Asia Pacific Education Review (2024) 25:703–717
https://doi.org/10.1007/s12564-024-09962-5

REVIEW ARTICLE

An introduction to causal mediation analysis

Xu Qin[1]

Received: 14 June 2023 / Revised: 5 March 2024 / Accepted: 20 March 2024 / Published online: 30 April 2024

Abstract

Causal mediation analysis has gained increasing attention in recent years. This article guides empirical researchers through the concepts and challenges of causal mediation analysis. I first clarify the difference between traditional and causal mediation analysis and highlight the importance of adjusting for the treatment-by-mediator interaction and confounders of the treatment–mediator, treatment–outcome, and mediator–outcome relationships. I then introduce the definition of causal mediation effects under the potential outcomes framework and different methods for the identification and estimation of the effects. After that, I highlight the importance of conducting a sensitivity analysis to assess the sensitivity of analysis results to potential unmeasured confounding. I also list various statistical software that can conduct causal mediation analysis and sensitivity analysis and provide suggestions for writing a causal mediation analysis paper. Finally, I briefly introduce some extensions that I made with my colleagues, including power analysis, multisite causal mediation analysis, causal moderated mediation analysis, and relaxing the assumption of no post-treatment confounding.

Keywords Causal · Mediation · Literature review

Introduction

Mediation analysis answers the question regarding how a treatment generates an impact on an outcome by uncovering the underlying pathways. It is crucial for advancing in-depth scientific understanding in various disciplines, such as education, psychology, and public health. In education research, mediation analysis is necessary when a researcher aims to investigate the mechanisms through which an educational program or intervention operates, i.e., to develop and test a theory explaining the educational processes that shape participants' learning and development. A hypothesized mediation mechanism characterizes the educational process, which often involves a change in cognitive or social-emotional behaviors induced by the program/intervention participation and a subsequent change in one's developmental outcomes. The variable that transmits the program/intervention impact on the outcome plays a role as a mediator. In the basic mediation framework, a treatment affects a focal mediator, which in turn affects an outcome. The total treatment effect can be decomposed into an indirect effect that transmits the treatment effect through the hypothesized focal mediator and a direct effect that works directly or through other unspecified mechanisms. In general, an average indirect effect in the desired direction and magnitude lends support to the theory about the central mechanism.

For example, in the United States National Evaluation of Welfare-to-Work Strategies (NEWWS) study, participants were randomly assigned to the labor force attachment (LFA) program and the control group. The LFA program provided participants with employment-focused incentives and services to transition low-income parents from welfare to work as rapidly as possible, while the control group received aid from the Aid to Families with Dependent Children (AFDC) program and was not given either incentives or mandates to work. Despite the potential effectiveness of LFA on employment, researchers (e.g., Morris, 2008) raised concerns about its potential harm to the long-term mental health of the participants, who tend to be low-income single mothers with young children, especially if they were not able to secure employment. To understand the mediation mechanism underlying the impact of LFA on maternal depression in the long run, Hong et al. (2015) hypothesized that, after

✉ Xu Qin
xuqin@pitt.edu

[1] Department of Health and Human Development at the School of Education, University of Pittsburgh, Office: 5312 Wesley W. Posvar Hall, 230 South Bouquet Street, Pittsburgh, PA 15260, USA

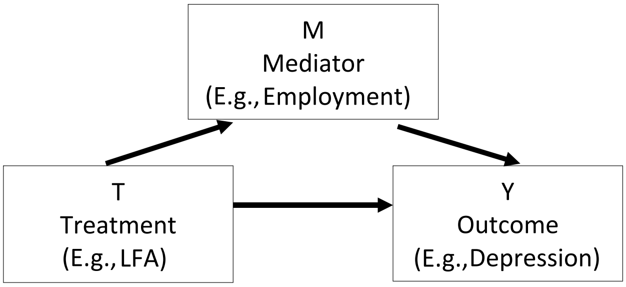

Fig. 1 Diagram of a mediation process

the random treatment assignment ($T = 1$ if assigned to the LFA program and $T = 0$ otherwise), whether a mother was employed in the following two years ($M = 1$ if employed and $M = 0$ otherwise) would mediate the LFA impact on maternal depression at the end of the second year (Y). The causal diagram in Fig. 1 depicts the mediation process. The arrow from T to M and that from M to Y represent how the treatment generates the impact on the outcome through the mediator. The arrow from T to Y captures all the other possible pathways that transmit the treatment effect on the outcome. The goal of a mediation analysis is to decompose the total treatment effect into an indirect effect transmitted through M and a direct effect that operates through all the other possible mechanisms.

Various methods have been developed for mediation analysis, including traditional mediation analysis methods developed within the framework of structural equation modeling (SEM) and causal mediation analysis methods developed within the counterfactual causal framework. The following two sections introduce these two frameworks, respectively, and discuss the advancements offered by causal mediation analysis methods over the traditional approaches.

Traditional mediation analysis and its limitations

In social science research, path analysis (Alwin & Hauser, 1975; Baron & Kenny, 1986; Duncan, 1966; Sobel, 1982; Wright, 1934) and SEM (Bollen, 1987; Jo, 2008; Jöreskog, 1970; MacKinnon, 2008; MacKinnon & Dwyer, 1993) have been the primary techniques for mediation analysis in the past decades. This technique regresses the mediator on the treatment and regresses the outcome on the mediator and the treatment:

$$M = \beta_0^m + \beta_t^m T + \varepsilon_m$$

$$Y = \beta_0^y + \beta_t^y T + \beta_m^y M + \varepsilon_y, \tag{1}$$

where β_t^m denotes the association between the treatment and the mediator, β_m^y indicates the association between the mediator and the outcome given the treatment condition, and β_t^y is the association between the treatment and the outcome given the mediator level. The indirect effect is represented as $\beta_t^m \times \beta_m^y$, and β_t^y represents the direct effect.

In presentations and applications of this technique, one major concern is that researchers are not able to make causal arguments of mediation effects due to a lack of clarifications of the underlying assumptions. In particular, $\widehat{\beta}_t^m \times \widehat{\beta}_m^y$ and $\widehat{\beta}_t^y$ estimated from mediator and outcome models in Eq. (1) would be biased for the indirect and direct effect estimation, if there were confounders of the treatment–mediator, treatment–outcome, and mediator–outcome relationships. Even if observed confounders can be adjusted for in Eq. (1), $\widehat{\beta}_t^m \times \widehat{\beta}_m^y$ and $\widehat{\beta}_t^y$ would still be biased in the presence of unobserved confounders that are hardly avoidable in reality. Another concern is that the definitions of the indirect and direct effects may vary as the mediator and outcome models change. For example, the indirect and direct effects are no longer $\beta_t^m \times \beta_m^y$ and β_t^y if the treatment interacts with the mediator when affecting the outcome, as shown in Eq. (8) or when the mediator and outcome models are nonlinear. Correspondingly, $\widehat{\beta}_t^m \times \widehat{\beta}_m^y$ and $\widehat{\beta}_t^y$ would be biased estimates of the indirect and direct effects if the models in Eq. (1) were misspecified.

Consequence of excluding confounders

Even if the treatment is randomized, mediator values are typically generated through a natural process rather than being experimentally manipulated. As a result, individuals displaying different mediator values tend to differ systematically in many aspects that would confound the relationship between the mediator and the outcome. As illustrated in Fig. 2, $\mathbf{X}$ is a vector of covariates that influence the mediator and the outcome and thus confound the mediator–outcome

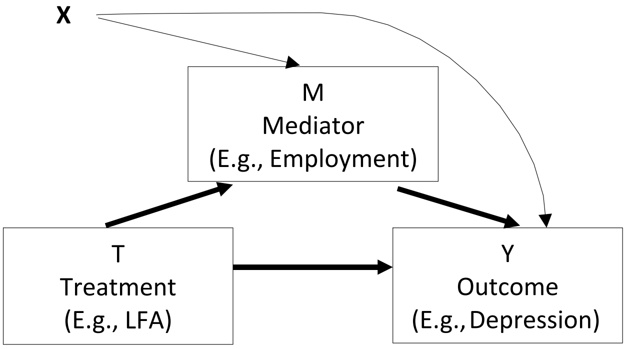

Fig. 2 Diagram of a mediation process with confounders of the M–Y relationship (if treatment is randomized)

relationship. In the NEWWS example, a mother who was less willing to accept a low-wage job might be less likely to be employed and tend to experience more depression if assigned to LFA. Hence, the observed association between employment and maternal depression might be partly attributable to the confounding of the willingness to accept a low-wage job. Failures to adjust for such confounders in the analysis would generate bias. Similarly, if treatment is not randomized, omitting confounders of the treatment–mediator and treatment–outcome relationships would also lead to bias. Therefore, it is important to include such confounders in the analysis. Otherwise, causal conclusions regarding the mediation mechanisms would be invalid.

Consequence of misspecified models such as ignoring treatment-by-mediator interaction

This method relies heavily on correct specifications of both the mediator model and the outcome model. Even if an analyst attempts to make statistical adjustments for all potential confounders, estimation of the indirect and direct effects would still be biased if the regression models were misspecified. Typically, an analyst may overlook a possible treatment-by-mediator interaction, ignoring the fact that the treatment effect may be generated not only by changing the mediator but also by changing the relationship between the mediator and the outcome (Judd & Kenny, 1981). For example, employment may reduce depression under the LFA condition but may not under the control condition, in which case a treatment-by-mediator interaction should be included in the outcome model but is usually ignored in the traditional mediation analysis. An analyst may also overlook a possible treatment-by-covariate interaction, a mediator-by-covariate interaction, a treatment-by-mediator-by-covariate interaction, a nonlinear covariate–mediator relationship, or a nonlinear covariate–outcome relationship (Hong, 2017). Any of these model misspecifications can result in bias.

Causal mediation analysis and its advantages

The mainstream literature on path analysis and SEM did not incorporate the causal inference framework until relatively recently (Holland, 1988; Jo, 2008; Sobel, 2008). This framework overcomes the major limitations of traditional mediation analysis as illustrated above and has gained increasing attention in the recent years (MacKinnon et al., 2020). It provides general definitions of the indirect and direct effects without relying on specific models and clarifies the underlying assumptions for making causal conclusions of mediation mechanisms. Details can be found in the sections of Definition and Identification.

Serious attempts have been made within this causal framework to reduce bias associated with confounders and relax the model-based assumptions including the no treatment-by-mediator interaction assumption. VanderWeele and Vansteelandt (2009) extended the mediator and outcome regressions in Eq. (1) by further adjusting for pretreatment confounders (i.e., confounders preceding the treatment) and the treatment-by-mediator interaction. Imai et al. (2010) proposed a more general procedure that allows mediator and outcome regressions to be semi-parametric or nonparametric and thus relaxes functional form assumptions. In addition to these regression-based methods, other methods were developed to rely on only one of the mediator and outcome models and thus further relax model-based assumptions. As introduced in the Estimation section, the weighting-based method (e.g., Hong et al., 2015) does not rely on an outcome model, while the imputation-based method does not require a mediator model (e.g., Vansteelandt et al., 2012). They are more robust to model misspecifications but less efficient than the regression-based methods, while multiply robust estimation combines their strengths (e.g., Tchetgen Tchetgen & Shpitser, 2012; Vansteelandt et al., 2012).

In real applications, it is almost impossible to measure and adjust for all the potential pretreatment confounders. It is also likely that a confounder of the mediator-outcome relationship is affected by the treatment, which is known as a post-treatment confounder. As explicated in the section of Identification, causal inference regarding a hypothesized mediation mechanism might be invalidated when some confounders of the treatment-mediator, treatment-outcome, and mediator-outcome relationships are omitted from the analysis. Different sensitivity analysis methods have been developed for assessing the sensitivity of causal mediation analysis results to omitted confounders. A detailed introduction can be found in the sections of Sensitivity Analysis and Extensions.

In the rest of the article, I introduce the definitions of the causal mediation effects under the potential outcomes framework and clarify the identification and estimation of the effects using the regression-based and weighting-based methods, including a discussion of correspondence and differences between traditional and causal frameworks of mediation analysis. After introducing the sensitivity analysis methods, I provide a general guideline for writing a causal mediation analysis paper. In the end, I discuss some important extended topics in the literature on causal mediation analysis.

Definition

In causal mediation analysis, indirect and direct effects are defined under the potential outcomes framework, which is also known as the counterfactual causal framework (Neyman & Iwaszkiewicz, 1935; Rubin, 1978). Unlike the definitions of the indirect and direct effects in traditional mediation analysis, which rely on mediator and outcome models and thus can vary under different scenarios (Preacher et al., 2007), the potential outcomes framework provides general definitions of the effects independent of specific statistical models.

Potential mediators and potential outcomes

Under the potential outcomes framework, we can define the potential mediator and potential outcome as functions of the treatment for each individual i, i.e., $M_i(t)$ and $Y_i(t)$. In the NEWWS example, individual i has two potential mediators, $M_i(1)$ and $M_i(0)$, which, respectively, represent individual i's potential employment status if assigned to the LFA program and that if assigned to the control group. Although every individual has two potential mediators, only one of them is observed. For example, if individual i was actually assigned to the LFA program, only $M_i(1)$ is observed, while $M_i(0)$ is counterfactual. Similarly, individual i has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, and only the one under the actual treatment condition is observed. Given this setup, we can define the total treatment effect on the outcome for each individual i as follows:

$$TE_i = Y_i(1) - Y_i(0) \tag{2}$$

In causal mediation analysis, to reflect the fact that the outcome is affected by both the treatment and the mediator, we can alternatively define individual i's potential outcome as $Y_i(t, m)$, which is the potential outcome that we would have observed if the treatment condition were set to t and the mediator value were set to m through intervention or manipulation. Among the multiple potential outcomes, only the one under the actual treatment condition and mediator value is observed.

These definitions rely on the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980, 1986, 1990), which consists of (1) the *no hidden variations in treatment* assumption (also known as the *no multiple versions of treatment* or *treatment-variation irrelevance* assumption) that there is only one version of each treatment condition and (2) the *no-interference* assumption that an individual's potential mediators do not depend on other individuals' treatment status, and an individual's potential outcomes do not depend on other individuals' treatment status or mediator values (Imai et al., 2010, p. 311). In the NEWWS example, the SUTVA assumption requires that individual i's potential employment status and depression level under the LFA or control condition are not affected by the mechanism used to assign the treatment and the treatments the other individuals receive. In addition, individual i's potential depression level is not affected by the other individuals' employment status.

The no-interference assumption would be violated in the presence of spillover effects, e.g., if the depression level of a mother assigned to the LFA program was affected by the employment status of her friend who was assigned to the same program. The no-interference assumption tends to be violated especially in studies that involve clusters such as neighborhoods or schools because the individuals from the same cluster are more likely to influence each other due to social interactions. Hong (2015) provides a solution to potential violations of the no-interference assumption by considering "spillover as an intermediate process that may constitute a part of the mediation mechanism through which a treatment exerts a causal impact on an outcome." For example, the LFA program may impact maternal depression through changing not only a mother's own employment status, but also the employment status of the mother's friends. Therefore, we can account for such spillover effects by decomposing the total treatment effect into the indirect effect mediated by a focal individual's mediator, the indirect effect mediated by peers' mediator, and the direct effect.

Controlled direct effect

Controlling the mediator at a specific level m while changing the treatment condition from 0 to 1, we can define the controlled direct effect of the treatment (Holland, 1988; Pearl, 2001; Robins & Greenland, 1992):

$$CDE_i(m) = Y_i(1, m) - Y_i(0, m) \tag{3}$$

In the NEWWS example, it represents the LFA impact on maternal depression when each individual i was employed if $m = 1$ or when each individual i was unemployed if $m = 0$.

Controlled direct effects are of great interest in policy evaluation (e.g., Pearl, 2001; Robins, 2003). Nevertheless, two major concerns were raised regarding the controlled direct effect (VanderWeele & Vansteelandt, 2009). First, the definition is based on the principle that both the treatment and the mediator are manipulable. However, it is often not conceivable to force everyone's mediator to be the same (e.g., every individual is forced to have the same employment status). Second, an important goal of mediation analysis is to decompose the total treatment effect into a direct effect and an indirect effect. However, the total treatment effect in Eq. (2) minus the controlled direct effect in Eq. (3) may not define the indirect effect. For example, if the LFA impact on maternal depression is not transmitted through employment, i.e., if the indirect effect via employment is 0, the total treatment effect should be equal to the direct effect, but the difference between (2) and (3) is nonzero if the treatment interacts with the mediator when affecting the outcome. In fact, the difference between the total treatment effect and the controlled direct effect is attributed to not only the indirect effect transmitted through the mediator but also the treatment-by-mediator interaction.[1] Therefore, the total treatment effect cannot be decomposed into the controlled direct effect and the indirect effect unless there is no treatment-by-mediator interaction. The natural direct and indirect effects, as introduced below, remove these concerns and are of greater interest in the evaluation of mediation mechanisms (e.g., Hafeman & Schwartz, 2009; Robins, 2003).

Natural direct effect

In consideration of the fact that the mediator value is a potential natural response to the treatment assigned, individual i's potential outcome $Y_i(t)$ can be alternatively expressed as a function of both the treatment and the potential mediator under the same treatment condition $Y_i(t, M_i(t))$. Defining the potential outcome as $Y_i(t, M_i(t))$ instead of $Y_i(t, m)$ allows mediator values under a given treatment condition to naturally vary among individuals rather than being fixed at an arbitrary level. When $M_i(t) = m$, $Y_i(t, M_i(t))$ is equivalent to $Y_i(t, m)$. In other words, when the naturally occurring level of the potential mediator under treatment condition t is equal to m, the two versions of the potential outcome are identical.

Given that $Y_i(t) = Y_i(t, M_i(t))$, the total treatment effect can be alternatively expressed as follows:

$$TE_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \tag{4}$$

Holding the potential mediator at the level that would be observed under the control condition while changing the treatment from 0 to 1, we can define the natural direct effect (Pearl, 2001), which is alternatively termed as pure direct effect by Robins and Greenland (1992):

$$NDE_i = Y_i\left(1, M_i(0)\right) - Y_i\left(0, M_i(0)\right) \tag{5}$$

where $Y_i\left(1, M_i(0)\right)$ denotes individual i's potential outcome if they were assigned to the experimental condition yet counterfactually having the mediator value as they would have under the control condition. In the NEWWS example, the natural direct effect represents the impact of LFA on individual i's maternal depression, while their employment status was kept at the level that would have been observed in the absence of the intervention.

Unlike the controlled direct effect, the natural direct effect maintains the natural relationship between the treatment and the mediator because the mediator is controlled at $M_i(0)$, which allows for natural variation from person to person.

Natural indirect effect

In mediation analysis, we are particularly interested in the indirect effect, which is the treatment effect that is solely transmitted through the mediator. The essence lies in assessing how the treatment-induced change in the mediator affects the outcome. It can be well captured by the difference between the total treatment effect and the natural direct effect, which represents how much the outcome would change if the treatment condition were controlled at the experimental level, while the mediator were changed from the level under the control condition to the level under the experimental condition. It is termed the natural indirect effect by Pearl (2001) and the total indirect effect by Robins and Greenland (1992):

$$NIE_i = Y_i\left(1, M_i(1)\right) - Y_i\left(1, M_i(0)\right) \tag{6}$$

In the NEWWS example, it represents the LFA impact on individual i's maternal depression that is solely attributable to the change in their employment status induced by LFA, when the treatment is fixed at the intervention condition. The individual i's natural indirect effect is zero if LFA does not change their employment status or if, despite a change in the employment status, their maternal depression remains unchanged.

The natural direct effect in Eq. (5) and the natural indirect effect in Eq. (6) add up to the total treatment effect in Eq. (4):

$$\begin{aligned} TE_i = NDE_i + NIE_i = & \left[Y_i\left(1, M_i(0)\right) - Y_i\left(0, M_i(0)\right)\right] \\ & + \left[Y_i\left(1, M_i(1)\right) - Y_i\left(1, M_i(0)\right)\right] \end{aligned}$$

which resolves a major concern of the controlled direct effect.

The above decomposition of the total treatment effect is not unique. Alternatively, the total treatment effect can be decomposed into the sum of the total direct effect

[1] A detailed reasoning can be found in VanderWeele (2014).

$Y_i\big(1, M_i(1)\big) - Y_i\big(0, M_i(1)\big)$ and the pure indirect effect $Y_i\big(0, M_i(1)\big) - Y_i\big(0, M_i(0)\big)$ (Robins & Greenland, 1992). The two ways of decomposition are not necessarily equivalent. A discrepancy between the two would exist if the treatment interacts with the mediator when affecting the outcome. For simplicity, I focus on decomposing the total treatment effect into the natural direct and indirect effects in this article. A more detailed discussion about various ways of decomposing the total treatment effect can be found in VanderWeele (2014).

The above definitions are illustrated with a binary treatment. If a treatment has more than two categories or is continuous, we can extend these definitions by replacing 1 and 0 with any two different values of the treatment, t and t' (Imai et al., 2010; VanderWeele & Vansteelandt, 2009). By taking an average of each individual-specific effect over all the individuals, we can define the population average effects as listed in Table 1.

Identification

The definition of the population average causal effects in Table 1 relies on four potential outcomes of every individual. However, as elaborated above, $Y_i\big(t, M_i(t')\big)$ and $Y_i\big(t', M_i(t)\big)$, where $t \neq t'$ is never observable and $Y_i\big(t, M_i(t)\big)$ (or $Y_i\big(t', M_i(t')\big)$) can be observed only if individual i's actual treatment condition was t (or t'). As illustrated in Table 2, among the eight participants of the NEWWS study, the first four were assigned to the LFA group, for whom the only observables are $M_i(1)$ and $Y_i\big(1, M_i(1)\big)$. The last four were assigned to the control group, for whom the only observables are $M_i(0)$ and $Y_i\big(0, M_i(0)\big)$. Therefore, causal inference is essentially a missing data problem (Ding & Li, 2018; Holland, 1986). The key to causal mediation analysis lies in how to infer the counterfactual quantities from the observed data so that the population average of each potential outcome and correspondingly the population average of each causal effect can be identified.

Table 1 Definitions of population average causal effects

	Notation	Definition
Total effect	$TE = E\big[Y_i\big(t, M_i(t)\big)\big] - E\big[Y_i\big(t', M_i(t')\big)\big]$	Overall average change in the potential outcome if the treatment condition is changed from t' to t
Natural direct effect	$NDE = E\big[Y_i\big(t, M_i(t')\big)\big] - E\big[Y_i\big(t', M_i(t')\big)\big]$	Average change in the potential outcome if the treatment condition is changed from t' to t, while the potential mediator is held at the level that would be observed under the treatment condition t'
Natural indirect effect	$NIE = E\big[Y_i\big(t, M_i(t)\big)\big] - E\big[Y_i\big(t, M_i(t')\big)\big]$	Average change in the potential outcome if the treatment is fixed at condition t, while the potential mediator is changed from the level that would be observed under the treatment condition t' to that under t

Table 2 Illustration of observable and counterfactual quantities

Individual	Treatment	Observed mediator	Observed outcome	Potential mediators		Potential outcomes		
	T_i	M_i	Y_i	$M_i(1)$	$M_i(0)$	$Y_i(1, M_i(0))$	$Y_i(0, M_i(1))$	$Y_i(0, M_i(0))$
1	1	0	25	0	?	?	?	?
2	1	1	5	1	?	?	?	?
3	1	0	10	0	?	?	?	?
4	1	0	2	0	?	?	?	?
5	0	0	30	?	0	?	?	30
6	0	1	2	?	1	?	?	2
7	0	0	8	?	0	?	?	8
8	0	0	10	?	0	?	?	10

Note: ? denotes unobserved

Sequential ignorability assumption

As proved by Hong et al. (2015), Imai et al. (2010), and VanderWeele and Vansteelandt (2009), the identification requires the sequential ignorability assumption, including.

Assumption 1 Strongly ignorable treatment assignment. This assumption can be interpreted as the treatment is as if randomized within levels of pretreatment covariates. In other words, there are no omitted or unmeasured pretreatment confounders of the treatment–mediator or treatment–outcome relationship. The assumption is guaranteed to be met if the treatment is randomized. If the treatment is not randomized, the confounders of the treatment–mediator and treatment–outcome relationships must be adjusted for in the analysis.

Assumption 2 Strongly ignorable mediator value assignment. This assumption can be interpreted as the mediator is as if randomized within the same treatment group or across treatment groups among individuals with the same levels of pretreatment covariates. In other words, (1) there are no omitted or unmeasured pretreatment confounders (i.e., confounders preceding the treatment) of the mediator–outcome relationship and (2) there are no post-treatment confounders (i.e., confounders affected by the treatment) of the mediator–outcome relationship within each treatment condition or across treatment conditions. As illustrated in Fig. 3, component (2) indicates that the covariates that affect both the mediator and the outcome must not be affected by the treatment. In the NEWWS example, the LFA participants who were less willing to accept a low-wage job before participating in the study (pretreatment) or in the first year after participation (post-treatment) might be less likely to be employed in the two years after participation and tend to experience more depression at the end of the second year. Component (1) of Assumption 2 requires that such a pretreatment confounder should be controlled for in the analysis, and component (2) requires that such a post-treatment confounder should not exist. Otherwise, Assumption 2 would be violated. Because mediator is usually generated through a natural process, and it is almost impossible to measure and control for all the confounders, this assumption tends to be violated even if the treatment is randomized. Solutions can be found in the sections of Sensitivity Analysis and Extensions.

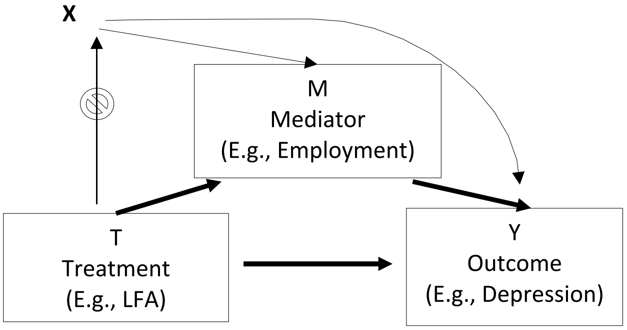

Fig. 3 Illustration of the no post-treatment confounders of the M–Y relationship

Assumption 3 Positivity assumption. Within levels of pretreatment covariates, every individual has a nonzero probability of receiving each treatment level and a nonzero probability of taking each mediator value within the response space of the mediator under each treatment condition. This assumption would be violated if one or more subgroups are never or rarely observed to receive a treatment level or take a certain mediator value under a treatment condition, in which case no or little information would be available for inference about those subgroups' counterfactual outcome. Consequently, as Petersen et al. (2012) argued, "the resulting sparsity in the data may increase bias with or without an increase in variance and can threaten valid inference." The assumption is more likely to be violated at a smaller sample. Petersen et al. (2012) discussed diagnosing violations of the positivity assumption and remedies for treatment effect estimation. Discussions around the positivity assumption are very limited in the literature of causal mediation analysis. Hong et al., (2015, p. 324–325) assessed if the positivity assumption is violated in the case of a discrete mediator by identifying common supports.

Under these assumptions, different identification strategies have been proposed, including.

(1) Regression-based method, which relies on both a mediator model and an outcome model;
(2) Weighting-based method, which constructs weights based on a mediator model and identifies the effects through weighted mean contrasts of the outcome without relying on an outcome model;
(3) Imputation-based method, which identifies the effects through mean contrasts of the potential outcomes that are imputed via an outcome model without relying on a mediator model;
(4) Multiply robust method, which combines (1) with (2) or (3).

Qin and Yang (2022) provided a brief review of the above methods. Due to limited space, I focus on the regression-based and weighting-based identification in this article, while details about the imputation-based and multiply robust identification can be found in Vansteelandt et al. (2012) and Tchetgen Tchetgen and Shpitser (2012), respectively.

Regression-based identification

As proved by Imai et al. (2010), under the sequential ignorability assumption, the distribution of each potential outcome given pretreatment covariates can be written as a function of the distributions of the observed data, i.e., the distribution of the outcome given the treatment, mediator, and pretreatment covariates and the distribution of the mediator given the treatment and pretreatment covariates:

$$f\Big(Y_i(t, M_i(t'))\Big|\mathbf{X}_i = x\Big) = \int_M f(Y_i|T_i = t, M_i = m, \mathbf{X}_i = x)\,dF_{M_i}(m|T_i = t', \mathbf{X}_i = x)$$

This enables us to relate counterfactual quantities to the observed data and thus identify the causal effects using any parametric, semi-parametric, or nonparametric regressions for any type of mediator or outcome. Qin (2023) summarized in detail the identification results based on parametric regressions when the mediator and/or the outcome are continuous or binary. Below I focus on parametric regressions for continuous mediator and outcome.

By further adjusting for the treatment-by-mediator interaction TM and pretreatment covariates $\mathbf{X}$, which contain pretreatment confounders of the mediator–outcome relationship, as well as confounders of the treatment–mediator and treatment–outcome relationships if treatment is not randomized, we can extend Eq. (1) to

$$M = \beta_0^m + \beta_t^m T + \boldsymbol{\beta}_x^{m\prime}\mathbf{X} + \varepsilon_m,$$

$$Y = \beta_0^y + \beta_t^y T + \beta_m^y M + \beta_{tm}^y TM + \boldsymbol{\beta}_x^{y\prime}\mathbf{X} + \varepsilon_y \tag{7}$$

Assuming that the underlying data-generating process is consistent with Eq. (7) (i.e., linear functional form and constant β's across individuals) and that Assumptions 1–3 hold, we can identify $E\big[Y_i(t, M_i(t'))\big]$ as follows:

$$E\big[Y_i(t, M_i(t'))\big] = \beta_0^y + \beta_t^y t + \big(\beta_m^y + \beta_{tm}^y t\big)\big(\beta_0^m + \beta_t^m t' + \boldsymbol{\beta}_x^{m\prime}E[\mathbf{X}]\big) + \boldsymbol{\beta}_x^{y\prime}E[\mathbf{X}].$$

Correspondingly, the causal mediation effects, as defined in Table 1, can be identified as follows:

$$NIE = \big(\beta_m^y + \beta_{tm}^y t\big)\beta_t^m\big(t - t'\big),$$

$$NDE = \big(\beta_t^y + \beta_{tm}^y(\beta_0^m + \beta_t^m t' + \boldsymbol{\beta}_x^{m\prime}E[\mathbf{X}]\big)\big(t - t'\big) \tag{8}$$

where t and t' are two different treatment levels to be contrasted. Proof can be found in VanderWeele and Vansteelandt (2009).

A comparison of the indirect and direct effect estimands between the traditional and regression-based causal mediation analyses. When the mediator and outcome are both continuous, despite the difference in the indirect and direct effect estimands in the presence of the treatment-by-mediator interaction (i.e., $\beta_{tm}^y \neq 0$), the traditional mediation analysis based on Eq. (1) and the regression-based causal mediation analysis based on Eq. (7) share the same estimands of the indirect and direct effects if there is no treatment-by-mediator interaction and $t - t' = 1$ (i.e., $NIE = \beta_t^m \times \beta_m^y$ and $NDE = \beta_t^y$). The correspondence may not hold for other types of mediator and outcome or when the mediator and outcome models are semi-parametric or nonparametric.

Weighting-based identification

To illustrate the weighting-based identification, I focus on a randomized treatment, a discrete mediator, and a discrete or continuous outcome. As proved by Hong (2010), under Assumptions 1–3, the expected potential outcome $E\big[Y_i(t, M_i(t'))\big]$ can be identified through the weighted mean outcome of the treatment group t,

$$E\big[Y_i(t, M_i(t'))\big] = E\big[W_{Mi}Y_i|T_i = t\big] \tag{9}$$

When $t = t'$, $W_{Mi} = 1$, i.e., the expected potential outcome under each treatment condition can be identified through the expected outcome of the corresponding treatment group, given that the treatment is randomized. When $t \neq t'$, $W_{Mi} = \frac{Pr(M_i=m|T_i=t', \mathbf{X}_i=\mathbf{x})}{Pr(M_i=m|T_i=t, \mathbf{X}_i=\mathbf{x})}$ is known as ratio of mediator-probability weighting (RMPW). In the NEWWS example, given $t = 1$ and $t' = 0$, W_{Mi} transforms a treated individual's employment rate to resemble that under the control condition within levels of pretreatment covariates, so that the counterfactual quantity $E\big[Y_i(1, M_i(0))\big]$ can be related to the observed outcome of the treated group.

If treatment is not randomized, we can identify $E\big[Y_i(t, M_i(t'))\big]$ by further multiplying W_{Mi} with $W_{Ti} = \frac{Pr(T_i=t)}{Pr(T_i=t|\mathbf{X}_i=\mathbf{x})}$ (Hong et al., 2018), which is known as inverse probability of treatment weighting (IPTW). It removes treatment selection by equalizing everyone's treatment assignment probability, as in a randomized experiment.

If the mediator is continuous, RMPW can be constructed based on the ratio of conditional densities of M or replaced with its mathematical equivalent based on conditional probabilities of T given M and $\mathbf{X}$ (Huber, 2014).

With the expectation of each potential outcome identified, each causal effect, as defined in Table 1, can be identified via weighted mean contrasts of the observed outcome, which does not require specifying the functional form of the outcome with regard to T, M, and $\mathbf{X}$.

In summary, the regression-based identification requires that both the mediator and outcome models are correctly specified. In contrast, the weighting-based identification only requires the mediator model to be correctly specified. Therefore, the weighting-based method is robust to outcome model misspecifications but is less efficient than the regression-based method if both the mediator and outcome models are correctly specified.

Estimation and software

Despite the various identification methods, estimation of the causal mediation effects shares the same essence, i.e., estimating each expected potential outcome and taking their contrasts to estimate the effects. The estimation and inference can be conducted with common algorithms, such as Bootstrapping method (Imai et al., 2010), the Monte Carlo confidence interval method (King et al., 2000), also known as the quasi-Bayesian Monte Carlo (Imai et al., 2010), and Bayesian method (Yuan & MacKinnon, 2009). Qin and Wang (2023) introduced step-by-step implementations of the three methods for causal moderated mediation analysis, which also applies to causal mediation analysis in general.

For the regression-based methods, given the estimands as in Eq. (8), we can alternatively estimate the standard errors of the effect estimates with the delta method and test the significance of the effects by assuming the sampling distributions of the effect estimates to be approximately normal (VanderWeele & Vansteelandt, 2009). However, an effect estimate is not normal at a relatively small sample size if it involves products of regression coefficient estimates. When the normality assumption does not hold, inference based on the delta method is inappropriate in most cases (Preacher & Selig, 2012).

For the weighting-based methods, the estimation involves two steps. The first step estimates the weight based on a mediator model, and the second step estimates the causal effects through weighted mean contrasts of the observed outcome. It is crucial to account for the estimation uncertainty of the weights when estimating the standard errors of the causal mediation effect estimates. Otherwise, the inference results might be misleading (Bein et al., 2018). In addition to the three common algorithms mentioned above, Bein et al. (2018) provided an alternative solution to the two-step estimation problem by stacking the moment functions from both steps. While the former is more flexible, the latter is less computationally intensive.

Various software programs have been developed for implementing different causal mediation analysis methods.

- Some implement the regression-based methods, but the regressions can only be parametric, including M*plus* Model Indirect command (Muthén & Muthén, 2017) and those that do not allow the pretreatment covariates to interact with the treatment, mediator, or treatment-by-mediator interaction, i.e., SAS and SPSS mediation macros (Valeri & VanderWeele, 2013, 2015), SAS PROC CAUSALMED procedure (SAS Institute, 2018), Stata PARAMED macro (Emsley & Liu, 2013), and Stata Med4Way macro that is focused on four-way decomposition of the treatment effect (Discacciati et al., 2019). R regmedint package extends the SAS mediation macro and the SAS PROC CAUSALMED procedure by allowing for treatment-by-covariate and mediator-by-covariate interactions.
- Some implement the regression-based methods and allow for parametric, semi-parametric, and nonparametric models and interactions of the pretreatment covariates with the treatment, mediator, or treatment-by-mediator interaction, including R mediation package (Tingley et al., 2014) and Stata medeff macro (Hicks & Tingley, 2011).
- R rmpw package (Qin et al., 2018) implements the weighting-based method.
- R medflex package (Steen et al., 2017) implements both the weighting-based and imputation-based methods.
- The R CMAverse package (Shi et al., 2021) implements the parametric regression-based, weighting-based, imputation-based method, and several other methods. They also allow for the analysis from the interventional perspective as introduced in the section of Extensions.

A detailed review and comparison of the programs except for CMAverse can be found in Valente et al. (2020).

Sensitivity analysis and software

The above causal mediation analysis relies on the sequential ignorability assumption. The assumption would be violated when there are post-treatment confounders or omitted pretreatment confounders, which is highly likely in real applications. It is important to assess if such a potential violation would easily alter the initial causal conclusions. The majority of the existing literature of sensitivity analysis for causal mediation analysis is focused on assessing the influence of omitted pretreatment confounding, assuming no post-treatment confounding. A brief review of sensitivity analysis methods for post-treatment confounding can be found in the section of Extensions. As reviewed by Qin and Yang (2022), there are two types of omissions of pretreatment confounding. Some pretreatment confounders are observed but omitted to avoid model overfitting, while others are unmeasured. To evaluate the influence of the former, one may simply compare the results before and after including the omitted covariates in the analysis. To evaluate the influence of the latter, different sensitivity analysis methods have been developed in the past decade (e.g., Cox et al., 2013; Imai et al., 2010; Qin & Yang, 2022). The basic idea is to evaluate how the point estimates and significance of the indirect and direct effects change with sensitivity parameters that imply the strength of unmeasured pretreatment confounding. Each method has its unique advantage. However, some ignore the treatment-by-mediator interaction, some lack intuitive interpretations of the sensitivity parameters, and some ignore the influence of unmeasured confounding on the standard errors of the causal effect estimates, while others are applicable only when the treatment is randomized.

Qin and Yang (2022) overcame the limitations through a simulation-based strategy that can assess the sensitivity of results obtained from different causal mediation analysis approaches. The method (1) allows applied researchers to quantify the strength of an unmeasured pretreatment confounder based on its conditional associations with the treatment, mediator, and outcome, which serve as sensitivity parameters; (2) simulates the pretreatment confounder from its conditional distribution; and (3) assesses the influence of the pretreatment confounder on both the estimation and inference of the causal effects by comparing the results before and after adjusting for the simulated confounder in the analysis. Step (2) is repeated multiple times to account for the uncertainty of the simulation of the unmeasured pretreatment confounder. The sensitivity analysis results can be visualized to ease the evaluation of sensitivity. The original analysis results obtained under the assumption of no unmeasured pretreatment confounding would be considered sensitive if the signs or significance of the effects can be altered by a slight violation of the identification assumption, i.e., by an omitted confounder that is merely weakly associated with the treatment, mediator, and outcome. The degree to which the assumption is violated, i.e., the strength of unmeasured confounding implied by conditional associations of unmeasured confounding with the treatment, mediator, and outcome, can be gauged by prior knowledge, theoretical reasoning, or the observed pretreatment confounders in the data, as illustrated in Qin and Yang (2022).

Sensitivity analysis is an essential component in causal mediation analysis. However, it has not received enough attention in most empirical studies. Software has been developed to ease implementations. The method can be implemented in R mediationsens package (Qin & Yang, 2020). It is applicable to both regression-based and weighting-based causal mediation analysis. A review of five R packages for sensitivity analysis in mediation analysis, including mediationsens, can be found in Kawabata et al. (2023). In addition, the M*plus* Model Indirect command and R packages mediation and rmpw, as mentioned above, have sensitivity analysis embedded.

Writing a causal mediation analysis paper

Mediation analysis is becoming increasingly popular. Articles with "mediation analysis" in the title or text have been growing exponentially in the past two decades (Nguyen et al., 2021). Yet the advancement of mediation analysis methods under the counterfactual causal framework has not been widely adopted by educational researchers. To promote applications of causal mediation analysis methods, I highlight key components of a causal mediation analysis paper in this section. Detailed guidelines can be found in Lee et al. (2021) and Montoya (2023).

1. **Introduction.** It is essential to clarify the hypothesized mediation mechanism based on a literature review of theoretical rationale or supporting evidence for why the treatment affects the mediator, which subsequently affects the outcome. It is also necessary to list research questions, such as whether the mediator significantly mediates the relationship between the treatment and the outcome and how much of the treatment effect is transmitted through the mediator (if the indirect effect and the total treatment effect have the same sign).
2. **Methods.** This section includes the study design, participants, sample size, and measures, including the treatment, mediator, outcome, and a set of pretreatment confounders of the mediator–outcome relationship (and the treatment–mediator and treatment–outcome relationships if the treatment is not randomized). It is particularly important to clarify when the variables were measured. Under the principle of temporal precedence (i.e., the cause precedes the effect), the treatment, mediator, and outcome should be measured in order, rather than simultaneously, to allow the treatment to generate an impact on the mediator and subsequently influence the outcome, as illustrated in the NEWWS example. The pretreatment covariates, as the name suggests, should be measured before the treatment. If a pretreatment covariate does not vary over time, it can be measured at any time. To minimize the possibility of violating the assumption of no post-treatment confounding, one may choose the time point of measurement for the mediator to be relatively close to that for the treatment (VanderWeele & Vansteelandt, 2009, p. 462). It is worth noting that the indirect effect may vary when the mediator is measured at a different time point. For example, the mediating role of whether a mother was employed during the first year after randomization underlying the LFA impact on maternal depression at the end of the second year may differ from that of whether a mother was employed during the two-year period after

randomization. The latter can better capture the entire process, but at a higher risk of violating the no post-treatment confounding assumption, as illustrated in the section of Identification. Therefore, when stating research questions, it is crucial to clarify when the mediator is measured. In the presence of potential post-treatment confounders, possible solutions are discussed in the following section of Extensions. To better capture the time-varying property of a mediator, researchers may also consider conducting a longitudinal mediation analysis, as briefly mentioned in the section of Extensions. In addition, in this section, it is essential to describe the chosen causal mediation analysis method, rationalize the choice of the method, clarify the underlying assumptions, discuss the plausibility of the assumptions, and justify the methods used to handle missing data.

3. **Analysis results.** This section reports the estimation and inference results of the causal mediation analysis. The effect sizes reflect the practical significance, while the p-values or confidence intervals reveal the statistical significance. In terms of the effect sizes, some researchers reported mediation effect estimates in the standard deviation of the outcome in the control group (Kraft, 2020), or the standard deviation of the outcome in the whole sample (e.g., Hong et al., 2015; MacKinnon, 2008), while some fully standardized the mediation effects by standardizing all the variables (e.g., Preacher and Hayes, 2008). As Kraft (2020) argued on page 245, for a binary treatment, "it is preferable to use the standard deviation of the control group outcome rather than the pooled sample because the intervention may have affected the variation in outcomes among the treatment group. Intuitive interpretations of the effects are needed to better answer the research questions proposed in the introduction section.
4. **Sensitivity analysis.** A causal mediation analysis is incomplete without a sensitivity analysis. It is crucial to report how robust the reported analysis results are to potential unmeasured pretreatment confounding.
5. **Discussion.** This section summarizes the findings, states the implications of the results for practice, policy, and science, and discusses limitations of the analysis, such as a possible violation of the SUTVA assumption, the positivity assumption, or the assumption of no post-treatment confounding, failures to account for measurement error, and vulnerability to model misspecifications. It is also necessary to clarify how these limitations would affect the validity of the conclusions.

Extensions

Many other advancements have emerged in causal mediation analysis over the past two decades but have not received enough attention among empirical researchers. This section briefly introduces some extensions that I made with my colleagues and several other extended topics.

Power analysis for causal mediation analysis

With the rapid development of causal mediation analysis, tools for power and sample size calculations are increasingly needed for study design or grant application. However, the development of such tools has lagged far behind the development of analytic methods. Qin (2023) developed a simulation-based method and the very first easy-to-use R Shiny app (https://xuqin.shinyapps.io/CausalMediationPowerAnalysis/) for power and sample size calculations for parametric regression-based causal mediation analysis under the single-level settings. It is compatible with the widely used R mediation package. Users can either calculate the power for detecting a causal mediation effect at a given sample size or determine the sample size required for achieving a specific power. The app is applicable to a wide range of scenarios, including a randomized or nonrandomized treatment, a mediator, and an outcome that can be either binary or continuous. The article also provided sample size suggestions under a wide range of scenarios and a detailed guideline for app implementation to facilitate study designs.

Multisite causal mediation analysis

Intervention programs in education are usually delivered in various sites, such as organizations or communities. Multisite trials that randomly assign individuals to different treatment conditions within each site have been pervasive in education in the past two decades. While most evaluation research in the past has focused solely on average intervention impacts, researchers have argued that the average alone is not sufficient for informing policy and practice. Investigations of between-site variations in causal mechanisms will allow researchers to assess the generalizability or a lack thereof of findings across a wide range of contexts. Qin and colleagues (Qin & Hong, 2017; Qin et al., 2019, 2021) enabled such evaluations by developing conceptual frameworks, statistical methods, and the R package MultisiteMediation. Sampling weights and nonresponse weights are incorporated into the analysis to enhance the external and internal validity of conclusions. The analysis results include the point estimates and significance of the population average and between-site variance of each causal mediation effect. Such evidence may generate important information for understanding why the total impacts vary across different contexts and reveal a need to revisit the program or intervention theory.

Causal moderated mediation analysis

In addition to the population average and between-site variance of causal mediation mechanisms, it is also important to evaluate how causal mediation mechanisms are moderated by individual and contextual characteristics. The evidence may suggest individual and/or site-specific modifications of the intervention practice and thus improve interventions to enhance participants' outcomes across various subpopulations and contextual settings.

Different moderated mediation analysis methods have been developed under the traditional path analysis/ structural equation modeling framework. There are three major concerns about this conventional line of research. First, the definitions of moderated mediation effects depend on specific mediator and the outcome regressions and can be very complex as multiple moderated relationships are involved. Second, when the mediator or outcome model is nonlinear, moderated mediation analysis is so challenging that no solutions have been provided. Third, analysis results may not have causal interpretations due to a lack of care about confounding.

Qin and Wang (2023) developed general definitions, identification, estimation, and sensitivity analysis for causal moderated mediation effects under the counterfactual causal framework. They also developed a user-friendly R package moderate.mediation that can provide numerical summaries and intuitive visualizations for the conditional and moderated causal mediation effects and the extent to which the results are sensitive to unmeasured pretreatment confounding. The package is applicable to a wide range of scenarios, including a binary or continuous treatment that is either randomized or nonrandomized, a binary or continuous mediator, a binary or continuous outcome, and one or more moderators of any scale. The article provided a step-by-step guide on how to use the package to conduct the entire analytic procedure. The method was developed under single-level settings. An extension to multilevel settings is being developed.

Relaxing the no post-treatment confounding assumption

A big challenge of causal mediation analysis lies in relaxing the assumption of no post-treatment confounders of the mediator–outcome relationship. Because a post-treatment confounder is affected by the treatment, directly controlling for it in the mediator and/or outcome models would bias the indirect and direct effect estimates. Intuitively speaking, the original direct effect estimand would no longer indicate the treatment effect that is not transmitted through the focal mediator. Instead, it would be the treatment effect that is transmitted through neither the focal mediator nor the post-treatment confounder. The essence of the problem is that a post-treatment confounder is only partially observed. Specifically, if an individual has been assigned to the experimental condition, the individual's potential post-treatment confounder value associated with the counterfactual control condition is unobserved. In this sense, the problem with statistical adjustment for a post-treatment confounder is a problem of missing data.

Hong et al. (2023) developed an imputation-based sensitivity analysis strategy for handling post-treatment confounding and incorporated it into weighting-based causal mediation analysis. Daniel et al. (2015) provided a similar solution for regression-based causal mediation analysis. An alternative solution is to conduct a causal mediation analysis from the interventional perspective (Nguyen et al., 2021; VanderWeele et al., 2014), which allows adjustment for post-treatment confounders. Park and colleagues (Park et al., 2022, 2023) developed sensitivity analysis strategies for such analysis to assess the influence of unmeasured confounding.

Other extensions

Beyond the above extensions, researchers have made many other important extensions to enrich the literature on causal mediation analysis. For example, researchers have developed methods for causal mediation analysis with multiple mediators (e.g., Daniel et al., 2015; Imai & Yamamoto, 2013) and causal mediation analysis that accounts for noncompliance with treatment randomization (e.g., Keele et al., 2015; Park & Kurum, 2020). Another emerging domain is longitudinal causal mediation analysis with time-varying treatment, mediator, and outcome (e.g., VanderWeele & Tchetgen Tchetgen, 2017). Because mediation consists of causal processes that unfold over time, longitudinal causal mediation analysis enables researchers to make more rigorous inferences about causal relations.

Summary

The literature on causal mediation analysis has been growing rapidly in the past two decades. This article aims at guiding empirical researchers who have little knowledge about this field through concepts and challenges of mediation analysis under the counterfactual causal framework. After reading the articles, readers are expected to be able to tell the difference between causal mediation analysis and traditional mediation analysis, understand the definitions of potential outcomes and the causal mediation effects, realize the importance of controlling for pretreatment confounders in the analysis and conducting sensitivity analysis to assess the influence of unmeasured confounding, know the difference among different causal mediation analysis methods and what software to use for the analysis, and learn how to write a causal mediation analysis paper. For more in-depth investigations of causal mediation mechanisms, readers may further read the references cited in the Extensions section and read two books written by leading scholars in causal mediation analysis, Hong (2015) and VanderWeele (2015).

Acknowledgements I thank the co-editors of this special issue, Drs. Peter Steiner and Yongnam Kim, for inviting me to write this introductory article on causal mediation analysis. I appreciate the detailed comments from them and an external reviewer, which have led to major improvements in this article. In addition, I thank Dr. Guanglei Hong for her insightful suggestions on the revision of the manuscript.

Declarations

Conflict of interest The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval Because this is a review paper with no identifying information, it was exempted from Institutional Review Board approval.

Consent for publication Some sentences in the Introduction section were published in the author's dissertation. Dissertation authors hold the copyright for their own dissertations and retain the right to publish elsewhere.

References

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review, 40*, 37–47.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173.

Bein, E., Deutsch, J., Hong, G., Porter, K., Qin, X., & Yang, C. (2018). Two-step estimation in rmpw analysis. *Statistics in Medicine, 37*(8), 1304–1324.

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology, 17*, 37–69.

Cox, M. G., Kisbu-Sakarya, Y., Miočević, M., & MacKinnon, D. P. (2013). Sensitivity plots for confounder bias in the single mediator model. *Evaluation Review, 37*(5), 405–431.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics, 71*(1), 1–14.

Ding, P., & Li, F. (2018). Causal inference. *Statistical Science, 33*(2), 214–237.

Discacciati, A., Bellavia, A., Lee, J. J., Mazumdar, M., & Valeri, L. (2019). Med4way: A Stata command to investigate mediating and interactive mechanisms using the four-way effect decomposition. *International Journal of Epidemiology, 48*, 15–20.

Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology, 72*, 1–16.

Emsley, R., & Liu, H. (2013). PARAMED: Stata module to perform causal mediation analysis using parametric regression models.

Hafeman, D. M., & Schwartz, S. (2009). Opening the Black Box: A motivation for the assessment of mediation. *International Journal of Epidemiology, 38*(3), 838–845.

Hicks, R., & Tingley, D. (2011). Causal mediation analysis. *The Stata Journal, 11*(4), 605–619.

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

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology, 18*, 449–484.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proceedings of the American Statistical Association, biometrics section* (pp. 2401–2415). American Statistical Association.

Hong, G. (2015). *Causality in a social world: Moderation, mediation and spill-over*. Wiley.

Hong, G. (2017). A review of "Explanation in causal inference: Methods of mediation and interaction." *Journal of Educational and Behavioral Statistics, 42*(4), 491–495.

Hong, G., Deutsch, J., & Hill, H. D. (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics, 40*, 307–340.

Hong, G., Qin, X., & Yang, F. (2018). Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics, 43*(1), 32–56.

Hong, G., Yang, F., & Qin, X. (2023). Post-treatment confounding in causal mediation studies: A cutting-edge problem and a novel solution via sensitivity analysis. *Biometrics, 79*, 1042.

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics, 29*, 920–943.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*, 309.

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis, 21*, 141–171.

Institute, S. A. S. (2018). *User's guide the CAUSALMED procedure*. SAS Institute Inc.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods, 13*(4), 314–336.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*, 239–251.

Judd, C. M., & Kenny, D. A. (1981). Process analysis estimating mediation in treatment evaluations. *Evaluation Review, 5*, 602–619.

Kawabata, E., Tilling, K., Groenwold, R. H., & Hughes, R. A. (2023). Quantitative bias analysis in practice: Review of software for regression with unmeasured confounding. *BMC Medical Research Methodology, 23*(1), 1–13.

Keele, L., Tingley, D., & Yamamoto, T. (2015). Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management, 34*(4), 937–963.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science, 44*(2), 347–361.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

Lee, H., Cashin, A. G., Lamb, S. E., Hopewell, S., Vansteelandt, S., VanderWeele, T. J., MacKinnon, D. P., Mansell, G., Collins, G. S., Golub, R. M., McAuley, J. H., Localio, A. R., van Amelsvoort, L., Guallar, E., Rijnhart, J., Goldsmith, K., Fairchild, A. J., Lewis, C. C., & McAuley, J. H. (2021). A guideline for reporting mediation analyses of randomized trials and observational studies: The agrema statement. *JAMA, 326*, 1045–1056.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Erlbaum.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review, 17*, 144–158.

MacKinnon, D. P., Valente, M. J., & Gonzalez, O. (2020). The correspondence between causal and traditional mediation analysis: The link is the mediator by treatment interaction. *Prevention Science, 21*, 147–157.

Montoya. (2023). Combining statistical and causal mediation analysis. In H. Reis, T. West, & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (3rd ed.). Cambridge University Press.

Morris, P. A. (2008). Welfare program implementation and parents' depression. *Social Service Review, 82*(4), 579–614.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide*. Muthén & Muthén.

Neyman, J., & Iwaszkiewicz, K. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society, 2*, 107–180.

Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods, 26*(2), 255.

Park, S., Kang, S., Lee, C., & Ma, S. (2023). Sensitivity analysis for causal decomposition analysis: Assessing robustness toward omitted variable bias. *Journal of Causal Inference, 11*(1), 20220031.

Park, S., & Kürüm, E. (2020). A two-stage joint modeling method for causal mediation analysis in the presence of treatment noncompliance. *Journal of Causal Inference, 8*(1), 131–149.

Park, S., Qin, X., & Lee, C. (2022). Estimation and sensitivity analysis for causal decomposition analysis in disparity research. *Sociological Methods & Research*. https://doi.org/10.1177/00491241211067516

Pearl, J. (2001). Direct and indirect effects. In J. Breese & D. Koller (Eds.), *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420). Morgan Kaufmann.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & Van Der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research, 21*(1), 31–54.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, *40*(3), 879–891.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*(1), 185–227.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*(2), 77–98.

Qin, X. (2023). Sample size and power calculations for causal mediation analysis. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02118-0

Qin, X., & Hong, G. (2017). A weighting method for assessing between-site heterogeneity in causal mediation mechanism. *Journal of Educational and Behavioral Statistics, 42*(3), 308–340.

Qin, X., & Wang, L. (2023). Causal moderated mediation analysis: methods and software. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02095-4

Qin, X. & Yang, F. (2020). Mediationsens: Simulation-based sensitivity analysis for causal mediation studies. R package version 0.0.2. https://CRAN.R-project.org/package=mediationsens

Qin, X., & Yang, F. (2022). Simulation-based sensitivity analysis for causal mediation studies. *Psychological Methods, 27*(6), 1000–1013.

Qin, X, Hong, G., & Yang, F. (2018). rmpw: Causal mediation analysis using weighting approach. R package version 0.0.4. https://CRAN.R-project.org/package=rmpw

Qin, X., Deutsch, J., & Hong, G. (2021). Unpacking complex mediation mechanisms and their heterogeneity between sites in a Job Corps evaluation. *Journal of Policy Analysis and Management, 40*(1), 158–190.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. J. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 70–81). Oxford University Press.

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*, 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association, 75*, 591–593.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961–962.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference, 25*, 279–292.

Shi, B., Choirat, C., Coull, B. A., VanderWeele, T. J., & Valeri, L. (2021). CMAverse: a suite of functions for reproducible causal mediation analyses. *Epidemiology*, *32*(5), e20–e22.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–312). Jossey-Bass.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics, 33*(2), 230–251.

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017). Medflex: An R package for flexible mediation analysis using natural effect models. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v076.i11

Tchetgen Tchetgen, E. J., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics, 40*, 1816.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v059.i05

Valente, M. J., Rijnhart, J. J., Smyth, H. L., Muniz, F. B., & MacKinnon, D. P. (2020). Causal mediation programs in R, M plus, SAS, SPSS, and Stata. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(6), 975–984.

Valeri, L., & Vanderweele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods, 18*, 137–150.

Valeri, L., & VanderWeele, T. J. (2015). SAS macro for causal mediation analysis with survival data. *Epidemiology, 26*, E23–E24.

VanderWeele, T. J. (2014). A unification of mediation and interaction: a 4-way decomposition. *Epidemiology, 25*, 749–761.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J., & Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society Series b: Statistical Methodology, 79*(3), 917–938.

VanderWeele, T. J., & Vansteelandt. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface, 2*, 457–468.

VanderWeele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology (cambridge, Mass), 25*(2), 300.

Vansteelandt, S., Bekaert, M., & Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods, 1*(1), 131–158.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics, 5*(3), 161–215.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14*(4), 301.