UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

WORKING PAPERS SES

A review of causal mediation analysis for assessing direct and indirect treatment effects

Martin Huber

N. 500
I.2019

Faculté des sciences économiques et sociales
Wirtschafts- und sozialwissenschaftliche Fakultät

A review of causal mediation analysis for assessing direct and indirect treatment effects

Martin Huber

University of Fribourg, Dept. of Economics

Abstract: Mediation analysis aims at evaluating the causal mechanisms through which a treatment or intervention affects an outcome of interest. The goal is to disentangle the total treatment effect into an indirect effect operating through one or several observed intermediate variables, the so-called mediators, as well as a direct effect reflecting any impact not captured by the observed mediator(s). This paper reviews methodological advancements with a particular focus on applications in economics. It defines the parameters of interest, covers various identification strategies, e.g. based on control variables or instruments, and presents sensitivity checks. Furthermore, it discusses several extensions of the standard mediation framework, such as multivalued treatments, mismeasured mediators, and outcome attrition.

Keywords: mediation, direct effect, indirect effect, sequential conditional independence, instrument.

JEL classification: C21.

Address for correspondence: Martin Huber, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland; martin.huber@unifr.ch.

1 Introduction

The majority of studies on treatment, policy, or impact evaluation confine themselves to assessing the total effect of a treatment on an outcome of interest, such as the average treatment effect (ATE). At the same time, however, a range of studies makes conjectures about potential mechanisms that may underly soem total effect. It is thus not only the 'effect of a cause', i.e. the treatment effect, that seems policy relevant in many problems, but also the 'cause of the effect', i.e. the mechanisms through which the total effect materializes, see Gelman and Imbens (2013). As an example, consider the employment or earnings effect of an active labor market program, such as a training. On top of the overall effect, policy makers might want to know to which extent the program's impact stems from a change in search effort, human capital increase, or other intermediate variables that are themselves influenced by training participation. In fact, a better understanding of the mechanisms that drive the effect may help improving the design of such programs.

Causal mediation analysis aims at disentangling a total treatment effect into an indirect effect operating through one or several intermediate variables – commonly referred to as mediators – as well as the direct effect, which includes any causal mechanisms not operating through the mediators of interest. Even under random treatment assignment, direct and indirect effects are generally not identified by bluntly controlling for mediators without accounting for their possible endogeneity, as this likely introduces selection bias, see Robins and Greenland (1992). Much of the earlier work on mediation analysis (see for instance the seminal papers of Cochran (1957), Judd and Kenny (1981), and Baron and Kenny (1986)) typically relied on linear models for the mediator and outcome equations and often neglected endogeneity issues.

An example for how careless conditioning on a mediator might flaw identification is the evaluation of the effect of mother's smoking behavior during pregnancy on post-natal infant mortality, see Wilcox (2001) and Hernandez-Diaz, Schisterman, and Hernan (2006). In general, the empirical literature finds a positive relationship between smoking and infant mortality. However, several studies point out that among those children with the lowest birth weight, smoking appears to decrease mortality. As acknowledged by Hernandez-Diaz, Schisterman, and Hernan (2006), this paradox is most likely due to not controlling for (important) confounders when conditioning on low birth weight as mediator. In fact, if smoking is a less lethal cause of having a low birth weight than other reasons like birth defects, then the mortality of children with a low birth weight due to birth defects is higher than among those whose mothers smoked during pregnancy.

More recent research in mediation analysis considers more general identification approaches based on the potential outcome framework commonly used in treatment evaluation and aims at tackling confounding. Examples include Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), VanderWeele (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Tchetgen Tchetgen and Shpitser (2012), and Vansteelandt, Bekaert, and Lange (2012). The vast majority of the literature obtains identification by assuming that the treatment and the mediator are conditionally exogenous given observed characteristics.

Such or related assumptions have also been used in empirical economic research. See for instance Simonsen and Skipper (2006), who evaluate the direct wage effect of motherhood, and Flores and Flores-Lagunes (2009), who evaluate the direct earnings effect of Job Corps when controlling for work experience as mediator. Heckman, Pinto, and Savelyev (2013) and Keele, Tingley, and Yamamoto (2015) investigate cognitive and non-cognitive mechanisms of the Perry Preschool Program. Conti, Heckman, and Pinto (2016) assess the effect of the Perry Preschool Program and Carolina Abecedarian Project on health and healthy behaviour mediated by personality traits. Bijwaard and Jones (2018) evaluate the effect of education on mortality, considering cognitive ability as mediator. Bellani and Bia (2018) examine education as mediator through which growing up poor affects economic outcomes in adulthood in the EU. Huber (2015) applies the causal mediation framework to the context of wage gap decompositions using data from the U.S. National Longitudinal Survey of Youth 1979. Huber, Lechner, and Mellace (2017) investigate whether the employment effect of more rigorous caseworkers in the counselling process of job seekers in Switzerland is mediated by placement into labor market programs. Huber, Lechner, and Strittmatter (2018) evaluate the employment effect of awarding vouchers for vocational training to unemployed individuals in Germany, and whether there exists a direct effect net of actual redemption, which may for instance be driven by preference shaping or learning about available programs.

For studies using instrumental variables for identification, see for instance Powdthavee, Lekfuangfu, and Wooden (2013), who estimate the indirect effect of education on life satisfaction running through the mediator income. Brunello, Fort, Schneeweis, and Winter-Ebmer (2016) investigate the effect of education on health mediated by health behaviors. Chen, Chen, and Liu (2017) assess the effect of family composition on the educational attainment of the first-born child.

This paper reviews the methodological advancements in the causal mediation literature. Section 2 defines the parameters of interest: (natural) direct and indirect effects, the controlled direct effect, and principal strata-specific effects. Section 3 discusses identification and estimation under sequential conditional exogeneity of the treatment and mediator given observed characteristics. It distinguishes between the case that the same set of control variables can be used to satisfy treatment and mediator exogeneity and the more challenging case of dynamic confounding, where some control variables for the mediator are functions of the treatment. Section 4 provides further evaluation strategies based on partial identification, randomization of the treatment and mediator, instrumental variables for the treatment and/or mediator, and difference-in-differences. Section 5 discusses several extensions to the standard framework: multivalued rather than binary treatments, target populations different to the total population, mismeasured mediators, and sample selection/outcome attrition. Section 6 concludes.

2 Parameters of interest

This section introduces various effects that have been considered in causal mediation analysis: Natural direct and indirect effects, controlled direct effects, and principal strata-specific effects.

2.1 Natural direct and indirect effect

Mediation analysis typically aims at decomposing the average treatment effect (ATE) of a binary treatment indicator, which we denote by D, on an outcome variable, Y, into a direct effect and an indirect effect operating through a mediator, M. The latter is assumed to have bounded support and may be discrete or continuous, scalar or a vector of variables. To define natural direct and indirect effects, we use the potential outcome framework, see for instance Rubin (1974), which has been considered in the mediation framework for instance by Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007) and Albert (2008). Let $Y(d)$ and $M(d)$ denote the potential outcome and the potential mediator state, respectively, under treatment value $d \in \{0,1\}$.[1] For each unit only one of the two potential outcomes or mediator states is observed, because the realized outcomes and mediators are $Y = D \cdot Y(1) + (1-D) \cdot Y(0)$ and $M = D \cdot M(1) + (1-D) \cdot M(0)$.

The ATE is given by $\Delta = E[Y(1) - Y(0)]$. For decomposing this total effect into a direct and

[1] In general, we will use capital letters for random variables and small letters for specific values of random variables in the subsequent discussion.

indirect impact (through M), we rewrite the potential outcome as a function of both the treatment and the potential mediator: $Y(d) = Y(d, M(d))$. This allows writing the (average) direct effect as

$$\theta(d) \quad = \quad E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}. \tag{1}$$

$\theta(d)$ corresponds to the change in mean potential outcomes when exogenously varying the treatment but keeping the mediator fixed at its potential value for $D = d$, which shuts down causal mechanisms via M. Similarly, the (average) indirect effects is given by

$$\delta(d) \quad = \quad E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}. \tag{2}$$

$\delta(d)$ corresponds to the change in mean potential outcomes when exogenously shifting the mediator to its potential values under treatment and non-treatment but keeping the treatment fixed at $D = d$ to shut down the direct effect. Robins and Greenland (1992) and Robins (2003) referred to these parameters as pure/total direct and indirect effects and Pearl (2001) as natural direct and indirect effects, which is the denomination used in this paper.

The ATE is the sum of the natural direct and indirect effects defined upon opposite treatment states:

$$\begin{aligned}
\Delta \quad &= \quad E[Y(1, M(1)) - Y(0, M(0))] \\
&= \quad E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0) \\
&= \quad E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1). \qquad (3)
\end{aligned}$$

This follows from adding and subtracting $E[Y(0, M(1))]$ after the first and $E[Y(1, M(0))]$ after the third equality in (3). Furthermore, the notation $\theta(1), \theta(0)$ and $\delta(1), \delta(0)$ points to potential effect heterogeneity with respect to the treatment state, i.e., the presence of interaction effects between the treatment and the mediator. For instance, the effectiveness of job search activities (M) for finding employment (Y) might depend on whether a job seeker has obtained a job application training (D). Or put differently, the direct effect of the training might depend on the level of job search activity (M).

Obviously, none of the effects are obtained without imposing some kind of identifying assumptions. First, only one of $Y(1, M(1))$ and $Y(0, M(0))$ is observed for any unit (i.e. both potential

outcomes cannot be observed at the same time), which is known as fundamental problem of causal inference. Second, $Y(1, M(0))$ and $Y(0, M(1))$ are never observed for any individual, as mediator and outcome values can only be observed for the same, factual treatment, rather than for opposite treatment states. Therefore, identification of direct and indirect effects hinges on exploiting exogenous variation in the treatment and the mediator.

It appears instructive to discuss the effects of interest and identification issues in the context of a simple structural model consisting of a system of linear equations for the outcome and a scalar mediator:

$$Y = \beta_D D + \beta_M M + U, \tag{4}$$

$$M = \alpha_D D + V. \tag{5}$$

β_D, β_M denote the coefficients on D and M in the outcome equation, α_D is the coefficient on D in the mediator equation, and U and V are unobserved terms. Exogenously switching on and off the treatment in the mediator equation identifies the potential mediators:

$$M(1) = \alpha_D + V, \quad M(0) = V.$$

Switching on and off the treatment in the outcome equation and plugging in the potential mediators yields all in all four potential outcomes:

$$\begin{aligned} Y(1, M(1)) &= \beta_D + \beta_M M(1) + U, \quad Y(0, M(0)) = \beta_M M(0) + U, \\ Y(1, M(0)) &= \beta_D + \beta_M M(0) + U, \quad Y(0, M(1)) = \beta_M M(1) + U. \end{aligned}$$

By subtracting potential outcomes appropriately, it follows from our simple model without treatment-mediator interaction effects that direct effects are homogenous and equal to the coefficient on D in the outcome equation: $\theta(1) = \theta(0) = \beta_1$. Secondly, the indirect effect corresponds to the effect of D on M times the effect of M on Y: $\delta(1) = \delta(0) = \beta_2 \cdot \alpha_1$.

Estimating equations (2.1) and (5) by OLS to compute the effects of interest is likely inconsistent in most empirical problems, as V may be correlated with the treatment and U with both the treatment and the mediator. Furthermore, the model imposes strong functional assumptions: Linearity in parameters and no interaction effects, neither between the mediator and the treat-

ment, nor between observed variables and the unobservables. This rules out any form of effect heterogeneity. To (at least) permit for heterogeneity of $\theta(d)$ and $\delta(d)$ with respect to d , equation can be augmented by mediator-treatment interactions:

$$Y = \beta_D D + \beta_M M + \beta_{DM} DM + U,$$

with β_3 being the coefficient on the interaction term. This implies the following potential outcomes:

$$\begin{aligned} Y(1, M(1)) &= \beta_D + \beta_M M(1) + \beta_{DM} M(1) + U, \quad Y(0, M(0)) = \beta_M M(0) + U, \\ Y(1, M(0)) &= \beta_D + \beta_M M(0) + \beta_{DM} M(0) + U, \quad Y(0, M(1)) = \beta_M M(1) + U. \end{aligned}$$

Also this model is, however, quite rigid as it still imposes additivity between the observed and unobserved terms, implying that total, direct, and indirect treatment effects are constant across individual characteristics.

A more general model is given by the nonparametric structural model

$$Y = \varphi(D, M, U), \quad M = \zeta(D, V), \tag{6}$$

where φ, ζ denote general functions such that arbitrary nonlinearities and interactions between variables are permitted. The potential mediators and outcomes are given by

$$M(d) = \zeta(d, V), \quad Y(d, M(d')) = \varphi(d, M(d'), U),$$

for $d, d' \in \{1, 0\}$. Obviously, natural direct and indirect effects cannot be conveniently represented by (a combination of) coefficients as in the simple linear model considered before. On the other hand, the nonparametric model is more flexible and thus more robust to misspecification. However, identification is not straightforward if unobservables (U, V) are not statistically independent of (D, M). Different strategies to tackle such treatment and mediator endogeneity are discussed further below.

2.2 Controlled direct effect

A further parameter considered in the mediation literature is the so-called controlled direct effect. It corresponds to the difference in mean potential outcomes when exogenously varying the

treatment and setting the mediator to a specific value (say m) for everyone:

$$\gamma(m) = E[Y(1,m) - Y(0,m)], \qquad \text{for } m \text{ in the support of } M. \tag{7}$$

For the general structural model (6), the potential outcome is given by $Y(d,m) = \varphi(d,m,U)$. In contrast to the natural direct effect, which is conditional on the mediator value that would 'naturally' occur under a particular treatment state (and may be different for different individuals), the controlled direct effect is based on enforcing the same mediator value for all individuals. The two parameters are only equivalent if the effects of D and M on the outcome do not interact, see the linear outcome model (2.1).

Whether the natural or controlled direct effect is of primary interest depends on the empirical problem. Suppose one aims at assessing the effectiveness of the first program in a sequence of two labor market programs (e.g., a job application training followed by a computer course) with respect to finding employment. The natural direct effect evaluates the first program (D) conditional on the value of the second one (M) that would in the current institutional context follow from (non-)participation in the first program. This is suitable for assessing the first program under status quo assignment rules for the second program. However, if such rules can be manipulated, evaluating whether the first program is effective conditional on enforcing (non-)participation in the second one appears interesting, too, for appropriately designing program sequences. Therefore, the controlled direct effect may provide policy guidance if the mediator can be prescribed, while the natural direct effect, which is defined on status quo mediator response to (non-)treatment, appears more pertinent if prescription is infeasible. See Pearl (2001) for more discussion of the 'descriptive' and 'prescriptive' natures of natural and controlled effects.

It is worth noting that the controlled, but not the natural direct effect fits the dynamic treatment effects framework for assessing sequences of prescribed treatments, see for instance Robins (1986), Robins, Hernan, and Brumback (2000), and Lechner (2009). In fact, the controlled direct effect relies on comparing sequences with different values in the first treatment, but with same values in second treatment. It can therefore be regarded as a special case of dynamic treatment evaluation. Finally, note that there is no indirect effect-pendant to the controlled direct effect. Specifically, the difference between the total effect and the controlled direct effect does in general not correspond to the indirect effect, unless there are no treatment-mediator interactions, see for instance the discussion in Kaufman, MacLehose, and Kaufman (2004).

2.3 Principal strata effects

Most mediation studies evaluate direct and indirect effects for the total population. A smaller strand of the literature uses the principal stratification framework of Frangakis and Rubin (2002) to assess effects in subpopulations defined upon potential mediator states under treatment and non-treatment, see for instance Rubin (2004). Assuming a binary mediator, the total population can be partitioned into four principal strata according to the individuals' potential mediator states. Among always mediated, the potential mediator is always one irrespective of the treatment: $M(1) = M(0) = 1$. Among never mediated, the potential mediator is always zero: $M(1) = M(0) = 0$. Within either group, the total treatment effect coincides with the direct effect, because the mediator is unaffected by the treatment such that the indirect effect is zero by the definition of the principal strata:

$$E[Y(1, M(1)) - Y(0, M(0))|M(1) = M(0) = 1] = E[Y(1, 1) - Y(0, 1)|M(1) = M(0) = 1],$$
$$E[Y(1, M(1)) - Y(0, M(0))|M(1) = M(0) = 0] = E[Y(1, 0) - Y(0, 0)|M(1) = M(0) = 0]. \quad (8)$$

For the remaining principal strata, potential mediators are non-constant, but react on the treatment. Among so-called mediator compliers, the mediator complies with the treatment in the sense that it equals the treatment: $M(1) = 1, M(0) = 0$. Among mediator defiers, however, the mediator opposes the treatment value: $M(1) = 0, M(0) = 1$.[2] Due to the variation of the potential mediator across treatment states, indirect effects cannot be ruled out a priori. For this reason, the total effect on mediator compliers or defiers does generally neither coincide with direct nor indirect effects. Principal stratification has therefore been criticized for typically not decomposing direct and indirect effects among groups with non-constant potential mediators and for focussing on subgroups rather than the total population, see VanderWeele (2008) and VanderWeele (2012a). The policy relevance of specific subpopulations like the always and never mediated should therefore be scrutinized in the empirical context at hand.

[2]The notation of compliers and defiers has been inspired by the local average treatment effect framework, see Angrist, Imbens, and Rubin (1996), which fits the principal stratification context, too.

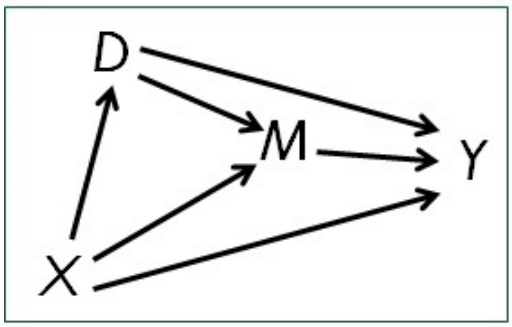
3 Mediation based on sequential conditional independence

This section considers the evaluation of direct and indirect effects based on sequential conditional independence of the treatment and mediator. It first discusses assumptions under which the same set of observed confounders is sufficient for tackling both treatment and mediator endogeneity and presents various approaches for identification and estimation. It then considers the case that some observed confounders of the mediator are themselves functions of the treatment, which is more challenging in terms of identification.

3.1 Assumptions with covariates not affected by treatment

A large part of the mediation literature invokes sequential conditional independence assumptions with respect to the treatment and the mediator for identification. We subsequently consider assumptions under which the same set of observed covariates, denoted by X, can be used to tackle both treatment and mediator endogeneity. Specifically, X must not be a function of D (notationally: $X(d) = X$), with the leading case being pre-treatment covariates evaluated prior to treatment assignment. Figure 1 illustrates the framework based on a directed acyclic graph, in which the arrows represent causal effects. It is worth noting that each of D, M, and Y might be causally affected by distinct and statistically independent sets of unobservables not displayed in Figure 1. However, what needs to be ruled out is that such unobservables jointly affect two or all three variables in (D, M, Y).

Figure 1: Causal paths under conditional exogeneity given pre-treatment covariates



The first assumption requires the treatment to be conditionally independent of any potential post-treatment variables, namely the potential mediators and outcomes, given X. This assumption is known as conditional independence, selection on observables, or exogeneity in the treatment evaluation literature, see for instance Imbens (2004).

Assumption 1 (conditional independence of the treatment):

$\{Y(d', m), M(d)\} \perp D|X$ for all $d', d \in \{0, 1\}$ and m in the support of M.

'$\perp$' denotes statistical independence. By Assumption 1, there are no unobserved confounders jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand conditional on X. In non-experimental data, the plausibility of this assumption critically hinges on the richness of X. In experimental data, the assumption holds if the treatment is either randomized within strata defined on X or randomized unconditionally, i.e., independent of X, such that even the stronger assumption $\{Y(d', m), M(d), X\} \perp D$ is satisfied.

The second assumption requires the mediator to be conditionally independent of the potential outcomes given the treatment and the covariates:

Assumption 2 (conditional independence of the mediator):

$Y(d', m) \perp M|D = d, X = x$ for all $d', d \in \{0, 1\}$ and m, x in the support of M, X.

By Assumption 2, there are no unobserved confounders jointly affecting the mediator and the outcome conditional on D and X. This generally rules out post-treatment confounders of the mediator-outcome relation not captured by X. The strength of this assumption cannot be overstated, in particular if the time window between the measurement of the treatment and the mediator is large, which makes the absence of post-treatment confounding less plausible in a world of time-varying variables.

The third assumption imposes common support on the conditional treatment probability across treatment states. **Assumption 3 (common support):**

$\Pr(D = d|M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and m, x in the support of M, X.

By Assumption 3, the conditional probability to receive or not receive the treatment given M, X, henceforth referred to as propensity score, is larger than zero. It implies (but is stronger than) the standard common support assumption in treatment evaluation that $\Pr(D = d|X = x) > 0$. That is, the treatment must not be a deterministic function of X, otherwise identification is infeasible due to a lack of comparable units in terms of X across treatment states. By Bayes' theorem, Assumption 3 also implies that $\Pr(M = m|D = d, X = x) > 0$ if M is discrete or that the conditional density of M given D, X is larger than zero if M is continuous. Conditional on X, the mediator state must not be a deterministic function of the treatment, otherwise no comparable units in terms of the mediator are available across treatment states. Assumptions 1 to 3 have been frequently imposed in causal mediation analyses, see for instance Imai, Keele, and

Yamamoto (2010), Tchetgen Tchetgen and Shpitser (2012), and Huber (2014). For closely related assumptions, see Pearl (2001), Petersen, Sinisi, and van der Laan (2006), and Hong (2010).

3.2 Identification with covariates not affected by treatment

Following Baron and Kenny (1986), many earlier mediation studies have evaluated direct and indirect effects based on a system of linear equations. When allowing for observed confounders X as in Assumptions 1 and 2, this amounts to assuming the following model:

$$Y = \beta_D D + \beta_M M + X'\beta_X + U, \tag{9}$$

$$M = \alpha_D D + X'\beta_X + V, \tag{10}$$

such that the conditional expectations of the outcome and the mediator are given by

$$E[Y|D, M, X] = \beta_0 + \beta_D D + \beta_M M + X'\beta_X, \tag{11}$$

$$E[M|D, X] = \alpha_0 + \alpha_D D + X'\alpha_X. \tag{12}$$

$\beta_0, \beta_D, \beta_M, \beta_X$ denote the constant (i.e. $E(U)$) and the coefficients on D, M, X in the outcome equation and $\alpha_0, \alpha_D, \alpha_X$ the constant (i.e. $E(V)$) and the coefficients on D, X in the mediator equation. As discussed in Section 2.1, natural direct and indirect effects are identified by β_D and $\alpha_D \cdot \beta_M$, respectively. This rather simplistic model, which does not allow for interactions of D and M or X, D, and M, could be made more flexible by including such interaction terms, see the discussion in Imai, Keele, and Yamamoto (2010) and Imai, Keele, and Tingley (2010). We also note that due to the linearity restrictions, Assumptions 1 and 2 could be relaxed to mean independence, while Assumption 3 is not required at all.

Nonparametric identification does not rely on functional form restrictions such as linearity, but instead requires Assumptions 1 to 3. We subsequently show identification of the mean potential outcome $E[Y(d, M(d'))]$ for $d, d' \in \{1, 0\}$ by parameters observed in the population, implying that natural direct and indirect effects are identified, too. To this end we introduce some notation. Let $f_{A=a}$ and $f_{A=a|B=b}$ denote the probability density functions of some random variable A, either unconditionally or conditional some random variable(s) $B = b$. We assume that M and X are continuously distributed. If M and/or X are discrete, then the respective densities and integrals below are to be replaced by probabilities and sums, so imposing continuity is only for ease of

exposition without substantial importance.

$$
\begin{aligned}
& E[Y(d, M(d'))] \\
= & \int\int E[Y(d,m)|M(d')=m, X=x] f_{M(d')=m|X=x} dm f_{X=x} dx \\
= & \int\int E[Y|D=d, M=m, X=x] f_{M=m|D=d',X=x} f_{X=x} dm dx \qquad (13) \\
= & \int\int E[Y|D=d, M=m, X=x] \cdot \frac{\Pr(D=d'|M=m, X=x)}{\Pr(D=d'|X=x)} f_{M=m|X=x} dm f_{X=x} dx \\
= & E\left[E\left[E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)}\middle| M, X\right] \cdot \frac{\Pr(D=d'|M,X)}{\Pr(D=d'|X)}\middle| X\right]\right] \\
= & E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \frac{\Pr(D=d'|M,X)}{\Pr(D=d'|X)}\right], \\
= & E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \frac{\Pr(D=d'|M,X)}{\Pr(D=d'|X)}\right], \qquad (14) \\
= & E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|X)} \cdot \frac{f(M=m|D=d',X)}{f(M=m|D=d,X)}\right]. \qquad (15)
\end{aligned}
$$

The first equality follows from the law of iterated expectations and from replacing the outer expectations by integrals, the second from Assumptions 1 and 2, the third from Bayes' theorem, the fourth from basic probability theory and from replacing the integrals by expectations, the fifth and sixth from the law of iterated expectations. We also see that Assumption 3 is required for guaranteeing that no expression goes to infinity due to division by zero.

(13) underlies the so-called mediation formula for identifying direct and indirect effects, see for instance equations (8) and (26) in Pearl (2001) and Theorem 1 in Imai, Keele, and Yamamoto (2010):

$$
\begin{aligned}
\theta(d) &= \int\int \{E[Y|D=1, M=m, X=x] - E[Y|D=0, M=m, X=x]\} f_{M=m|D=d,X=x} dm f_{X=x} dx, \\
\delta(d) &= \int\int E[Y|D=d, M=m, X=x] \{f_{M=m|D=d,X=x} - f_{M=m|D=0,X=x}\} dm f_{X=x} dx. \qquad (16)
\end{aligned}
$$

(14) is the base for identification using inverse probability weighting by the treatment propensity score, see Huber (2014):

$$
\begin{aligned}
\theta(d) &= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d|X)}\right], \\
\delta(d) &= E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right]. \qquad (17)
\end{aligned}
$$

Finally, (15) motivates inverse mediator density weighting, see Hong (2010) and Tchetgen Tchetgen

and Shpitser (2012):

$$\begin{aligned}\theta(d) &= E\left[\left(\frac{Y\cdot D}{\Pr(D=1|X)}-\frac{Y\cdot(1-D)}{1-\Pr(D=1|X)}\right)\cdot\frac{f_{M=m|D=d,X}}{f_{M=m|D,X}}\right],\\ \delta(d) &= E\left[\frac{Y\cdot I\{D=d\}}{f_{M=m|D=d,X}\cdot\Pr(D=d|X)}\cdot\left(f_{M=m|D=1,X}-f_{M=m|D=0,X}\right)\right].\end{aligned} \tag{18}$$

As in the conventional treatment evaluation framework, the parameters of interest can be identified by various strategies relying on structural models for the outcome and the mediator, or on reweighting outcomes based on conditional probabilities/densities, or on combinations thereof.

The controlled direct effect is obtained under somewhat weaker restrictions than natural effects, as the distribution of potential mediators need not be identified. Therefore, Assumption 1 can be relaxed because conditional independence between $M(d)$ and D is not required, see for instance the discussion in Petersen, Sinisi, and van der Laan (2006). Assuming that M is discrete, the potential outcome $E[Y(d,m)]$ is is identified in the following way:

$$\begin{aligned}E[Y(d,m)] &= E[E[Y(d,m)|X]] = E[E[Y|D=d,M=m,X]]\\ &= E\left[E\left[\frac{Y\cdot I\{D=d\}\cdot I\{M=m\}}{\Pr(D=d,M=m|X)}\middle|X\right]\right]=E\left[\frac{Y\cdot I\{D=d\}\cdot I\{M=m\}}{\Pr(D=d,M=m|X)}\right]\end{aligned} \tag{19}$$

The first equality follows from the law of iterated expectations, the second from Assumptions 1 (specifically, from the conditional independence of potential outcomes and treatment) and 2, the third from basic probability theory and the fourth from iterated expectations. This implies that the controlled direct effect is identified by both a regression representation or reweighting:

$$\begin{aligned}\gamma(m) &= E[E[Y|D=1,M=m,X]-E[Y|D=0,M=m,X]]\\ &= E\left[\frac{Y\cdot D\cdot I\{M=m\}}{\Pr(D=1,M=m|X)}-\frac{Y\cdot(1-D)\cdot I\{M=m\}}{\Pr(D=0,M=m|X)}\right].\end{aligned} \tag{20}$$

3.3 Estimation with covariates not affected by treatment

This section presents various estimators of natural direct and indirect effects, some of them directly using the identification results of Section 3.2. We to this end assume the availability of an i.i.d. sample of size n in which i is the index of an observation ($i \in \{1,...,n\}$) and denote by (Y_i, M_i, D_i, X_i) the sample realizations of the respective random variables (Y, M, D, X).

Estimation based on the mediation formula (16) requires plug-in estimates for the conditional

mean outcomes and the conditional mediator densities. One popular approach is g-computation going back to Robins (1986), which obtains these parameters based on maximum likelihood estimation (MLE). Let $\hat{\mu}_Y(d,m,x)$, $\hat{f}(m|d,x)$ denote the estimates of the conditional mean outcome $E[Y|D=d,M=m,X=x]$ and the conditional mediator density $f_{M=m|D=d,X=x}$ (or conditional probability $\Pr(M=m|D=d,X=x)$ if the mediator is discrete). The g-computation estimators of the direct and indirect effects are given by

$$\begin{aligned}\hat{\theta}(d) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\hat{\mu}_Y(1,M_i,X_i)-\hat{\mu}_Y(0,M_i,X_i)\right]\hat{f}(M_i|d,X_i)\right\},\\ \hat{\delta}(d) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_Y(d,M_i,X_i)\left[\hat{f}(M_i|1,X_i)-\hat{f}(M_i|0,X_i)\right]\right\},\end{aligned} \tag{21}$$

where $\hat{\theta}(d)$, $\hat{\delta}(d)$ are estimates of the direct and indirect effects. In general, both parametric models for $E[Y|D=1,M=m,X=x]$ and $f(M=m|D=d,X=x)$ need to be correctly specified for consistency. Alternatively and as discussed in Imai, Keele, and Yamamoto (2010), the plug-in parameters $\hat{\mu}_Y(D,M,X)$ and $\hat{f}(M|D,X)$ can be estimated nonparametrically to safeguard against misspecification. This might, however, be cumbersome in finite samples if X and/or M are high dimensional, a problem known as curse of dimensionality.

Concerning weighting expressions (17), natural direct and indirect effects can be estimated by their normalized sample analog, where normalization guarantees that the weights underlying the estimators sum up to one in each treatment group. For instance, the direct effect under non-treatment is given by

$$\hat{\theta}(0) = \frac{\sum_{i=1}^{n} Y_i D_i(1-\hat{\rho}(M_i,X_i))/[\hat{\rho}(M_i,X_i)(1-\hat{p}(X_i))]}{\sum_{i=1}^{n} D_i(1-\hat{\rho}(M_i,X_i))/[\hat{\rho}(M_i,X_i)(1-\hat{p}(X_i))]} - \frac{\sum_{i=1}^{n} Y_i(1-D_i)/(1-\hat{p}(X_i))}{\sum_{i=1}^{n}(1-D_i)/(1-\hat{p}(X_i))} \tag{22}$$

where $\hat{\rho}(m,x)$ and $\hat{p}(x)$ denote the respective estimates of the propensity scores $\Pr(D=1|M=m,X=x)$ and $\Pr(D=1|X=x)$. A practical advantage of this approach is that the estimation of conditional mediator densities is not required, which is particularly relevant when M is multi-dimensional and/or continuously distributed. Treatment propensity scores might be estimated by probit or logit specifications, see for instance Huber (2014) and Tchetgen Tchetgen (2013), as implemented in the 'causalweight' package for R by Bodory and Huber (2018). Hsu, Huber, and Lai (2018) show that also for nonparametrically estimated propensity scores, effect estimation is root-n-consistent under particular regularity conditions. In addition, this approach is asymptotically semiparametrically efficient, i.e. it has the smallest possible asymptotic variance and thus attains

the semiparametric efficiency bounds for causal mediation analysis derived in Tchetgen Tchetgen and Shpitser (2012). In finite samples, however, the curse of dimensionality might kick in.

Also for the weighting expression (18), a sample analog estimator can be constructed with $\hat{p}(X)$ $\hat{f}(M|D,X)$ representing parametric or nonparametric plug-in estimates, see Hong, Deutsch, and Hill (2015) for examples. Lange, Vansteelandt, and Bekaert (2012) combine weighting with imputation-based estimation of potential outcomes. Finally, Chan, Imai, Yam, and Zhang (2016) suggest a nonparametric weighting approach that does not require plug-in estimates of propensity scores or conditional mediator densities, but applies an empirical calibration approach. That is, it algorithmically derives the weights using specific moment conditions that reflect the true weights' property to balance the distributions of X and M across treatment groups. The method is root-n-consistent and asymptotically semiparametrically efficient.

Tchetgen Tchetgen and Shpitser (2012) estimate the effects based on the sample analogue of the score or efficient influence function for computing potential outcomes, which relies on estimating conditional mean outcomes, mediator densities, and treatment probabilities. The direct effect under non-treatment, for instance, is obtained by

$$\begin{aligned}\hat{\theta}(0) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{D_i\hat{f}(M_i|0,X_i)}{\hat{p}(X_i)\hat{f}(M_i|1,X_i)}-\frac{1-D_i}{1-\hat{p}(X_i)}\right][Y_i-\hat{\mu}_Y(D_i,M_i,X_i)]\right.\\ &+ \left.\frac{1-D_i}{1-\hat{p}(X_i)}\left[\hat{\mu}_Y(1,M_i,X_i)-\hat{\mu}_Y(0,M_i,X_i)-\hat{\theta}(0,X_i)\right]+\hat{\theta}(0,X_i)\right\}. \qquad (23)\end{aligned}$$

$\hat{\theta}(d,x)$ denotes an estimate of

$$\theta(d,x) = E\left[E[Y|D=1,M=m,X=x]-E[Y|D=0,M=m,X=x]|D=d,X=x\right],$$

which may for instance be obtained by regressing $\hat{\mu}_Y(1,M,X)-\hat{\mu}_Y(0,M,X)$ on X among those with $D=d$.

One attractive feature of this estimator is that it is 'multiply robust' in the sense that it remains consistent if only particular subspecifications of the model are correct. Namely, it needs to hold that at least either (i) $E[Y|D,M,X]$ and $f_{M|D,X}$ (or, alternatively and somewhat weaker, $E[Y|D,M,X]$ and $\theta(D,X)$), (ii) $E[Y|D,M,X]$ and $\Pr(D=1|X)$, or (iii) $\Pr(D=1|X)$ and $f_{M|D,X}$ are correctly specified, see Tchetgen Tchetgen and Shpitser (2012) and Zheng and van der Laan (2012). If all three conditions hold, multiply robust estimation is asymptotically semiparametrically efficient.

The same properties hold for the targeted maximum likelihood approach of Zheng and van der Laan (2012) which iteratively optimizes maximum likelihood estimation of the target (direct or indirect) effect based on initial plug-in estimates of $E[Y|D,M,X]$, $f_{M|D,X}$, and $\Pr(D=1|X)$. For this reason, both estimators improve upon the 'doubly robust' estimator of Van der Laan and Petersen (2008), which allows for a misspecification of either the outcome or the treatment model, while consistency requires the conditional mediator density to be correctly specified. Vansteelandt, Bekaert, and Lange (2012) propose an estimator based on imputation of potential outcomes which possesses a doubly robust property, too. It remains consistent under outcome misspecification, provided that models for both the mediator and the treatment are correct.

A range of further (mostly parametric) estimators that allow for effect heterogenity across d has been proposed in the literature. For linear models, VanderWeele and Vansteelandt (2009) and Preacher, Rucker, and Hayes (2007) (among others) discuss which combinations of coefficients yield the direct and indirect effects when including treatment-mediator interaction terms and control variables. VanderWeele and Vansteelandt (2010) provide a related discussion in the context of nonlinear binary outcome models. VanderWeele (2009) considers so-called marginal structural models for modelling mean potential outcomes (rather than observed outcomes). The approach combines regression with reweighting to estimate controlled direct effects and natural effects. Van der Laan and Petersen (2008) directly model the direct effects of interest, rather than potential outcomes.

Tingley, Yamamoto, Hirose, Imai, and Keele (2014) suggest a simulation approach based on the linear or non-linear estimation of the mediator and outcome models and the simulation of potential mediators and outcomes according to the imposed models for computing direct and indirect effects. This method is available in the 'mediation' package for R by Tingley, Yamamoto, Hirose, Imai, and Keele (2014). As an alternative, the 'medflex' package by Steen, Loeys, Moerkerke, and Vansteelandt (2017b) implements potential outcome imputation as discussed in Vansteelandt, Bekaert, and Lange (2012) and weighting as in Lange, Vansteelandt, and Bekaert (2012). We also refer to Hong (2015) and VanderWeele (2016) for further reviews on estimation approaches. Finally, we note that Huber, Lechner, and Mellace (2016) investigate the finite sample performance of a range of different estimators relying on sequential conditional independence.

3.4 Effects under dynamic confounding and multiple mediators

In many empirical problems, it might seem unlikely that pre-treatment covariates suffice to control for the endogeneity of the mediator M, as the latter is a post-treatment variable. Just as the control variables of the treatment are typically measured shortly before treatment assignment, it seems reasonable to control for possible confounders of the mediator-outcome relation just prior to selection into the mediator. It then appears likely that at least some of these confounders are influenced by the treatment (in particular if the time lag between D and M is non-negligible), such that they are themselves mediators that affect the mediator M of interest. For this reason, Robins (2003) suspects that Assumptions 1 and 2, which imply that mediator confounders are not a function of D, are of limited practical relevance.

We subsequently consider the case that the treatment may influence observed post-treatment confounders of the mediator-outcome relation, denoted by W. To account for this extension in our notation, we rewrite the potential mediators and outcomes as functions of W: $M(d) = M(d, W(d))$ and $Y(d, M(d)) = Y(d, M(d, W(d)), W(d))$, where $W(d)$ is a vector of potential post-treatment covariates for $D = d$. The direct and indirect effect then correspond to

$$\begin{aligned} \theta^M(d) &= E[Y(1, M(d, W(d)), W(1)) - Y(0, M(d, W(d)), W(0))], \\ \delta^M(d) &= E[Y(d, M(1, W(1)), W(d)) - Y(d, M(0, W(0)), W(d))]. \end{aligned} \tag{24}$$

$\theta^M(d)$ contains any effect of D on Y not operating through M. In addition to the inherently direct causal path from D to Y, it also includes the causal mechanism from D to W to Y. Therefore, the term direct effect might appear ambiguous in this context. $\delta^M(d)$, on the other hand, consists of any effect via M which either directly comes from D or 'takes a devious route' via W. This effect takes into consideration that D may influence M either directly or indirectly through W.

Alternatively, one might consider the path-specific indirect effect directly going from D to M, but not operating through W:

$$\delta^{Mp}(d) = E[Y(d, M(1, W(d)), W(d)) - Y(d, M(0, W(d)), W(d))]. \tag{25}$$

$\delta^{Mp}(d)$ represents a partial indirect effect when keeping W fixed at its level implied by d, such that any effect from D to W to M is switched off. Arguably, $\delta^M(d)$ is more interesting than $\delta^{Mp}(d)$, but

unfortunately also more difficult to identify, as discussed below. Finally, one could be interested in the joint indirect effects via M and/or W:

$$\delta^{M,W}(d) = E[Y(d, M(1, W(1)), W(1)) - Y(d, M(0, W(0)), W(0))]. \tag{26}$$

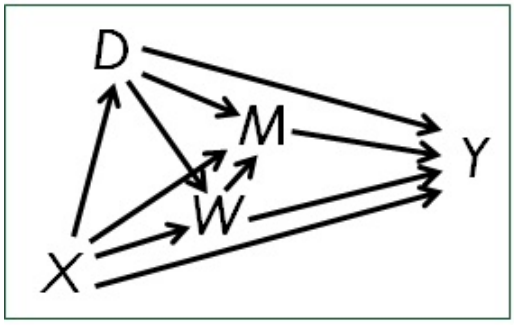
$\delta^{M,W}(d)$ comprises any causal mechanism going from D to Y that operates through either M, or W, or both. Accordingly, the direct effect $\theta^{M,W}(d)$ corresponds to the impact of D on Y operating neither through M, nor through W, which distinguishes it from $\theta^{M}(d)$ that may include indirect mechanisms via W (but not M):

$$\theta^{M,W}(d) = E[Y(1, M(d, W(d)), W(d)) - Y(0, M(d, W(d)), W(d))]. \tag{27}$$

3.5 Identification under dynamic confounding and multiple mediators

It is easy to see that $\delta^{M,W}(d)$ and $\theta^{M,W}(d)$ are identified based on Assumptions 1 and 2 and the results (13), (14), and (15) in Section 3.2 when replacing M by (M, W) everywhere. Notably, one could allow for unobserved confounders of the W-M relation, as Assumtpions 1 and 2 do not rule out confounding within the set of mediators. The subsequent discussion focusses on the identification of $\delta^{Mp}(d)$, $\delta^{M}(d)$, and $\theta^{M}(d)$. The assumptions considered match with the causal graph in Figure 2, which provides a framework where the treatment affects the observed confounders W of the mediator M. Importantly, the use of W as controls for M implies that

Figure 2: Causal paths with pre-treatment and post-treatment confounders



there must not exist any unobserved confounders that jointly influence W on the one hand and M and/or Y on the other hand. This conditional independence of W imposes strong requirements concerning X. Not only must X contain all pre-treatment variables confounding the D-Y, D-M, or M-Y relationship, but also any factors confounding the W-M or W-Y relationship (apart from D). Otherwise, conditioning on W would introduce an association between D and those factors

affecting M or Y and thus, treatment endogeneity. As an example, suppose that Y is wealth, M is employment, D is college education, and W includes the pre-mediator health state, which affects Y and is a function of D and pre-treatment health. In this case, X must include pre-treatment health if the latter affects Y also through other channels than pre-mediator health. This seems plausible, as wealth is typically determined by previous income streams that may themselves be influenced by previous health.

Assumptions 4, 5, and 6 provide formal conditions for identifying $\delta^{Mp}(d)$ in the presence of post-treatment confounders of the mediator.

Assumption 4 (conditional independence of the treatment):

$\{Y(d'', m, w'), M(d', w), W(d)\} \perp D | X = x$ for all $d'', d', d \in \{0, 1\}$ and m, w', w, x in the support of M, W, X.

Similar to Assumption 1, Assumption 4(a) requires D to be conditionally independent of potential post-treatment variables, namely potential outcomes, mediators, and confounders of the mediator-outcome relation.

Assumption 5 (conditional independence of the post-treatment variables):

(a) $\{Y(d'', m, w'), M(d', w)\} \perp W | D = d, X = x$ for all $d'', d', d \in \{0, 1\}$ and m, w', w, x in the support of M, W, X,

(b) $Y(d', m, w') \perp M | D = d, W = w, X = x$ for all $d', d \in \{0, 1\}$ and m, w', w, x in the support of M, W, X.

Assumption 5(a) states that W is conditionally independent of the potential mediators and outcomes given X and D. This implies that all pre-treatment covariates affecting both W and M or Y are included in X. Assumption 5(b) is somewhat weaker than Assumption 2, as it imposes conditional independence of the mediator and the potential outcomes given X and W (rather than X alone).

Assumption 6 (common support):

$\Pr(D = d | M = m, W = w, X = x) > 0$ for all $d \in \{0, 1\}$ and m, w, x in the support of M, W, X.

Finally, Assumption 6 is somewhat stronger than Assumption 3, as it requires the common support restriction on the treatment propensity to hold when conditioning on M and both W and X (rather than X alone).

Under Assumptions 4-6, the partial indirect effect is identified, for instance, by the following

weighting expression using three different treatment propensity scores, see Huber (2014) (where Assumption 5(a) is, however, stated somewhat differently):

$$\delta^{Mp}(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,W,X)} \cdot \frac{\Pr(D=d|W,X)}{\Pr(D=d|X)} \cdot \left(\frac{\Pr(D=1|M,W,X)}{\Pr(D=1|W,X)} - \frac{1-\Pr(D=1|M,W,X)}{1-\Pr(D=1|W,X)}\right)\right]. \tag{28}$$

We refer to Steen, Loeys, Moerkerke, and Vansteelandt (2017a) for a more comprehensive discussion on the identification and estimation of $\delta^{Mp}(d)$ and further path-specific effects. $\delta^{M}(d)$ and $\theta^{M}(d)$ are, however, not identified without further assumptions, see the impossibility result in Avin, Shpitser, and Pearl (2005). Intuitively, identifying $E[Y(d, M(1-d, W(1-d)), W(d))]$ (and the natural effects requires that conditional on X, the distribution of M given $D = d$ is exogenously adjusted to match that of M given $D = 1 - d$, while keeping the distribution of W fixed for $D = d$. Doing both at the same time is, however, impossible if W affects M. For this reason, the identification of $\delta^{M}(d)$ and $\theta^{M}(d)$ necessarily requires parametric restrictions.

One restricting yielding identification is ruling out interactions between D and M, see Robins and Greenland (1992) and Robins (2003). This strong condition even permits easing the previous assumptions somewhat by avoiding conditional independence between potential outcomes and potential mediators defined upon opposite treatment states. While Assumption 5(b), for instance, imposes conditional independence between $Y(d', m, W(d'))$ and $M(d)$ also for $d \neq d'$, this is only required for $d = d'$ under the no interactions condition. Economists might, however, find it hard to come up with a realistic case in which conditional independence likely holds for $d = d'$, but not for $d \neq d'$, even though hypothetical examples are provided in Robins and Richardson (2010). The absence of treatment-mediator interactions, on the other hand, appears to be quite unattractive, as it restricts effect heterogeneity.

As an improvement, Imai and Yamamoto (2013) allow for treatment-mediator interactions, but require them to be homogeneous in the population. Huber (2014) presents a related assumption that restricts the average treatment-mediator-interaction to be homogeneous conditional on X and W and assumes the outcome to be linear in M. Alternatively, Tchetgen Tchetgen and VanderWeele (2012) show that $\delta^{M}(d)$ and $\theta^{M}(d)$ are identified (under particular conditional independence assumptions) if the average interaction effects of W and M on Y are zero or all elements in W are binary and monotonic in D. Yet another restriction is that potential confounders $W(1)$ and $W(0)$ are independent of each other or that their dependence follows a known distribution, see Robins and Richardson (2010) and Albert and Nelson (2011), who

consider general models with sequences of multiple mediators. For a discussion of identification in parametric mediation models we refer to De Stavola, Daniel, Ploubidis, and Micali (2015), who show that the assumption of no unobserved confounders of the $W - Y$ relationship can be relaxed under specific assumptions. While some of the discussed restrictions appear more or less attractive than others, all of them share the caveat that they impose specific functional form constraints that reduce model generality.

Finally, we note that if W is a function of D, does (in contrast to Figure 2) not affect M, but still influences Y, then the mediation model consists of two independent indirect mechanisms operating through M and W, respectively. In this case, either natural indirect effect can be identified by imposing Assumptions 1 to 3 of Section (3.1) for both M and W separately, see Imai and Yamamoto (2013) and Lange, Rasmussen, and Thygesen (2014) for further details. We refer to VanderWeele and Vansteelandt (2014) for a comprehensive discussion of different mediation models involving multiple mediators, including the ones considered here.

We have so far focussed on natural effects. While $\delta^M(d)$ and $\theta^M(d)$ are not identified nonparametrically in the presence of post-treatment confounders of the mediator-outcome relation that are a function of the treatment, the controlled direct effect is obtained even under somewhat weaker conditions than Assumptions 4 to 6. Specifically, Assumption 5(a) and the independence of D and the potential mediators and post-treatment confounders postulated in Assumption 4 are not required. Assuming that M is discrete, the potential outcome $E[Y(d,m,W)] = E[Y(d,m)]$, which assumes that one is interested in a particular mediator value $M = m$ while being agnostic about the values of confounders W, corresponds to:

$$\begin{aligned}
E[Y(d,m)] &= E[E[Y(d,m)|D=d,X]] = \int\int E[Y(d,m)|D=d,W,X] f_{W=w|D=d,X=x} dw f_{X=x} dx \\
&= \int\int E[Y|D=d,M=m,W,X] f_{W=w|D=d,X=x} dw f_{X=x} dx \\
&= \int\int E\left[\frac{Y \cdot I\{M=m\}}{\Pr(M=m|D=d,W,X)}\middle| D=d,W,X\right] f_{W=w|D=d,X=x} dw f_{X=x} dx \\
&= E\left[E\left[\frac{Y \cdot I\{M=m\}}{\Pr(M=m|D=d,W,X)}\middle| D=d,X\right]\right] \\
&= E\left[E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\}}{\Pr(M=m|D=d,W,X)\Pr(D=d|X)}\middle| X\right]\right] \\
&= E\left[\frac{Y \cdot I\{D=d\} \cdot I\{M=m\}}{\Pr(M=m|D=d,W,X)\Pr(D=d|X)}\right]. \qquad (29)
\end{aligned}$$

The first equality follows from the law of iterated expectations and the conditional independence of

D and the potential outcomes, see Assumption 4, the second from the law of iterated expectations and replacing expectations by integrals, the third from Assumption 5(b), the fourth from basic probability theory, the fifth from integrating out W and replacing the integral by an expectation, the sixth from basic probability theory, and the last from the law of iterated expectations. This proof is equivalent to those in the context of dynamic treatment evaluation, see for instance Lechner (2009).

The controlled direct effect is therefore identified by

$$\begin{aligned}
\gamma(m) &= \int\int E[Y|D=1,M=m,W=w,X=x] f_{W=w|D=1,X=x} dw f_{X=x} dx \\
&- \int\int E[Y|D=0,M=m,W=w,X=x] f_{W=w|D=0,X=x} dw f_{X=x} dx \qquad (30) \\
&= E\left[\frac{Y \cdot D \cdot I\{M=m\}}{\Pr(M=m|D=1,W,X)\cdot\Pr(D=1|X)} - \frac{Y\cdot(1-D)\cdot I\{M=m\}}{\Pr(M=m|D=0,W,X)\cdot\Pr(D=0|X)}\right].
\end{aligned}$$

The expression after the first equality in (30) might for instance be estimated by g-computation, see Robins (1986) and Vansteelandt (2009), or matching, see Lechner and Miquel (2010), the one after the second equality by weighting, see Robins, Hernan, and Brumback (2000).

4 Further identification approaches

This section reviews approaches that relax sequential conditional independence or rely on alternative assumptions. First, it discusses partial identification of effects based on sensitivity checks and bounds under weaker conditions than imposed in Section 3.2. Second, it considers the identifying power of randomizing both the treatment and the mediator. Third, it presents various instrumental variable methods for tackling the endogeneity of the mediator and/or treatment. Finally, it briefly presents a difference-in-differences approach.

4.1 Partial identification based on sensitivity checks and bounds

As the sequential conditional independence assumptions outlined in Section 3.1 are quite strong, several sensitivity checks have been suggested that permit assessing the robustness of direct and indirect effects to deviations from these assumptions. This amounts to identifying intervals or sets rather than point values of the parameters of interest. VanderWeele (2010), for instance, provides a general formula for the bias (denoted by B) of the controlled direct effect as well as

natural effects in the presence of an unobserved mediator-outcome confounder (denoted by U) that violates Assumption 2:

$$\begin{aligned}
B(\theta(d)) &= \int\int\int \big\{E[Y|D=1-d,M=m,X=x,U=u] - E[Y|D=1-d,M=m,X=x,U=u']\big\} \\
&\quad [f_{U=u|D=1-d,M=m,X=x} - f_{U=u|D=d,M=m,X=x}]du f_{M=m|D=d,X=x}dm f_{X=x}dx, \\
B(\delta(1-d)) &= -B(\theta(d)), \\
B(\gamma(m)) &= \int\int \big\{E[Y|D=1-d,M=m,X=x,U=u] - E[Y|D=1-d,M=m,X=x,U=u']\big\} \\
&\quad [f_{U=u|D=1-d,M=m,X=x} - f_{U=u|D=1-d,X=x}]du f_{X=x}dx \\
&- \int\int \big\{E[Y|D=d,M=m,X=x,U=u] - E[Y|D=d,M=m,X=x,U=u']\big\} \\
&\quad [f_{U=u|D=d,M=m,X=x} - f_{U=u|D=d,X=x}]du f_{X=x}dx. \qquad (31)
\end{aligned}$$

$u \neq u'$ are two values or sets of values in the support of the supposedly omitted U, that might be scalar or a vector.

Under the restrictions that U is scalar and binary, $E[Y|D=1-d,M=m,X=x,U=1] - E[Y|D=1-d,M=m,X=x,U=0]$ is homogeneous across values of D, M, and X, and $\Pr(U=1|D=1-d,M=m,X=x) - \Pr(U=1|D=d,M=m,X=x)$ is homogeneous across values of M and X, the expressions in (31) simplify to

$$B(\theta(d)) = B(\gamma(m)) = \phi\omega, \quad B(\delta(1-d)) = -\phi\omega, \qquad (32)$$

with $E[Y|D=1-d,M=m,X=x,U=1] - E[Y|D=1-d,M=m,X=x,U=0] = \phi$ and $\Pr(U=1|D=1-d,M=m,X=x) - \Pr(U=1|D=d,M=m,X=x) = \omega$. By considering sensible values for the conditional difference in Y across values of U given by ϕ and the conditional difference in U across treatment states given by ω, researches may correct their estimates for the hypothetical bias. The use of the more general formulas in (31), on the other hand, requires encoding such conditional differences for any combinations of D, M, and X or M and X, respectively, which might be cumbersome.

Imai, Keele, and Yamamoto (2010) propose a sensitivity analysis that can be implemented in parametric (both linear and nonlinear) mediation models by specifying the correlation of unobserved terms in the mediation and outcome equations, e.g. U and V in (9) and (10). A non-zero correlation of U and V, denoted by $\rho_{U,V}$, implies a violation of Assumption 2. Therefore, varying

$\rho_{U,V}$ between -1 and 1 and investigating how the regression-based estimates vary permits analyzing sensitivity to mediator-outcome confounding and is implemented in the 'mediation' package by Tingley, Yamamoto, Hirose, Imai, and Keele (2014). The sensitivity analysis, however, assumes that mediator-outcome confounders are not a function of the treatment.

In contrast, Tchetgen Tchetgen and Shpitser (2012) suggest a procedure that allows for confounders of the mediator-outcome relation which are affected by D, see Section 3.5. They suggest a semiparametric method based on specifying and calibrating the selection bias function $E[Y(1,m)|D=d,M=m,X=x]-E[Y(1,m)|D=d,M\neq m,X=x]$, which is agnostic about the dimension of unobserved confounders. See VanderWeele and Chiba (2014) and Vansteelandt and VanderWeele (2012) for further methods based on alternative selection bias functions. A conceptually different approach is offered in Albert and Nelson (2011), who consider the correlation of counterfactual values of post-treatment variables as sensitivity parameter.

Hong, Qin, and Yang (2018) provide a method tailored to weighting estimators based on (15), which is applicable under the omission of both pre- and post-treatment confounders. The idea is that such confounders create a discrepancy between the correct weight an observation should obtain and the one actually used. The resulting bias can be represented by the covariance between the weight discrepancy and the outcome conditional on treatment, which serves as base for conducting sensitivity analyses. Finally, Imai and Yamamoto (2013) propose a robustness check for violations of the no-treatment-mediator interaction assumption for instance discussed in Robins (2003). It relies on two sensitivity parameters: the correlation between the mediator and the individual-level effect of the treatment–mediator interaction in the outcome equation and the standard deviation of the individual-level effect of the treatment-mediator interaction.

All studies mentioned so far investigate the robustness of direct and indirect effects to prespecified deviations from the identifying assumptions. Alternatively, worst case bounds on the effects of interest may be derived, which are based on the possibly most extreme forms of violations of specific assumptions. However, this generality typically comes at the cost of a rather wide range of admissible effect values. We refer to Kaufman, Kaufman, MacLenose, Greenland, and Poole (2005), Cai, Kuroki, Pearl, and Tian (2008), and Sjölander (2009) for methodological contributions in this field, as well as to Flores and Flores-Lagunes (2010), who investigate the employment and earnings effects of the Job Corps program mediated by the achievement of a formal degree. Typically, the treatment is assumed to be randomized in such studies, while either no or weaker

assumptions than exogeneity are imposed on the mediator, e.g. monotonicity of the latter in the treatment, to attain upper and lower bounds on direct and indirect effects.

4.2 Experimental randomization of treatment and mediator

This section discusses experimental designs for evaluating causal mechanisms based on double randomization of both the treatment and the (scalar) mediator. Imai, Tingley, and Yamamoto (2013) consider a so-called parallel design which is based on two experiments with distinct subjects. In the first experiment, the treatment alone is randomized, implying that Assumption 1 holds unconditionally, i.e. even without controlling for X. In the second experiment, the treatment and the mediator are jointly randomized, such that both Assumptions 1 and 2 hold without conditioning on X. Furthermore, it is assumed that the assignment to one or the other experiment does itself not directly affect the outcomes, an assumption Imai, Tingley, and Yamamoto (2013) call consistency. Then, the first experiment allows assessing the ATE by taking mean differences in outcomes across treatment groups, because $\Delta = E[Y|D=1] - E[Y|D=0]$ by the randomization of D. The second experiment permits evaluating the controlled direct effect because $\gamma(m) = E[Y|D=1, M=m] - E[Y|D=0, M=m]$ by the randomization of D and M.

Natural direct and indirect effects are, however, not point identified without further assumptions (even though they can be bounded, see Imai, Tingley, and Yamamoto (2013)), as the counterfactual $E[Y(d, M(1-d)]$ is unknown even when combining information from both experiments. For instance, for observations in the first experiment with actual treatment assignment $D = d$, one cannot infer whether they would satisfy $M(1-d) = m$, because their behavior under the alternative assignment $D = 1 - d$ is unknown. Therefore, information on $Y(d, m)$ from the second experiment can generally not be used to recover $E[Y(d, M(1-d))]$. One restriction that yields identification is ruling out treatment-mediator interactions, see for instance Robins (2003). This implies that $\theta(d) = \gamma(m) = E[Y|D=1, M=m] - E[Y|D=0, M=m]$ for any d and m, such that direct effects are constant: $\theta = \gamma$. A testable implication of the latter is that the controlled direct effect is the same across different choices of m. In this special case, the indirect effect is identified by $\delta = \Delta - \gamma$.

As an alternative approach, Wunsch and Strobl (2018) assume homogeneous average effects of a binary M on Y given $D = d$ across principal strata defined on potential mediator states, see Section 2.3. Formally, $E[Y(d,1) - Y(d,0)|M(1) = m, M(0) = m'] = E[Y(d,1) - Y(d,0)]$ for

$m, m' \in \{1, 0\}$. The indirect effect then simplifies to $\delta(d) = E[Y(d,1) - Y(d,0)] \cdot E[M(1) - M(0)]$. The first and second experiments identify $E[M(1) - M(0)]$ and $E[Y(d,1) - Y(d,0)]$ by $E[M|D = 1] - E[M|D = 0]$ and $E[Y|D = d, M = 1] - E[Y|D = d, M = 0]$, respectively. Furthermore, in the presence of observed covariates X, the homogeneity assumption can be tested by verifying whether $E[Y|D = d, M = 1, X = x] - E[Y|D = d, M = 0, X = x]$ is constant across values x. Importantly, the availability of covariates also permits relaxing the assumption to hold conditional on X:

Assumption 7 (conditional independence of the mediator):
$E[Y(d,1) - Y(d,0)|M(1) = m, M(0) = m', X = x] = E[Y(d,1) - Y(d,0)|X = x]$ for all x in the support of X and $m, m' \in \{1, 0\}$.

Combining Assumption 7 with consistency, the randomization of D and of D and M, respectively, in the two experiments identifies the indirect effect: $\delta(d) = E[A \cdot B|D = d]$, where $A = E[Y|D = d, M = 1, X] - E[Y|D = d, M = 0, X]$ comes from the second experiment and $B = E[M|D = 1, X] - E[M|D = 0, X]$ from the first. Wunsch and Strobl (2018) discuss identification also for setups different to the parallel design, e.g. when only M (but not D) is randomized in the second experiment. See also Pirlott and MacKinnon (2016) for a survey of alternative approaches to randomization.

Imai, Tingley, and Yamamoto (2013) suggest yet another experimental method, the cross-over design, which allows identifying natural effects based on involving the same subjects in either experiment. In the first experiment, the treatment is randomized and the resulting mediator and outcome values are measured. In the second experiment, all subjects obtain the treatment opposite to their status in the first experiment, while the mediator is fixed to the same value as observed in the first experiment. Under consistency and no carry-over effects from the first experiment that contaminate the potential outcomes in the second experiment, natural effects are straightforwardly identified. In fact, $Y(d, M(d))$ and $Y(1-d, M(d))$ are observed in the first and second experiment, respectively.

4.3 Mediation based on separate instruments for treatment and mediator

For many empirical problems, experimental randomization is not feasible and sequential conditional independence assumptions based on observational data do not appear plausible either, as treatment/mediator endogeneity might be related to unobserved characteristics. In this case, in-

strumental variables (IV) are an alternative potential source of identification, given that specific conditions hold. To gain some intuition, consider the following system of linear equations for the outcome, mediator, and treatment.

$$Y = \beta_D D + \beta_M M + U, \tag{33}$$

$$M = \alpha_D D + \alpha_{Z_2} Z_2 + V, \tag{34}$$

$$D = \sigma_{Z_1} Z_1 + Q. \tag{35}$$

This scenario bears some similarity with the simple mediation model given by (2.1) and (5). However, a notable difference is that D and M are now assumed to be functions of instruments Z_1 and Z_2, respectively, which satisfy an exclusion restriction in the sense that they do not directly affect Y. Furthermore, the unobserved terms U, V, and Q are allowed to be arbitrarily associated, which renders M and D endogenous.

Identification is obtained if Z_1 and Z_2 are independent of (or in the linear model at least uncorrelated with) the unobservables (U, V, Q). For ease of exposition, we also assume that they are independent of each other. Similar to two stage least squares, replacing the original treatment variable D by the exogenous prediction $E(D|Z_1)$ in the mediator and outcome equations and replacing M in the outcome equation by $E(M|E(D|Z_1), Z_2)$ permits identifying α_D, β_D, and β_M, respectively. The reason is that only the exogenous variation in the treatment and the mediator, which is unrelated to the unobservables, is exploited. Therefore, the probability limits of regressions of M and Y on the respective predictions identifies the coefficients of interest required for computing direct and indirect effects. If the IV exclusion restrictions and independence assumptions appear plausible only conditional on observed covariates X, the regressions are to be augmented by these control variables.

Such parametric IV approaches to mediation have been used in the context of Mendelian randomization, which uses genetic variants as IVs (see for instance Burgess, Daniel, Butterworth, and Thompson (2015)), and in comparably few economic studies. Considering Australian HILDA survey data, Powdthavee, Lekfuangfu, and Wooden (2013) estimate a positive indirect effect of education on life satisfaction running via the mediator income, using regional differences in changes of schooling laws as IV for education and income shocks (inheritance, severance pay, lottery wins) as IVs for income. Chen, Hsu, and Wang (2018) apply the IV mediation framework to stochastic frontier modelling, in order to disentangle the total effect of a government policy for enhancing

productivity (e.g. dam construction to increase agricultural output) into technology and efficiency components based on topographic instruments.

In contrast to parametric mediation models, the following nonparametric model allows for arbitrary effect heterogeneity through interactions of D, M, X, and/or the unobservables:

$$\begin{aligned} Y &= \varphi(D, M, X, U), && (36) \\ M &= \zeta(D, Z_2, X, V), && (37) \\ D &= \chi(Z_1, X, Q). && (38) \end{aligned}$$

φ, ζ, χ are unknown functions. As in , M is assumed to be scalar, while extensions to vector valued mediators would require additional instrumental variables for identification. We subsequently assume both Z_1 and D to be binary, which matches the case of a randomized experiment with imperfect compliance, where random treatment assignment is the instrument and actual take up is the treatment. The potential mediators and outcomes are defined as $M(d) = \zeta(d, Z_2, X, V)$ and $Y(d, M(d') = \varphi(d, M(d'), X, U) = \varphi(d, \zeta(d', Z_2, X, V), X, U)$, respectively, for $d, d' \in \{1, 0\}$. Similarly, we consider the potential treatment state as a function of the instrument, denoted as $D(z_1)$. Based on (38), the latter corresponds to $D(z_1) = \chi(z_1, X, Q)$ for $z_1 \in \{0, 1\}$.

Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) discuss IV-based identification of the total average effect on the subpopulation complying with treatment assignment in terms of treatment take up, i.e. satisfying $D(1) = 1$ and $D(0) = 0$. This effect is known as local average treatment effect (LATE) or complier average causal effect (CACE) and formally defined as follows:

$$\Delta_c = E[Y(1) - Y(0)|D(1) = 1, D(0) = 1] = E[Y(1, M(1)) - Y(0, M(0))|D(1) = 1, D(0) = 0].$$

In analogy to Sections 2.1 and 2.2, the natural direct and indirect effects as well as the controlled effect on treatment compliers (rather than the total population) are defined as

$$\begin{aligned} \theta_c(d) &= E[Y(1, M(d)) - Y(0, M(d))|D(1) = 1, D(0) = 0], \\ \delta_c(d) &= E[Y(d, M(1)) - Y(d, M(0))|D(1) = 1, D(0) = 0], \\ \gamma_c(m) &= E[Y(1, m) - Y(0, m)|D(1) = 1, D(0) = 0]. && (39) \end{aligned}$$

Frölich and Huber (2017) discuss nonparametric identification of $\theta_c(d)$, $\delta_c(d)$, and $\gamma_c(m)$ when IVs are conditionally valid given observables. While Z_1 and D are assumed to be binary, the authors consider scenarios in which (i) both M and Z_2 are continuous, (ii) M is discrete and Z_2 is continuous, and (iii) M is continuous and Z_2 is discrete. The required assumptions vary across these scenarios and we subsequently briefly explain the conditions yielding identification when both M and Z_2 are continuous. Firstly, instruments (Z_1, Z_2) must be independent of unobservables (U, V, W) conditional on covariates X and satisfy the exclusion restrictions as postulated in the nonparametric model above. Secondly, Z_1 must be independent of Z_2 given X. Both assumptions are satisfied under a separate (i.e. independent) randomization of the instruments. Furthermore and in analogy to Imbens and Angrist (1994), Z_1 must weakly increase D for everyone (a condition known as monotonicity) and strictly increase D in a subpopulation, implying that compliers exist (known as first stage relevance). A further condition is the strict monotonicity of the mediator in V, which is assumed to be a continuously distributed scalar unobservable or index of unobservables. Finally, the common support restriction that $\Pr(Z_1 = z_1|M, V, X, D(1) = 1, D(0) = 0)$ is larger than zero for $z_1 \in \{1, 0\}$ must hold, too.

Under these conditions, mean potential outcomes among compliers are identified. For instance,

$$E\left[Y(0, M(1))|D(1) = 1, D(0) = 0\right] = \frac{E\left[Y \cdot (D-1) \cdot \frac{1}{\Omega} \cdot (Z_1/\pi(X) - (1-Z_1)/(1-\pi(X)))\right]}{E\left[D \cdot (Z_1/\pi(X) - (1-Z_1)/(1-\pi(X)))\right]}. \quad (40)$$

$\Omega = \frac{E[(D-1)\cdot\{Z_1 - \Pr(Z_1=1)\}|M,C]}{E[D\cdot\{Z_1 - \Pr(Z_1=1)\}|M,C]}$ is a weighting function and $\pi(X) = \Pr(Z_1 = 1|X)$ is the instrument propensity score. C is a nonparametric control function, see for instance Imbens and Newey (2009), which identifies the distribution of V and thus controls for mediator endogeneity. This approach as well as fully parametric IV-based estimation is implemented in the 'causalweight' package by Bodory and Huber (2018).

Finally, Miquel (2002) considers a nonparametric framework with binary instruments Z_1, Z_2 and endogenous variables D, M. She discusses the identification of controlled direct effects (and dynamic treatment effects in general) for subpopulations defined upon the compliance in either endogenous variable.

4.4 Single instrument approaches

Several identification strategies suggested in the mediation literature rely on a single instrument. They therefore consider less general problems or methods than the double IV approach discussed Section (4.3), but might be more feasible in practice, because credible instruments are typically rare in empirical data. One strand of the literature assumes treatment take up to be exogenous such that it need not be instrumented, as under randomized treatment assignment with perfect compliance, and focusses on mediator endogeneity only. In this context, Robins and Greenland (1992) assume a 'perfect' instrument Z_2 for M that is strong enough to force the latter to take any desired value, just as direct exogenous manipulation of M. Such perfect instruments appear, however, hard to find in economic applications.

Imai, Tingley, and Yamamoto (2013) consider experiments with a randomized treatment (and perfect treatment compliance) and an 'imperfect' and discrete instrument Z_2. The latter induces a subset of subjects to change the assumably binary mediator M, the so-called mediator compliers (now defined with respect to the instrument rather than the treatment as it was the case in Section 2.3). In spite of the random assignment of Z_2 and D, natural direct and indirect effects are (in contrast to the controlled direct effect) not identified nonparametrically. While treatment randomization identifies the distributions of $M(d)$ and $Y(d, M(d))$ in the total population, making use of Z_2 to exogenously vary M identifies the distribution of $Y(d, m)$ for mediator compliers in either treatment group. However, the distribution of $Y(d, M(1-d))$ is not identified in any population, for the same reasons as discussed for the experimental parallel design in Section 4.2: For the group with $D = d$, it remains unknown which individuals satisfy $M(1-d) = m$, such that information on $Y(d, m)$ among mediator compliers can not be used for learning the distribution of $Y(d, M(1-d))$.

One might therefore turn to interval rather than point identification, see Imai, Tingley, and Yamamoto (2013) and Mattei and Mealli (2011) for bounding natural and principal strata direct effects, respectively, under a random D and a discrete instrument Z_2 for a binary M. A second approach is invoking functional form restrictions about the mediation model, for instance the absence of treatment-mediator interactions in (33) or assuming a continuous and powerful Z_2 as in Frölich and Huber (2017), see Section 4.3. An example for a fully parametric IV mediation model is provided in Chen, Chen, and Liu (2017). They investigate the direct effect of the supposedly random gender of second siblings (D) on first siblings' education (Y), as well as the indirect effect

via family size (M), which is instrumented by twin births (Z_2). Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007) use (arguably strong) parametric restrictions directly as instruments for the mediator. Specifically, treatment-covariate interactions serve as IVs for M while imposing the absence of treatment-mediator, mediator-covariate, and treatment-covariate interactions in the outcome model. See Dunn and Bentall (2007) and Albert (2008) for related approaches as well as Small (2012), who in contrast to the previous studies allows for some forms of effect heterogeneity.

A nonparametric approach to identification is the experimental cross-over design of Imai, Tingley, and Yamamoto (2013) already discusses in Section 4.2. In the first experiment, the treatment is randomized and the resulting mediator and outcome values are measured. In the second experiment, all subjects obtain the treatment opposite to their first period status and are randomly selected to have M encouraged by instrument Z_2 to equal the observed mediator value in the first period. Under specific IV assumptions and the absence of carry-over effects from the first experiment that contaminate potential outcomes in the second experiment, natural effects are identified for the mediator compliers.

A further strand of the literature assumes that both D and M are endogenous, but only uses one instrument for solving both issues. Yamamoto (2013) considers nonparametric identification when the treatment is endogenous and an instrument Z_1 is available for D, as in the standard LATE framework. To control for mediator endogeneity despite the absence of a second instrument, a latent ignorability assumption similar to Frangakis and Rubin (1999) is invoked with respect to the mediator. Conditional on the treatment compliance type (and observed covariates), the mediator is assumed to be exogenous, implying that the compliance type is a sufficient statistic to tackle unobserved confounders of the mediator. Brunello, Fort, Schneeweis, and Winter-Ebmer (2016) assume an instrument Z_1 for the treatment, while they combine a selection on observables assumption with specific structural restrictions to control for mediator endogeneity. Joffe, Small, Have, Brunelli, and Feldman (2008) assumes a single instrument that jointly affects the treatment and the mediator and show identification under arguably strong parametric restrictions. Dippel, Gold, Heblich, and Pinto (2017) identify direct and indirect effects by a single instrument based on the assumption that the unobserved confounders of the treatment-mediator and mediator-outcome relations are independent of each other. They apply this approach to investigate whether the effect of import exposure on voting in Germany is mediated by labor market adjustments.

4.5 Mediation based on difference-in-differences

In contrast to standard treatment evaluation, approaches based on so-called natural experiments have rarely been considered in causal mediation analysis. One exception is Deuchert, Huber, and Schelker (2017), who use a difference-in-differences strategy to identify direct and indirect effects within principal strata, see Section2.3. They to this end assume a randomized treatment, monotonicity of the (binary) mediator in the treatment, and particular common trend assumptions on mean potential outcomes across principal strata. The latter imply that mean potential outcomes under specific treatment and mediator states evolve in the same way over time across specific subpopulations. It is for instance imposed that the mean potential outcomes without treatment and mediator across the strata (i) of the never mediated and (ii) of those whose mediator complies with the treatment assignment follow a common trend over time (while outcome levels might differ):

$$E[Y_{t=1}(0,0) - Y_{t=0}(0,0)|M(1) = M(0) = 0] = E[Y_{t=1}(0,0) - Y_{t=0}(0,0)|M(1) = 1, M(0) = 0], \quad (41)$$

where $t = 0$ denotes a pre-treatment period, i.e. prior to the measurement of D and M, and $t = 1$ the post-mediator period in which the effects on the outcome are to be estimated. Under this assumption, the direct principal strata effect on the never mediated is obtained by

$$\begin{aligned} & E[Y_{t=1}(1,0) - Y_{t=1}(0,0)|M(1) = M(0) = 0] \quad (42) \\ = \; & E[Y_{t=1}|D = 1, M = 0] - E[Y_{t=0}|D = 1, M = 0] - E[Y_{t=1}|D = 0, M = 0] - E[Y_{t=0}|D = 0, M = 0]. \end{aligned}$$

Invoking further common trend and effect homogeneity assumptions eventually permits identifying direct and indirect effects on mediator compliers and (if identification is obtained in all strata) on the total population, see Deuchert, Huber, and Schelker (2017).

5 Extensions

We subsequently discuss some extensions of the standard framework with sequential conditional independence in Section (3.1) to adapt the analysis to target groups different than the total population, functions of outcomes, and multivalued (rather than binary) treatments. We also consider issues of measurement error in the mediator and sample selection due to missing outcomes.

5.1 Different populations and outcome functions

While most mediation studies evaluate effects on the total population, alternative target groups as for instance the treated population might be of policy interest, too. In analogy to the concept of weighted treatment effects in Hirano, Imbens, and Ridder (2003), direct and indirect effects are identified for a particular target group under Assumptions 1 to 3 by reweighing observations according to the distribution of observables X in the target group. To this end, we define $\omega(X)$ to be a well-behaved weighting function depending on X. Including $\frac{\omega(X)}{E[\omega(X)]}$ in the expectation operators of (17) identifies the direct and indirect effects on the target group:

$$\begin{aligned}
\theta_{\omega(X)}(d) &= E\left[\frac{\omega(X)}{E[\omega(X)]} \cdot \left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d|X)}\right], \\
\delta_{\omega(X)}(d) &= E\left[\frac{\omega(X)}{E[\omega(X)]} \cdot \frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right] \quad (43)
\end{aligned}$$

The subscript $\omega(X)$ highlights that the effects refer to a specific target group defined by a function of X. Important examples are $\omega(X) = \Pr(D=1|X)$ and $\omega(X) = 1 - \Pr(D=1|X)$, yielding the direct and indirect effects on the treated and non-treated, respectively. The identifying assumptions might be weakened somewhat when considering these groups, see the discussion in Vansteelandt and VanderWeele (2012).

Furthermore, the identification results can be extended to well-behaved functions of the outcome, rather than Y itself. For instance, replacing Y by the indicator function that Y is not larger than some value a, $I\{Y \leq a\}$, in (13), (14), or (15) yields the cumulative distribution function of potential outcomes. The inversion of the latter allows identifying direct and indirect quantile treatment effects at specific quantiles of the (unconditional) potential outcome distribution. See also Schmidpeter (2018), who identifies direct and indirect quantile treatment effects on the treated based on a weighted version of the so-called check function, a loss function for quantile regression originally suggested by Koenker and Bassett (1978). He applies the method to decompose wage losses after displacement into an indirect effect operating via interim jobs and a direct channel accounting for any other factors. Further contributions in the context of quantile mediation analysis assess direct and indirect effects at conditional (rather than unconditional) quantiles of the outcome and/or the mediator, see Dominici, Zeger, Parmigiani, Katz, and Christian (2006), Imai, Keele, and Tingley (2010), Shen, Chou, Pentz, and Berhane (2014), Bind, VanderWeele, Schwartz, and Coull (2017), and Geraci and Mattei (2017).

Conditional direct and indirect quantile effects do, however, not necessarily add up to the total effect at some unconditional quantile of potential outcomes in the population.

5.2 Multivalued treatments

Most contributions to causal mediation analysis that rely on the potential outcome framework assume a binary treatment. In empirical applications, treatments might, however, also be multivalued discrete, representing e.g. alternative labor market programs (e.g. no training, job application training, language course, computer course), or continuous, for instance time spent in training. For a discussion on multivalued discrete treatments in linear models, see Hayes and Preacher (2014). In nonparametric models, the evaluation of multivalued discrete treatments can be straightforwardly implemented based on the results of Section 3.2, be they categorical or ordered. For any pair of values $d \neq d'$ in the discrete support of D, the expressions for potential outcomes provided in (13), (14), and (15) directly apply, given that Assumptions 1 to 3 are satisfied with respect to non-binary d and d'. Direct and indirect effects based on pairwise comparisons of appropriately defined potential outcomes (using, for instance, $d = 1$, $d' = 0$, or $d = 2$, $d' = 0$, or $d = 2$, $d' = 1$ if $d, d' \in \{0, 1, 2\}$) are then identified analogously to the binary treatment case.

If D is continuous, identification remains unchanged when compared to a binary treatment for the case of a linear model as described by equations (9) to (12). For the nonparametric case, however, the results of Section (3.2) need to be modified to account for the fact that continuous treatments do (in contrast to discrete ones) not have mass points. Hsu, Huber, Lee, and Pipoz (2018) adapt weighting-based identification of potential outcomes as given by (14) to a continuously distributed treatment, assuming that Assumptions 1 to 3 hold with respect to the latter. Specifically, any indicator functions for treatment values are replaced by kernel functions. Treatment propensity scores are substituted by conditional density functions, also known as generalized propensity score, see Hirano and Imbens (2005) and Imai and van Dyk (2004). For any pair of treatment values $d \neq d'$, the expression for the mean potential outcome becomes

$$E[Y(d, M(d'))] \;=\; \lim_{h \to 0} E\left[\frac{Y\omega(D; d, h)}{E[\omega(D; d, h)|M, X]} \cdot \frac{E[\omega(D; d', h)|M, X]}{E[\omega(D; d', h)|X]}\right]. \tag{44}$$

The weighting function $\omega(D; d, h) = K\left((D - d)/h\right)/h$, where K is a symmetric second order kernel function assigning more weight to observations closer to d, and h is a bandwidth operator.

For h going to zero, i.e. $\lim_{h\to 0}$, the conditional means $E[\omega(D;d',h)|X]$ and $E[\omega(D;d',h)|M,X]$ correspond to the generalized propensity scores $f(D=d|X)$ and $f(D=d|M,X)$, respectively.

We refer to Hsu, Huber, Lee, and Pipoz (2018) for a discussion of weighting-based estimation using either parametric or nonparametrically estimated generalized propensity scores, with the former approach being available in 'causalweight' package by Bodory and Huber (2018). Alternatively to weighting, imputation-based estimation of potential outcomes as suggested by Vansteelandt, Bekaert, and Lange (2012) can be applied to both multivalued discrete and continuous treatments. This has been implemented in the 'medflex' package by Steen, Loeys, Moerkerke, and Vansteelandt (2017b). Finally, also the regression-based 'mediation' package by Tingley, Yamamoto, Hirose, Imai, and Keele (2014) allows for non-binary treatments in both linear and nonlinear models.

5.3 Mismeasured mediators and missing outcomes

VanderWeele (2012b) points to a subtle but important issue concerning a too 'coarse' measurement of the mediator, as it likely occurs in many empirical problems. For illustration, we consider employment as mediator of interest. We further assume that the data provide a binary indicator for employment at the extensive margin, i.e. whether an individual provides zero or positive hours of work, while no information at the intensive margin is available, i.e. on the hours actually supplied by the working. This is an issue if the effect of the treatment on the mediator does not exclusively operate through the extensive margin, i.e. inducing a switch from not working to working or vice versa, but also at the intensive margin, i.e. inducing some working individuals to change the hours worked. In the latter case, the measured indirect effect only accounts for impacts related to treatment-induced mediator changes at the extensive margin. In contrast, any changes at the intensive margin that occur without switches at the extensive margin are credited to the direct effect. This issue should be taken into account when interpreting the effects in the presence of a coarse measure of the mediator. In general, any form of mediator mismeasurement entails problems for properly assessing direct and indirect effects. Heckman and Pinto (2015) provide a procedure to correct for measurement error in the mediator under specific structural assumptions, requiring for instance that the mediation model is parametric and that the measurement error is independent of potential mediators.

A further complication arises when outcomes are only observed for a subpopulation due to

sample selection or outcome attrition. Such issues frequently occur in empirical applications, as for instance wage gap decompositions, where wages are only observed for those employed, or whenever individuals refuse to participate in follow up surveys measuring the outcome. Huber and Solovyeva (2018) discuss the identification of natural effects as well as the direct controlled effect when combining sequential conditional independence for the treatment and the mediator with either selection on observables or IV assumptions with respect to the outcome attrition process. The authors provide weighting-based expressions for the parameters of interest that make use of specific treatment, mediator, and/or selection propensity scores.

6 Conclusion

This paper provided a survey on methodological developments in causal mediation analysis, with a specific focus on applications in economics. After defining the direct and indirect effects of interest, we discussed identification and estimation based on conditional independence assumptions with respect to treatment and the mediator selection. We considered both static and dynamic confounding, i.e. when at least some of the confounders of the mediator-outcome relationship are a function of the treatment. We reviewed further evaluation strategies based on partial identification, randomization of the treatment and the mediator, instrumental variables for the treatment and/or mediator, and difference-in-differences. Finally, we sketched several extensions to the standard framework, like multivalued rather than binary treatments, target populations that differ from the total population, mismeasured mediators, and sample selection/outcome attrition.

References

Albert, J. M. (2008): "Mediation analysis via potential outcomes models," *Statistics in Medicine*, 27, 1282–1304.

Albert, J. M., and S. Nelson (2011): "Generalized causal mediation analysis," *Biometrics*, 67, 1028–1038.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472 (with discussion).

Avin, C., I. Shpitser, and J. Pearl (2005): "Identifiability of path-specific effects," in *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pp. 357–363, Edinburgh, UK.

Baron, R. M., and D. A. Kenny (1986): "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51, 1173–1182.

Bellani, L., and M. Bia (2018): "The long-run effect of childhood poverty and the mediating role of education," *forthcoming in the Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Bijwaard, G. E., and A. M. Jones (2018): "An IPW estimator for mediation effects in hazard models: with an application to schooling, cognitive ability and mortality," *Empirical Economics*, pp. 1–47.

Bind, M.-A., T. J. VanderWeele, J. D. Schwartz, and B. A. Coull (2017): "Quantile causal mediation analysis allowing longitudinal data," *Statistics in Medicine*, 36, 4182–4195.

Bodory, H., and M. Huber (2018): "The causalweight package for causal inference in R," *SES Working Paper 493, University of Fribourg*.

Brunello, G., M. Fort, N. Schneeweis, and R. Winter-Ebmer (2016): "The Causal Effect of Education on Health: What is the Role of Health Behaviors?," *Health Economics*, 25, 314–336.

Burgess, S., R. M. Daniel, A. S. Butterworth, and S. G. Thompson (2015): "Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways," *International Journal of Epidemiology*, 44, 484–495.

Cai, Z., M. Kuroki, J. Pearl, and J. Tian (2008): "Bounds on Direct Effects in the Presence of Confounded Intermediate Variables," *Biometrics*, 64, 695–701.

Chan, K., K. Imai, S. Yam, and Z. Zhang (2016): "Efficient nonparametric estimation of causal mediation effects," *working paper arXiv:1601.03501*.

Chen, S. H., Y. C. Chen, and J. T. Liu (2017): "The impact of family composition on educational achievement," *forthcoming in the Journal of Human Resources*.

Chen, Y.-T., Y.-C. Hsu, and H.-J. Wang (2018): "A Stochastic Frontier Model with Endogenous Treatment Status and Mediator," *forthcoming in the Journal of Business & Economic Statistics*.

Cochran, W. G. (1957): "Analysis of Covariance: Its Nature and Uses," *Biometrics*, 13, 261–281.

Conti, G., J. J. Heckman, and R. Pinto (2016): "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour," *The Economic Journal*, 126, F28–F65.

De Stavola, B. L., R. M. Daniel, G. B. Ploubidis, and N. Micali (2015): "Mediation Analysis With Intermediate Confounding: Structural Equation Modeling Viewed Through the Causal Inference Lens," *American Journal of Epidemiology*, 181, 64–80.

Deuchert, E., M. Huber, and M. Schelker (2017): "Direct and indirect effects based on difference-in-differences with an application to political preferences following the Vietnam draft lottery," *forthcoming in the Journal of Business & Economic Statistics*.

Dippel, C., R. Gold, S. Heblich, and R. Pinto (2017): "Instrumental variables and causal mechanisms: Unpacking the effect of trade on workers and voters," *National Bureau of Economic Research (No. w23209)*.

Dominici, F., S. L. Zeger, G. Parmigiani, J. Katz, and P. Christian (2006): "Estimating Percentile-Specific Treatment Effects in Counterfactual Models: A Case-Study of Micronutrient Supplementation, Birth Weight and Infant Mortality," *Journal of the Royal Statistical Society Series C*, 55, 261–280.

Dunn, G., and R. Bentall (2007): "Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments)," *Statistics in Medicine*, 26, 4719–4745.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA DP No. 4237*.

Flores, C. A., and A. Flores-Lagunes (2010): "Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects," *mimeo, University of Florida*.

Frangakis, C., and D. Rubin (1999): "Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes," *Biometrika*, 86, 365–379.

Frangakis, C., and D. Rubin (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Frölich, M., and M. Huber (2017): "Direct and Indirect Treatment Effects – Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society: Series B*, 79(5), 1645–1666.

Gelman, A., and G. W. Imbens (2013): "Why ask Why? Forward Causal Inference and Reverse Causal Questions," *NBER Working Paper No. 19614*.

Geraci, M., and A. Mattei (2017): "A novel quantile-based decomposition of the indirect effect in mediation analysis with an application to infant mortality in the US population," *arXiv working paper 1710.00720v2*.

Hayes, A. F., and K. J. Preacher (2014): "Statistical mediation analysis with a multicategorical independent variable," *British Journal of Mathematical and Statistical Psychology*, 67, 451–470.

Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103, 2052–2086.

Heckman, J. J., and R. Pinto (2015): "Econometric Mediation Analyses: Identifying the Sources of Treatment Effects from Experimentally Estimated Production Technologies with Unmeasured and Mismeasured Inputs," *Econometric Reviews*, 34, 6–31.

Hernandez-Diaz, S., E. F. Schisterman, and M. A. Hernan (2006): "The Birth Weight "Paradox" Uncovered?," *American Journal of Epidemiology*, 164(1115-1120).

Hirano, K., and G. W. Imbens (2005): *The Propensity Score with Continuous Treatments*chap. 7, pp. 73–84. Wiley-Blackwell.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Hong, G. (2010): "Ratio of mediator probability weighting for estimating natural direct and indirect effects," in *Proceedings of the American Statistical Association, Biometrics Section*, p. 2401–2415. Alexandria, VA: American Statistical Association.

Hong, G. (2015): *Causality in a Social World: Moderation, Meditation and Spill-over*. John Wiley & Sons, Ltd, West Sussex, UK.

Hong, G., J. Deutsch, and H. D. Hill (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., X. Qin, and F. Yang (2018): "Weighting-Based Sensitivity Analysis in Causal Mediation Studies," *Journal of Educational and Behavioral Statistics*, 43, 32–56.

Hsu, Y., M. Huber, Y. Lee, and L. Pipoz (2018): "Direct and indirect effects of continuous treatments based on generalized propensity score weighting," *SES Working Paper 495, University of Fribourg.*

Hsu, Y. C., M. Huber, and T. C. Lai (2018): "Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting," *forthcoming in the Journal of Econometric Methods.*

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

Huber, M. (2015): "Causal pitfalls in the decomposition of wage gaps," *Journal of Business and Economic Statistics*, 33, 179–191.

Huber, M., M. Lechner, and G. Mellace (2016): "The finite sample performance of estimators for mediation analysis under sequential conditional independence," *Journal of Business & Economic Statistics*, 34, 139–160.

Huber, M., M. Lechner, and G. Mellace (2017): "Why Do Tougher Caseworkers Increase Employment? The Role of Program Assignment as a Causal Mechanism," *The Review of Economics and Statistics*, 99, 180–183.

Huber, M., M. Lechner, and A. Strittmatter (2018): "Direct and indirect effects of training vouchers for the unemployed," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 441–463.

Huber, M., and A. Solovyeva (2018): "Direct and indirect effects
under sample selection and outcome attrition," *Working paper, University of Fribourg.*

Imai, K., L. Keele, and D. Tingley (2010): "A General Approach to Causal Mediation Analysis," *Psychological Methods*, 15, 309–334.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imai, K., D. Tingley, and T. Yamamoto (2013): "Experimental Designs for Identifying Causal Mechanisms," *Journal of the Royal Statistical Society, Series A*, 176, 5–51.

Imai, K., and D. A. van Dyk (2004): "Causal Inference With General Treatment Regimes," *Journal of the American Statistical Association*, 99, 854–866.

Imai, K., and T. Yamamoto (2013): "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," *Political Analysis*, 21, 141–171.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

Joffe, M. M., D. Small, T. T. Have, S. Brunelli, and H. I. Feldman (2008): "Extended Instrumental Variables Estimation for Overall Effects," *The International Journal of Biostatistics*, 4.

Judd, C. M., and D. A. Kenny (1981): "Process Analysis: Estimating Mediation in Treatment Evaluations," *Evaluation Review*, 5, 602–619.

Kaufman, J. S., R. F. MacLehose, and S. Kaufman (2004): "A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation," *Epidemiologic Perspectives & Innovations*, 1, 4.

Kaufman, S., J. Kaufman, R. MacLenose, S. Greenland, and C. Poole (2005): "Improved Estimation of Controlled Direct Effects in the Presence of Unmeasured Confounding of Intermediate Variables," *Statistics in Medicine*, 24, 1683–1702.

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

Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46(1), 33–50.

Lange, T., M. Rasmussen, and L. C. Thygesen (2014): "Assessing Natural Direct and Indirect Effects Through Multiple Pathways," *American Journal of Epidemiology*, 179(4), 513–518.

Lange, T., S. Vansteelandt, and M. Bekaert (2012): "A Simple Unified Approach for Estimating Natural Direct and Indirect Effects," *American Journal of Epidemiology*, 176, 190–195.

Lechner, M. (2009): "Sequential Causal Models for the Evaluation of Labor Market Programs," *Journal of Business and Economic Statistics*, 27, 71–83.

Lechner, M., and R. Miquel (2010): "Identification of the effects of dynamic treatments by sequential conditional independence assumptions," *Empirical Economics*, 39, 111–137.

Mattei, A., and F. Mealli (2011): "Augmented Designs to Assess Principal Strata Direct Effects," *Journal of Royal Statistical Society Series B*, 73, 729–752.

Miquel, R. (2002): "Identification of Dynamic Treatment Effects by Instrumental Variables," *University of St. Gallen Economics Discussion Paper Series*, 2002-11.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006): "Estimation of Direct Causal Effects," *Epidemiology*, 17, 276–284.

Pirlott, A. G., and D. P. MacKinnon (2016): "Design approaches to experimental mediation," *Journal of Experimental Social Psychology*, 66, 29 – 38.

Powdthavee, N., W. N. Lekfuangfu, and M. Wooden (2013): "The Marginal Income Effect of Education on Happiness: Estimating the Direct and Indirect Effects of Compulsory Schooling on Well-Being in Australia," *IZA Discussion Paper No. 7365*.

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

Robins, J. (1986): "A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

Robins, J. M. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects," in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.

Robins, J. M., and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Robins, J. M., M. A. Hernan, and B. Brumback (2000): "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560.

Robins, J. M., and T. Richardson (2010): "Alternative graphical causal models and the identification of direct effects," in *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, ed. by P. Shrout, K. Keyes, and K. Omstein. Oxford University Press.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B. (2004): "Direct and Indirect Causal Effects via Potential Outcomes," *Scandinavian Journal of Statistics*, 31, 161–170.

Schmidpeter, B. (2018): "Involuntary Unemployment and the Labor Market Returns to Interim Jobs," *working paper, Institute for Social and Economic Research, University of Essex.*

Shen, E., C.-P. Chou, M. A. Pentz, and K. Berhane (2014): "Quantile Mediation Models: A Comparison of Methods for Assessing Mediation Across the Outcome Distribution, Multivariate Behavioral Research," *Multivariate Behavioral Research*, 49, 471–485.

Simonsen, M., and L. Skipper (2006): "The Costs of Motherhood: An Analysis Using Matching Estimators," *Journal of Applied Econometrics*, 21, 919–934.

Sjölander, A. (2009): "Bounds on natural direct effects in the presence of confounded intermediate variables," *Statistics in Medicine*, 28, 558–571.

Small, D. S. (2012): "Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables," *Journal of Statistical Research*, 46, 91–103.

Steen, J., T. Loeys, B. Moerkerke, and S. Vansteelandt (2017a): "Flexible Mediation Analysis With Multiple Mediators," *American Journal of Epidemiology*, 186, 184–193.

Steen, J., T. Loeys, B. Moerkerke, and S. Vansteelandt (2017b): "Medflex: an R package for flexible mediation analysis using natural effect models," *Journal of Statistical Software*, 76.

Tchetgen Tchetgen, E. J. (2013): "Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis," *Statistics in Medicine*, 32, 4567–4580.

Tchetgen Tchetgen, E. J., and I. Shpitser (2012): "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis," *The Annals of Statistics*, 40, 1816–1845.

Tchetgen Tchetgen, E. J., and T. J. VanderWeele (2012): "On Identification of Natural Direct Effects when a Confounder of the Mediator is Directly Affected by Exposure," *forthcoming in Epidemiology.*

Ten Have, T. R., M. M. Joffe, K. G. Lynch, G. K. Brown, S. A. Maisto, and A. T. Beck (2007): "Causal mediation analyses with rank preserving models," *Biometrics*, 63, 926–934.

Tingley, D., T. Yamamoto, K. Hirose, K. Imai, and L. Keele (2014): "Mediation: R package for causal mediation analysis," *Journal of Statistical Software*, 59, 1–38.

Van der Laan, M. J., and M. L. Petersen (2008): "Direct Effect Models," *The International Journal of Biostatistics*, 4, 1–27.

VanderWeele, T. J. (2008): "Simple relations between principal stratification and direct and indirect effects," *Statistics & Probability Letters*, 78, 2957–2962.

VanderWeele, T. J. (2009): "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26.

VanderWeele, T. J. (2010): "Bias formulas for sensitivity analysis for direct and indirect effects," *Epidemiology*, 21, 540–551.

VanderWeele, T. J. (2012a): "Comments: Should Principal Stratification Be Used to Study Mediational Processes?," *Journal of Research on Educational Effectiveness*, 5(3), 245–249.

VanderWeele, T. J. (2012b): "Mediation analysis with multiple versions of the mediator," *Epidemiology*, 23, 454–463.

VanderWeele, T. J. (2016): "Mediation Analysis: A Practitioner's Guide," *Annual Review of Public Health*, 37(1), 17–32.

VanderWeele, T. J., and Y. Chiba (2014): "Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders," *Epidemiology, biostatistics, and public health*, 11.

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

VanderWeele, T. J., and S. Vansteelandt (2010): "Odds Ratios for Mediation Analysis for a Dichotomous Outcome," *American Journal of Epidemiology*, 172, 1339–1348.

VanderWeele, T. J., and S. Vansteelandt (2014): "Mediation analysis with multiple mediators," *Epidemiologic methods*, 2, 95–115.

Vansteelandt, S. (2009): "Estimating Direct Effects in Cohort and Case–Control Studies," *Epidemiology*, 20(6), 851–860.

Vansteelandt, S., M. Bekaert, and T. Lange (2012): "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects," *Epidemiologic Methods*, 1, 129–158.

Vansteelandt, S., and T. J. VanderWeele (2012): "Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions," *Biometrics*, 68, 1019–1027.

Wilcox, A. J. (2001): "On the importance-and the unimportance-of birthweight," *International Journal of Epidemiology*, 30, 1233–1241.

Wunsch, C., and R. Strobl (2018): "Identification of Causal Mechanisms Based on Between-Subject Double Randomization Designs," *IZA Discussion Paper No. 11626*.

Yamamoto, T. (2013): "Identification and Estimation of Causal Mediation Effects with Treatment Noncompliance," *unpublished manuscript, MIT Department of Political Science*.

Zheng, W., and M. J. van der Laan (2012): "Targeted Maximum Likelihood Estimation of Natural Direct Effects," *The International Journal of Biostatistics*, 8, 1–40.

Working Paper 500 january 2019

Author
Martin HUBER
University of Fribourg, Department of Economics, Bd. de Pérolles 90, 1700 Fribourg, Switzerland. Phone: +41 26 300 8274;
Email: martin.huber@unifr.ch; Website: http://www3.unifr.ch/appecon/de/lehrstuhl/team/prof

Abstract
Mediation analysis aims at evaluating the causal mechanisms through which a treatment or intervention affects an outcome of interest. The goal is to disentangle the total treatment effect into an indirect effect operating through one or several observed intermediate variables, the so-called mediators, as well as a direct effect reflecting any impact not captured by the observed mediator(s). This paper reviews methodological advancements with a particular focus on applications in economics. It defines the parameters of interest, covers various identification strategies, e.g. based on control variables or instruments, and presents sensitivity checks. Furthermore, it discusses several extensions of the standard mediation framework, such as multivalued treatments, mismeasured mediators, and outcome attrition.

Citation proposal
Martin Huber. 2019. «A review of causal mediation analysis for assessing direct and indirect treatment effects». Working Papers SES 500, Faculty of Economics and Social Sciences, University of Fribourg (Switzerland)

Jel Classification
C21

Keywords
Mediation, direct effect, indirect effect, sequential conditional independence, instrument.

Working Papers SES collection

Last published
493 Bodory H., Huber M.: The causalweight package for causal inference in R; 2018
494 Huber M., Imhof D.: Machine Learning with Screens for Detecting Bid-Rigging Cartels; 2018
495 Hsu Y.-C., Huber M., Lee Y.-Y.: Direct and indirect effects of continuous treatments based on generalized propensity score weighting; 2018
496 Huber M., Solovyeva A.: Direct and indirect effects under sample selection and outcome attrition; 2018
497 Huber M., Solovyeva A.: On the sensitivity of wage gap decompositions; 2018
498 Isakov D., Pérignon C., Weisskopf J.-P.: What if dividends were tax-exempt? Evidence from a natural experiment; 2018
499 Denisova-Schmidt E., Huber M., Prytula Y.: The effects of anti-corruption videos on attitudes towards corruption in a Ukrainian online survey. 2019

Catalogue and download links
http://www.unifr.ch/ses/wp
http://doc.rero.ch/collection/WORKING_PAPERS_SES

Publisher
Université de Fribourg, Suisse, Faculté des sciences économiques et sociales
Universität Freiburg, Schweiz, Wirtschafts- und sozialwissenschaftliche Fakultät
University of Fribourg, Switzerland, Faculty of Economics and Social Sciences

Bd de Pérolles 90, CH-1700 Fribourg
Tél.: +41 (0) 26 300 82 00
decanat-ses@unifr.ch www.unifr.ch/ses