptions of these various estimands will help researchers conduct appropriate mediation analyses and interpret the results with appropriate caution given the plausibility of the assumptions.

Keywords: identification, assumptions, mediation, causal inference, causal mediation analysis

MSC 2020: 62D20

1 Introduction

Causal inference analyses, explicitly or implicitly, generally involve three steps: define the target causal effect (also known as the *estimand*, i.e., what we wish to estimate); assess its identifiability (what assumptions are required to learn this causal effect from observed data, and whether they are likely to hold); and then estimate it (i.e., learn it from data) [1]. If there is a concern that an identification assumption may not hold, this issue should be dealt with if the analysis were to proceed, e.g., via adding a sensitivity analysis as a fourth step after estimation, or building the uncertainty about the assumption into the estimation procedure. While assumptions are part of most statistical analyses, they are especially important when inferring causal effects from observational data, because some assumptions are untestable, and if they do not

* **Corresponding author: Trang Quynh Nguyen**, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, e-mail: trang.nguyen@jhu.edu

Ian Schmid: Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Elizabeth L. Ogburn: Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Elizabeth A. Stuart: Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

hold effects may even be not interpretable. It is important that the researcher conducting an analysis understand the assumptions and judge their plausibility. This article clarifies for applied researchers the identifying assumptions often invoked in a *causal mediation analysis* – using a simple setting with a binary exposure A , a single mediator M , and a single outcome Y , with appropriate temporal ordering. Even here things are more complicated than in the non-mediation situation. Our goal is to unpack the complicated in a way that is digestible and thus helpful for practice.

Two comments before we proceed. First, this article focuses on the single mediator case, where the mediator may be univariate or multivariate (but considered en bloc). This leaves quite a few cases outside the scope of the article, such as settings with multiple mediators where the effects through each mediator are of interest [2–4] and settings with repeated exposure and mediator over a longitudinal process [5,6]. Second, as mediation analysis concerns causal effects, exposure-mediator-outcome temporal ordering is required. Unfortunately, reviews [7,8] continue to find many mediation analyses not satisfying this minimal requirement, making it all the more important to reiterate. Without appropriate temporal ordering, our effect identification exercise here would be nonsense.

1.1 Estimands

In this article, we talk about estimands using the language of potential outcomes [9] and potential mediator values. There are a variety of causal estimands to choose from, each being a contrast of potential outcomes under two conditions. The estimands addressed in this article are defined by conditions where the exposure and in most cases the mediator are (hypothetically) manipulated – an approach championed by Pearl [10].¹ These include well-known direct and indirect effects of several types, and a range of effects that do not fit a direct or indirect effect label. We give a brief introduction of these estimands here, and return to each one when discussing identification. As the current focus is on identifying (not defining) effects, we refer the reader to our companion article [12] for a detailed discussion of the meaning and relevance of these effect types.

Direct effects reflect the notion of the exposure’s influence on the outcome that does not go through the mediator. They are each defined based on some manner of “blocking” the influence that goes through the mediator. With *controlled direct effects* [10,13], this blocking is done by fixing the mediator to one value (not letting it change in response to the exposure), so a controlled direct effect is the effect of the exposure on the outcome when the mediator is fixed, and it depends on the mediator control value. With *natural direct effects* [10,13], the blocking is instead done by holding the mediator at the individual’s own potential mediator value under one exposure condition. While there are as many controlled direct effects as there are possible mediator values, there are only two natural direct effects, depending on which of the two potential mediator values (under exposure and nonexposure) is used. For the *interventional direct effects* [2,14], it is the mediator distribution (rather than the individual’s mediator value) that is fixed, and it is fixed to be the same as a potential mediator distribution conditional on covariates. For a set of covariates, there is one potential mediator distribution for exposure and another for nonexposure, so there are two interventional direct effects. The *generalized direct effects* [12,14,15] generalize the different types of direct effects by letting the mediator distribution be held at any relevant distribution, not just a value or a potential mediator distribution.

¹ Separate from the common approach of defining causal effects based on conditions where the mediator is manipulated, Robins et al. [11] take a different approach anchored on the idea of splitting exposure into components that affect the outcome through different pathways. They define estimands as effects of the split exposures, and develop a theory with graphical rules for identification. That theory is relevant to the research questions addressed in our illustrative example in several ways: the notion of split exposure effects speaks directly to the trimmed intervention question; some insights from the theory are indirectly helpful for the consideration of other interventional effects; and a result from the theory sheds additional light on natural (in)direct effects identification. The same research questions from the current illustrative example here could alternatively be analyzed using this split exposure theory plus a transportability lens.

Indirect effects reflect the notion of the exposure's influence on the outcome that goes through the mediator. An indirect effect is defined as the effect on the outcome of a switch in the mediator value or distribution from the potential value/distribution under nonexposure to that under exposure (as if in response to a switch in the exposure), while keeping exposure unchanged. When the switch involves the individual specific potential mediator values, we have *natural indirect effects*; when the switch involves the potential mediator distributions (conditional on covariates), we have *interventional indirect effects*. Each natural indirect effect pairs with a natural direct effect in summing up to the total causal effect; in fact the motivation for the original establishment of natural (in)direct effects is to split the total effect into path-specific components. Interventional (in)direct effects, in contrast, are not made to decompose the total effect. There are no controlled indirect effects.

All the aforementioned effects, except the natural (in)direct effects, are part of a class of effects we call *interventional effects*. An effect in this class is a contrast between an *interventional condition* where the exposure and/or mediator are manipulated and a *comparison condition* with a different manipulation (or no manipulation); the sort of manipulation referenced here is one that sets the variable or its distribution to a value/distribution that is known or can be determined. This is a broad class that contains many effects that do not fit the notion of direct or indirect effects. An example is where the exposure is an existing intervention program, but researchers are also interested in the effect of a hypothetically modified intervention program that no longer targets a mediator, relative to the no intervention condition. For more examples, see the companion article [12].

For simplicity, causal effects are represented in this article on the additive scale and in average form – as differences in potential outcome mean between contrasted conditions. Alternatively, other effect scales (e.g., ratio of means) could be used, and other features of the potential outcome distribution (e.g., median) could be contrasted; the same identification assumptions apply.

1.2 Nonparametric point identification

The type of identification discussed here – the type commonly encountered in the causal mediation literature – is *nonparametric point identification*. Let us clarify what this means.

Since each of our estimands (an average causal effect) is a contrast of potential outcomes under two conditions, things would be simple if we were to observe both of those potential outcomes for each individual. Unfortunately, at most one potential outcome may be observed for each individual; this is called *the fundamental problem of causal inference* [16]. For each individual, we observe the one actual outcome (Y) plus the exposure (A), the mediator (M), and perhaps some other variables including pre-exposure covariates (C) and other covariates that are affected by exposure (L). The key idea is, if certain assumptions hold, the estimand can be connected to the *observed data distribution*, i.e., the distribution of $\{C, A, L, M, Y\}$. Specifically, the estimand is equated to a function of features of the observed data distribution (e.g., marginal or conditional means and densities). We then call the estimand *identified*, or more precisely *point identified*.²

As a well-known example, in a perfect two-arm randomized controlled trial (RCT), the relevant identifying assumptions (discussed later) hold by design. The average total effect, i.e., the difference between the means of the two *potential outcomes* (under treatment and control), is identified: it is equal to the difference in mean *observed* outcome between the two RCT arms. But the RCT does not guarantee identification of the various other effects mentioned earlier, because the mediator is not randomized. For those effects, identification requires untestable assumptions that should be carefully considered by the researcher.

² There are cases when a causal estimand is not point identified, but is bounded by functions of the observed data distribution (see, e.g., [17]). Then it is called *partially identified*. This topic is outside the scope of the current article.

The identifying assumptions we make do not place any restrictions on the observed data distribution. This means no parametric assumptions such as the type of distribution (normal or other) of variables or the functional form (linear or other) for the associations among variables. The identifying assumptions we make are about equating certain conditional means or densities of potential outcomes (or potential mediators) with conditional means or densities of observed variables, the latter being free to be what they are. This type of identification is thus called *nonparametric identification*.³

Note that this article addresses identification, not estimation. Questions such as what models should be fit, or how much can be learned from a sample of a certain size, belong in the estimation step. To put them aside, it may be helpful to imagine having infinite data.

1.3 Effect identification via five potential outcome types and three assumptions

Much has been written about assumptions for identification of specific effects in causal mediation analysis (e.g., [2,11, 10,18,19]). The current paper focuses on a systematic explanation of the assumptions, so that the logic of why an assumption is required for one estimand and not another is clear, and the reader can pick out which assumptions are required for the causal effect they target.

Causal effect identification amounts to identifying the two potential outcome means in the contrast. We organize the potential outcomes involved in the various causal effects mentioned earlier into five types and explain the identifying assumptions required for each, starting with the type that requires the weakest and gradually building toward the one that requires the strongest, assumptions, clarifying connections from one type to the next.

We show that identification of the mean (or distribution) of each potential outcome requires three different types of assumptions. We refer to them as *consistency*, *conditional independence*, and *positivity*, but note that they have been discussed under various names in the literature. *Consistency* assumptions [20,21] are closely related to Rubin's *stable unit treatment value assumption* (SUTVA) [9]. *Positivity* is also known as *overlap* or *common support* in the context of identifying causal contrasts. Assumptions in our *conditional independence* category have been called *unconfoundedness*, *ignorability*, *exchangeability*, *conditional randomization*, etc. [20,22–25]. As assumptions in this category are formally stated using conditional independence statements – of certain variables with potential outcomes (potential mediators) – we adopt the shorthand label *conditional independence*, which allows concise reference to specific assumption components.

A note for readers not familiar with the concept of conditional independence: that variables A and B are independent *conditional on* (or *given*) variable C , formally $A \perp\!\!\!\perp B|C$, means that within levels of C (or within each subpopulation that shares the same value on C), knowing A does not tell us anything about B and *vice versa*. In our current problem, the place of B is occupied by a potential outcome or potential mediator; the A place in most assumptions is occupied by the observed exposure or mediator, except in one assumption it is occupied by a potential mediator; and the C place is occupied by covariates.

1.4 Illustrative example

After establishing the identifying assumptions specific to each potential outcome type, we apply them to examine the full set of assumptions needed to identify one or more relevant causal effects. We illustrate this using a running fictional example. In this example, of interest is the health of people who have both a

³ There are cases where effects are nonparametrically not identified, but the researcher uses a parametric assumption as a remedy. Examples include the no exposure-mediator interaction assumption when the cross-world independence assumption required by natural (in)direct effects is violated, or the use of model extrapolation to compensate for violation of positivity. Such *parametric identification* is outside the scope of the current article.

psychiatric disorder and chronic medical problems, and the issue is that psychiatric symptoms may pose challenges to the patient's access to and effective use of medical care [26]. We restrict our attention to people who are members of a (more or less) organized system for the provision of health care; in the United States, this could be a health maintenance organization or an accountable care organization. The members all have health insurance coverage, and the system has some ability to enact certain system-wide change in the practice of care.

Suppose a local (e.g., state) branch of the system offers an intervention program that aims to improve outcomes for members with a bipolar diagnosis and a chronic medical problem, at no additional charge. (This intervention is loosely based on the bipolar disorder medical care model in [27].) It consists of (i) a self-management education component that teaches patients (in group sessions over a 3 month period) how to manage chronic psychiatric and medical conditions, improve dietary behavior and physical activity, and communicate better with medical providers; and (ii) a care-management component (starting at about four months) that involves help from a case manager who facilitates the patient's communication with their medical providers and oversees the patient's health care.

With this *intervention*, we take the outcome to be the patient's health-related quality of life at 18 months after program enrollment (*quality of life*). Two variables are theorized to be on the causal path: a measure of proficient self-management of psychiatric symptoms at three months and a measure of effective use of medical care services at 12 months. We label these variables *symptom management* and *service use* for conciseness, noting that the second variable is about effectiveness (not quantity) of service use. We illustrate causal contrasts of several types, some focusing on the intervention itself, others in consideration of context changes or system-wide practice adjustments.

2 Five potential outcome types

2.1 First type: Y_a

Y_a is the potential outcome in a world where exposure is set to a , where a may be 1 (exposed) or 0 (unexposed). We denote its mean by $E[Y_a]$, with the general notation $E[\cdot]$ indicating *expectation* (or *mean*). This potential outcome type is relevant to the *average total effect*, defined as $TE = E[Y_1] - E[Y_0]$, the difference between mean potential outcome under exposure and mean potential outcome under non-exposure. $E[Y_a]$ is also involved in the natural (in)direct effects mentioned earlier, which decompose the total effect. (These will be formally defined shortly.) In addition, $E[Y_a]$ is relevant to effects of all (hypothetical) intervention conditions that are assessed relative to the existing nonexposure condition.

2.2 Second type: Y_{am}

This is the potential outcome in the hypothetical world where exposure is set to a and mediator is set to a specific value m . Y_{am} is relevant to controlled direct effects, as formally the controlled direct effect for a mediator control level m is $CDE(m) = E[Y_{1m}] - E[Y_{0m}]$. In addition, $E[Y_{am}]$ is the building block for identification of the next potential outcome means, where we consider not just one mediator value but a range of values under a distribution.

2.3 Third type: $Y_{a\mathcal{M}}$ where \mathcal{M} is a known distribution

This is the potential outcome in a hypothetical world where the exposure is set to a and the mediator is intervened upon and set to a distribution \mathcal{M} (which we call the *interventional mediator distribution*), where

\mathcal{M} is a distribution that is either known or is defined based on data that are observed. Seen from the individual perspective, each individual is assigned a mediator value randomly drawn from the distribution \mathcal{M} . This is thus a *stochastic* intervention [28] on the mediator, whereas the intervention corresponding to the second potential outcome type is *deterministic*.

$Y_{a\mathcal{M}}$ is relevant to generalized direct effects where the mediator distribution is fixed to a distribution \mathcal{M} , i.e., $GDE(\mathcal{M}) = E[Y_{1\mathcal{M}}] - E[Y_{0\mathcal{M}}]$. In addition, it is relevant to a wide range of interventional effects where in the active intervention condition (or in both conditions contrasted) the potential outcome is of the $Y_{a\mathcal{M}}$ type. Applications 3 and 4 in our illustrative example (see Sections 6.5 and 6.6) concern such effects.

Depending on the specific condition of interest, the interventional mediator distribution \mathcal{M} may be defined unconditionally (i.e., the same distribution applies to everyone) or conditional on pre-exposure covariates (e.g., different distributions for men and women) or on post-exposure covariates. In the latter case, we require a *same-world* rule: the distribution \mathcal{M} may be defined conditional on L_a (which arises after exposure has been set to a), but not on $L_{a'}$ (where a' is the other exposure condition) or on the observed L (a mixture of L_a and $L_{a'}$). We can think of this as a plausible interventional world: after exposure is set to a , only L_a arises, so an intervention on the mediator may condition on C and L_a .

2.4 Fourth type: $Y_{a\mathcal{M}}$ with \mathcal{M} defined based on potential mediator distribution(s)

This is similar to the third type, except that \mathcal{M} is defined based on potential mediator distribution(s). This type is involved in interventional (in)direct effects. Recall that an interventional direct effect is the effect of exposure on outcome had the mediator distribution been fixed to be the same as a potential mediator distribution, and an interventional indirect effect is the effect on the outcome of a switch in the mediator distribution from the potential distribution under nonexposure to that under exposure while exposure itself is fixed. These are formally $IDE_{a^*} = E[Y_{1\mathcal{M}_{a^*|C}}] - E[Y_{0\mathcal{M}_{a^*|C}}]$ and $IIE_a = E[Y_{a\mathcal{M}_{1|C}}] - E[Y_{a\mathcal{M}_{0|C}}]$, where $\mathcal{M}_{a^*|C}$ (a^* being either 0 or 1) is convenient notation indicating that the interventional mediator distribution \mathcal{M} is defined to be the same distribution as that of the potential mediator M_{a^*} given C .

Like the previous potential outcome type, this fourth type is relevant to a wide range of interventional effects, not limited to interventional (in)direct effects. Applications 5 and 6 in the illustrative example (Sections 7.7 and 7.8) concern such effects.

With this potential outcome type, we make an important subtle differentiation between the interventional mediator distribution \mathcal{M} and the potential mediator distribution(s): the former is defined based on the latter, the latter informs the former, but the two are not the same. This allows us to consider a simple potential mediator type, M_{a^*} , the potential mediator if exposure were set to a^* (a^* being just a separate index that is not tied to a), but have ample flexibility in defining the distribution \mathcal{M} . For example, \mathcal{M} could be defined based on the distribution of M_1 only, or of M_0 only (as in the interventional (in)direct effects mentioned earlier), or a mixture of both. \mathcal{M} could be defined unconditionally, e.g., to be the same distribution as the marginal distribution of M_{a^*} , or could be defined conditional on C to be the same as the distribution of M_{a^*} given C . \mathcal{M} could also be defined conditional on (C, L_a) to be the same distribution as that of M_{a^*} given (C, L_a) . Note that in all these cases, the interventional mediator distribution \mathcal{M} respects the same-world rule.

2.5 Fifth type: the cross-world potential outcome $Y_{aM_{a'}}$, with $a \neq a'$

This is the potential outcome in a completely imaginary world where the exposure is set to condition a , and then the mediator is set, for each individual, to its potential value under the opposite condition a' . There are two potential outcomes of this type, Y_{1M_0} and Y_{0M_1} . They decompose the total effect into two pairs of natural (in)direct effects:

$$TE = \overbrace{E[Y_1] - E[Y_{1M_0}]}^{NIE_1} + \overbrace{E[Y_{1M_0}] - E[Y_0]}^{NDE_0} \quad \text{and} \quad TE = \overbrace{E[Y_1] - E[Y_{0M_1}]}^{NDE_1} + \overbrace{E[Y_{0M_1}] - E[Y_0]}^{NIE_0}.$$

Additional note: The five aforementioned potential outcomes are not exhaustive. These types treat the exposure differently from the mediator: the exposure is set to one value (with no randomness), while the mediator is set either to one value or to a distribution (i.e., with randomness). There may be cases in which we are interested in a condition where exposure is set to a certain distribution \mathcal{A} instead of a single value 1 or 0 (e.g., where an outreach campaign helps substantially increase the rate of enrollment in the intervention program but not make it 100%). We do not consider this potential outcome type separately, because identification for the five listed types renders this type identified, e.g., if the distribution \mathcal{A} is that of 1/3 exposed and 2/3 unexposed, then $E[Y_{\mathcal{A}}] = (1/3)E[Y_1] + (2/3)E[Y_0]$.

3 Identification of $E[Y_a]$

3.1 Consistency assumption

Consistency is the type of assumption that connects potential variables to observed variables. Here, the assumption is that $Y = Y_a$ if $A = a$. That is, for individuals with actual exposure $A = a$, the observed outcome Y reveals the potential outcome Y_a . This seems like an obvious fact, but it is an assumption; the idea is that the potential outcome Y_a is well defined [21], no matter how exposure value a is assigned to the individual and no matter what exposure is assigned to other individuals [9]. This assumption would be violated if one person's exposure affects others' outcomes (a typical example being vaccination), or, say, if a person's potential outcome under an exposure varies depending on whether they self-select or are assigned the exposure.

3.2 Leveraging covariates

Consistency takes us one step toward our goal; it says that we observe Y_a in *some* individuals. But there is a missing data problem, as we want the mean of Y_a over the *whole* population. To handle this problem, the strategy is to leverage a set of observed covariates C such that conditional on C (i.e., within each subpopulation that shares the same C values), the mean of Y_a is identified from the partially observed data. Consequently, the population mean is identified, as it is basically a weighted average of the subpopulation means, where the weights reflect the distribution of the covariates. This is written formally as follows:

$$E[Y_a] = E_C\{E[Y_a|C]\},$$

where the right-hand side is a *double expectation*, with the inner expectation $E[Y_a|C]$ representing the mean of Y_a *conditional on* covariates C , and the outer expectation averaging the conditional mean over the covariate distribution.⁴ Essentially we identify $E[Y_a]$ by identifying the conditional mean $E[Y_a|C]$. To identify this conditional mean, in addition to *consistency*, we need *conditional independence* and *positivity* assumptions.

To understand these assumptions (formalized shortly), we need to take a close look at what is meant by covariates. For simplicity, we take covariates to be *confounders* and use the common cause definition: A *confounder* of two variables (an exposure and an outcome) is a cause they share. As a common cause, it induces an association between the two variables, which *confounds*, or confuses, their true causal

⁴ Readers from some disciplines may be more used to writing expectations in probability-weighted average formats, so $E_C\{E[Y_a|C]\}$ may be written as $\sum_c E[Y_a|C=c]P(C=c)$ or $\sum_c \sum_y yP(Y_a=y|C=c)P(C=c)$. For complicated identification results, these formats become quite visually taxing and confusing. The expectation notation, however, remains succinct and clear on two key pieces of information: averaging what and averaging over which distribution.

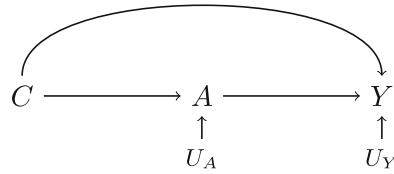


Figure 1: A simple DAG with variables relevant to $E[Y_a]$.

relationship. For example, education is likely a confounder of the occupation–happiness relationship, as it influences what occupation people have, and may influence happiness in ways other than through occupation. Confounders are only a subset of variables that can be used to remove confounding, which are called *deconfounders* [29]. The use of deconfounders that are not simply confounders is an advanced topic we leave out of this article.

We draw a *causal directed acyclic graph* (DAG) [1] in Figure 1 representing the relationships among the relevant variables. Because the potential outcome Y_a is agnostic to mediators, we can leave mediators out of the DAG and just include exposure A , outcome Y , common causes C of these two variables, and arrows representing the causal influences among those variables.⁵ (In the special case where exposure is randomized, there will be no C , as A and Y do not have common causes.) Also shown are U_A and U_Y , causes of A and Y that are not shared; such unique causes are often omitted from DAGs. An important note: The arrow from A to Y in this DAG captures all the influence of A on Y (inclusive of influence through and not through the mediator M); the arrow from C to Y captures all the influence of C on Y except the part that goes through A .

3.3 Conditional independence assumption

In most applications, we are not privy to the truth, or the full truth, about confounders. What we try to do is to guess, based on prior knowledge and theory, what the important confounders are, and collect data on them. Then we resort to making the untestable assumption that the covariates C we observe capture all the confounders (of the relationship between A and Y_a). This is the gist of the *conditional independence* (also known as *unconfoundedness*, *exchangeability*, or *ignorability*) assumption.

Let us be a bit more formal here to build clarity that will help with the later potential outcome types. For ease of reference, we label this assumption (I_a) , with I for “independence” and the subscript a indexing potential outcome Y_a . This assumption is that Y_a is independent of exposure status conditional on covariates, formally

$$A \perp\!\!\!\perp Y_a | C, \quad (I_a)$$

where $\perp\!\!\!\perp$ is the symbol for independent. Intuitively, *within each subpopulation that shares the same values on covariates C* , individuals are similar enough that whether an individual happens to be exposed or not does not carry any additional information about their Y_a . This means we can *ignore* exposure status when considering Y_a . Put another way, within each such subpopulation, exposed and unexposed individuals are *exchangeable* in the sense that they share the same distribution of variable Y_a . This allows equating the subpopulation mean of Y_a with the mean of Y_a in the $A = a$ group in the subpopulation, formally,

$$E[Y_a | C] = E[Y_a | C, A = a].$$

⁵ The order in which a DAG is drawn is to start with variables of interest and arrows representing their causal relationships, and then add variables that are common causes of included variables.

What conditional independence allows us to do, here and later in the article, is what we informally call *going from whole to part*, or *vice versa*. Here, the *whole* is the (C -value-specific) subpopulation, and the *part* is the $A = a$ group in the subpopulation. The beauty of this move is that we do not need to worry about the unobserved Y_a values of the individuals whose actual exposure is not a .

What we now want to do is to replace the potential outcome Y_a on the right-hand side of the aforementioned equation with the observed Y . *Consistency* apparently suggests doing so, since on the right-hand side, we are considering Y_a only among those with $A = a$. It turns out, though, that in addition to *consistency*, we also need the third assumption, *positivity*.

3.4 Positivity assumption

A replacement of $E[Y_a|C, A = a]$ with $E[Y|C, A = a]$ is only legitimate if the latter is well defined for all values of C . This requires the assumption that there is a positive chance of $A = a$ for all C values, formally, $P(A = a|C) > 0$, where $P(\cdot)$ is the notation for probability or probability density. Combined with *consistency*, this means that for all C values, there is a positive chance of observing Y_a . If this is not the case for certain C values, then the mean of Y_a given those values is unidentified, and thus, $E[Y_a]$ is unidentified.

Unlike the other two assumptions, *positivity* is testable, in the sense that given data that have been collected, one could check whether there are parts of the observed covariate distribution where there are no individuals with exposure condition a .

For ease of reference, these three assumptions for all of the five potential outcome types are collected in Table 1.

3.5 Identification result

The three assumptions combined help identify the conditional mean of Y_a ,

$$E[Y_a|C] \stackrel{\text{conditional independence}}{=} E[Y_a|C, A = a] \stackrel{\text{consistency and positivity}}{=} E[Y|C, A = a].$$

And then using double expectation, we identify the population mean:

$$E[Y_a] = E_C\{E[Y|C, A = a]\}, \quad (R_a)$$

where the inner expectation is the mean *observed* outcome given C among those in the $A = a$ condition, and the outer expectation averages this over the distribution of C . (In identification result labels, R stands for “result.”)

3.6 Application 1: the total effect

Identifying the average total effect, $TE = E[Y_1] - E[Y_0]$, involves identifying both potential outcome means. The assumptions required, collected in Table 2, include (i) consistency of both potential outcomes, (ii) conditional independence of exposure with both potential outcomes; and (iii) positivity of both exposure conditions. (i) and (iii) are often written concisely as $Y = AY_1 + (1 - A)Y_0$ and $0 < P(A = 1|C) < 1$, respectively. Under these assumptions, the identification result is $TE = E_C\{E[Y|C, A = 1]\} - E_C\{E[Y|C, A = 0]\}$.

In the example, we consider all effects as average effects over the population of people with a bipolar diagnosis and a chronic medical problem who are in the local branch of the health care system. The total effect is the difference between the potential outcome means under intervention and under usual care, where the means are taken over this population.

Table 1: Three identifying assumptions for five potential outcome means

Consistency assumption			
Potential outcome mean	Consistency of the potential outcome	Consistency of potential intermediate confounder	Consistency of relevant potential mediator(s)
$E[Y_a]$	$Y = Y_a$ if $A = a$		
$E[Y_{am}]$	$Y = Y_{am}$ if $A = a$ and $M = m$		
$E[Y_{a\mathcal{M}}]$, where \mathcal{M} is a known distribution	For all values m in dist. \mathcal{M} , $Y = Y_{am}$ if $A = a$ and $M = m$	If dist. \mathcal{M} conditions on L_a , also require: $L = L_a$ if $A = a$	
$E[Y_{a\mathcal{M}}]$, where \mathcal{M} is defined based on dist. of M_{a^*}	Same as above	Same as above, and If definition of \mathcal{M} relies on info. about dist. of M_{a^*} given L_{a^*} , also require: $L = L_{a^*}$ if $A = a^*$	$M = M_{a^*}$ if $A = a^*$
$E[Y_{aM_{a'}}]$, where $a \neq a'$	For all m in dist. of $M_{a'} C$, $Y = Y_{am}$ if $A = a$ and $M = m$; $Y_{aM_{a'}} = Y_{am}$ if $M_{a'} = m$		$M = M_{a'}$ if $A = a'$
Conditional independence assumption			
Potential outcome mean	Exposure-outcome conditional independence	Mediator-outcome conditional independence	Exposure-mediator conditional independence
$E[Y_a]$	$A \perp\!\!\!\perp Y_a C$		
$E[Y_{am}]$	$A \perp\!\!\!\perp Y_{am} C$	$M \perp\!\!\!\perp Y_{am} C, L, A = a$	
$E[Y_{a\mathcal{M}}]$, where \mathcal{M} is a known distribution	For all values m in dist. \mathcal{M} , $A \perp\!\!\!\perp Y_{am} C$	For all values m in dist. \mathcal{M} , $M \perp\!\!\!\perp Y_{am} C, L, A = a$	
$E[Y_{a\mathcal{M}}]$, where \mathcal{M} is defined based on dist. of M_{a^*}	Same as above	Same as above	$A \perp\!\!\!\perp M_{a^*} C$
$E[Y_{aM_{a'}}]$, where $a \neq a'$	For all m in dist. of $M_{a'} C$, $A \perp\!\!\!\perp Y_{am} C$	For all m in dist. of $M_{a'} C$, $M \perp\!\!\!\perp Y_{am} C, A = a$ and $M_{a'} \perp\!\!\!\perp Y_{am} C$	$A \perp\!\!\!\perp M_{a'} C$
Positivity assumption			
Potential outcome mean	Positivity of exposure conditions	Positivity of mediator values	
$E[Y_a]$	$P(A = a C) > 0$		
$E[Y_{am}]$	Same as above	$P(M = m C, L, A = a) > 0$	
$E[Y_{a\mathcal{M}}]$, where \mathcal{M} is a known distribution	Same as above	For all values m in dist. \mathcal{M} , $P(M = m C, L, A = a) > 0$	
$E[Y_{a\mathcal{M}}]$, where \mathcal{M} is defined based on dist. of M_{a^*}	Same as above, and if $a^* \neq a$, also require $P(A = a^* C) > 0$ if $a^* \neq a$ and dist. \mathcal{M} conditions on (C, L_a) and is defined to be the same as dist. of M_{a^*} given (C, L_{a^*}) , also require $P(A = a^* C, L) > 0$	Same as above	
$E[Y_{aM_{a'}}]$, where $a \neq a'$	$0 < P(A = a C) < 1$	For all m in dist. of $M_{a'} C$, $P(M = m C, A = a) > 0$	

Note: dist. = distribution.

Table 2: Identifying assumptions for TE

	Assumptions	Relevant to	
		$E[Y_1]$	$E[Y_0]$
Consistency	$Y = Y_1$ if $A = 1$ $Y = Y_0$ if $A = 0$	✓	✓
Conditional independence	$A \perp\!\!\!\perp Y_1 C$ $A \perp\!\!\!\perp Y_0 C$	✓	✓
Positivity	$P(A = 1 C) > 0$ $P(A = 0 C) > 0$	✓	✓

The consistency assumption simply says that the observed quality of life in a patient in the intervention group is the same as their potential quality of life under intervention (Y_1), and the observed quality of life in a patient in the usual care group is the same as their potential quality of life under usual care (Y_0).

We assemble the set C of likely confounders (i.e., common causes of intervention participation and quality of life): age, sex, education, occupation, income, psychiatric and medical diagnoses, baseline psychiatric and medical symptoms, baseline quality of life, and baseline measures of self-management of symptoms and effective medical care use. The conditional independence assumption says that individuals that share the same values on these C variables also share the same Y_1 distribution and the same Y_0 distribution, regardless of whether they actually receive the intervention or usual care. If an important confounder, say baseline quality of life, is not included in C , this assumption is violated.

The positivity assumption means that for individuals with any given realization of C , there is a positive chance of receiving the intervention and a positive chance of receiving usual care. If the sample includes individuals with some specific covariate value none of whom participated in the intervention (e.g., patients who could not attend group sessions due to physical mobility challenges), positivity is violated.

A couple of comments: First, in most practical settings, the distinction between the double conditional independence assumption for TE (which involves both Y_1 and Y_0) and the single version specific to one potential outcome Y_a may not matter, for we would arrive at roughly the same set of covariates. This is fortunate, as substantive considerations tend to be imprecise – asking what are the common causes of A and Y , instead of A and Y_a for a specific value a . Note though that when we need to identify the mean of Y_0 or of Y_1 but not both (e.g., the effect of a modified intervention relative to usual care involves Y_0 but not Y_1), only the single version relevant to the specific potential outcome is required. Second, the reason why the double positivity assumption here is called *covariate overlap* or *common support* is that positivity of both exposure conditions implies that the support of the covariate distribution (i.e., range of covariate values) is shared between the exposed and unexposed.

4 Several types of confounders in the mediation setting

Unlike Y_a , the other four potential outcome types correspond to conditions where both the exposure and the mediator are manipulated. Since the mediator is manipulated, the relevant DAG is expanded from the one shown in Figure 1. It includes exposure A , outcome Y , mediator of interest M , and covariates that are common causes of any two of these variables. Intuitively, we can think of the covariates as consisting of confounders of the exposure-mediator, exposure-outcome, and mediator-outcome relationships. (And in the special case where exposure is randomized, there are only mediator-outcome confounders.) For any application, the actual DAG (the one that informs the analysis) may get rather complicated, with multiple confounders and complex causal relationships. For the sake of explaining the identifying assumptions, the causal mediation literature commonly uses a shorthand DAG of the form in Figure 2.

Here, covariates are represented by C and L , which differentiate whether they are influenced by exposure – L is but C is not. Relating this to the relationships to be confounded, exposure-mediator and

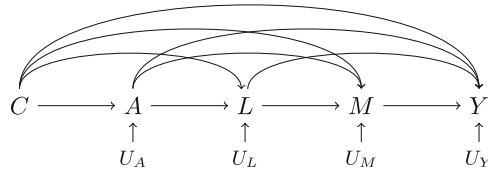


Figure 2: A shorthand DAG: C is a collection of variables, each of which has at least two of the four depicted arrows emitted from C .

exposure-outcome confounders all belong in C – they influence exposure rather than the other way around. Mediator-outcome confounders are split between C and L : those not influenced by exposure (regardless of whether they influence exposure or not) are in C , while those influenced by exposure are in L . We call this DAG shorthand because C is a collection of different (although overlapping) types of variables. Precisely, of the four arrows depicted as emitting from the node C in the DAG to the A , L , M , Y , each variable in C needs to have a minimum of only two.⁶

Note that the current set C is larger than the set C in Figure 1. The set in Figure 1 consists of exposure-outcome confounders (which include exposure-mediator confounders⁷), but does not include mediator-outcome confounders that do not influence exposure. While an abuse of notation, the reuse of the label C here is not problematic because the current set C also works for the identifying assumptions of $E[Y_a]$. In fact, there are a couple of other places where either the current set C or a subset of it could be used in an assumption. To keep presentation simple, we simply use C when stating the assumptions, but also note subsets that could replace C where appropriate.

These C variables are commonly referred to as *pre-exposure confounders* or *pre-exposure covariates*. This terminology is somewhat imprecise, as mediator-outcome confounders in C do not necessarily precede exposure either in time or in the causal structure; the key point is that they are not influenced by exposure. L variables are often called *post-exposure confounders* or *intermediate confounders*. The latter label signals that they are intermediate variables, i.e., also mediators of the effect of A on Y , albeit not the mediator of interest (M). Another way to think of L variables is that they are variables on the causal pathway from A to M that happen to also influence Y in ways that are not through M .

In the illustrative example with the intervention for bipolar patients, depending on the specific research question, one may take symptom management, or service use, or the combination of both, as the mediator of interest M . Whichever the choice, the covariates need to be expanded to cover the different types of confounders. In principle, exposure-mediator confounders should already be part of the exposure-outcome confounders selected when considering the total effect, but it is helpful to double-check. Also, mediator-outcome confounders may need to be added (as C and L variables), and this should be thought through for the specific mediator being considered.

If symptom management is taken as M , baseline general health-related self-efficacy may be an important common cause of symptom management and quality of life. We thus add this variable to the pre-exposure covariate set C . Since symptom management is measured early on (at completion of the self-management intervention component), we are confident that there are no important post-exposure confounders.

If service use is taken as M , then symptom management is likely an intermediate confounder (an L variable). In addition, we add to the set C a variable indicating whether the patient had an annual checkup in the previous year, which is believed to reflect a baseline tendency to use medical services for self-care, a

⁶ Each exposure-mediator confounder has an arrow into A and at least an arrow into L or M . Each exposure-outcome confounder has an arrow into A and at least an arrow into L , M or Y . Each mediator-outcome confounder in C has at least two of the three arrows into L , M , and Y .

⁷ Exposure-mediator confounders form a subset of exposure-outcome confounders, because a cause of the mediator is also a cause of the outcome (via the mediator).

likely mediator-outcome confounder. A variable indicating the patient's specific health insurance plan is also added to the set C .

5 Identification of $E[Y_{am}]$

5.1 Consistency assumption

For $E[Y_{am}]$, this assumption is simply that $Y = Y_{am}$ if $A = a$ and $M = m$ (i.e., in individuals with actual exposure a and actual mediator value m , their observed Y reveals their potential outcome Y_{am}), for a and m being the specific exposure and mediator values that define the potential outcome.

5.2 Conditional independence assumption

Recall that $E[Y_a]$ identification requires *exposure-outcome* conditional independence. Identification of $E[Y_{am}]$, which corresponds to a condition where not only the exposure but also the mediator is manipulated, requires an assumption of both *exposure-outcome* and *mediator-outcome* conditional independence, where *outcome* refers to Y_{am} . Specifically,

$$A \perp\!\!\!\perp Y_{am} | C, \quad (I_{am-AY})$$

$$M \perp\!\!\!\perp Y_{am} | C, L, A = a, \quad (I_{am-MY})$$

where the exposure-outcome component (I_{am-AY}) says that within levels of C , individuals are similar enough that their actual exposure condition provides no information about Y_{am} . The mediator-outcome component (I_{am-MY}) says that among those with exposure $A = a$, within levels of the combination of covariates $\{C, L\}$, individuals are similar enough that the actual mediator value M does not provide any additional information about Y_{am} . (In other words, with appropriate conditioning, exposure is *as good as randomized* and so is mediator value.) The exposure-outcome component is similar to assumption (I_a) , and like (I_a) , it is satisfied by design if exposure is randomized. Since the mediator is not randomized, the mediator-outcome component is always an untestable assumption.

Relating to other terminology, this assumption (essentially *unconfoundedness of Y_{am}*) includes *no unobserved exposure-outcome and mediator-outcome confounding*, where C captures all A - Y_{am} confounders, and C and L capture all M - Y_{am} confounders for those with $A = a$. The two components of this assumption could also be referred to as *ignorability of exposure assignment*, and *ignorability of mediator value assignment*, for Y_{am} . In *exchangeability* terms, the distribution of Y_{am} is exchangeable between exposure conditions conditional on C , and is exchangeable across mediator values within the $A = a$ group conditional on C, L .

5.3 Positivity assumption

Because Y_{am} corresponds to a condition in which both exposure and mediator are manipulated, the positivity assumption includes both *positivity of exposure condition a* and *positivity of mediator value m* . The latter is formally $P(M = m | C, L, A = a) > 0$. That is, among those with $A = a$, the chance of $M = m$ is positive for any combination of values that covariates $\{C, L\}$ may take.

A side note about L : In both the mediator conditional independence and mediator positivity assumptions above, L is part of the conditioning set. We will see that for this and the next two potential outcome types (and all the interventional effects that involve them), identification is possible with intermediate confounders L , but for the fifth potential outcome type (and the natural (in)direct effects that involve it), the presence of L generally results in nonidentifiability.

5.4 Identification

Under *consistency*, Y_{am} values are observed in individuals with $A = a$ and $M = m$. To see how the other assumptions then bridge to the population mean of Y_{am} , it is easier to first consider the special case with no intermediate confounders. In this special case, (I_{am-MY}) reduces to $M \perp\!\!\!\perp Y_{am}|C, A = a$, and the argument involves *going from whole to part* twice. Again, consider a subpopulation of individuals that share the same values of C . In this subpopulation, the overall mean of Y_{am} is equal to the mean of Y_{am} in those with $A = a$ (because exposure status is ignorable for Y_{am} under (I_{am-AY})), which in turn is equal to the mean of Y_{am} among those with $A = a$ and $M = m$ (because mediator value is ignorable for Y_{am} among those with $A = a$ under (I_{am-MY})). Formally,

$$E[Y_{am}|C] \stackrel{\substack{\text{conditional} \\ \text{independence} \\ (I_{am-AY})}}{=} E[Y_{am}|C, A = a] \stackrel{\substack{\text{conditional} \\ \text{independence} \\ (I_{am-MY})}}{=} E[Y_{am}|C, A = a, M = m].$$

Similar to previous reasoning, under *consistency* and *positivity*, the right-hand side is replaced with $E[Y|C, A = a, M = m]$, thus identifying the conditional mean $E[Y_{am}|C]$. Then using the double expectation trick, we obtain $E[Y_{am}] = E_C\{E[Y|C, A = a, M = m]\}$, where the inner expectation is the mean *observed* outcome among those with $A = a, M = m$ within levels of C , and the outer expectation averages over the distribution of C . This is very much in the same spirit as (R_a) , the identification result of $E[Y_a]$.

In the general case with intermediate confounders (L), the bridging involves additional steps similar in nature to those mentioned earlier. We leave the details to the Appendix and just consider the result here. As the outcome depends on L in addition to exposure, mediator, and C , the inner expectation now conditions on both C and L , becoming $E[Y|C, L, A = a, M = m]$, the mean observed outcome for those with $A = a, M = m$ within levels of the combination of $\{C, L\}$. And instead of the double expectation, we have a triple expectation that averages over L before averaging over C , where the distribution of L that is averaged over is that of those with $A = a$, within levels of C . Our general identification result is

$$E[Y_{am}] = E_C(E_{L|C, A=a}\{E[Y|C, L, A = a, M = m]\}). \quad (R_{am})$$

There are two observations that are technical (and non-essential) but may provide additional insight. First, since $E[Y_{am}]$ concerns only one mediator value, m , (I_{am-MY}) may be simplified by replacing M with a dichotomized version of this variable indicating whether it is equal to m or not. We can thus think about Y_{am} as the potential outcome in a condition that intervenes on two binary variables. In most applications, this likely does not make a difference as to which variables are included in C and L . However, this clarity might help make the assumption more meaningful⁸ as it provides a parallelism: (I_{am-AY}) says within levels of C , individuals are similar enough that whether they receive $A = a$ or not does not carry any information about Y_{am} ; and the simplified (I_{am-MY}) says that in the $A = a$ condition, within levels of the combination of $\{C, L\}$, individuals are similar enough that whether they receive $M = m$ or not does not tell us anything about Y_{am} .

Second, C here may be replaced by a subset of C consisting of variables that directly influence L and/or Y (labeled C^{LY}), leaving out variables with no arrows to L and Y (labeled C^{LX}). The reason is simple: when targeting Y_{am} , we need to deal with confounders of the relationship between the combination $\{A, M\}$ and Y_{am} , and variables in C^{LX} influence the former but do not influence the latter (other than through the former) so they do not confound this relationship. Intuitively, variables that are included in C *only because* they are exposure-mediator confounders may be ignored for the purpose of identifying $E[Y_{am}]$.⁹

⁸ For some reason, it seems that the idea of ignorability of exposure assignment is quite intuitive, but the idea of ignorability of mediator value assignment is illusive. This puts them both on equal footing.

⁹ If we consider the two components of (I_{am}) and the three elements of (R_{am}) separately, the answer is: (i) C in (I_{am-AY}) and in the outer expectation in (R_{am}) can be replaced with C^{LY} ; (ii) C in (I_{am-MY}) and in the inner expectation in (R_{am}) can be replaced with a subset of C^{LY} that additionally leaves out variables with an arrow to L that without this arrow would not have been included in C ; and (iii) C in the middle expectation in (R_{am}) can be replaced with another subset of C^{LY} that contains only A - L confounders.

5.5 Application 2: a controlled direct effect

Now suppose that our local branch of the health care system receives communication from headquarters that leadership is considering a system-wide revamping of standard operating procedures, which would incorporate substantial support for the care of medical problems in people with psychiatric disorders, through the use of a range of case management, provider education, integrated patient records, and enhanced linkage network solutions. The communication also says that the expected result of this practice change is to obtain the “maximal effectiveness in use of medical services for care of chronic medical conditions.”

Such a system change would have many implications which branch management has to consider. One of the first questions asked is, if left as is, what would be the effect of the intervention for bipolar patients in the new context, where (based on the expected result of the system change) the service use variable takes the highest value (5 on a 0-to-5 scale) for all bipolar patients. Treating this variable as the mediator (M), the question points to the controlled direct effect $CDE(5) = E[Y_{1,5}] - E[Y_{0,5}]$ for mediator control level 5. (There is, however, doubt about whether this high level is realistic.)

$CDE(5)$ identification requires identifying the means of $Y_{1,5}$ and $Y_{0,5}$, with assumptions collected in Table 3. Note that within the conditional independence assumption, the mediator-outcome component is exposure congruent. It is among patients in the *intervention* program ($A = 1$) that we assume service use (M) is ignorable for the potential outcome under *intervention*-and-highest-service-use ($Y_{1,5}$) given baseline covariates (C) and symptom management (L). Similarly, conditional independence of M with $Y_{0,5}$ is required among patients in usual care ($A = 0$) only.

The positivity of mediator value assumption, written concisely $P(M = 5|C, A, L) > 0$, means that in either exposure condition (intervention or usual care) patients with any realized values of $\{C, L\}$ had a positive chance of scoring 5 on service use. This assumption is violated if there is any subpopulation defined by C, L, A values whose range of this variable does not include level 5.

These assumptions in Table 3 are weaker than assumptions often stated for controlled direct effects in the literature (e.g., [20]) in two ways. First, our assumptions involve only the specific mediator control level ($m = 5$), while assumptions in the literature cover all possible mediator values. Such blanket assumptions are less likely to hold, and are only needed to identify the collection of controlled direct effects corresponding to every mediator level. Second, the mediator-outcome conditional independence assumption in the literature is not exposure specific, e.g., $M \perp\!\!\!\perp Y_{am}|C, L, A$ (for $a = 1, 0$). For Y_{1m} for example, this statement means both $M \perp\!\!\!\perp Y_{1m}|C, L, A = 0$ and $M \perp\!\!\!\perp Y_{1m}|C, L, A = 1$. We only require the latter.

Under the assumptions in Table 3, $CDE(5)$ is identified by $E_C(E_{L|C, A=1}\{E[Y|C, L, A = 1, M = 5]\}) - E_C(E_{L|C, A=0}\{E[Y|C, L, A = 0, M = 5]\})$, a straightforward application of (R_{am}) .

Table 3: Identifying assumptions for $CDE(5)$

	Assumptions	Relevant to	
		$E[Y_{1,5}]$	$E[Y_{0,5}]$
Consistency	$Y = Y_{1,5}$ if $A = 1, M = 5$	✓	
	$Y = Y_{0,5}$ if $A = 0, M = 5$		✓
Conditional independence Exposure-outcome	$A \perp\!\!\!\perp Y_{1,5} C$	✓	
	$A \perp\!\!\!\perp Y_{0,5} C$		✓
Mediator-outcome	$M \perp\!\!\!\perp Y_{1,5} C, L, A = 1$	✓	
	$M \perp\!\!\!\perp Y_{0,5} C, L, A = 0$		✓
Positivity Exposure condition	$P(A = 1 C) > 0$	✓	
	$P(A = 0 C) > 0$		✓
Mediator value	$P(M = 5 C, L, A = 1) > 0$	✓	
	$P(M = 5 C, L, A = 0) > 0$		✓

6 Identification of $E[Y_{a\mathcal{M}}]$ where \mathcal{M} is a known distribution

The assumptions that identify $E[Y_{a\mathcal{M}}]$ for a known distribution \mathcal{M} build on those that identify $E[Y_{am}]$. The key extension is that the assumptions are now required to hold for all values m in the support of the distribution \mathcal{M} . Note that if the interventional mediator distribution \mathcal{M} is defined conditional on covariates, the range of m values depends on the covariates. For example, if \mathcal{M} represents an intervention that differentiates by sex, then the range of m for which the assumptions must hold may differ between men and women.

6.1 Consistency assumption

Consistency of the potential outcome now means $Y = Y_{am}$ if $A = a$ and $M = m$ for all relevant values m in the support of the distribution \mathcal{M} . In addition, if the distribution \mathcal{M} is defined conditional on post-exposure covariates L_a , then *consistency of L_a* is also required, that is, $L = L_a$ if $A = a$.

6.2 Conditional independence assumption

This assumption¹⁰ is that for all the relevant values m in the support of the distribution \mathcal{M} ,

$$\begin{aligned} A &\perp\!\!\!\perp Y_{am}|C, & (I_{a\mathcal{M}}-AY) \\ M &\perp\!\!\!\perp Y_{am}|C, L, A = a. & (I_{a\mathcal{M}}-MY) \end{aligned}$$

This assumption calls for conditional independence of exposure and mediator with not just one potential outcome corresponding to a single mediator value, but a collection of potential outcomes corresponding to the range of mediator values from the distribution \mathcal{M} . Like the previous case, C here may be replaced by the subset C^{LY} .

6.3 Positivity assumption

This assumption includes both *positivity of exposure condition a* and *positivity of relevant mediator values*. The latter is $P(M = m|C, L, A = a) > 0$ for all values m in the support of \mathcal{M} . This means among those with exposure a , within each subpopulation that shares the same $\{C, L\}$ values, the actual range of mediator values has to cover the range of values prescribed by the interventional distribution \mathcal{M} . If not, this assumption is violated.

6.4 Identification result

The identification result is an extension of the triple expectation (R_{am}) to a quadruple expectation that involves averaging over the distribution \mathcal{M} in addition to averaging over L and C .

$$E[Y_{a\mathcal{M}}] = E_C[E_{L|C, A=a}(E_M\{E[Y|C, L, M, A = a]\})]. \quad (R_{a\mathcal{M}})$$

¹⁰ If the distribution \mathcal{M} is defined conditional on L_a , strictly speaking, the first component of the assumption is $A \perp\!\!\!\perp (L_a, Y_{am})|C$. Conditional independence between A and L_a is needed because, since L_a is observed only in those with $A = a$, identification of $E[Y_{a\mathcal{M}}]$ (where \mathcal{M} conditions on L_a) requires borrowing information about L_a across exposure conditions. Fortunately this distinction does not make any practical difference, because the C variables we are considering are confounders. A set of confounders C that satisfies $A \perp\!\!\!\perp Y_{am}|C$ also satisfies $A \perp\!\!\!\perp L_a|C$. This follows from the fact that L_a is a cause of Y_{am} , which means the confounders of the A - Y_{am} relationship (a subset of C) contain the confounders of the A - L_a relationship.

6.5 Application 3: a generalized direct effect

Additional communication from headquarters later clarified that the earlier statement about “maximal effectiveness in use of medical services” meant matching the effective service use level of otherwise similar patients who are without psychiatric disorders, not the highest possible level 5. This means that we are drawing information from the observed effective service use distribution in non-psychiatric patients, conditional on key covariates (age, sex, education, occupation, income, health insurance plan, medical diagnoses, and previous year annual checkup). Instead of the controlled direct effect $CDE(5)$, we are now considering the generalized direct effect $GDE(\mathcal{M}) = E[Y_{1\mathcal{M}}] - E[Y_{0\mathcal{M}}]$, where \mathcal{M} is defined to be the same as that observed distribution.

The identifying assumptions for $GDE(\mathcal{M})$ are the same as those for $CDE(5)$ in Table 3, except that the service use score 5 is replaced with all values m from the support of the distribution \mathcal{M} (the range of variable service use observed in non-psychiatric patients). This range is covariate-dependent, e.g., it may be different for people on different health insurance plans, or for people who did versus did not have a checkup in the previous year.

While the assumptions here are more complex than those for $CDE(m)$ where m is a single value, practical considerations of the conditional independence assumption tend to be similar – seeking in broad terms exposure-mediator, exposure-outcome, and mediator-outcome confounders. The positivity of relevant mediator values assumption, however, deserves attention. It requires that for any (C, L) pattern, regardless of exposure condition, the observed range of mediator values covers the range given that (C, L) pattern in the distribution \mathcal{M} . In the current example, since \mathcal{M} is defined based on the distribution in the non-psychiatric population, this means that within (C, L) levels, (i) the effective service use score range in bipolar patients in intervention and (ii) the corresponding range in bipolar patients in usual care *both* cover (iii) the range of this variable in non-psychiatric patients.

Under these assumptions, $GDE(\mathcal{M})$ is identified by $E_C[E_{L|C,A=1}(E_M\{E[Y|C, L, M, A = 1]\})] - E_C[E_{L|C,A=0}(E_M\{E[Y|C, L, M, A = 0]\})]$, a straightforward application of $(R_{a\mathcal{M}})$.

A side note: While this $GDE(\mathcal{M})$ is a useful contrast to consider, branch management notes that, as an approximation of an anticipated situation, it has an important limitation. If the branch’s intervention is effective in helping bipolar patients with symptom management, one would expect that the effective service use distribution would not be exactly the same with or without the intervention. (A patient with better managed bipolar symptoms might benefit more from support, e.g., because they are more likely to answer the calls of a case manager). This is a general limitation of controlled/generalized direct effects. It is hard to match them to plausible situations where in both the exposed and unexposed conditions the mediator (a variable that naturally is affected by exposure) could be fixed to one value or set to the same distribution.

Generalized direct effects are just one of many types of contrasts that involve this potential outcome type. Let us examine another simple example.

6.6 Application 4: effect of a not yet implemented program

After a round of consultation between headquarters and branches, it is decided that more research needs to be done before deciding whether to adopt the sweeping system change. One question is what would be its effect on health and well-being, assuming the aforementioned result of eliminating the difference between psychiatric and non-psychiatric patients in terms of effective use of services for chronic medical problems. Our local branch decides to look into this potential effect for our population of bipolar patients. For a rough answer, we use data from the usual care condition, and consider the contrast $\tau_1 = E[Y_{0\mathcal{M}}] - E[Y_0]$, where \mathcal{M} is defined to be the same as in Application 3.

Identification of τ_1 requires identifying the means of $Y_{0\mathcal{M}}$ and Y_0 . Table 4 collects all the required assumptions (from the relevant sections above, or relevant rows of Table 1). These assumptions are arguably weaker than those for $GDE(\mathcal{M})$, because for τ_1 we do not need to identify $E[Y_{1\mathcal{M}}]$.

Table 4: Identifying assumptions for τ_1

	Assumptions	Related to	
		$E[Y_{0M}]$	$E[Y_0]$
Consistency	$Y = Y_{0m}$ if $A = 0, M = m$ $Y = Y_0$ if $A = 0$	✓	✓
Conditional independence			
Exposure-outcome	$A \perp\!\!\!\perp Y_{0m} C$ $A \perp\!\!\!\perp Y_0 C$	✓	✓
Mediator-outcome	$M \perp\!\!\!\perp Y_{0m} C, L, A = 0$	✓	
Positivity			
Exposure condition	$P(A = 0 C) > 0$	✓	✓
Mediator values	$P(M = m C, L, A = 0) > 0$	✓	
Range of m values	Support of distribution \mathcal{M} (C -dependent)	✓	

Table 5: Identification results for the effects in the applications (top panel, see assumptions in relevant sections) and their simplification in the special case with no L (bottom panel)

Application	General case with intermediate confounders L
1	$TE = E_C\{E[Y C, A = 1]\} - E_C\{E[Y C, A = 0]\}$
2	$CDE(5) = E_C(E_{L C,A=1}\{E[Y C, L, A = 1, M = 5]\}) - E_C(E_{L C,A=0}\{E[Y C, L, A = 0, M = 5]\})$
3	$GDE(\mathcal{M}) = E_C[E_{L C,A=1}(E_{\mathcal{M}}\{E[Y C, L, M, A = 1]\})] - E_C[E_{L C,A=0}(E_{\mathcal{M}}\{E[Y C, L, M, A = 0]\})]$
4	$\tau_1 = E_C[E_{L C,A=0}(E_{\mathcal{M}}\{E[Y C, L, M, A = 0]\})] - E_C\{E[Y C, A = 0]\}$
5	$IDE_0 = E_C[E_{L C,A=1}(E_{\mathcal{M}_{0C}}\{E[Y C, L, M, A = 1]\})] - E_C[E_{L C,A=0}(E_{\mathcal{M}_{0C}}\{E[Y C, L, M, A = 0]\})],$ $IIE_1 = E_C[E_{L C,A=1}(E_{\mathcal{M}_{1C}}\{E[Y C, L, M, A = 1]\})] - E_C[E_{L C,A=1}(E_{\mathcal{M}_{0C}}\{E[Y C, L, M, A = 1]\})]$ where \mathcal{M}_{0C} is identified as: $P(\mathcal{M}^{draw} = m C, L) = P(M = m C, A = 0),$ and \mathcal{M}_{1C} is identified as: $P(\mathcal{M}^{draw} = m C, L) = P(M = m C, A = 1)$
6	$\tau_2 = E_C[E_{L C,A=1}(E_{\mathcal{M}_{0C}}\{E[Y C, L, M, A = 1]\})] - E_C\{E[Y C, A = 0]\}$ where \mathcal{M}_{0C} is identified the same as in Application 5, $\tau_3 = E_C[E_{L C,A=1}(E_{\mathcal{M}_{ C,0,L_1}}\{E[Y C, L, M, A = 1]\})] - E_C\{E[Y C, A = 0]\}$ where $\mathcal{M}_{ C,0,L_1}$ is identified as: $P(\mathcal{M}^{draw} = m C, L_1 = l) = P(M = m C, L = l, A = 0)$
7	natural (in)direct effects are not identified
Application	Special case with no intermediate confounders L
1	same as in the general case
2	$CDE(5) = E_C\{E[Y C, A = 1, M = 5]\} - E_C\{E[Y C, A = 0, M = 5]\}$
3	$GDE(\mathcal{M}) = E_C(E_{\mathcal{M}}\{E[Y C, M, A = 1]\}) - E_C(E_{\mathcal{M}}\{E[Y C, M, A = 0]\})$
4	$\tau_1 = E_C(E_{\mathcal{M}}\{E[Y C, M, A = 0]\}) - E_C\{E[Y C, A = 0]\}$
5	$IDE_0 = E_C(E_{M C,A=0}\{E[Y C, M, A = 1]\}) - E_C\{E[Y C, A = 0]\},$ $IIE_1 = E_C\{E[Y C, A = 1]\} - E_C(E_{M C,A=0}\{E[Y C, M, A = 1]\})$
6	$\tau_2 = E_C(E_{M C,A=0}\{E[Y C, M, A = 1]\}) - E_C\{E[Y C, A = 0]\},$ not applicable to τ_3 , because τ_3 is defined specifically in the context with L
7	$NDE_0 = E_C(E_{M C,A=0}\{E[Y C, M, A = 1]\}) - E_C\{E[Y C, A = 0]\},$ $NIE_1 = E_C\{E[Y C, A = 1]\} - E_C(E_{M C,A=0}\{E[Y C, M, A = 1]\})$

Under these assumptions, $\tau_1 = E_C[E_{L|C,A=0}(E_M\{E[Y|C, L, M, A = 0]\})] - E_C\{E[Y|C, A = 0]\}$, where the first term is the same as the second term in the result for $GDE(\mathcal{M})$. To help the reader easily spot this and several other connections among the identification results of the various effects considered in the illustrative example, we gather all those results in Table 5. The table also shows simplified results in the special case with no intermediate confounders.

After examining data from and consulting with branches, and after serious consideration of logistics, costs, and benefits, leadership drops the plan for the system change. This concludes a chapter of our story. In the next sections, we will pay more attention to the intervention program for bipolar patients at our local branch.

7 Identification of $E[Y_{a\mathcal{M}}]$ where \mathcal{M} is defined based on potential mediator distribution(s)

This case inherits the same assumptions and the same result ($R_{a\mathcal{M}}$) from the previous case. The only complication is that ($R_{a\mathcal{M}}$) involves averaging over the distribution \mathcal{M} , which in the current case is not known a priori. This means the distribution \mathcal{M} itself needs to be identified, which requires *identification of the potential mediator distribution(s)* used in the definition of \mathcal{M} . This adds components to the three assumptions.

As a notation reminder, we use M_{a^*} with a generic index a^* to denote a potential mediator whose distribution informs the definition of the interventional mediator distribution \mathcal{M} . If \mathcal{M} is informed by both a distribution of M_1 and a distribution of M_0 , then both distributions need to be identified, and we simply combine the assumptions. The definition of the distribution \mathcal{M} determines the form of the distribution of M_{a^*} that needs to be identified. We consider three cases that may be of interest in applications (and that can serve as building blocks for more complex scenarios):

- (i) \mathcal{M} is defined unconditionally to be the same as the marginal distribution of M_{a^*} ;
- (ii) \mathcal{M} is defined conditional on C to be the same as the distribution of M_{a^*} given C ;
- (iii) \mathcal{M} is defined conditional on (C, L_a) to be the same distribution as that of M_{a^*} given (C, L_a) , precisely, $P(\mathcal{M}^{\text{draw}} = m|C, L_a = l) = P(M_{a^*} = m|C, L_{a^*} = l)$, where $\mathcal{M}^{\text{draw}}$ denotes a draw from the distribution \mathcal{M} .

7.1 Identifying first the relevant M_{a^*} distribution(s) and then the \mathcal{M} distribution

Identification of the M_{a^*} marginal distribution and the conditional distribution given C (cases (i) and (ii)), not surprisingly, requires assumptions similar to those that identify $E[Y_a]$: consistency of M_{a^*} , conditional independence $A \perp\!\!\!\perp M_{a^*}|C$, and positivity of exposure condition a^* . Identification of the distribution of M_{a^*} given (C, L_{a^*}) (case (iii)) additionally requires consistency of L_{a^*} .¹¹

Under these assumptions, the conditional distribution of M_{a^*} is identified by the corresponding conditional distribution of the observed M under exposure condition a^* : $P(M_{a^*} = m|C) = P(M = m|C, A = a^*)$ for case (ii); and $P(M_{a^*} = m|C, L_{a^*} = l) = P(M = m|C, L = l, A = a^*)$ for case (iii). The result for case (i) is obtained by averaging case (ii) result over the distribution of C , so $P(M_{a^*} = m) = E_C[P(M = m|C, A = a^*)]$.

Connecting from here to the distribution \mathcal{M} is straightforward. The assumptions remain the same, except in case (iii), if $a^* \neq a$, the positivity assumption $P(A = a^*|C)$ is replaced with $P(A = a^*|C, L)$. This is to

¹¹ Strictly speaking, it also requires $A \perp\!\!\!\perp (L_{a^*}, M_{a^*})|C$, but this is implied by the assumption $A \perp\!\!\!\perp M_{a^*}|C$ and the fact that the C variables we are considering are confounders. The reasoning here is similar to that in footnote 10. Because L_{a^*} is a cause of M_{a^*} , the confounders of the A - M_{a^*} relationship contain the confounders of the A - L_{a^*} relationship.

ensure that the range of covariate values that the distribution \mathcal{M} conditions on is covered by the corresponding range in the relevant observed mediator distribution. The distribution \mathcal{M} in the three cases is identified as follows:

$$\begin{aligned} \text{Case (i): } P(\mathcal{M}^{\text{draw}} = m) &= E_C[P(M = m|C, A = a^*)], \\ \text{Case (ii): } P(\mathcal{M}^{\text{draw}} = m|C) &= P(M = m|C, A = a^*), \\ \text{Case (iii): } P(\mathcal{M}^{\text{draw}} = m|C, L_a = l) &= P(M = m|C, L = l, A = a^*). \end{aligned}$$

Now we are ready to assemble the full set of identifying assumptions for $E[Y_{a\mathcal{M}}]$. Note that the following assumption statements all refer to m values in the support of the distribution \mathcal{M} . What specific range this is for the three cases will be clarified at the end.

7.2 Consistency assumption

In all cases, this assumption includes *consistency of the potential outcome* Y_{am} for all values m in the support of the distribution \mathcal{M} , plus *consistency of the potential mediator* M_{a^*} (i.e., $M = M_{a^*}$ if $A = a^*$). In case (iii), *consistency of L_a and of L_{a^*} is also assumed*.

7.3 Conditional independence assumption

Like before, *exposure-outcome* and *mediator-outcome* conditional independence is assumed: for all relevant values m in the support of the distribution \mathcal{M} ,

$$\begin{aligned} A \perp\!\!\!\perp Y_{am}|C, & \quad (I_{a\mathcal{M}} - AY) \\ M \perp\!\!\!\perp Y_{am}|C, L, A = a. & \quad (I_{a\mathcal{M}} - MY) \end{aligned}$$

In addition *exposure-mediator* conditional independence is assumed:

$$A \perp\!\!\!\perp M_{a^*}|C. \quad (I_{a\mathcal{M}} - AM)$$

Unlike when the distribution \mathcal{M} was known, here no subset of C can replace C in all three components of this conditional independence assumption.¹²

7.4 Positivity assumption

Like before, *positivity of exposure condition a* and *positivity of relevant mediator values* are required, i.e., $P(A = a|C) > 0$ and $P(M = m|C, L, A = 1) > 0$ for all m values in the support of the distribution \mathcal{M} . In addition, *positivity of exposure condition a^** (for all levels of C) is required, i.e., $P(A = a^*|C) > 0$. In case (iii), if $a^* \neq a$, this is replaced by the stronger assumption of *positivity of exposure condition a^* for all levels of (C, L)* , i.e., $P(A = a^*|C, L) > 0$.

¹² The reason is that the subset C^{LY} that satisfies the first two components may leave out some exposure-mediator confounders, but the third component requires including all such confounders. Yet if we consider components one at a time, C in the first two components can be replaced by C^{LY} , and C in the third component can be replaced by the subset of variables that have (i) an arrow to A and (ii) an arrow to L or an arrow to M or both.

7.5 Identification result

The identification result is of the same form as that for when \mathcal{M} is a known distribution:

$$E[Y_{aM}] = E_C[E_{L|C,A=a}(E_{\mathcal{M}}\{E[Y|C, L, M, A = a]\})], \quad (\mathbf{R}_{aM})$$

except the distribution \mathcal{M} here is not known but is identified as a function of the observed data distribution. In cases (ii) and (iii), this result could be written out in a simple format:

$$\text{Case (ii): } E[Y_{aM}] = E_C\left[E_{L|C,A=a}\left(E_{M|C,A=a^*}\{E[Y|C, L, M, A = a]\}\right)\right],$$

$$\text{Case (iii): } E[Y_{aM}] = E_C\left[E_{L|C,A=a}\left(E_{M|C,L,A=a^*}\{E[Y|C, L, M, A = a]\}\right)\right].$$

For case (i), the expression¹³ is complicated.

7.6 The support of the distribution \mathcal{M}

To be complete, we now clarify the support of the distribution \mathcal{M} mentioned in the assumptions. In case (i), the support of the distribution \mathcal{M} is the same as the support of variable M given $A = a^*$. In case (ii), it is the same as the support of variable M given $C, A = a^*$. In case (iii), it is the same as the support of variable M given $C, A = a^*, L = l$, where l is the actual value of L_a . These details will become more salient in the applications.

7.7 Application 5: interventional (in)direct effects

Even though headquarters' plan was dropped, the debate about that plan raised branch management's interest in taking a closer look at the local intervention for bipolar patients, with special attention to the measure of effective service use. The interest is in quantifying how much of the intervention's effect is through this variable (M) and how much otherwise. This points to natural (in)direct effects (which decompose the total effect), but these effects are not identified in the current setting (this will be clear when we visit the fifth potential outcome type). Some suggest using interventional (in)direct effects as an approximation; others have reservations about this – see arguments in [12]. Although there is no agreement, for the moment, we consider identification of interventional (in)direct effects.

Consider a specific effect pair, $\text{IDE}_0 = E[Y_{1M_0|C}] - E[Y_{0M_0|C}]$ and $\text{IIE}_1 = E[Y_{1M_1|C}] - E[Y_{1M_0|C}]$, which involves three potential outcomes. $Y_{1M_1|C}$ ($Y_{0M_0|C}$) is the potential outcome in a hypothetical world where the patient receives the intervention (usual care), but the service use variable, instead of arising naturally and revealing the individual specific M_1 (M_0) value, is assigned a value drawn from the population distribution of M_1 (M_0) given C . $Y_{1M_0|C}$ is the potential outcome in a hypothetical world where the patient receives the intervention but is assigned a mediator value drawn from the population distribution of M_0 given C . The identifying assumptions are collected in Table 6.

While the statement of the conditional independence assumption here is complex, satisfying it in applications boils down to capturing all exposure-mediator and exposure-outcome confounders (in C), and capturing all mediator-outcome confounders within each exposure condition (in C, L). The positivity assumption, however, deserves more attention than it has been paid in the literature, as it may often be violated.

Note that there is asymmetry in the ranges of mediator values for which the assumptions need to hold, which reflects the asymmetry of $Y_{1M_0|C}$. We represent this asymmetry via separate expressions of m and n value ranges. The m value range that defines the collection of potential outcomes Y_{1m} is larger than the n

¹³ Case (i): $E[Y_{aM}] = E_C\left\{E_{L|C,A=a}\left[\int E[Y|C, L, M = m, A = a]E_C[P(M = m|C, A = a^*)]dm\right]\right\}$.

Table 6: Identifying assumptions for $IDE_0 = E[Y_{1M_0|C}] - E[Y_{0M_0|C}]$ and $IIE_1 = E[Y_{1M_1|C}] - E[Y_{1M_0|C}]$

	Assumptions	Related to		
		$E[Y_{1M_1 C}]$	$E[Y_{1M_0 C}]$	$E[Y_{0M_0 C}]$
Consistency				
Potential outcome	$Y = Y_{1m}$ if $A = 1, M = m$ $Y = Y_{0n}$ if $A = 0, M = n$	✓	✓	✓
Potential mediator	$M = M_0$ if $A = 0$		✓	✓
Conditional independence				
Exposure-mediator	$A \perp M_1 C$ $A \perp M_0 C$	✓	✓	✓
Exposure-outcome	$A \perp Y_{1m} C$ $A \perp Y_{0n} C$	✓	✓	✓
Mediator-outcome	$M \perp Y_{1m} C, L, A = 1$ $M \perp Y_{0n} C, L, A = 0$	✓	✓	✓
Positivity:				
Exposure condition	$P(A = 1 C) > 0$ $P(A = 0 C) > 0$	✓	✓	✓
Mediator values	$P(M = m C, L, A = 1) > 0$ $P(M = n C, L, A = 0) > 0$	✓	✓	✓
Ranges of m and n	m : the support of distribution of M given C n : the support of distribution of M given $C, A = 0$	✓*	✓*	✓

* $E[Y_{1M_1|C}]$ and $E[Y_{1M_0|C}]$ identifying assumptions involve m values in the support of M given $C, A = 1$ and n values in the support of M given $C, A = 0$, respectively. The combined range is the support of M given C .

value range that defines the collection of potential outcomes Y_{0n} . The former is the support of the observed M given C (or the combined support of M_1 and M_0 given C); the latter is the support of M given C and $A = 0$ (or the support of M_0 given C).

Consider the n -specific component, $P(M = n|C, L, A = 0) > 0$ for all n values in the support of M given $C, A = 0$. To simplify reasoning, we condition on C and $A = 0$, and consider a subpopulation of patients in usual care that share the same values of C . Within such a subpopulation, this assumption means that the range of M values does not depend on L , otherwise there are L values for which the range of M does not fully cover the support of M in the subpopulation. In our example, if there is such a subpopulation (patients with a certain profile defined by baseline covariates C who receive usual care) for whom the range of effective service use score (M) depends on the level of symptom management (L) (e.g., patients with poor symptom management have this score in the range of 0 to 3, while the full range for this subpopulation is 0–5), then this assumption is violated. The m -related component, $P(M = m|C, L, A = 1) > 0$ for all m values in the support of M given C , is even more restrictive in that it requires not only that the range of M given C in the exposed not depend on L , but also that it cover the corresponding range in the unexposed.

Putting the two components together, the *positivity of relevant mediator values* assumption for identification of this effect pair means that within levels of C , (i) the range of M if unexposed does not depend on L , (ii) the range of M if exposed does not depend on L , and (iii) the latter covers the former. For any application, this stringent positivity assumption should be checked against data.

In the special case with no intermediate confounders, then L is an empty set, which substantially simplifies the positivity requirement. For example, if instead of service use we take symptom management to be the mediator of interest M , and define IDE_0 and IIE_1 the same way as mentioned earlier, except changing the mediator variable, then *positivity of relevant mediator values* simply means that (iii) within levels of baseline covariates (C), the range of symptom management (M) under intervention covers the range under usual care.

Under the assumptions in Table 6, the identification result is:

$$\begin{aligned} \text{IDE}_0 &= E_C[E_{L|C,A=1}(E_{\mathcal{M}_{0|C}}\{E[Y|C, L, M, A = 1]\})] - E_C[E_{L|C,A=0}(E_{\mathcal{M}_{0|C}}\{E[Y|C, L, M, A = 0]\})], \\ \text{IIE}_1 &= E_C[E_{L|C,A=1}(E_{\mathcal{M}_{1|C}}\{E[Y|C, L, M, A = 1]\})] - E_C[E_{L|C,A=1}(E_{\mathcal{M}_{0|C}}\{E[Y|C, L, M, A = 1]\})], \end{aligned}$$

where the interventional mediator distributions $\mathcal{M}_{0|C}$ and $\mathcal{M}_{1|C}$ are identified as $P(\mathcal{M}^{\text{draw}} = m|C, L) = P(M = m|C, A = 0)$ and $P(\mathcal{M}^{\text{draw}} = m|C, L) = P(M = m|C, A = 1)$, respectively (both independent of L given C). In the special case with no L (see Table 5, bottom panel), this simplifies and coincides with the result for natural (in)direct effects (in Application 7, Section 8.5).

7.8 Application 6: effect of a modified intervention

We continue to treat the service use variable as the mediator M . Now we have a different problem. Anticipating funding cuts we need to trim the intervention down to a lighter version, so a question is whether to remove the care-management component from the intervention and keep only the self-management component. With such a modified intervention, we expect that effective service use scores may be lower compared to levels under the original intervention, but would not be lower than levels under usual care.

What would be the effect on quality of life of a modified intervention removing care management? Our first answer is this effect could be conservatively approximated by $\tau_2 = E[Y_{1\mathcal{M}_{0|C}}] - E[Y_0]$, where $Y_{1\mathcal{M}_{0|C}}$ is the potential outcome in a hypothetical world where the exposure is set to 1 (intervention) and the mediator is assigned a value drawn from the distribution of M_0 (effective service use distribution under usual care) given C . This assumes that the effective service use distribution under the modified intervention is the same as that under usual care. The identifying assumptions are collected in Table 7.

As τ_2 and IDE_0 share the same active intervention condition, but differ in the comparison condition, let us compare their identifying assumptions. There are differences in the consistency and conditional independence assumptions between the two contrasts, but these are not likely to matter in most applications. For example, with a rich enough collection of mediator-outcome confounders that we are willing to assume $M \perp\!\!\!\perp Y_{1m}|C, L, A = 1$ (which is required for both effects), it is likely that we are also willing to assume $M \perp\!\!\!\perp Y_{0m}|C, L, A = 0$ (which is additionally required by IDE_0), simply because there is a limit to how deeply we can realistically think about these assumptions. But again, the difference in the *positivity of relevant mediator values* assumption has practical implications. This assumption, for τ_2 , is simply that within levels of baseline covariates (C), the effective service use (M) range under intervention, regardless of symptom management (L) value, has to cover the full range under usual care. For IDE_0 , however, the assumption also requires that within levels of C , the M range if unexposed does not depend on L values. This means positivity is more likely to hold for τ_2 than for IDE_0 .

Under the assumptions in Table 7, τ_2 is identified by $E_C[E_{L|C,A=1}(E_{\mathcal{M}_{0|C}}\{E[Y|C, L, M, A = 1]\})] - E_C\{E[Y|C, A = 0]\}$, where the first term is the same as the first term in the result for IDE_0 , but the second term is simpler (see Table 5). In the special case with no L , the two results simplify and coincide with each other, and coincide with the result for NDE_0 .

As τ_2 is a conservative approximation of the effect of the modified intervention, we also consider a closer approximation that does not fix the mediator distribution at $\mathcal{M}_{0|C}$. The rationale is that improvement in symptom management (which results from the self-management component of the intervention) may itself lead to more effective service use. We thus consider $\tau_3 = E[Y_{1\mathcal{M}_{|C,0,L_1}}] - E[Y_0]$, where $Y_{1\mathcal{M}_{|C,0,L_1}}$ is the potential outcome where everything occurs as if in the original intervention condition, except that the effective service use variable is shifted to a distribution $\mathcal{M}_{|C,0,L_1}$ that conditions on covariate values (C, L_1) but is defined to be the same as the distribution of M_0 given (C, L_0). That is, $P(\mathcal{M}_{|C,0,L_1}^{\text{draw}} = m|C, L_1 = l) = P(M_0 = m|C, L_0 = l)$. Intuitively, this distribution allows the change in symptom management to influence effective service use.

Table 7: Identifying assumptions for $\tau_2 = E[Y_{1M_0|C}] - E[Y_0]$

	Assumptions	Related to	
		$E[Y_{1M_0 C}]$	$E[Y_0]$
Consistency			
Potential outcome	$Y = Y_{1m}$ if $A = 1, M = m$ $Y = Y_0$ if $A = 0$	✓	✓
Potential mediator	$M = M_0$ if $A = 0$	✓	
Conditional independence			
Exposure-mediator	$A \perp M_0 C$	✓	
Exposure-outcome	$A \perp Y_{1m} C$ $A \perp Y_0 C$	✓	✓
Mediator-outcome	$M \perp Y_{1m} C, L, A = 1$	✓	
Positivity			
Exposure condition	$P(A = 1 C) > 0$ $P(A = 0 C) > 0$	✓ ✓	✓
Mediator values	$P(M = m C, L, A = 1) > 0$	✓	
Range of m values	the support of distribution of M given $C, A = 0$	✓	

The identifying assumptions for τ_3 are collected in Table 8. Comparing to τ_2 , there are two differences in the positivity assumption. The *positivity of the unexposed condition* assumption is more restrictive for τ_3 than for τ_2 : for τ_3 it requires that for any realized values of (C, L) combined, the probability of receiving usual care is positive; for τ_2 it only requires this for all C values. On the other hand, the *positivity of relevant mediator values* assumption is more restrictive for τ_2 : for τ_2 it requires that within levels of C , the M range under intervention for any L value covers the full M range under usual care; for τ_3 it requires that within levels of (C, L) , the M range under intervention covers the corresponding range under usual care. Footnote¹⁴ clarifies this using numeric values.

Under the assumptions in Table 8, $\tau_3 = E_C[E_{L|C,A=1}(E_{\mathcal{M}_{|C,0,L_1}}\{E[Y|C, L, M, A = 1]\})] - E_C\{E[Y|C, A = 0]\}$, where the interventional mediator distribution $\mathcal{M}_{|C,0,L_1}$ is identified as $P(\mathcal{M}^{\text{draw}} = m|C, L_1 = l) = P(M = m|C, L = l, A = 0)$.

7.9 Recap so far

We have seen that as the condition that defines the potential outcome gets more complex, the assumptions required for identifying its mean get more complicated. The conditional independence assumption evolved from a single component (*exposure-outcome* conditional independence) for $E[Y_a]$ identification, to pairs of components (*exposure-outcome* and *mediator-outcome*) for identification of $E[Y_{am}]$ and of $E[Y_{aM}]$ where M is a known distribution, to a trio of components (*exposure-mediator*, *exposure-outcome* and *mediator-outcome*) for identification of $E[Y_{aM}]$ where M is defined based on potential mediator distribution(s). As we

¹⁴ Consider a subpopulation homogeneous in C . For simplicity, symptom management (L) has two levels, high and low. First, suppose that the service use (M) range under intervention is 0–3 if low L and 0–5 if high L , and the M range in usual care is 0 to 3 for both low and high L ; in this case the *positivity of relevant mediator values* assumption holds for both τ_2 and τ_3 because both of the ranges under intervention cover the full range under usual care. Second, if instead the M range in usual care is 0–3 if low L and 0–4 if high L , the assumption holds for τ_3 but not τ_2 , because the intervention range covers the usual care range in low and high L strata, but one stratum-specific intervention range does not cover the full usual care range. Third, if one (or both) of the ranges under intervention is shifted up and does not include value 0, then the assumption does not hold either for τ_2 or τ_3 .

Table 8: Identifying assumptions for $\tau_3 = E[Y_{1M|C,0,l_1}] - E[Y_0]$

	Assumptions	Related to	
		$E[Y_{1M C,0,l_1}]$	$E[Y_0]$
Consistency			
Potential outcomes	$Y = Y_{1m}$ if $A = 1, M = m$ $Y = Y_0$ if $A = 0$	✓	✓
Potential mediator	$M = M_0$ if $A = 0$	✓	
Potential int. confnders	$L = AL_1 + (1 - A)L_0$	✓	
Conditional independence			
Exposure-mediator	$A \perp\!\!\!\perp M_0 C$	✓	
Exposure-outcome	$A \perp\!\!\!\perp Y_{1m} C$ $A \perp\!\!\!\perp Y_0 C$	✓	✓
Mediator-outcome	$M \perp\!\!\!\perp Y_{1m} C, L, A = 1$	✓	
Positivity			
Exposure condition	$P(A = 1 C) > 0$ $P(A = 0 C, L) > 0$	✓	✓
Mediator values	$P(M = m C, L, A = 1) > 0$	✓	
Range of m values	The support of distribution of M given $C, L = l, A = 0$, where l is the actual value of L_1	✓	

shall see, the mean of the cross-world potential outcome requires the most complex conditional independence assumption.

But before diving into that last one, it helps to pause and take a high-level view of the types of reasoning we have engaged in so far, and consolidate two key moves we use. The first move, which we have used repeatedly, is *going from whole to part* (or *vice versa*), made possible by conditional independence. The gist is that within levels of *observed* covariates, when the *observed* A (or M) is independent of a *potential* variable (mediator or outcome), that allows us to learn about the overall distribution (and mean) of that potential variable just from the values observed in a relevant subset of individuals. For example, within levels of C , we learn about the Y_a distribution from the observed Y of individuals with $A = a$ under (I_a) , and about the M_{a^*} distribution from the observed M of individuals with $A = a^*$ under (I_{aM-AM}) . Within levels of C, L , we learn about the distribution of Y_{am} from individuals with $A = a, M = m$ under (I_{am-AY}) and (I_{am-MY}) combined. The second move, which we use when considering a condition where the mediator is set to a certain distribution rather than is let to occur naturally, is *swapping of mediator distributions*. When targeting the mean of Y_{aM} , going from whole to part means we only need to consider the outcome in those with $A = a$. But this only gets us halfway there, because each outcome follows a mediator value, so the mean outcome in those with $A = a$ is tied to the corresponding observed mediator distribution. We need to swap this distribution out for the interventional mediator distribution \mathcal{M} . This is done above by averaging over the distribution \mathcal{M} instead of over the observed mediator distribution – see (R_{aM}) . Equipped with these two moves, we now tackle the cross-world potential outcome.

8 Identification of $E[Y_{aM_{a'}}]$ where $a \neq a'$

It is important to note that this potential outcome is fundamentally different from the ones considered so far, in that it belongs in a completely counterfactual world where the mediator is set, not to a fixed value for all, not to a draw from a certain distribution, but to the specific value that the individual would experience under a different exposure. This cross-world potential outcome is unobservable. Yet its distribution (and mean) is identified under a set of assumptions. Let us now build some intuitive appreciation for the

assumptions. (Interested readers are referred to the Appendix for precise proof, and to the relevant literature for extensive discussion of this potential outcome [10,19,20,30,31].)

For clarity, let us sketch a picture of this hypothetical world. In this world, at first things happen naturally, including the pre-exposure covariates C . Then exposure is set to a . After that everything follows naturally from a for a while; any intermediate confounder L reveals the individual's L_a . Then right at the point where the mediator M is about to realize (as M_a), it is magically intervened upon and set to the individual's $M_{a'}$ value. After that again everything, including the outcome, follows naturally. The outcome $Y_{aM_{a'}}$ arises based on the combination of $(C, a, L_a, M_{a'})$, where L_a and $M_{a'}$ are from different worlds.

We need to somehow connect the mean of this unobservable $Y_{aM_{a'}}$ to observable data. First, we use the *whole-to-part* move to narrow down to considering this potential outcome only among those in the $A = a$ condition. By conditioning on C and invoking an *exposure-outcome* conditional independence assumption similar to (I_{aM-AY}) , within levels of C , we replace the mean of $Y_{aM_{a'}}$ with its mean among those who experience $A = a$. That is, $E[Y_{aM_{a'}}|C] = E[Y_{aM_a}|C, A = a]$. This move does not make $Y_{aM_{a'}}$ any more observable (it is not) but matches the exposure condition; the mediator remains mismatched.

Next, consider only those in the $A = a$ condition. For each of these individuals, the observed mediator is M_a (under consistency). Ideally we want to swap this M_a value for the individual's $M_{a'}$ value in order to obtain knowledge about the target potential outcome $Y_{aM_{a'}}$, but unfortunately $M_{a'}$ is unobserved. Our strategy is to use as a proxy for $M_{a'}$ a distribution (to which $M_{a'}$ belongs) that captures all the information about $M_{a'}$ that is relevant for the purpose of identifying $E[Y_{aM_{a'}}]$, and try to identify that distribution. Using the proxy distribution, we could then apply the *swapping of mediator distributions* move to identify the target potential outcome mean. But what should be the proxy distribution? It turns out that to capture all the relevant information about $M_{a'}$, it has to be a distribution of $M_{a'}$ that conditions on a set of *observed* variables that removes the confounding of the relationship between $M_{a'}$ and $Y_{aM_{a'}}$.

In the special case with no intermediate confounders, C is the only common cause of $M_{a'}$ and $Y_{aM_{a'}}$: C causes $M_{a'}$ in the world with exposure a' , and causes $Y_{aM_{a'}}$ directly in the hypothetical world we are considering. This simple confounding structure is shown in Figure 3b. The appropriate proxy for $M_{a'}$ in those with $A = a$ is thus the distribution of $M_{a'}$ given $C, A = a$. Although $M_{a'}$ is not observed for those with $A = a$, under the same *exposure-mediator* conditional independence assumption as (I_{aM-AM}) (except replacing $M_{a'}$ with $M_{a'}$), this distribution is identified to be equal to the distribution of the observed M given $C, A = a'$. This means that, assuming all the other assumptions hold, the result for $E[Y_{aM_{a'}}]$ is a simple adaptation of the result for $E[Y_{aM}]$, replacing M with this proxy distribution and removing L .

In the general case, there are likely intermediate confounders. Now the relationship of $M_{a'}$ and $Y_{aM_{a'}}$ is confounded not only by C but also by the unique cause U_L of L : U_L causes $M_{a'}$ through $L_{a'}$, and also causes $Y_{aM_{a'}}$ through L_a (see Figure 3a). This means that the set of variables that a proxy distribution (if one exists) conditions on must include not only C but also at least one of the three variables $U_L, L_a, L_{a'}$ (to remove confounding by U_L). Among those with $A = a$ (whose outcomes we are counting on using to learn about the target potential outcome mean), only L_a is observed. This suggests using the distribution of $M_{a'}$ given $C, L_a, A = a$ as the proxy distribution. Unfortunately, any distribution of $M_{a'}$ that conditions on the other-world L_a is unidentified. Hence, $E[Y_{aM_{a'}}]$ is unidentified. Consequently, identification of $E[Y_{aM_{a'}}]$ requires that there are no intermediate confounders.

8.1 Consistency assumption

Similar to the previous case, this assumption here includes the regular (i) consistency of potential outcomes, $Y = Y_{am}$ if $A = a, M = m$ for m in the support of the observed M given $C, A = a'$; and (ii) consistency of the potential mediator, $M = M_{a'}$ if $A = a'$. In addition, it includes (iii) consistency of the cross-world potential outcome, $Y_{aM_{a'}} = Y_{am}$ if $M_{a'} = m$. The latter belongs in a different category that connect different types of potential variables, rather than connecting potential to observed variables.

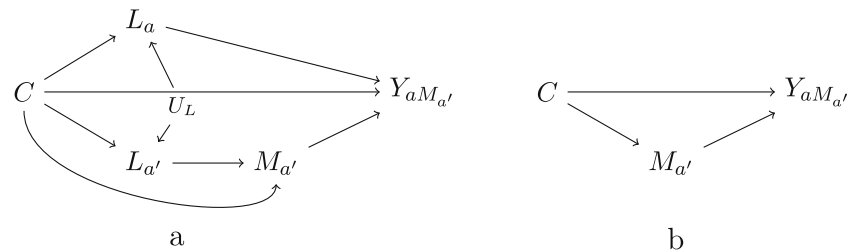


Figure 3: Examining the common causes of $M_{a'}$ and $Y_{aM_{a'}}$. (a) The general case and (b) the special case with no L .

8.2 Conditional independence assumption

To show how this assumption compares to the one for the previous potential outcome type, we present it in four elements as follows:

$$A \perp\!\!\!\perp M_{a'} | C, \quad (I_{aM_{a'}}\text{-AM})$$

and for all values m in the support of M given $C, A = a'$,

$$A \perp\!\!\!\perp Y_{am} | C, \quad (I_{aM_{a'}}\text{-AY})$$

$$M \perp\!\!\!\perp Y_{am} | C, A = a, \quad (I_{aM_{a'}}\text{-MY1})$$

$$M_{a'} \perp\!\!\!\perp Y_{am} | C. \quad (I_{aM_{a'}}\text{-MY2})$$

The first three elements are similar to the assumption from the previous potential outcome type, except replacing a^* with a' and removing L (since we require an empty L).¹⁵ The key difference (from all the previous sets of assumptions) is the fourth element, which says that conditioning on C is sufficient to remove confounding between $M_{a'}$ and $Y_{aM_{a'}}$. This fourth element is well known as the *cross-world independence assumption*.

8.3 Positivity assumption

This assumption includes *positivity of both exposure conditions*, $0 < P(A = 1|C) < 1$, and *positivity of relevant mediator values*, $P(M = m|C, A = a) > 0$ for all m values in the support of M given $C, A = a'$.

8.4 Identification result

The identification result is a rather simple triple expectation:

$$E[Y_{aM_{a'}}] = E_C(E_{M|C, A=a'}\{E[Y|C, M, A = a]\}). \quad (R_{aM_{a'}})$$

Here, when moving from the inner to the middle expectation in this result, we have swapped out (i) the mediator distribution in those whose outcome data we use for (ii) the mediator distribution in those with $A = a'$ within levels of C , which identifies (iii) the distribution of $M_{a'}$ given $C, A = a$, which is the proxy distribution.

¹⁵ Strictly speaking, $(I_{aM_{a'}}\text{-MY1})$ can be relaxed to condition on some L in addition to C , but it is required that $(I_{aM_{a'}}\text{-MY2})$ holds conditional on C only. This applies to the very special case where there exist variables L that are influenced by A that influence the mediator and outcome under exposure a , but do not influence mediator under exposure a' ; this amounts to removing the arrow from $L_{a'}$ to $M_{a'}$ in Figure 3a.

It is noteworthy that (R_{aM_a}) bears some resemblance to the identification result of $E[Y_{aM}]$ where the distribution \mathcal{M} is defined to be the same as the distribution of $M_{a'}$ given C ; the difference is that that previous result conditions on L in the inner expectation and averages over L in a middle expectation. In the special case with no intermediate confounders, that result reduces exactly to the current result, which is not surprising given that the proxy distribution we use above is equivalent to that definition of the distribution \mathcal{M} . Note, though, that the assumptions that lead from two different potential outcome means to the same identification result are not the same; we return to this point shortly.

8.5 Application 7: a pair of natural (in)direct effects

The cross-world potential outcome type is involved in the natural (in)direct effects. Here, we consider one pair of these effects, which results from using Y_{1M_0} to split the total effect:

$$TE = \overbrace{E[Y_1] - E[Y_{1M_0}]}^{NIE_1} - \overbrace{E[Y_{1M_0}] - E[Y_0]}^{NDE_0}.$$

These contrasts are interpreted as (in)direct effects because we could re-express Y_1 as Y_{1M_1} and Y_0 as Y_{0M_0} , so one contrast is the effect of switching exposure while the mediator is fixed, and the other is the effect of switching mediator value while exposure is fixed. The re-expression $Y_a = Y_{aM_a}$ is actually another consistency assumption that connects the two potential outcomes, commonly known as the *composition* assumption [20]. This assumption is only needed for the interpretation.

Identification of this effect pair requires identifying the means of Y_0 , Y_1 and Y_{1M_0} . The required assumptions are collected in Table 9.

Before looking more closely at these assumptions, we note that in our example, with the service use variable as the mediator (M), natural (in)direct effects are not identified due to the presence of the intermediate confounder symptom management (L); this violates the key cross-world independence assumption required to identify $E[Y_{1M_0}]$.

While natural (in)direct effects are not identified, recall from Application 5 that if the assumptions in Table 6 hold, interventional (in)direct effects are identified. This is perhaps a common situation, as the absence of intermediate confounders may be a special situation; it has been noted that unless the mediator comes closely in time after the exposure, there are likely intermediate confounders [32]. In such a situation, it might be tempting to want to switch the estimand from natural to interventional (in)direct effects (as in our story in Application 5), but then that effectively changes the question being asked. An alternative is to opt for estimating bounds on natural (in)direct effects [17] or to do a sensitivity analysis on the unidentified cross-world associations [3].

Let us put our bipolar intervention example with the intermediate confounder problem aside and examine the assumptions in Table 9 as generic assumptions. We note that these assumptions are asymmetric, which means they are weaker than the symmetric assumptions often found in the literature. While the symmetric assumptions allow identification of both pairs of natural (in)direct effects, it seems that often researchers are interested in only one of the two pairs, which means only the asymmetric assumptions relevant to the pair are required. Admittedly it may be hard to think in practical terms about when the conditional independence assumptions may hold for one but not the other pair of natural (in)direct effects. But this is more clear with the *positivity of mediator values* assumption. Its symmetric statement in the literature implies that within levels of C , the mediator range is the same between the two exposure conditions, which is unnecessarily restrictive if we are interested in only one pair of effects. For the current pair of effects, it is only required that within levels of C , the mediator range in the exposed condition covers the mediator range in the unexposed condition. For the other pair of effects, the opposite is required. The practical implication is that in some cases, one pair of natural (in)direct effects may be identified but the other is not due to positivity violation.

Under the assumptions in Table 9, the current pair of natural (in)direct effects are identified as follows:

Table 9: Identifying assumptions for NDE_0 and NIE_1

	Assumptions	Related to	
		$E[Y_1]$ and $E[Y_0]$	$E[Y_{1M_0}]$
Consistency			
Potential outcomes	$Y = AY_1 + (1 - A)Y_0$ $Y_a = Y_{aM_a}$ for $a = 0, 1$ $Y = Y_{1m}$ if $A = 1, M = m$ $Y_{aM_{a'}} = Y_{am}$ if $M_{a'} = m$	✓	✓
Potential mediator	$M = M_0$ if $A = 0$		✓
Conditional independence			
Exposure-outcome	$A \perp\!\!\!\perp Y_a C$ for $A = 0, 1$	✓	
	$A \perp\!\!\!\perp Y_{1m} C$		✓
Exposure-mediator	$A \perp\!\!\!\perp M_0 C$		✓
Mediator-outcome	$M \perp\!\!\!\perp Y_{1m} C, A = 1$		✓
	$M_0 \perp\!\!\!\perp Y_{1m} C$		✓
Positivity			
Exposure condition	$0 < P(A = 0 C) < 1$	✓	✓
Mediator values	$P(M = m C, A = 1) > 0$		✓
Range of m values	the support of distribution of M given $C, A = 0$	✓	

$$NDE_0 = E_C(E_{M|C,A=0}\{E[Y|C, M, A = 1]\}) - E_C\{E[Y|C, A = 0]\},$$

$$NIE_1 = E_C\{E[Y|C, A = 1]\} - E_C(E_{M|C,A=0}\{E[Y|C, M, A = 1]\}).$$

8.5.1 A foot note on extension of natural effects in the case with L

The presence of intermediate confounders L means that the natural (in)direct effects are unidentified. In this case, an alternative is to treat both L and M as mediators and target path-specific effects [2]. This would take us to the general topic of multiple mediators analysis, which is outside the scope of the current article. Here we just note a few key points without explication. Path-specific effects are extensions of natural (in) direct effects for the multiple causally ordered mediators case. Definition of these effects requires a different kind of nested potential outcome, $Y_{aL_a'M_{a'}L_{a'}}$ (for the two mediators case). For example, one decomposition of the total effect is into three components: a direct effect ($E[Y_{1L_0M_0L_0}] - E[Y_0]$), an effect through the first mediator ($E[Y_{1L_1M_0L_1}] - E[Y_{1L_0M_0L_0}]$), and an effect through the second but not the first mediator ($E[Y_1] - E[Y_{1L_1M_0L_1}]$). Roughly speaking, identification of these three effects requires that there are no unobserved exposure-mediator, exposure-outcome, mediator-mediator (L - M), and mediator-outcome confounders, and that there are no exposure-induced (observed or unobserved) confounders of the relationship between (L, M) and Y , plus relevant consistency and positivity assumptions.

9 Concluding remarks

We have shown that identification of a wide range of causal effects in the single mediator case boils down to identification of the mean (distribution) of potential outcomes of five types, and how the assumptions required are connected, getting more complex only as the condition that defines the potential outcome gets more complex. We provide Table 1 as a menu that the substantive researcher can use to assemble identifying assumptions for their target causal estimand. We demonstrate the plausibility consideration of such assumptions for several estimands of common interest through an illustrative example.

We recommend using this article alongside the companion “estimands” paper [12]. The combination of the two papers aim to help the applied researcher first flexibly define causal mediation effects to match their research question and then to assess the effects’ identifiability.

This article did not cover more complex cases such as multiple causally ordered or unordered mediators and repeated exposure and/or mediator over a longitudinal process. Each of these settings comes with a range of effect definitions and identification strategies. The same kind of exercises we conducted in the simple case (connecting effect definitions to real-world research questions and systematic examination of their identification assumptions) is highly recommended for these more complex cases – for the purpose of making advanced methods more accessible and meaningful to applied researchers, facilitating their appropriate use and promoting quality research.

Acknowledgments: This work was supported in a large part by National Institute of Mental Health (NIMH) grant R01MH115487 (PI Stuart), and in small parts by NIMH grant T32MH122357 (PI Stuart) and Office of Naval Research grant N00014-21-1-2820 (PIs Shpitser and Ogburn). All viewpoints and errors belong to the authors. This work has benefited from the helpful feedback we received from several reviewers and from students who participated in our mediation analysis course at Johns Hopkins in 2021 and 2022.

Conflict of interest: Prof. Elizabeth Stuart is a member of the Editorial Board in the Journal of Causal Inference but was not involved in the review process of this article.

References

- [1] Pearl J. *Causality: models, reasoning, and inference*. 2nd edn. New York, NY: Cambridge University Press; 2018.
- [2] VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*. 2014;25(2):300–6. ISSN 1531-5487. doi: 10.1097/EDE.0000000000000034.
- [3] Daniel R, De Stavola BL, Cousens SN, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics*. 2015;71:1–14. ISSN 15410420. doi: 10.1111/biom.12248.
- [4] Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology*. 2017;28(2):258–65. ISSN 1044-3983. doi: 10.1097/EDE.0000000000000596.
- [5] Zheng W, van der Laan M. Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *J Causal Infer*. 2017;5(2). ISSN 2193-3677. doi: 10.1515/jci-2016-0006.
- [6] VanderWeele TJ, Tchetgen Tchetgen EJ. Mediation analysis with time varying exposures and mediators. *J R Statist Soc B Statist Methodol*. 2017;79(3):917–38. ISSN 14679868. doi: 10.1111/rssb.12194.
- [7] Vo TT, Superchi C, Boutron I, Vansteelandt S. The conduct and reporting of mediation analysis in recently published randomized controlled trials: Results from a methodological systematic review. *J Clinical Epidemiol*. January 2020;117:78–88. ISSN 08954356. doi: 10.1016/j.jclinepi.2019.10.001.
- [8] Stuart EA, Schmid I, Nguyen T, Sarker E, Pittman A, Benke K, et al. Assumptions not often assessed or satisfied in published mediation analyses in psychology and psychiatry. *Epidemiol Rev*. September 2021;page mxab007. ISSN 1478-6729. doi: 10.1093/epirev/mxab007.
- [9] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educat Psychol*. 1974;66(5):688–701. ISSN 0022-0663. doi: 10.1037/h0037350.
- [10] Pearl J. Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*; 2001. p. 411–20.
- [11] Robins JM, Richardson TS, Shpitser I. An Interventionist Approach to Mediation Analysis. Vol. 2020; 2020. p. 1–43. <http://arxiv.org/abs/2008.06019>.
- [12] Nguyen TQ, Schmid I, Stuart EA. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychol Meth*. 2021;26(2):255–71. doi: 10.1037/met0000299.
- [13] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;3(2):143–55.
- [14] Didelez V, PhilipDawid A, Geneletti S. Direct and Indirect Effects of Sequential Treatments. In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*; 2006. p. 138–46. doi: 10.1.1.107.4710.
- [15] Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *J R Statist Soc B Statist Methodol*. 2007;69(2):199–215. ISSN 13697412. doi: 10.1111/j.1467-9868.2007.00584.x.
- [16] Holland PW. *Statistics and Causal Inference*. *J Am Statist Assoc*. 1986;81(396):945. doi: 10.2307/2289064.

- [17] Miles C, Kanki P, Meloni S, Tchetgen Tchetgen E. On partial identification of the natural indirect effect. *J Causal Infer.* 2017;5(2):1–12. ISSN 2193-3685. doi: 10.1515/jci-2016-0004.
- [18] Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology.* 2006;17(3):276–84. ISSN 1044-3983. doi: 10.1097/01.ede.0000208475.99429.2d.
- [19] Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Statist Sci.* 2010;25(1):51–71. doi: 10.1214/10-STS321.
- [20] Vander Weele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statist Interface.* 2009;2:457–68.
- [21] Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology* 2009;20(1):3–5. doi: 10.1097/EDE.0b013e31818ef366.
- [22] Hernán MA, Robins JM. *Causal Inference.* Boca Raton, FL: Chapman & Hall/CRC; 2019.
- [23] Imbens GW. The role of propensity score in estimating dose-response functions. *Biometrika* 2000;87(3):706–10.
- [24] Imbens GW, Rubin DB. Rubin causal model. In: Durlauf SN, Blume LE, editors. *The New Palgrave Dictionary of Economics.* 2nd edn. London, UK: Palgrave MacMillan; 2008. ISBN 978-1-349-95121-5. doi: 10.1057/978-1-349-95121-5.
- [25] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70(1):41. doi: 10.2307/2335942.
- [26] Zeber JE, Copeland LA, McCarthy JF, Bauer MS, Kilbourne AM. Perceived access to general medical and psychiatric care among veterans with bipolar disorder. *Am J Public Health* 2009;99(4):720–7. ISSN 00900036. doi: 10.2105/AJPH.2007.131318.
- [27] Kilbourne AM, Post EP, Nosseck A, Drill L, Cooley S, Bauer MS. Improving medical and psychiatric outcomes among individuals with bipolar disorder: a randomized controlled trial. *Psychiat Services.* 2008;59(7):760–8. ISSN 15579700. doi: 10.1176/ps.2008.59.7.760.
- [28] DiiiazMunnnoz I, van der Laan M. Population intervention causal effects based on stochastic interventions. *Biometrics* 2008;68:541–9. ISSN 0006341X. doi: 10.1111/j.1541-0420.2011.01685.x.
- [29] Pearl J. Interpretation and identification of causal mediation. *Psychol Meth.* 2014;19(4):459–81. ISSN 1939-1463. doi: 10.1037/a0036434.
- [30] Pearl J. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevent Sci* 2012;13(4):426–36. doi: 10.1007/s11121-011-0270-1.
- [31] Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green P, Hjort N, Richardson S, editors. *Highly Structured Stochastic Systems, Chapter 2B.* Oxford, UK: Oxford University Press; 2003. p. 70–81. <http://www.hsph.harvard.edu/james-robins/files/2013/03/semantics.pdf>.
- [32] Vansteelandt S, VanderWeele TJ. Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics* 2012;68:1019–27. doi: 10.1111/j.1541-0420.2012.01777.x.
- [33] Koller D, Friedman N. *Probabilistic graphical models: principles and techniques.* Cambridge, Massachusetts: The MIT Press; 2009. ISBN 9780262013192.

Appendix

We derive the identification results for the five potential outcome means based on the corresponding consistency, conditional independence and positivity assumptions.

Identification of $E[Y_a]$

$$\begin{aligned} E[Y_a] &= E\{E[Y_a|C]\} && \text{(iterated expectation)} \\ &= E\{E[Y_a|C, A = a]\} && \text{(conditional independence } A \perp\!\!\!\perp Y_a|C) \\ &= E\{E[Y|C, A = a]\}. && \text{(consistency and positivity)} \end{aligned}$$

In the main text, for an expression with multiple layers of expectations, we adopt an index notation for all the outer layers (excluding the innermost one). This result is thus stated,

$$E[Y_a] = E_C\{E[Y|C, A = a]\}.$$

Identification of $E[Y_{am}]$

$$\begin{aligned} E[Y_{am}] &= E\{E[Y_{am}|C]\} && \text{(iterated expectation)} \\ &= E\{E[Y_{am}|C, A = a]\} && \text{(conditional independence } A \perp\!\!\!\perp Y_{am}|C) \\ &= E\{E\{E[Y_{am}|C, A = a, L]|C, A = a\}\} && \text{(iterated expectation)} \\ &= E\{E\{E[Y_{am}|C, A = a, L, M = m]|C, A = a\}\} && \text{(conditional independence } M \perp\!\!\!\perp Y_{am}|C, A = a, L) \\ &= E\{E\{E[Y|C, A = a, L, M = m]|C, A = a\}\}. && \text{(consistency and positivity)} \end{aligned}$$

Using the index notation, this result is stated in the main text as follows:

$$E[Y_{am}] = E_C(E_{L|C, A=a}\{E[Y|C, A = a, L, M = m]\}).$$

In the following, for simplicity, the notation we use treats the mediator as a discrete variable. The reasoning is the same in the case where the mediator is not discrete, but requires more complicated notation.

Identification of $E[Y_{a\mathcal{M}}]$ where \mathcal{M} is a known distribution

If the interventional mediator distribution \mathcal{M} is unconditional. Denote the probability of mediator value m in the distribution \mathcal{M} by $p_{\mathcal{M}}(m)$. Since \mathcal{M} is defined to be the same distribution for all units, the probability $p_{\mathcal{M}}(m)$ applies regardless of what values C, A, L take.

$$\begin{aligned} E[Y_{a\mathcal{M}}] &= \sum_m p_{\mathcal{M}}(m) E[Y_{am}] && \text{(iterated expectation)} \\ &= \sum_m p_{\mathcal{M}}(m) E\{E\{E[Y|C, A = a, L, M = m]|C, A = a\}\} && \text{(plugging in the result for } E[Y_{am}]) \\ &= E\left\{E\left\{\sum_m p_{\mathcal{M}}(m) E[Y|C, A = a, L, M = m]|C, A = a\right\}\right\} && \text{(linearity of expectation)} \\ &= E\{E_{\mathcal{M}}\{E[Y|C, A = a, L, M]|C, A = a, L\}|C, A = a\}. && \text{(rewriting in short form)} \end{aligned}$$

In the last line, $E_{\mathcal{M}}\{\cdot\}$ indicates that the expectation is taken with respect to the interventional mediator distribution \mathcal{M} instead of the observed mediator distribution.

If \mathcal{M} conditions on C but not L_a . Denote the probability of mediator value m given C in the distribution \mathcal{M} by $p_{\mathcal{M}}(m|C)$. Since \mathcal{M} is defined to be the same distribution for all units that share the same value of C , within levels of C , the probability $p_{\mathcal{M}}(m|C)$ applies regardless of what values A, L take.

$$\begin{aligned}
E[Y_{aM}] &= E\{E[Y_{aM}|C]\} && \text{(iterated expectation)} \\
&= E\left\{\sum_m p_M(m|C)E[Y_{am}|C]\right\} && \text{(iterated expectation)} \\
&= E\left(\sum_m p_M(m|C)E\{E[Y|C, A = a, L, M = m]|C, A = a\}\right) && \text{(plugging in the result for } E[Y_{am}|C]) \\
&= E\left(E\left\{\sum_m p_M(m|C)E[Y|C, A = a, L, M = m]|C, A = a\right\}\right) && \text{(linearity of expectation)} \\
&= E[E(E_M\{E[Y|C, A = a, L, M]|C, A = a, L\}|C, A = a)]. && \text{(rewriting in short form)}
\end{aligned}$$

If \mathcal{M} conditions on C and L_a (or just L_a). By the decomposition and weak union rules of conditional probability [33], the conditional independence assumption $A \perp\!\!\!\perp (L_a, Y_{am})|C$ implies

$$\begin{aligned}
A \perp\!\!\!\perp L_a|C, \text{ and } A \perp\!\!\!\perp Y_{am}|C, &&& \text{(decomposition)} \\
A \perp\!\!\!\perp Y_{am}|C, L_a. &&& \text{(weak union)}
\end{aligned}$$

Denote the probability of mediator value m given C, L_a in the distribution \mathcal{M} by $p_M(m|C, L_a)$. Since \mathcal{M} is defined to be the same distribution for all units that share the same (C, L_a) value combination, within levels of (C, L_a) , the probability $p_M(m|C, L_a)$ applies regardless of what value A takes.

$$\begin{aligned}
E[Y_{aM}] &= E\{E[Y_{aM}|C, L_a|C]\} && \text{(iterated expectation)} \\
&= E\left\{E\left\{\sum_m p_M(m|C, L_a)E[Y_{am}|C, L_a]|C\right\}\right\} && \text{(iterated expectation)} \\
&= E\left\{E\left\{\sum_m p_M(m|C, L_a)E[Y_{am}|C, A = a, L_a]|C\right\}\right\} && (A \perp\!\!\!\perp Y_{am}|C, L_a) \\
&= E\left\{\sum_l P(L_a = l|C) \sum_m p_M(m|C, L_a = l)E[Y_{am}|C, A = a, L_a = l]\right\} && \text{(writing in long form)} \\
&= E\left\{\sum_l P(L_a = l|C) \sum_m p_M(m|C, L_a = l)E[Y_{am}|C, A = a, L = l]\right\} && \text{(consistency)} \\
&= E\left\{\sum_l P(L_a = l|C) \sum_m p_M(m|C, L_a = l)E[Y_{am}|C, A = a, L = l, M = m]\right\} && (M \perp\!\!\!\perp Y_{am}|C, A = a, L) \\
&= E\left\{\sum_l P(L_a = l|C) \sum_m p_M(m|C, L_a = l)E[Y_{am}|C, A = a, L_a = l, M = m]\right\} && \text{(consistency)} \\
&= E\left\{\sum_l P(L_a = l|C) \sum_m p_M(m|C, L_a = l)E[Y|C, A = a, L_a = l, M = m]\right\} && \text{(consistency and positivity)} \\
&= E\left\{\sum_l P(L_a = l|C, A = a) \sum_m p_M(m|C, L_a = l)E[Y|C, A = a, L_a = l, M = m]\right\} && (A \perp\!\!\!\perp L_a|C) \\
&= E\left\{E\left\{\sum_m p_M(m|C, L_a)E[Y|C, A = a, L_a, M = m]|C, A = a\right\}\right\} && \text{(writing in short form)} \\
&= E[E(E_M\{E[Y|C, A = a, L_a, M]|C, A = a, L_a\}|C, A = a)] && \text{(writing in short form)} \\
&= E[E(E_M\{E[Y|C, A = a, L, M]|C, A = a, L\}|C, A = a)]. && \text{(consistency)}
\end{aligned}$$

To sum up, in all three cases of \mathcal{M} , the result has the same form in the last line above. Of course, what varies is the definition of \mathcal{M} . Using the index notation, we write this in the main text as follows:

$$E[Y_{aM}] = E_C[E_{L|C, A=a}(E_M\{E[Y|C, A = a, L, M]\})].$$

Identification of $E[Y_{a\mathcal{M}}]$ where \mathcal{M} is defined based on a distribution of potential mediator M_{a^*}

This case inherits the result above, so we only show derivation of distributions of M_{a^*} .

First, the distribution conditional on C .

$$\begin{aligned} P(M_{a^*} = m|C) &= P(M_{a^*} = m|C, A = a^*) && \text{(conditional independence } A \perp\!\!\!\perp M_{a^*}|C) \\ &= P(M = m|C, A = a^*). && \text{(consistency and positivity)} \end{aligned}$$

Second, the marginal distribution.

$$\begin{aligned} P(M_{a^*} = m) &= E_C[P(M_{a^*} = m|C)] && \text{(iterated expectation)} \\ &= E_C[P(M = m|C, A = a^*)]. && \text{(plugging in result for } P(M_{a^*} = m|C)) \end{aligned}$$

Third, the distribution conditional on C, L_{a^*} . By the weak union rule of conditional independence, the assumption $A \perp\!\!\!\perp (L_{a^*}, M_{a^*})|C$ implies that $A \perp\!\!\!\perp M_{a^*}|C, L_{a^*}$.

$$\begin{aligned} P(M_{a^*} = m|C, L_{a^*} = l) &= P(M_{a^*} = m|C, L_{a^*} = l, A = a^*) && (A \perp\!\!\!\perp M_{a^*}|C, L_{a^*}) \\ &= P(M = m|C, L = l, A = a^*). && \text{(consistency and positivity)} \end{aligned}$$

Identification of $E[Y_{aM_{a'}}]$

$$\begin{aligned} E[Y_{aM_{a'}}] &= E(E\{E[Y_{aM_{a'}}|M_{a'}, C]|C\}) && \text{(iterated expectation)} \\ &= E\left(\sum_m P(M_{a'} = m|C)E[Y_{aM_{a'}}|M_{a'} = m, C]\right) && \text{(writing middle expectation out by definition)} \\ &= E\left(\sum_m P(M_{a'} = m|C)E[Y_{am}|M_{a'} = m, C]\right) && \text{(consistency)} \\ &= E\left(\sum_m P(M_{a'} = m|C)E[Y_{am}|C]\right) && \text{(conditional independence } M_{a'} \perp\!\!\!\perp Y_{am}|C) \\ &= E\left(\sum_m P(M_{a'} = m|C)E[Y_{am}|C, A = a]\right) && \text{(conditional independence } A \perp\!\!\!\perp Y_{am}|C) \\ &= E\left(\sum_m P(M_{a'} = m|C)E[Y_{am}|C, A = a, M = m]\right) && \text{(conditional independence } M \perp\!\!\!\perp Y_{am}|C, A = a) \\ &= E\left(\sum_m P(M_{a'} = m|C)E[Y|C, A = a, M = m]\right) && \text{(consistency and positivity)} \\ &= E\left(\sum_m P(M_{a'} = m|C, A = a')E[Y|C, A = a, M = m]\right) && \text{(conditional independence } A \perp\!\!\!\perp M_{a'}|C) \\ &= E\left(\sum_m P(M = m|C, A = a')E[Y|C, A = a, M = m]\right) && \text{(consistency and positivity)} \\ &= E(E\{E[Y|C, A = a, M]|C, A = a'\}). && \text{(rewriting in short form)} \end{aligned}$$

Using the index notation, this result is stated in the main text as follows:

$$E[Y_{aM_{a'}}] = E_C(E_{M|C, A=a'}\{E[Y|C, A = a, M]\}).$$