Mediationsanalyse

Martin Huber

Inhaltsverzeichnis

Zusammenfassung

Die Mediationsanalyse befasst sich mit der Evaluation von kausalen Mechanismen, durch die eine Intervention (oder „treatment“) eine Ergebnisvariable beeinflusst. Ziel ist es, den gesamten Interventionseffekt in einen indirekten Effekt, der über eine oder mehrere beobachtbare Variablen, sogenannte Mediatoren, wirkt, sowie in einen direkten Effekt, der nicht über den/die Mediator(en) wirkt, zu zerlegen. Dieses Kapitel gibt einen Überblick zu den methodischen Fortschritten in der Mediationsanalyse mit besonderem Schwerpunkt auf Anwendungen in den Wirtschaftswissenschaften. Es definiert die interessierenden kausalen Parameter, diskutiert verschiedene Identifikationsstrategien, z. B. basierend auf Kontrollvariablen oder Instrumenten, und stellt mögliche Sensitivitätsanalysen vor. Darüber hinaus werden verschiedene Erweiterungen des Standardmodells erörtert, wie z. B. mehrfache Interventionen, Messfehler im Mediator und selektiver Datenschwund (wodurch die Ergebnisvariariable nicht für die gesamte Stichprobe beobachtbar ist).

M. Huber (✉)
Departement für Volkswirtschaftslehre, Universität Fribourg/Freiburg, Freiburg, Schweiz
E-Mail: martin.huber@unifr.ch

J. Gertheiss et al. (Hrsg.), *Moderne Verfahren der Angewandten Statistik*,
https://doi.org/10.1007/978-3-662-63496-7_21-1

1 Einleitung

Der Großteil der Studien zur Evaluierung von Interventionen („treatments“) oder politischen Maßnahmen beschränkt sich auf die Messung des Gesamteffekts auf ein bestimmtes Ergebnis, z. B. den durchschnittlichen Interventionseffekt („average treatment effect“ – ATE) einer Bildungsmaßnahme auf den Lohn. Gleichzeitig werden jedoch in einer Reihe von Studien Vermutungen über mögliche Mechanismen angestellt, die dem Gesamteffekt zugrunde liegen könnten. Es ist also nicht nur die „Wirkung einer Ursache“, d. h. der Interventionseffekt, der häufig von Interesse ist, sondern auch die „Ursache der Wirkung“, d. h. die Mechanismen, durch die der Gesamteffekt zustande kommt, siehe Gelman und Imbens (2013). Betrachten wir als Beispiel den Beschäftigungs- oder Einkommenseffekt eines aktiven Arbeitsmarktprogramms, wie z. B. einer Schulung. Neben dem Gesamteffekt würden die politischen Entscheidungsträger eventuell auch gerne wissen, inwieweit die Auswirkungen des Programms auf eine Veränderung der Suchanstrengungen, einen Anstieg des Humankapitals oder andere intermediäre Variablen zurückzuführen sind, die ihrerseits durch die Teilnahme an der Schulung beeinflusst werden. Ein besseres Verständnis der Mechanismen, die für die Auswirkung der Intervention verantwortlich sind, kann dazu beitragen, die Gestaltung solcher Programme zu verbessern.

Die kausale Mediationsanalyse hat zum Ziel, den Gesamteffekt einer Intervention aufzuspalten, und zwar in einen indirekten Effekt, der über eine oder mehrere intermediäre Variablen – gemeinhin als Mediatoren bezeichnet – wirkt, sowie in einen direkten Effekt, der alle kausalen Mechanismen umfasst, die nicht über die Mediatoren wirken. Selbst bei zufälliger Interventionszuweisung werden direkte und indirekte Effekte in der Regel nicht durch naives Kontrollieren für die Mediatoren ohne Berücksichtigung ihrer möglichen Endogenität identifiziert, da dies wahrscheinlich zu Selektionsverzerrungen führt, siehe Robins und Greenland (1992). Ein Großteil der früheren Arbeiten zur Mediationsanalyse (siehe z. B. Cochran 1957; Judd und Kenny 1981; Baron und Kenny 1986) stützte sich typischerweise auf lineare Modelle für die Mediator- und Ergebnisgleichungen und vernachlässigte häufig Endogenitätsprobleme.

Ein Beispiel dafür, wie ein unüberlegtes Kontrollieren für einen Mediator die Identifikation von Effekten beeinträchtigen kann, ist die Analyse der Auswirkungen des Rauchverhaltens von Müttern während der Schwangerschaft auf die postnatale Kindersterblichkeit, siehe Wilcox (2001) und Hernandez-Diaz et al. (2006). Im Allgemeinen wird in der empirischen Literatur ein positiver Zusammenhang zwischen Rauchen und Gesundheitsproblemen bei Kindern festgestellt. Einige Studien suggerieren jedoch, dass das Rauchen die Sterblichkeit von Kindern mit einem sehr geringen Geburtsgewicht zu reduzieren scheint. Wie in Hernandez-Diaz et al. (2006) beschrieben, ist dieses Paradoxon höchstwahrscheinlich darauf zurückzuführen, dass bei der Konditionierung auf das niedrige Geburtsgewicht als Mediator nicht für (wichtige) Störfaktoren („confounders“) kontrolliert wird. Wenn

nämlich Rauchen eine weniger tödliche Ursache für ein niedriges Geburtsgewicht ist als andere Faktoren wie etwa Geburtsfehler, dann ist die Sterblichkeit von Kindern mit geringem Geburtsgewicht aufgrund von Geburtsfehlern höher als bei Kindern mit geringem Geburtsgewicht aufgrund von Müttern, die während der Schwangerschaft rauchten.

Neuere Studien auf dem Gebiet der Mediationsanalyse verwenden allgemeinere Identifikationsansätze, häufig unter der Verwendung der (nichtparametrischen) Notation der hypothetischen Ergebnisse („potential outcomes"). Beispiele hierfür sind Robins und Greenland (1992), Pearl (2001), Robins (2003), Petersen et al. (2006), VanderWeele (2009), Imai et al. (2010b), Hong (2010), Albert und Nelson (2011), Imai und Yamamoto (2013), Tchetgen Tchetgen und Shpitser (2012) und Vansteelandt et al. (2012). In der überwiegenden Mehrheit der Studien basiert die Identifikation auf der Annahme, dass die Intervention und der Mediator bedingt exogen sind, wenn für beobachtbare Merkmale kontrolliert wird.

Solche oder ähnliche Annahmen finden auch in der empirischen Wirtschaftsforschung Anwendung. Siehe z. B. Simonsen und Skipper (2006), welche den direkten Lohneffekt einer Mutterschaft untersuchen, und Flores und Flores-Lagunes (2009), die den direkten Einkommenseffekt eines Weiterbildungsprogramms schätzen, wenn die Berufserfahrung als Mediator verwendet wird. Heckman et al. (2013) und Keele et al. (2015) untersuchen die kognitiven und nichtkognitiven Effekte eines Förderprogramms für Kinder im Vorschulalter. Conti et al. (2016) evaluieren die Wirkung zweier Vorschulprogramme auf die Gesundheit und gesundheitsbewusstes Verhalten, wobei Persönlichkeitsmerkmale als Mediatoren dienen. Bijwaard und Jones (2018) analysieren die Wirkung von Bildung auf die Sterblichkeit unter Berücksichtigung kognitiver Fähigkeiten als Mediator. Bellani und Bia (2019) untersuchen die Rolle von Bildung als Mediator für die Frage, wie die Erfahrung von Armut im Kindesalter den wirtschaftlichen Erfolg im Erwachsenenalter beeinflusst. Huber (2015) verwendet die Mediationsanalyse zur Dekomposition von Lohnunterschieden zwischen verschiedenen ethnischen Gruppen. Huber et al. (2017) analysieren den Beratungsprozess von Stellensuchenden in der Schweiz, insbesondere ob sich der Beschäftigungseffekt, den Sachbearbeiterinnen und Sachbearbeitern mit einem bestimmten Beratungsstil aufweisen, auf die Vermittlung bestimmter Weiterbildungsprogramme zurückführen lässt.

Manche Studien verwenden Instrumentenvariablen für die Identifikation, siehe z. B. Powdthavee et al. (2013), die den indirekten Effekt von Bildung über das Einkommen als Mediator auf die Lebenszufriedenheit schätzen. Brunello et al. (2016) untersuchen die Wirkung von Bildung auf die Gesundheit mit dem Gesundheitsverhalten als Mediator. Chen et al. (2017) analysieren die Auswirkungen der Familienzusammensetzung auf den Bildungsstand des erstgeborenen Kindes.

Dieses Kapitel gibt einen Überblick über die methodischen Fortschritte in der Literatur zur kausalen Mediation. Abschn. 2 definiert die kausalen Parameter: (natürliche) direkte und indirekte Effekte, den kontrollierten direkten Effekt und sogenannte Stratum-spezifische Effekte (für bestimmte Subpopulationen).

Abschn. 3 diskutiert die Identifikation und Schätzung unter sequentieller bedingter Exogenität der Intervention und des Mediators, wenn für beobachtete Variablen kontrolliert wird. Dabei werden zwei Szenarien unterschieden. Das erste Szenario unterstellt, dass derselbe Satz an Kontrollvariablen sowohl für die Exogenität der Intervention als auch des Mediators ausreichend ist. Das zweite, kompliziertere Szenario erlaubt es, dass manche Kontrollvariablen für den Mediator selbst von der Intervention beeinflusst werden („dynamic confounding"), so dass sich die Variablensätze für die Exogenität der Intervention und des Mediators zumindest teilweise unterscheiden. Abschn. 4 beleuchtet weitere Evaluationsstrategien, die auf partieller Identifikation, Randomisierung der Intervention und des Mediators, Instrumentenvariablen für die Intervention und/oder den Mediator und Differenz-in-Differenzen-Ansätzen basieren. Abschn. 5 erörtert mehrere Erweiterungen des Standardmodells: multiple statt binäre Interventionen, Zielpopulationen, die sich von der Gesamtpopulation unterscheiden, Messfehler in Mediatoren und endogene Stichprobenselektion/Nichtbeobachtbarkeit des Ergebnisses. Abschn. 6 gibt eine Zusammenfassung.

2 Kausale Parameter

In diesem Abschnitt werden verschiedene kausale Effekte vorgestellt, die in der Mediationsanalyse von Interesse sind: natürliche direkte und indirekte Effekte, kontrollierte direkte Effekte und die Stratum-spezifischen Effekte.

2.1 Natürliche direkte und indirekte Effekte

Die Mediationsanalyse beschäftigt sich typischerweise damit, den durchschnittlichen Interventionseffekt (average treatment effect – ATE) einer binären Variable, die mit D bezeichnet wird, auf eine Ergebnisvariable Y in einen direkten Effekt und einen indirekten Effekt, der über einen Mediator M wirkt, aufzuspalten. Der Mediator kann dabei diskret oder stetig, eine einzelne Variable oder ein Vektor von Variablen sein. Um natürliche direkte und indirekte Effekte zu definieren, wird die Notation der hypothetischen Ergebnisse („potential outcomes") verwendet, siehe z. B. Rubin (1974), wie sie im Rahmen der Mediationsanalyse z. B. in Ten Have et al. (2007) und Albert (2008) zur Anwendung kommt. Bezeichnen wir mit $Y(d)$ und $M(d)$ das hypothetische Ergebnis bzw. den hypothetischen Mediator, also die Werte, die diese Variablen annehmen würden, wenn die Intervention auf den Wert $d \in \{0, 1\}$ gesetzt würde. Wir verwenden also Großbuchstaben für Zufallsvariablen und Kleinbuchstaben für bestimmte Werte dieser Zufallsvariablen. Für jede Einheit wird nur eines der beiden hypothetischen Ergebnisse oder Mediatorenzustände beobachtet, da die tatsächlich realisierten Ergebnisse und Mediatoren anhand von $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$ und $M = D \cdot M(1) + (1 - D) \cdot M(0)$ gegeben sind.

Der ATE entspricht der Differenz der Mittelwerte der hypothetischen Ergebnisse mit und ohne Intervention, $\Delta = E[Y(1) - Y(0)]$. Um diesen Gesamteffekt in

einen direkten und einen indirekten Effekt (via M) aufzuspalten, wählen wir eine alternative Definition des hypothetischen Ergebnisses als Funktion sowohl der Intervention als auch des hypothetischen Mediators: $Y(d) = Y(d, M(d))$. Damit lässt sich der (durchschnittliche) direkte Effekt wie folgt definieren:

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

$\theta(d)$ entspricht der Veränderung des mittleren hypothetischen Ergebnisses, wenn die Intervention exogen variiert wird, der Mediator aber auf seinem hypothetischen Wert für $D = d$ verharrt, wodurch der indirekte Effekt über M ausgeschaltet wird. In analoger Weise ist der (durchschnittliche) indirekte Effekt definiert als

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

$\delta(d)$ entspricht der Veränderung der mittleren hypothetischen Ergebnisse, wenn der Mediator exogen auf seine hypothetischen Werte mit und ohne Intervention gesetzt wird, die Intervention aber mit $D = d$ konstant gehalten wird, um den direkten Effekt auszuschalten. Robins und Greenland (1992) und Robins (2003) bezeichnen diese Parameter als reine/totale direkte und indirekte Effekte und Pearl (2001) bezeichnet sie als natürliche direkte und indirekte Effekte und letztere Notation wird nachfolgend beibehalten.

Der ATE ist die Summe der natürlichen direkten und indirekten Effekte, die für entgegengesetzte Interventionszustände definiert sind:

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

Dieser Zusammenhang folgt aus der Addition und Subtraktion von $E[Y(0, M(1))]$ bzw. $E[Y(1, M(0))]$ nach der ersten Gleichheit in (3). Darüber hinaus weist die Schreibweise $\theta(1)$, $\theta(0)$ und $\delta(1)$, $\delta(0)$ auf eine mögliche Effektheterogenität in Bezug auf die Intervention hin, d. h. auf Interaktionseffekte zwischen der Intervention und dem Mediator. Beispielsweise könnte die Effektivität der Stellensuche (M) für das Finden einer Beschäftigung (Y) davon abhängen, ob ein(e) Stellensuchende(r) ein Bewerbungstraining (D) absolviert hat. Oder anders ausgedrückt, der direkte Effekt des Trainings könnte vom Umfang der Stellensuche (M) abhängen.

Es liegt auf der Hand, dass keiner der Effekte ohne identifizierende Annahmen gemessen werden kann. Erstens wird für jedes Individuum nur eines von $Y(1, M(1))$ und $Y(0, M(0))$ beobachtet (d. h., es können nicht beide hypothetischen Ergebnisse gleichzeitig beobachtet werden), was als grundlegendes Problem der kausalen Inferenz bekannt ist. Zweitens werden $Y(1, M(0))$ und $Y(0, M(1))$ für

niemanden beobachtet, da Mediator- und Ergebniswerte nur für dieselbe faktische Intervention beobachtet werden können, nicht aber für entgegengesetzte Interventionszustände. Daher hängt die Identifikation von direkten und indirekten Effekten von der Verwendung exogener Variation in der Intervention und dem Mediator ab.

Es erscheint aufschlussreich, die interessierenden Effekte und Identifikationsprobleme im Kontext eines einfachen Strukturmodells zu erörtern, das aus einem System von linearen Gleichungen für das Ergebnis und den Mediator (ohne Interaktionseffekte zwischen Intervention und Mediator) besteht:

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

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

β_D, β_M bezeichnen die Koeffizienten von D und M in der Ergebnisgleichung, α_D ist der Koeffizient von D in der Mediatorgleichung und U und V sind unbeobachtete Terme. Durch exogenes Ein- und Ausschalten der Intervention in der Mediatorgleichung werden die hypothetischen Mediatoren identifiziert:

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

Das Ein- und Ausschalten der Intervention in der Ergebnisgleichung und das Einsetzen der hypothetischen Mediatoren impliziert insgesamt vier hypothetische Ergebnisse:

$$Y(1, M(1)) = \beta_D + \beta_M M(1) + U, \quad Y(0, M(0)) = \beta_M M(0) + U,$$
$$Y(1, M(0)) = \beta_D + \beta_M M(0) + U, \quad Y(0, M(1)) = \beta_M M(1) + U.$$

Durch entsprechende Subtraktion der hypothetischen Ergebnisse ergibt sich aus diesem einfachen Modell, dass die direkten Effekte homogen und gleich dem Koeffizienten von D in der Ergebnisgleichung sind: $\theta(1) = \theta(0) = \beta_D$. Ferner entspricht der indirekte Effekt dem Effekt von D auf M mal dem Effekt von M auf Y: $\delta(1) = \delta(0) = \beta_M \cdot \alpha_D$.

Die Schätzung der Gleichungen (4) und (5) mittels OLS („ordinary least squares“, der Methode der kleinsten Quadrate) zur Berechnung der interessierenden Effekte ist in den meisten empirischen Problemen vermutlich inkonsistent, da der unbeobachtete Term V mit der Intervention und U sowohl mit der Intervention als auch mit dem Mediator korreliert sein könnte. Außerdem unterstellt das Modell starke funktionale Annahmen: Linearität der Parameter und keine Interaktionseffekte, weder zwischen dem Mediator und der Intervention, noch zwischen den beobachteten Variablen und den unbeobachtbaren Termen. Dies schließt jede Form von Effektheterogenität aus. Um (zumindest) die Heterogenität von $\theta(d)$ und $\delta(d)$ hinsichtlich d zu berücksichtigen, kann die Gleichung (4) durch Interaktionen zwischen dem Mediator und der Intervention ergänzt werden:

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

wobei β_{DM} der Koeffizient des Interaktionsterms ist. Dies impliziert die folgenden hypothetischen Ergebnisse:

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

Auch dieses Modell ist jedoch recht statisch, da es immer noch Additivität zwischen den beobachteten und unbeobachteten Termen unterstellt, was bedeutet, dass die gesamten, direkten und indirekten Interventionseffekte über individuelle Merkmale hinweg konstant sind.

Das nachfolgende nichtparametrische Strukturmodell ist demgegenüber wesentlich allgemeiner:

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

wobei φ und ζ allgemeine Funktionen bezeichnen, so dass beliebige Nichtlinearitäten und Wechselwirkungen zwischen Variablen zulässig sind. Die hypothetischen Mediatoren und Ergebnisse sind in diesem Fall wie folgt definiert:

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

für $d, d' \in \{1, 0\}$. Offensichtlich lassen sich die natürlichen direkten und indirekten Wirkungen nicht so einfach durch (eine Kombination von) Koeffizienten darstellen wie im zuvor betrachteten einfachen linearen Modell. Andererseits ist das nichtparametrische Modell flexibler und damit robuster gegenüber Fehlspezifikationen. Die Identifikation ist jedoch nicht trivial, wenn die unbeobachtbaren Variablen (U, V) nicht statistisch unabhängig von (D, M) sind. Verschiedene Strategien zum Umgang mit solcher Interventions- und Mediatorenendogenität werden weiter unten diskutiert.

2.2 Kontrollierter direkter Effekt

Ein weiterer kausaler Parameter ist der sogenannte kontrollierte direkte Effekt. Er entspricht dem Unterschied in den mittleren hypothetischen Ergebnissen, wenn die Intervention exogen variiert und der Mediator für alle auf einen bestimmten Wert m gesetzt wird:

$$\gamma(m) = E[Y(1, m) - Y(0, m)], \qquad \text{für alle } m \text{ im Träger von } M. \tag{7}$$

Unter dem allgemeinen Strukturmodell in (6) ist das hypothetische Ergebnis wie folgt definiert: $Y(d, m) = \varphi(d, m, U)$. Im Gegensatz zum natürlichen direkten Effekt, der vom Mediatorenwert abhängt, welcher unter einem bestimmten Interventionszustand „natürlich“ folgen würde (und für verschiedene Individuen un-

terschiedlich sein kann), basiert der kontrollierte direkte Effekt auf der Auferlegung des gleichen Mediatorenwerts für alle Individuen. Die beiden kausalen Parameter sind nur dann äquivalent, wenn es keine Interaktionseffekte von D und M auf das Ergebnis Y gibt, siehe das lineare Ergebnismodell (4). Ob der natürliche oder der kontrollierte direkte Effekt von primärem Interesse ist, hängt von der empirischen Fragestellung ab. Angenommen, wir wollen den Effekt des ersten Programms in einer Sequenz von zwei Arbeitsmarktprogrammen (z. B. ein Bewerbungstraining, gefolgt von einem Computerkurs) auf die Beschäftigungswahrscheinlichkeit evaluieren. Der natürliche direkte Effekt evaluiert das erste Programm (D) konditional auf den Wert des zweiten Programms (M), der im aktuellen institutionellen Kontext aus der (Nicht-)Teilnahme am ersten Programm folgen würde. Dies ist geeignet, um das erste Programm unter Status-quo-Zuweisungsregeln für das zweite Programm zu analysieren. Wenn solche Regeln jedoch seitens der Entscheidungsträger verändert werden können, erscheint auch die Evaluation des ersten Programms unter der Bedingung, dass die (Nicht-)Teilnahme am zweiten Programm auferlegt wird, zwecks einer effektiven Ausgestaltung von Programmsequenzen interessant. Daher kann der kontrollierte direkte Effekt dann für Entscheidungsträger relevant sein, wenn der Mediator (z. B. Programmteilnahme im Computerkurs) prinzipiell vorgeschrieben werden kann. Demgegenüber erscheint der natürliche direkte Effekt, bei dem der Mediatorenwert eine natürliche Reaktion auf die (Nicht-)Intervention ist, relevanter, wenn ein Vorschreiben des Mediators nicht durchführbar oder nicht erwünscht ist. Siehe Pearl (2001) für eine weitere Diskussion über die beschreibende und vorschreibende Art von natürlichen und kontrollierten Effekten.

Es erscheint erwähnenswert, dass der kontrollierte, aber nicht der natürliche direkte Effekt in den Bereich der sogenannten dynamischen Interventionsseffekte („dynamic treatment effects“) zur Evaluation von unterschiedlichen Interventionssequenzen fällt, siehe z. B. Robins (1986), Robins et al. (2000) und Lechner (2009). Der kontrollierte direkte Effekt beruht auf dem Vergleich von Sequenzen mit unterschiedlichen Werten in der ersten Intervention, aber gleichen Werten in der zweiten Intervention. Er kann daher als ein Spezialfall der dynamischen Interventionseffekte angesehen werden. Zu guter Letzt ist zu beachten, dass es keinen indirekten Effekt gibt, der mit dem kontrollierten direkten Effekt einhergeht. Insbesondere entspricht die Differenz zwischen dem Gesamteffekt und dem kontrollierten direkten Effekt im Allgemeinen nicht irgendeinem indirekten Effekt, es sei denn, es gibt keine Interaktionseffekte zwischen Intervention und Mediator, siehe beispielsweise die Diskussion in Kaufman et al. (2004).

2.3 Stratum-spezifische Effekte

Die meisten Mediationsstudien evaluieren direkte und indirekte Effekte innerhalb der Grundgesamtheit (oder Gesamtpopulation). Ein kleinerer Teil der Literatur verwendet den Stratum-spezifischen („principal strata“) Ansatz von Frangakis und Rubin (2002) zwecks Evaluation in bestimmten Subpopulationen, die auf der Grundlage der hypothetischen Mediatorenzustände mit und ohne Intervention

definiert sind, siehe beispielsweise Rubin (2004). Unter der Annahme eines binären Mediators kann die Gesamtpopulation in vier Strata als Funktion der hypothetischen Mediatorenwerte der Individuen unterteilt werden. Für die „immer Mediierten" („always mediated") nimmt der hypothetische Mediator unabhängig von der Intervention immer den Wert eins an: $M(1) = M(0) = 1$. Für die „niemals Mediierten" („never mediated") ist der hypothetische Mediator immer null: $M(1) = M(0) = 0$. In beiden Gruppen stimmt der Gesamteffekt der Intervention mit dem direkten Effekt überein, da der Mediator von der Intervention nicht beeinflusst wird, so dass der indirekte Effekt per Definition gleich null ist:

$$\begin{aligned} &E[Y(1, M(1)) - Y(0, M(0))|M(1) = M(0) = 1] \\ &\quad = E[Y(1, 1) - Y(0, 1)|M(1) = M(0) = 1], \\ &E[Y(1, M(1)) - Y(0, M(0))|M(1) = M(0) = 0] \\ &\quad = E[Y(1, 0) - Y(0, 0)|M(1) = M(0) = 0]. \end{aligned} \tag{8}$$

In den restlichen beiden Subpopulationen (oder Strata) sind die hypothetischen Mediatoren nicht konstant, sondern reagieren auf die Intervention. Für die sogenannten „Übereinstimmer" („mediator compliers") stimmt der hypothetische Mediator mit der Intervention in dem Sinne überein, dass der Mediatorenwert immer dem Wert der Intervention entspricht: $M(1) = 1, M(0) = 0$. Für die sogenannten „Widersetzer" („mediator defiers") hingegen entspricht der Mediator niemals dem Wert der Intervention: $M(1) = 0, M(0) = 1$. Aufgrund der Variation des hypothetischen Mediators in Abhängigkeit der Intervention können indirekte Effekte für diese beiden Subpopulationen nicht ausgeschlossen werden. Deshalb deckt sich der Gesamteffekt auf Übereinstimmer und Widersetzer im Allgemeinen weder mit direkten noch mit indirekten Effekten. Aus diesem Grund ist der Stratum-spezifische Ansatz kritisiert worden, weil er in der Regel direkte und indirekte Effekte in Subpopulationen mit nichtkonstanten hypothetischen Mediatoren nicht aufschlüsselt. Ein weiterer Kritikpunkt ist der Fokus auf Subpopulationen anstelle der Gesamtpopulation, siehe die Diskussion in VanderWeele (2008, 2012a). Die Relevanz derartiger Subpopulationen sollte daher im jeweiligen empirischen Kontext klar motiviert werden.

3 Mediation unter sequentieller bedingter Unabhängigkeit

Dieser Abschnitt befasst sich mit der Evaluation direkter und indirekter Effekte auf der Grundlage der sequentiellen bedingten Unabhängigkeit der Intervention und des Mediators. Zunächst werden die Annahmen erörtert, unter denen derselbe Satz beobachteter Kontrollvariablen (Kovariaten) ausreicht, um sowohl für die Endogenität der Intervention als auch des Mediators zu kontrollieren. Ferner werden verschiedene Ansätze für die Identifikation und Schätzung der Effekte vorgestellt. Anschließend wird der Fall betrachtet, dass einige Kontrollvariablen des Mediators

selbst von der Intervention beeinflusst werden, was hinsichtlich der Identifikation eine größere Herausforderung darstellt.

3.1 Annahmen unter nicht von der Intervention beeinflussten Kovariaten

Ein großer Teil der Mediationsliteratur basiert auf sequentiellen bedingten Unabhängigkeitsannahmen in Bezug auf die Intervention und den Mediator. Die nachfolgende Diskussion beleuchtet Annahmen, unter denen derselbe Satz beobachteter Kovariaten, bezeichnet mit X, verwendet werden kann, um sowohl für die Endogenität der Intervention als auch des Mediators zu kontrollieren. Insbesondere darf X keine Funktion von D sein (das bedeutet formal: $X(d) = X$), was insbesondere der Fall ist, wenn die Kovariaten vor der Intervention gemessen werden. Abb. 1 veranschaulicht dies anhand eines kausalen Graphen, in dem die Pfeile kausale Effekte zwischen den verschiedenen Variablen darstellen. Es ist erwähnenswert, dass jede der Variablen D, M und Y zusätzlich durch unterschiedliche und statistisch unabhängige unbeobachtbare Variablen, die in Abb. 1 nicht dargestellt sind, kausal beeinflusst werden könnte. Zwecks Identifikation der direkten und indirekten Effekte muss jedoch ausgeschlossen werden, dass solche unbeobachtbaren Variablen zwei oder gar alle drei Variablen in (D, M, Y) gemeinsam beeinflussen.

Die erste Annahme setzt voraus, dass die Intervention konditional auf X bedingt unabhängig von allen hypothetischen Postinterventionsvariablen, d. h. den hypothetischen Mediatoren und Ergebnissen ist. Diese Annahme ist in der Literatur zur Politikevaluation als bedingte Unabhängigkeit oder Exogenität bekannt, siehe z. B. Imbens (2004).

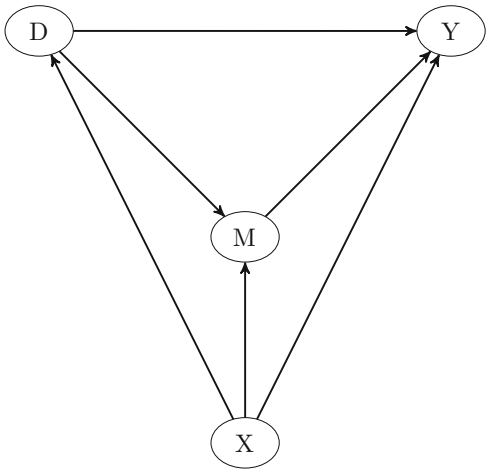

Abb. 1 Kausale Pfade unter bedingter Exogenität bei Vorinterventionskovariaten

Annahme 1 (bedingte Unabhängigkeit der Intervention). $\{Y(d', m), M(d)\} \perp D|X$ *für alle* $d', d \in \{0, 1\}$ *und* m *im Träger von* M.

Das Symbol $\perp$ steht für statistische Unabhängigkeit. Gemäß Annahme 1 gibt es keine unbeobachteten Störfaktoren, die unter der Bedingung auf X gleichzeitig die Intervention sowie den Mediator und/oder das Ergebnis beeinflussen. Bei nichtexperimentellen Daten hängt die Plausibilität dieser Annahme entscheidend vom Informationsgehalt von X ab. In experimentellen Daten ist die Annahme erfüllt, wenn die Intervention entweder innerhalb anhand von X definierten Subgruppen oder sogar unkonditional, d. h. unabhängig von X, randomisiert wird (im letzteren Fall ist sogar die stärkere Annahme $\{Y(d', m), M(d), X\} \perp D$ erfüllt).

Die zweite Annahme besagt, dass der Mediator bedingt unabhängig von den hypothetischen Ergebnissen ist, gegeben die Intervention und die Kovariaten:

Annahme 2 (bedingte Unabhängigkeit des Mediators). $Y(d', m) \perp M|D = d, X = x$ *für alle* $d', d \in \{0, 1\}$ *und* m, x *im Träger von* M, X.

Gemäß Annahme 2 gibt es keine unbeobachteten Störfaktoren, die gemeinsam den Mediator und das Ergebnis konditional auf D und X beeinflussen. Dies schließt im Allgemeinen postinterventionale Störfaktoren aus, die gleichzeitig den Mediator und das Ergebnis beeinflussen und nicht in X erfasst werden. Die Stärke dieser Annahme ist nicht zu unterschätzen, insbesondere wenn das Zeitfenster zwischen der Messung der Intervention und des Mediators groß ist, was den Ausschluss von postinterventionalen Störfaktoren in einer Welt mit zeitabhängigen Variablen (z. B. ein sich ändernder Gesundheitszustand) weniger plausibel macht.

Die dritte Annahme betrifft die bedingte Interventionswahrscheinlichkeit für alle möglichen Interventionszustände und postuliert, dass diese jeweils positiv ist, bekannt als „gemeinsamer Träger" („common support").

Annahme 3 (gemeinsamer Träger). $\Pr(D = d|M = m, X = x) > 0$ *für alle* $d \in \{0, 1\}$ *und* m, x *im Träger von* M, X.

Gemäß Annahme 3 ist die bedingte Wahrscheinlichkeit, die Intervention gegeben M, X zu erhalten oder nicht zu erhalten, im Folgenden als Propensity Score bezeichnet, größer als null. Sie impliziert (ist aber stärker als) die Standardannahme in der Politikevaluation, dass $\Pr(D = d|X = x) > 0$. Das bedeutet, dass die Intervention keine deterministische Funktion von X sein darf, da ansonsten die Identifikation nicht möglich ist, aufgrund des Mangels an hinsichtlich X vergleichbaren Individuen in unterschiedlichen Interventionszuständen. Gemäß dem Satz von Bayes impliziert Annahme 3 auch, dass $\Pr(M = m|D = d, X = x) > 0$, wenn M diskret ist oder dass die bedingte Dichte von M gegeben D, X größer als null ist, wenn M stetig ist. Gegeben X darf der Mediator keine deterministische Funktion der Intervention sein, da ansonsten keine hinsichtlich des Mediators vergleichbaren Individuen in unterschiedlichen Interventionszuständen existieren. Annahmen 1 bis 3 wurden häufig in kausalen Mediationsanalysen verwendet, siehe z. B. Imai

et al. (2010b), Tchetgen Tchetgen und Shpitser (2012) und Huber (2014). Siehe auch Pearl (2001), Petersen et al. (2006) und Hong (2010) für ähnliche Annahmen.

3.2 Identifikation unter nicht von der Intervention beeinflussten Kovariaten

Dem Artikel von Baron und Kenny (1986) folgend, evaluieren viele, insbesondere ältere Mediationsstudien direkte und indirekte Effekte anhand eines Systems linearer Gleichungen. Wenn man gemäß Annahmen 1 und 2 für die Kovariaten X kontrolliert, läuft dies auf das folgende Modell hinaus:

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

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

so dass die bedingten Erwartungen des Ergebnisses und des Mediators wie folgt definiert sind:

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

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

$\beta_0, \beta_D, \beta_M$ und β_X bezeichnen die Konstante (d. h. $E(U)$) und die Koeffizienten für D, M, X in der Ergebnisgleichung. α_0, α_D und α_X bezeichnen die Konstante (d. h. $E(V)$) und die Koeffizienten für D, X in der Mediatorengleichung. Wie in Abschn. 2.1 erörtert, werden die natürlichen direkten und indirekten Effekte durch β_D bzw. $\alpha_D \cdot \beta_M$ identifiziert. Dieses eher simplistische Modell, das keine Interaktionen zwischen D und M oder X, D und M zulässt, könnte durch die Einbeziehung eben solcher Interaktionsterme flexibler gestaltet werden, siehe die Diskussion in Imai et al. (2010a, b). Aufgrund der Linearitätsbeschränkungen könnten übrigens Annahmen 1 und 2 zu bedingten Unabhängigkeiten der Mittelwerte (anstatt der ganzen Verteilungen) abgeschwächt werden, während Annahme 3 überhaupt nicht erforderlich wäre.

Eine nichtparametrische Identifikation stützt sich hingegen nicht auf funktionale Restriktionen wie Linearität und erfordert deshalb im Allgemeinen alle drei Annahmen. Die nachfolgende Diskussion zeigt die Identifikation des mittleren hypothetischen Ergebnisses $E[Y(d, M(d'))]$ für $d, d' \in \{1, 0\}$ unter Annahmen 1 bis 3 durch in der Grundgesamtheit beobachtete Parameter. Das bedeutet, dass auch natürliche direkte und indirekte Effekte identifiziert werden. Zu diesem Zweck seien $f_{A=a}$ und $f_{A=a|B=b}$ die Dichtefunktionen einer Zufallsvariablen A, entweder unbedingt oder konditional auf eine oder mehrere Zufallsvariable(n) $B = b$. Ferner nehmen wir an, dass M und X stetig verteilt sind. Wenn M und/oder X hingegen diskret sind, dann sind die respektiven Dichten und Integrale im nachfolgenden Ausdruck ganz einfach durch Wahrscheinlichkeiten und Summen zu ersetzen.

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

Die erste Gleichheit folgt durch das Gesetz der iterierten Erwartungswerte und die Substitution der äußeren Erwartungen durch Integrale, die zweite von Annahmen 1 und 2, die dritte vom Satz von Bayes, die vierte von der Wahrscheinlichkeitstheorie und der Substitution der Integrale durch Erwartungen, die fünfte und sechste vom Gesetz der iterierten Erwartungswerte. Annahme 3 ist erforderlich, um zu gewährleisten, dass in keinem Ausdruck eine bedingte Wahrscheinlichkeit von null im Nenner steht, wodurch der respektive Ausdruck unendlich wäre.

Gleichung (13) liegt der sogenannten Mediationsformel zur Identifikation direkter und indirekter Effekte zugrunde, siehe z. B. Gleichungen (8) und (26) in Pearl (2001) und Theorem 1 in Imai et al. (2010b):

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

Gleichung (14) ist die Basis für die Identifikation anhand der inversen Wahrscheinlichkeitsgewichtung („inverse probability weighting“) mit dem Propensity Score, siehe Huber (2014):

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

Gleichung (15) schließlich basiert auf der Gewichtung mit der inversen Mediatorendichte, siehe Hong (2010) und Tchetgen Tchetgen und Shpitser (2012):

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

Wie in der konventionellen Politikevaluation können in der Mediationsanalyse also verschiedene Evaluationsstrategien zum Einsatz kommen, z. B. Regressionsmodelle für das Ergebnis und den Mediator, auf bedingten Wahrscheinlichkeiten/Dichten basierende Gewichtungsansätze oder sogar Kombinationen von Regressionen und Gewichtungen.

Der kontrollierte direkte Effekt wird bereits unter etwas schwächeren Annahmen identifiziert als die natürlichen Effekte. Dies folgt aufgrund der Tatsache, dass die Verteilung der hypothetischen Mediatoren nicht identifiziert werden muss. Daher kann Annahme 1 abgeschwächt werden, da eine bedingte Unabhängigkeit zwischen $M(d)$ und D nicht erforderlich ist, siehe z. B. die Diskussion in Petersen et al. (2006). Unter der Annahme, dass M diskret ist, entspricht das potenzielle Ergebnis $E[Y(d,m)]$ dem folgenden Ausdruck:

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

Die erste Gleichheit folgt vom Gesetz der iterierten Erwartungswerte, die zweite von Annahmen 1 (insbesondere aus der bedingten Unabhängigkeit der hypothetischen Ergebnisse und der Intervention) und 2, die dritte von der Wahrscheinlichkeitstheorie und die vierte vom Gesetz der iterierten Erwartungswerte. Dies bedeutet, dass der kontrollierte direkte Effekt sowohl durch Regression als auch durch Gewichtung identifiziert wird:

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

3.3 Schätzung unter nicht von der Intervention beeinflussten Kovariaten

In diesem Abschnitt werden verschiedene Schätzer für natürliche direkte und indirekte Effekte vorgestellt, von denen einige direkt die Identifikationsergebnisse aus Abschn. 3.2 anwenden. Nehmen wir zwecks Schätzung eine unabhängig gezogene Stichprobe der Größe n an, in der i den Index einer Beobachtung ($i \in \{1, ..., n\}$) bezeichnet und (Y_i, M_i, D_i, X_i) die Stichprobenrealisierungen der jeweiligen Zufallsvariablen (Y, M, D, X). Die Schätzung auf der Grundlage der Mediationsformel (16) erfordert „Plug-in"-Schätzungen für die bedingten mittleren Ergebnisse und die bedingten Mediatorendichten. Ein populärer Ansatz ist „g-computation", welcher auf Robins (1986) zurückgeht und diese Parameter anhand der Methode der maximalen Mutmaßlichkeit („maximum likelihood estimation" – MLE) berechnet. Seien $\hat{\mu}_Y(d, m, x)$, $\hat{f}(m|d, x)$ die Schätzungen des bedingten mittleren Ergebnisses $E[Y|D = d, M = m, X = x]$ und der bedingten Mediatorendichte $f_{M=m|D=d,X=x}$ (oder der bedingten Wahrscheinlichkeit $\Pr(M = m|D = d, X = x)$, wenn der Mediator diskret ist). Der g-computation-basierte Schätzer der direkten und indirekten Effekte entspricht

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

wobei $\hat{\theta}(d)$, $\hat{\delta}(d)$ Schätzungen der direkten und indirekten Effekte sind. Im Allgemeinen müssen beide parametrischen Modelle für $E[Y|D = 1, M = m, X = x]$ und $f(M = m|D = d, X = x)$ korrekt spezifiziert werden, damit das Schätzverfahren konsistent ist. Alternativ und wie in Imai et al. (2010b) beschrieben, können die Plug-in-Parameter $\hat{\mu}_Y(D, M, X)$ und $\hat{f}(M|D, X)$ nichtparametrisch geschätzt werden, um Fehlspezifikationen zu vermeiden. Dies kann jedoch bei endlichen Stichproben in der Praxis schwierig sein, wenn X und/oder M hochdimensional sind, ein Problem, das als „Fluch der Dimensionalität" bekannt ist.

Was die Gewichtungsausdrücke (17) betrifft, so können die natürlichen direkten und indirekten Effekte durch ihr normalisiertes Stichprobenanalogon geschätzt werden, wobei die Normalisierung gewährleistet, dass sich die den Schätzern zugrunde liegenden Gewichte in jeder Interventionsgruppe auf eins summieren. Der direkte Effekt bei Nicht-Intervention entspricht z. B.

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

wobei $\hat{\rho}(m, x)$ und $\hat{p}(x)$ die jeweiligen Schätzungen der Propensity Scores $\Pr(D = 1|M = m, X = x)$ und $\Pr(D = 1|X = x)$ bezeichnen. Ein praktischer Vorteil dieses Ansatzes besteht darin, dass er keine Schätzung von bedingten Mediatorendichten erfordert, was besonders relevant ist, wenn M mehrdimensional und/oder stetig verteilt ist. Propensity Scores der Intervention können mit Probit- oder Logit-Spezifikationen geschätzt werden, siehe z. B. Huber (2014) und Tchetgen Tchetgen (2013), wie es im „causalweight"-Paket von Bodory und Huber (2018) für die statistische Software „R" implementiert ist. Hsu et al. (2019) zeigen, dass auch für nichtparametrisch geschätzte Propensity Scores die Effektschätzung unter bestimmten Regularitätsbedingungen Wurzel-n-konsistent ist. Darüber hinaus ist dieser Ansatz asymptotisch semiparametrisch effizient, d. h. er hat die kleinstmögliche asymptotische Varianz und erreicht damit die in Tchetgen Tchetgen und Shpitser (2012) hergeleiteten semiparametrischen Effizienzschranken für kausale Mediationsanalysen. Insbesondere in kleineren Stichproben könnte jedoch der Fluch der Dimensionalität eine Rolle spielen.

Auch für den Gewichtungsausdruck (18) kann ein stichprobenanaloger Schätzer mit parametrisch oder nichtparametrisch geschätzten Plug-in-Parametern $\hat{p}(X)$ und $\hat{f}(M|D, X)$ konstruiert werden, siehe Hong et al. (2015). Lange et al. (2012) kombinieren die Gewichtung mit einer imputationsbasierten Schätzung der hypothetischen Ergebnisse. Ferner schlagen Chan et al. (2016) einen nichtparametrischen Gewichtungsansatz vor, der keine Plug-in-Schätzungen von Propensity Scores oder bedingten Mediatorendichten erfordert, sondern einen empirischen Kalibrierungsansatz anwendet. Dieser berechnet die Gewichte algorithmisch unter Verwendung spezifischer Momentengleichungen, die auf der Eigenschaft basieren, dass die wahren Gewichte Unterschiede in den Verteilungen von X und M zwischen den Gruppen mit und ohne Intervention ausgleichen. Die Methode ist Wurzel-n-konsistent und asymptotisch semiparametrisch effizient.

Tchetgen Tchetgen und Shpitser (2012) schätzen die Effekte anhand des Stichprobenanalogons der sogenannten effizienten Einflussfunktion („efficient influence function") zur Berechnung hypothetischer Ergebnisse, die auf der Schätzung bedingter mittlerer Ergebnisse, Mediatorendichten und Interventionswahrscheinlichkeiten beruht. Der direkte Effekt bei Nicht-Intervention entspricht z. B.

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

$$\left. + \frac{1 - D_i}{1 - \hat{p}(X_i)}\left[\hat{\mu}_Y(1, M_i, X_i) - \hat{\mu}_Y(0, M_i, X_i) - \hat{\theta}(0, X_i)\right] + \hat{\theta}(0, X_i)\right\}.$$

$\hat{\theta}(d, x)$ bezeichnet eine Schätzung von

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

welche man z. B. durch Regression von $\hat{\mu}_Y(1, M, X) - \hat{\mu}_Y(0, M, X)$ auf X unter Beobachtungen mit $D = d$ erhält.

Eine attraktive Eigenschaft dieses Schätzers ist seine „mehrfache Robustheit" in dem Sinne, dass er sogar dann konsistent ist, wenn nur bestimmte Unterspezifikationen des Modells korrekt sind. Es muss gelten, dass entweder (i) $E[Y|D, M, X]$ und $f_{M|D,X}$ (oder, alternativ und etwas schwächer, $E[Y|D, M, X]$ und $\theta(D, X)$), (ii) $E[Y|D, M, X]$ und $\Pr(D = 1|X)$ oder (iii) $\Pr(D = 1|X)$ und $f_{M|D,X}$ korrekt spezifiziert sind, siehe Tchetgen Tchetgen und Shpitser (2012) und Zheng und van der Laan (2012). Sind alle drei Bedingungen erfüllt, ist die mehrfach robuste Schätzung asymptotisch semiparametrisch effizient. Die Robustheit des Schätzers macht ihn auch für den Einsatz von Algorithmen des maschinellen Lernens („machine learning") interessant, um datenbasiert wichtige Störfaktoren unter den (potenziell zahlreichen) Kovariaten X zu identifizieren und für diese zu kontrollieren, wie in Farbmacher et al. (2022) implementiert. Dieselben Eigenschaften gelten übrigens auch für die „zielgerichtete Methode der maximalen Mutmaßlichkeit" („targeted maximum likelihood") von Zheng und van der Laan (2012), der die Schätzung des direkten oder indirekten Effekts auf der Grundlage vorausgehender Plug-in-Schätzungen von $E[Y|D, M, X]$, $f_{M|D,X}$ und $\Pr(D = 1|X)$ iterativ optimiert. Somit stellen beide Verfahren eine Weiterentwicklung gegenüber dem „doppelt robusten" Schätzer von Van der Laan und Petersen (2008) dar. Letzterer lässt eine Fehlspezifikation entweder des Ergebnis- oder des Interventionsmodells zu, erfordert aber eine korrekte Spezifikation der bedingten Mediatorendichte. Vansteelandt et al. (2012) schlagen einen weiteren Schätzer vor, der auf der Imputation potenzieller Ergebnisse basiert und ebenfalls eine doppelt robuste Eigenschaft besitzt. Er bleibt bei einer Fehlspezifikation des Ergebnisses konsistent, vorausgesetzt, die Modelle für den Mediator und die Intervention sind korrekt.

In der Literatur wurde eine Reihe weiterer (meist parametrischer) Schätzer vorgeschlagen, die Effektheterogenität in d erlauben. Für lineare Modelle erörtern z. B. VanderWeele und Vansteelandt (2009) und Preacher et al. (2007), welche Kombinationen von Koeffizienten den direkten und indirekten Effekten entsprechen, wenn Interaktionsterme zwischen der Intervention und dem Mediator sowie Kontrollvariablen berücksichtigt werden. Siehe auch VanderWeele und Vansteelandt (2010) für eine derartige Diskussion im Kontext von nichtlinearen binären Ergebnismodellen. VanderWeele (2009) analysiert sogenannte marginale strukturelle Modelle zur Modellierung mittlerer hypothetischer Ergebnisse. Der Ansatz kombiniert Regression mit Gewichtung, um kontrollierte direkte Effekte und natürliche Effekte zu schätzen. Alternativ modellieren Van der Laan und Petersen (2008) gezielt die interessierenden direkten Effekte anstatt der hypothetischen Ergebnisse.

Imai et al. (2010a) schlagen einen Simulationsansatz vor, der auf der linearen oder nichtlinearen Schätzung der Mediator- und Ergebnismodelle und der Simulation hypothetischer Mediatoren und Ergebnisse (gemäß der angenommenen Modelle) zwecks der Berechnung direkter und indirekter Effekte beruht. Diese Methode ist im „mediation"-Paket für die Software R von Tingley et al. (2014) verfügbar.

Als Alternative implementiert das Paket „medflex“ von Steen et al. (2017b) die Imputation hypothetischer Ergebnisse wie in Vansteelandt et al. (2012) und die Gewichtung wie in Lange et al. (2012). Siehe Hong (2015) und VanderWeele (2016) für eine weitere Übersicht über alternative Schätzverfahren und Huber et al. (2016) für eine Simulationsstudie zum Verhalten verschiedener Schätzer in endlichen Stichproben.

3.4 Effekte unter postinterventionalen Kovariaten und mehreren Mediatoren

In vielen empirischen Fragestellungen erscheint es unwahrscheinlich, dass vor der Intervention gemessene Kovariaten ausreichen, um für die Endogenität des Mediators M zu kontrollieren, da dieser eine postinterventionale Variable darstellt. So wie die Kontrollvariablen der Intervention typischerweise kurz vor der Intervention gemessen werden, erscheint es sinnvoll, für mögliche Störfaktoren, die den Mediator und das Ergebnis beeinflussen, kurz vor der Messung des Mediators zu kontrollieren. Es erscheint in diesem Fall wahrscheinlich, dass zumindest manche dieser Störfaktoren durch die Intervention beeinflusst werden (insbesondere, wenn die zeitliche Distanz zwischen D und M beträchtlich ist), so dass sie selbst Mediatoren sind, die den interessierenden Mediator M beeinflussen. Aus diesem Grund vermutet Robins (2003), dass Annahmen 1 und 2, die implizieren, dass Störfaktoren des Mediators keine Funktion von D sind, von begrenzter praktischer Relevanz sind.

In der nachfolgenden Diskussion untersuchen wir den Fall, dass die Intervention nachgelagerte (also postinterventionale) Kovariaten beeinflussen kann, welche sowohl einen Effekt auf den Mediator als auch auf das Ergebnis haben können und mit W bezeichnet werden. Um diese Erweiterung in der Notation zu berücksichtigen, drücken wir fortan die hypothetischen Mediatoren und Ergebnisse auch als Funktionen von W aus: $M(d) = M(d, W(d))$ und $Y(d, M(d)) = Y(d, M(d, W(d)), W(d))$, wobei $W(d)$ ein Vektor hypothetischer postinterventionaler Kovariaten unter $D = d$ ist. Der direkte und der indirekte Effekt entsprechen dann

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

$\theta^M(d)$ inkludiert jeglichen Effekt von D auf Y, der nicht über M wirkt. Zusätzlich zu dem inhärent direkten kausalen Pfad von D nach Y beinhaltet dies auch den Kausalmechanismus von D über W nach Y. Vor diesem Hintergrund könnte der Begriff des direkten Effekts ambivalent erscheinen. $\delta^M(d)$ hingegen umfasst jeglichen Effekt über M, der entweder direkt von D ausgeht oder einen „Umweg“ über W nimmt. Dieser Effekt berücksichtigt, dass die Intervention D den Mediator M entweder direkt oder indirekt über W beeinflussen kann.

Alternativ könnten wir auch den pfadspezifischen indirekten Effekt untersuchen, der direkt von D nach M geht, aber nicht über W wirkt:

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

$\delta^{Mp}(d)$ stellt einen partiellen indirekten Effekt dar, bei dem W auf seinem von d implizierten Niveau verharrt, so dass kausale Mechanismen von D über W nach M ausgeschaltet sind. Der indirekte Effekt $\delta^{M}(d)$ erscheint interessanter als der partielle indirekte Effekt $\delta^{Mp}(d)$, ist aber leider auch schwieriger zu identifizieren, wie weiter unten erläutert wird. Ferner könnten wir uns auch für die gemeinsamen indirekten Effekte über M und/oder W interessieren, die wie folgt definiert sind:

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

$\delta^{M,W}(d)$ umfasst jeglichen kausalen Mechanismus, der von D nach Y führt und entweder über M oder W oder beide wirkt. Dementsprechend entspricht der direkte Effekt $\theta^{M,W}(d)$ der Wirkung von D auf Y, die weder über M noch über W läuft (was ihn von $\theta^{M}(d)$ unterscheidet, der auch indirekte Mechanismen über W, aber nicht M inkludiert):

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

3.5 Identifikation unter postinterventionalen Kovariaten und mehreren Mediatoren

Die kausalen Parameter $\delta^{M,W}(d)$ und $\theta^{M,W}(d)$ können basierend auf den Annahmen 1 und 2 und den Ergebnissen (13), (14) und (15) in Abschn. 3.2 identifiziert werden, wenn M in allen Ausdrücken durch (M, W) ersetzt werden kann. Dies erlaubt übrigens unbeobachtete Faktoren, die gleichzeitig W und M beeinflussen, da Annahmen 1 und 2 Störfaktoren innerhalb der Mediatoren nicht ausschließen. Die nachfolgende Diskussion konzentriert sich deshalb auf die Identifikation von $\delta^{Mp}(d)$, $\delta^{M}(d)$ und $\theta^{M}(d)$. Die präsentierten Annahmen sind in Abb. 2 erfüllt, in der die Intervention die beobachteten Störfaktoren W des Mediators M und des Ergebnisses Y beeinflusst.

Die Verwendung von W als Kontrollvariablen für M impliziert, dass es keine unbeobachteten Störfaktoren geben darf, die gleichzeitig W einerseits sowie M und/oder Y andererseits beeinflussen. Diese bedingte Unabhängigkeit von W stellt hohe Anforderungen an X. Nicht nur muss X alle präinterventionalen Variablen enthalten, die D und Y, D und M oder M und Y beeinflussen, sondern auch alle Faktoren (außer D), die W und M oder W und Y beeinflussen. Andernfalls würde die Konditionierung auf W eine Assoziation zwischen D und den Faktoren, die M oder Y beeinflussen, und somit die Endogenität der Intervention zur Folge haben. Nehmen wir als Beispiel an, dass Y als Vermögen, M als Beschäftigung, D als Hochschulbildung und W als vor dem Mediator gemessener Gesundheitszustand,

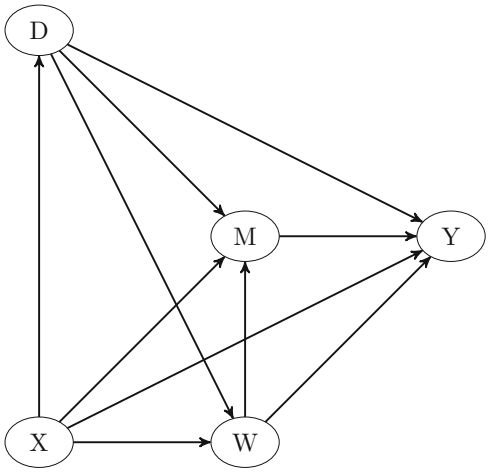

Abb. 2 Kausale Pfade mit vor und nach der Intervention gemessenen Kovariaten

welcher Y beeinflusst und eine Funktion von D sowie dem Gesundheitszustand vor der Intervention ist, definiert sind. In diesem Fall muss X den Gesundheitszustand vor der Intervention einschließen, wenn Letzterer Y nicht nur über W, sondern auch direkt beeinflusst. Dies erscheint plausibel, da das Vermögen in der Regel durch frühere Einkommensströme mitbestimmt wird, die ihrerseits durch den früheren Gesundheitszustand beeinflusst werden können. Die Annahmen 4, 5 und 6 postulieren formale Bedingungen für die Identifikation von $\delta^{Mp}(d)$ unter Einbeziehung postinterventionaler Kovariaten.

Annahme 4 (bedingte Unabhängigkeit der Intervention). $\{Y(d'', m, w'), M(d', w), W(d)\} \perp D|X = x$ für alle $d'', d', d \in \{0, 1\}$ und m, w', w, x im Träger von M, W, X. Ähnlich wie bei Annahme 1 erfordert Annahme 4, dass D bedingt unabhängig von hypothetischen postinterventionalen Variablen ist, d. h. von hypothetischen Ergebnissen, Mediatoren und nachgelagerten Kovariaten, die den Mediator und das Ergebnis beeinflussen.

Annahme 5 (bedingte Unabhängigkeit der postinterventionalen Kovariaten).

(a) $\{Y(d'', m, w'), M(d', w)\} \perp W|D = d, X = x$ für alle $d'', d', d \in \{0, 1\}$ und m, w', w, x im Träger von M, W, X,
(b) $Y(d', m, w') \perp M|D = d, W = w, X = x$ für alle $d', d \in \{0, 1\}$ und m, w', w, x im Träger von M, W, X.

Annahme 5(a) besagt, dass W konditional auf X und D bedingt unabhängig von den potenziellen Mediatoren und Ergebnissen ist. Dies impliziert, dass alle vor der Intervention gemessenen Variablen, die sowohl W als auch M oder Y beeinflussen, in X enthalten sind. Annahme 5(b) ist etwas schwächer als Annahme 2, da sie

die bedingte Unabhängigkeit des Mediators und der hypothetischen Ergebnisse konditional auf X und W (und nicht auf X allein) unterstellt.

Annahme 6 (gemeinsamer Träger). $\Pr(D = d|M = m, W = w, X = x) > 0$ *für alle* $d \in \{0, 1\}$ *und* m, w, x *im Träger von* M, W, X.

Annahme 6 ist etwas stärker als Annahme 3, da sie verlangt, dass der gemeinsame Träger hinsichtlich der Intervention auch dann gegeben ist, wenn neben M und X auch noch auf W konditioniert wird.

Annahmen 4–6 identifizieren den partiellen indirekten Effekt, z. B. anhand des folgenden Gewichtungsausdrucks unter Verwendung von drei verschiedenen Propensity Scores, siehe Huber (2014) (wo Annahme 5(a) allerdings etwas anders formuliert wird):

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

Siehe auch Steen et al. (2017a) für eine umfassendere Diskussion zur Identifikation und Schätzung von $\delta^{Mp}(d)$ und weiteren pfadspezifischen Effekten. Leider werden jedoch $\delta^{M}(d)$ und $\theta^{M}(d)$ nicht ohne weitere Annahmen identifiziert, wie es vom Unmöglichkeitstheorem in Avin et al. (2005) folgt. Die Identifikation von $E[Y(d, M(1 - d, W(1 - d)), W(d))]$ und den natürlichen Effekten erfordert nämlich, dass konditional auf X (i) die Verteilung von M für $D = d$ exogen angepasst wird, um mit jener von M für $D = 1 - d$ übereinzustimmen, während (ii) die Verteilung von W für $D = d$ konstant gehalten wird. Die gleichzeitige Durchführung von (i) und (ii) ist jedoch unmöglich, wenn W einen Effekt auf M aufweist. Aus diesem Grund erfordert die Identifikation von $\delta^{M}(d)$ und $\theta^{M}(d)$ notwendigerweise parametrische Annahmen.

Eine mögliche Annahme ist der Ausschluss von Interaktionen zwischen D und M, siehe Robins und Greenland (1992) und Robins (2003). Diese starke Bedingung erlaubt es sogar, die vorher formulierten Annahmen etwas abzuschwächen. Konkret müssen keine bedingten Unabhängigkeiten zwischen hypothetischen Ergebnissen und hypothetischen Mediatoren mehr unterstellt werden, die anhand von entgegengesetzten Interventionszuständen (also d und $1 - d$) definiert sind. Während Annahme 5(b) z. B. die bedingte Unabhängigkeit zwischen $Y(d', m, W(d'))$ und $M(d)$ auch für $d \neq d'$ impliziert, ist dies beim Ausschluss von Interaktionen zwischen D und M ausschließlich für $d = d'$ erforderlich. Es erscheint jedoch fragwürdig, ob viele empirische Fragestellungen existieren, in denen die bedingte Unabhängigkeit plausiblerweise für $d = d'$, aber nicht für $d \neq d'$ erfüllt ist, auch wenn in Robins und Richardson (2010) mehrere hypothetische Beispiele gegeben werden. Der Ausschluss von Interaktionen zwischen der Intervention und dem

Mediator erscheint hingegen sehr unattraktiv, da er die Heterogenität der Effekte einschränkt.

Diesbezüglich stellt die Methode von Imai und Yamamoto (2013) eine Verbesserung dar, da sie Interaktionen zwischen der Intervention und dem Mediator unter der Voraussetzung zulässt, dass die Interaktionseffekte in der Grundgesamtheit homogen sind. Huber (2014) schlägt eine verwandte Annahme vor, welche besagt, dass der durchschnittliche Interaktionseffekt zwischen der Intervention und dem Mediator konditional auf X und W homogen ist und dass die Ergebnisvariable linear in M ist. Als alternative Strategie zeigen Tchetgen Tchetgen und VanderWeele (2014), dass $\delta^M(d)$ und $\theta^M(d)$ (unter bestimmten bedingten Unabhängigkeitsannahmen) identifiziert werden, falls die durchschnittlichen Interaktionseffekte von W und M auf Y gleich null sind oder alle Variablen in W binär und monoton in D sind. Eine weitere mögliche Annahme besagt, dass die hypothetischen Kovariaten $W(1)$ und $W(0)$ unabhängig voneinander sind oder dass ihre Abhängigkeit einer bekannten Verteilung folgt, siehe Robins und Richardson (2010) und Albert und Nelson (2011) (Letztere untersuchen sogar allgemeinere Modelle mit Sequenzen von mehreren Mediatoren). Für eine Diskussion der Identifikation in parametrischen Mediationsmodellen verweisen wir auf De Stavola et al. (2015). Letztere zeigen, dass die Restriktion, dass es keine unbeobachteten Störfaktoren gibt, welche gleichzeitig W und Y beeinflussen, unter bestimmten Annahmen abgeschwächt werden kann.

Während einige der diskutierten Annahmen mehr oder weniger attraktiv erscheinen als andere, so teilen sie alle den Vorbehalt, dass sie dem Mediationsmodell spezifische Einschränkungen auferlegen, und somit dessen Allgemeingültigkeit reduzieren. Zu guter Letzt erscheint noch der Fall erwähnenswert, dass W eine Funktion von D ist, aber im Gegensatz zu Abb. 2 nicht M beeinflusst, sondern ausschließlich Y. Dann besteht das Mediationsmodell aus zwei unabhängigen indirekten Mechanismen, die über M bzw. W wirken. In diesem Szenario kann einer der beiden natürlichen indirekten Effekte identifiziert werden, indem die Annahmen 1 bis 3 des Abschn. (3.1) sowohl für M als auch für W separat angewandt werden, siehe Imai und Yamamoto (2013) und Lange et al. (2014) für weitere Details. Siehe auch VanderWeele und Vansteelandt (2014) für eine umfassende Diskussion verschiedener Mediationsmodelle mit mehreren Mediatoren, einschließlich der hier betrachteten.

Unsere Diskussion konzentrierte sich bislang auf natürliche Effekte. Während eine nichtparametrische Identifikation von $\delta^M(d)$ und $\theta^M(d)$ anhand von postinterventionalen Kovariaten, die von der Intervention beeinflusst werden, unmöglich ist, wird der kontrollierte direkte Effekt sogar unter etwas schwächeren Bedingungen als Annahmen 4 bis 6 identifiziert. Insbesondere sind Annahme 5(a) und die in Annahme 4 postulierte Unabhängigkeit von D und den hypothetischen Mediatoren und postinterventionalen Kovariaten nicht erforderlich. Unter der Annahme, dass M diskret ist, entspricht das hypothetische Ergebnis $E[Y(d, m, W)] = E[Y(d, m)]$, was impliziert, dass wir am Ergebnis für einen bestimmten Mediatorenwert $M = m$ interessiert sind, hinsichtlich der Werte der Kovariaten W aber agnostisch sind:

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

Die erste Gleichheit folgt vom Gesetz der iterierten Erwartungswerte und der bedingten Unabhängigkeit von D und den hypothetischen Ergebnissen, siehe Annahme 4, die zweite vom Gesetz der iterierten Erwartungswerte und der Substitution von Erwartungen durch Integrale, die dritte von Annahme 5(b), die vierte von der Wahrscheinlichkeitstheorie, die fünfte von der Integration von W und der Substitution des Integrals durch einen Erwartungswert, die sechste von der Wahrscheinlichkeitstheorie und die Letzte vom Gesetz der iterierten Erwartungswerte. Der kontrollierte direkte Effekt wird daher wie folgt identifiziert:

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

Der Ausdruck nach der ersten Gleichheit in (30) könnte z. B. durch g-computation geschätzt werden, siehe Robins (1986) und Vansteelandt (2009), oder durch Matching, siehe Lechner und Miquel (2010), der nach der zweiten Gleichheit durch Gewichtung, siehe Robins et al. (2000).

4 Weitere Identifikationsansätze

Dieser Abschnitt stellt Ansätze vor, welche die sequentielle bedingte Unabhängigkeit aufweichen oder sich auf alternative Annahmen stützen. Zunächst wird die partielle Identifikation von Effekten anhand von Sensitivitätsanalysen und sogenannten Effektschranken („bounds“) unter schwächeren Bedingungen als in Abschn. 3.2 behandelt. Zweitens wird die Randomisierung (oder experimentelle Zuteilung) sowohl der Intervention als auch des Mediators diskutiert. Drittens werden verschiedene Instrumentenvariablen-Methoden zwecks Kontrolle der Endogenität des Mediators und/oder der Intervention vorgestellt. Zu guter Letzt wird kurz ein sogenannter Differenz-von-Differenzen-Ansatz („difference-in-differences“) skizziert.

4.1 Partielle Identifikation basierend auf Sensitivitätsanalysen und Schranken

Da die in Abschn. 3.1 diskutierten, sequentiellen bedingten Unabhängigkeitsannahmen sehr streng sind, wurden in der Mediationsliteratur mehrere Sensitivitätsanalysen vorgeschlagen. Diese ermöglichen es, die Robustheit der direkten und indirekten Effekte hinsichtlich Abweichungen von diesen Annahmen zu untersuchen. Dies impliziert, dass nicht mehr ein einzelner Wert (oder Punkt) für den interessierenden kausalen Parameter, sondern ein bestimmtes Intervall an Werten identifiziert wird. So liefert beispielsweise VanderWeele (2010) eine allgemeine Formel für die Verzerrung (mit B bezeichnet) des kontrollierten direkten Effekts sowie der natürlichen Effekte bei Vorhandensein unbeobachteter Störterme, die den Mediator und das Ergebnis beeinflussen und somit Annahme 2 verletzen:

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

$u \neq u'$ sind zwei Werte im Träger des unbeobachteten Störterms U, der aus einer oder mehreren Variablen bestehen kann.

Unter den Einschränkungen, dass U einer einzigen binären Variable entspricht, $E[Y|D = 1 - d, M = m, X = x, U = 1] - E[Y|D = 1 - d, M = m, X = x, U = 0]$ homogen für alle Werte von D, M, und X ist und $\Pr(U = 1|D = 1-d, M = m, X = x) - \Pr(U = 1|D = d, M = m, X = x)$ homogen für alle Werte von M und X ist, vereinfachen sich die Ausdrücke in (31) zu

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

wobei $E[Y|D = 1 - d, M = m, X = x, U = 1] - E[Y|D = 1 - d, M = m, X = x, U = 0] = \phi$ und $\Pr(U = 1|D = 1 - d, M = m, X = x) - \Pr(U = 1|D = d, M = m, X = x) = \omega$. Durch die Annahme plausibler Werte für ϕ, die bedingte mittlere Differenz in Y zwischen $U = 1$ und $U = 0$, sowie ω, die bedingte mittlere Differenz in U zwischen $D = d$ und $D = 1 - d$, kann eine potenzielle Verzerrung in den Schätzungen korrigiert werden. Die Verwendung der allgemeineren Formeln in (31) erfordert hingegen die Festlegung solcher bedingten mittleren Differenzen für beliebige Kombinationen von D, M und X bzw. M und X, was aus praktischer Sicht aufwändig erscheint.

Imai et al. (2010b) schlagen eine Sensitivitätsanalyse vor, die für parametrische (sowohl lineare als auch nichtlineare) Mediationsmodelle verwendet werden kann, indem eine bestimmte Korrelation zwischen unbeobachteten Termen in den Mediatoren- und Ergebnisgleichungen, z. B. U und V in (9) und (10), angenommen wird. Eine von null verschiedene Korrelation von U und V, bezeichnet durch $\rho_{U,V}$, impliziert eine Verletzung von Annahme 2. Daher können wir durch eine Variation von $\rho_{U,V}$ zwischen -1 und 1 untersuchen, wie robust die Schätzungen der direkten und indirekten Effekte gegenüber Störtermen sind, die den Mediator und das Ergebnis beeinflussen. Die Methode ist im Paket „mediation“ von Tingley et al. (2014) für die Software R implementiert. Allerdings wird dabei unterstellt, dass diese Störterme keine Funktion der Intervention sind.

Im Gegensatz dazu schlagen Tchetgen Tchetgen und Shpitser (2012) ein Verfahren vor, das Störterme des Mediators und des Ergebnisses berücksichtigt, welche von D beeinflusst werden, siehe Abschn. 3.4. Diese semiparametrische Methode basiert auf der Spezifizierung und Kalibrierung einer Verzerrungsfunktion, nämlich $E[Y(1, m)|D = d, M = m, X = x] - E[Y(1, m)|D = d, M \neq m, X = x]$, die bezüglich der Anzahl der unbeobachteten Störfaktoren agnostisch ist. Siehe VanderWeele und Chiba (2014) und Vansteelandt und VanderWeele (2012) für weitere Methoden, die auf alternativen Verzerrungsfunktionen basieren. Einen konzeptionell unterschiedlichen Ansatz bieten Albert und Nelson (2011) an, der die Korrelation der hypothetischen Werte von postinterventionalen Variablen als zu kalibrierenden Sensitivitätsparameter betrachtet.

Hong et al. (2018) schlagen eine Methode vor, die auf Gewichtungsverfahren basierend auf (15) zugeschnitten ist und sowohl unter der Annahme von prä- als auch postinterventionalen unbeobachteten Störtermen anwendbar ist. Die Idee dabei ist, dass derartige Störterme eine Diskrepanz zwischen dem korrekten Gewicht,

das eine Beobachtung erhalten sollte, und dem tatsächlich in der Schätzung verwendeten Gewicht erzeugen. Die daraus resultierende Verzerrung kann konditional auf D anhand der Kovarianz zwischen der Gewichtsdiskrepanz und dem Ergebnis charakterisiert werden, welche somit als Grundlage für die Sensitivitätsanalyse dient. Imai und Yamamoto (2013) schlagen ferner eine Sensitivitätsanalyse für Verstöße gegen den in Robins (2003) diskutierten Ausschluss von Interaktionseffekten zwischen der Intervention und dem Mediator vor. Diese stützt sich auf zwei Sensitivitätsparameter: die Korrelation zwischen dem Mediator und dem Effekt der Intervention-Mediator-Interaktion in der Ergebnisgleichung sowie die Standardabweichung des Effekts der Intervention-Mediator-Interaktion, jeweils über alle Individuen in der Grundgesamtheit.

In allen bisher erwähnten Studien wurde die Robustheit der direkten und indirekten Effekte in Bezug auf vordefinierte Abweichungen von den identifizierenden Annahmen untersucht. Alternativ können auch obere und untere „worst case"-Werte oder Schranken („bounds") für die Effekte abgeleitet werden, die sich aus den extremst möglichen Verstößen bestimmter Annahmen ergeben. Die Allgemeingültigkeit dieser Schranken impliziert jedoch in der Regel eine relativ große Bandbreite zulässiger Effektwerte. Siehe z. B. Kaufman et al. (2005), Cai et al. (2008), Sjölander (2009) und Flores und Flores-Lagunes (2010) für methodische Beiträge auf diesem Gebiet. Üblicherweise nehmen solche Studien an, dass die Intervention randomisiert ist, während für den Mediator entweder keine oder schwächere Annahmen als Exogenität gelten, z. B. Monotonie des Mediators in der Intervention, um obere und untere Schranken für direkte und indirekte Effekte zu erhalten.

4.2 Randomisierung der Intervention und des Mediators

Dieser Abschnitt behandelt experimentelle Designs zur Bewertung kausaler Mechanismen, die auf einer „doppelten" Randomisierung sowohl der Intervention als auch des Mediators basieren. Imai et al. (2013) untersuchen ein sogenanntes paralleles Design, das auf zwei Experimenten mit unterschiedlichen Probanden basiert. Im ersten Experiment wird nur die Intervention randomisiert, was bedeutet, dass Annahme 1 unkonditional gilt, d. h. auch ohne Konditionierung auf X. Im zweiten Experiment werden die Intervention und der Mediator gemeinsam randomisiert, so dass Annahmen 1 und 2 ohne Konditionierung auf X gelten. Darüber hinaus wird angenommen, dass die Zuordnung zum einen oder anderen Experiment selbst keinen direkten Einfluss auf die Ergebnisse hat, eine Annahme, die Imai et al. (2013) Konsistenz nennen. In diesem Fall erlaubt das erste Experiment die Evaluation des ATE, basierend auf der mittleren Differenz in den Ergebnissen mit und ohne Intervention, weil $\Delta = E[Y|D = 1] - E[Y|D = 0]$ aufgrund der Randomisierung von D gilt. Das zweite Experiment erlaubt die Evaluation des kontrollierten direkten Effekts, weil $\gamma(m) = E[Y|D = 1, M = m] - E[Y|D = 0, M = m]$ aufgrund der Randomisierung von D und M gilt.

Natürliche direkte und indirekte Effekte sind jedoch ohne weitere Annahmen nicht eindeutig identifizierbar (auch wenn obere und untere Schranken be-

stimmt werden können, siehe Imai et al. 2013), da das hypothetische Ergebnis $E[Y(d, M(1-d))]$ selbst bei einer Kombination der Informationen aus beiden Experimenten unbekannt bleibt. Beispielsweise kann man für Beobachtungen im ersten Experiment mit der tatsächlichen Intervention $D = d$ nicht ergründen, ob $M(1-d) = m$ gelten würde oder nicht, weil deren Verhalten unter der alternativen Intervention $D = 1-d$ unbeobachtbar ist. Daher können Informationen zu $Y(d, m)$ aus dem zweiten Experiment im Allgemeinen nicht verwendet werden, um $E[Y(d, M(1-d))]$ zu ermitteln, es sei denn, $Y(d, m)$ ist unabhängig von $M(1-d)$. Eine Einschränkung, die eine Identifikation von natürlichen Effekten erlaubt, ist der Ausschluss von Interaktionen zwischen der Intervention und dem Mediator, wie z. B. in Robins (2003) diskutiert. Dies impliziert, dass $\theta(d) = \gamma(m) = E[Y|D = 1, M = m] - E[Y|D = 0, M = m]$ für jedes d und m gilt, so dass die direkten Effekte homogen sind: $\theta = \gamma$. Eine überprüfbare Implikation dieser Effekthomogenität ist, dass der kontrollierte direkte Effekt für verschiedene Werte von m immer gleich ist. In diesem speziellen Fall wird der indirekte Effekt durch $\delta = \Delta - \gamma$ identifiziert.

Ein alternativer Ansatz von Wunsch und Strobl (2018) unterstellt, dass der mittlere Effekt eines binären M auf Y konditional auf $D = d$ homogen über unterschiedliche Strata (oder Subpopulationen) ist, die anhand hypothetischer Mediatorenwerte definiert werden, siehe Abschn. 2.3. Formal gilt: $E[Y(d, 1) - Y(d, 0)|M(1) = m, M(0) = m'] = E[Y(d, 1) - Y(d, 0)]$ für $m, m' \in \{1, 0\}$. Der indirekte Effekt vereinfacht sich dann zu $\delta(d) = E[Y(d, 1) - Y(d, 0)] \cdot E[M(1) - M(0)]$. Das erste und zweite Experiment identifizieren $E[M(1) - M(0)]$ und $E[Y(d, 1) - Y(d, 0)]$ anhand von $E[M|D = 1] - E[M|D = 0]$ bzw. $E[Y|D = d, M = 1] - E[Y|D = d, M = 0]$. Darüber hinaus kann im Fall der Verfügbarkeit von Kovariaten X die Homogenitätsannahme getestet werden, indem überprüft wird, ob $E[Y|D = d, M = 1, X = x] - E[Y|D = d, M = 0, X = x]$ für unterschiedliche Werte x konstant ist. Ferner erlauben Kovariaten auch eine Abschwächung der Homogenitätsannahme, nämlich, dass diese (nur) konditional auf X gilt:

Annahme 7 (bedingte Effekthomogenität des Mediators). $E[Y(d, 1) - Y(d, 0)|M(1) = m, M(0) = m', X = x] = E[Y(d, 1) - Y(d, 0)|X = x]$ *für alle x im Träger von X und $m, m' \in \{1, 0\}$.*

Kombiniert man Annahme 7 mit der Konsistenz, so identifiziert die Randomisierung von D bzw. von D und M in den beiden Experimenten den indirekten Effekt: $\delta(d) = E[A \cdot B|D = d]$, wobei $A = E[Y|D = d, M = 1, X] - E[Y|D = d, M = 0, X]$ aus dem zweiten Experiment und $B = E[M|D = 1, X] - E[M|D = 0, X]$ aus dem ersten Experiment stammt. Wunsch und Strobl (2018) diskutieren die Identifikation auch für andere Designs, z. B. wenn nur M (aber nicht D) im zweiten Experiment randomisiert wird. Siehe auch Pirlott und MacKinnon (2016) für einen Überblick über unterschiedliche Ansätze zur Randomisierung.

Imai et al. (2013) schlagen eine weitere experimentelle Methode vor, das Kreuz-Design („cross-over design"), welches natürliche Effekte identifiziert, indem diesel-

ben Versuchspersonen an beiden Experimenten teilnehmen. Im ersten Experiment wird die Intervention randomisiert und die resultierenden Mediator- und Ergebniswerte werden gemessen. Im zweiten Experiment erhalten alle Versuchspersonen jenen Interventionsstatus, der dem jeweiligen Interventionsstatus im ersten Experiment entgegengesetzt ist, während der Mediator auf denselben Wert festgelegt wird, der im ersten Experiment beobachtet wurde. Unter der Voraussetzung, dass Konsistenz hält und keine Effekte aus dem ersten Experiment die hypothetischen Ergebnisse des zweiten Experiments kontaminieren, lassen sich natürliche Effekte ohne Weiteres identifizieren, weil $Y(d, M(d))$ und $Y(1-d, M(d))$ im ersten bzw. zweiten Experiment direkt beobachtet werden.

4.3 Mediation unter verschiedenen Instrumenten für die Intervention und den Mediator

Für viele empirische Fragestellungen ist eine experimentelle Randomisierung nicht durchführbar, während auch sequentielle bedingte Unabhängigkeitsannahmen in Beobachtungsdaten aufgrund einer wahrscheinlichen Endogenität der Intervention und/oder des Mediators unplausibel erscheinen. In diesem Fall stellen Instrumentenvariablen (IV) eine alternative potenzielle Identifikationsquelle dar, sofern bestimmte Bedingungen erfüllt sind. Um die Idee der Instrumentenvariablen zu diskutieren, betrachten wir das nachfolgende System linearer Gleichungen für das Ergebnis, den Mediator und die Intervention:

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

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

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

Dieses Szenario hat eine gewisse Ähnlichkeit mit dem einfachen Mediationsmodell, das in (4) und (5) gegeben ist. Ein bemerkenswerter Unterschied ist jedoch, dass D und M nun Funktionen von den Instrumenten Z_1 bzw. Z_2 sind, die eine Ausschlussrestriktion in dem Sinne erfüllen, dass sie Y nicht direkt beeinflussen. Außerdem dürfen die unbeobachteten Terme U, V und Q nun beliebig assoziiert sein, wodurch M und D endogen sind.

Direkte und indirekte Effekte können identifiziert werden, falls Z_1 und Z_2 unabhängig von (oder im linearen Modell zumindest unkorreliert mit) den unbeobachtbaren Variablen (U, V, Q) sind. Der Einfachheit halber nehmen wir an, dass die Instrumente auch voneinander unabhängig sind. Ähnlich wie in der zweistufigen Methode der kleinsten Quadrate („two stage least squares“) ermöglicht das Substituieren der ursprünglichen Interventionsvariablen D durch die exogene Vorhersage $E(D|Z_1)$ in den Mediator- und Ergebnisgleichungen (34) und (33) sowie das Substituieren von M durch $E(M|E(D|Z_1), Z_2)$ in der Ergebnisgleichung (33) die Identifikation von α_D, β_D bzw. β_M. Der Grund dafür ist, dass nur die exogene Variation in der Intervention und dem Mediator, die nicht mit den unbeobachtbaren

Variablen zusammenhängt, verwendet wird. Daher liefern Regressionen von M und Y auf die jeweiligen Vorhersagen der Intervention und des Mediators die Koeffizienten, die für die Berechnung der direkten und indirekten Effekte von Interesse sind. Wenn die IV-Ausschlussrestriktionen und Unabhängigkeitsannahmen allerdings nur konditional auf beobachtete Kovariaten X plausibel erscheinen, sind die Regressionen um diese Kontrollvariablen zu ergänzen.

Solche parametrischen IV-Ansätze für die Mediationsanalyse wurden im Kontext der Mendel'schen Randomisierung, bei der genetische Variablen als Instrumente verwendet werden (siehe z. B. Burgess et al. 2015), und in manchen ökonomischen Studien verwendet. Beispielsweise analysieren Powdthavee et al. (2013) australische Umfragedaten und schätzen einen positiven indirekten Effekt von Bildung auf die Lebenszufriedenheit, der über das Einkommen als Mediator läuft. Dazu verwenden die Autoren regionale Unterschiede in Schulgesetzänderungen als Instrument für die Bildung sowie Einkommensschocks (Erbschaft, Abfindung, Lotteriegewinn) als Instrument für das Einkommen. Chen et al. (2018) spalten den Gesamteffekt einer staatlichen Politik zur Produktivitätssteigerung (z. B. Staudammbau zur Steigerung der landwirtschaftlichen Produktion) in Technologie- und Effizienz-basierte Teileffekte auf, unter der Benutzung von topografischen Instrumentenvariablen.

Im Gegensatz zu parametrischen Mediationsmodellen erlaubt das folgende nichtparametrische Modell arbiträre Effektheterogenitäten aufgrund von Interaktionen zwischen D, M, X und/oder den unbeobachtbaren Variablen:

$$Y = \varphi(D, M, X, U), \tag{36}$$

$$M = \zeta(D, Z_2, X, V), \tag{37}$$

$$D = \chi(Z_1, X, Q). \tag{38}$$

φ, ζ, χ bezeichnen unbekannte Funktionen. Nehmen wir der Einfachheit halber an, dass sowohl Z_1 als auch D binär sind, was z. B. einem Experiment entspricht, in dem die zufällige Zuteilung der Intervention (z. B. Zugang zu einer Weiterbildung) das Instrument darstellt und die tatsächliche Inanspruchnahme (z. B. Teilnahme an der Weiterbildung) der Intervention entspricht. Die hypothetischen Mediatoren und Ergebnisse sind definiert als $M(d) = \zeta(d, Z_2, X, V)$ und $Y(d, M(d') = \varphi(d, M(d'), X, U) = \varphi(d, \zeta(d', Z_2, X, V), X, U)$, jeweils für $d, d' \in \{1, 0\}$. In analoger Weise definieren wir den hypothetischen Interventionszustand als Funktion des Instruments, bezeichnet als $D(z_1)$. Basierend auf (38) entspricht die hypothetische Intervention $D(z_1) = \chi(z_1, X, Q)$ für $z_1 \in \{0, 1\}$. Imbens und Angrist (1994) und Angrist et al. (1996) diskutieren die IV-basierte Identifikation des mittleren Gesamteffekts auf jene Subpopulation, deren hypothetischer Interventionsstatus mit dem Instrument übereinstimmt, so dass $D(1) = 1$ und $D(0) = 0$. Dieser Effekt wird als lokaler mittlerer Interventionseffekt („local average treatment effect" – LATE) oder „complier average causal effect" (CACE) bezeichnet, da er sich auf die Gruppe der „Interventionsübereinstimmer" („treatment compliers") bezieht, und ist formal wie folgt gegeben:

$$\begin{aligned}\Delta_c &= E[Y(1) - Y(0)|D(1) = 1, D(0) = 0] \\ &= E[Y(1, M(1)) - Y(0, M(0))|D(1) = 1, D(0) = 0].\end{aligned}$$

Analog zu den Abschn. 2.1 und 2.2 lassen sich auch die natürlichen direkten und indirekten Effekte sowie der kontrollierte Effekt für diese Subpopulation definieren:

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

Frölich und Huber (2017) diskutieren die nichtparametrische Identifikation von $\theta_c(d)$, $\delta_c(d)$ und $\gamma_c(m)$, wenn die Instrumente bedingt gültig sind, d. h. konditional auf Kovariaten X. Während Z_1 und D als binär angenommen werden, untersuchen die Autoren Szenarien, in denen (i) sowohl M als auch Z_2 stetig sind, (ii) M diskret und Z_2 stetig sind und (iii) M stetig und Z_2 diskret sind. Die erforderlichen Annahmen variieren in diesen Szenarien, und im Folgenden werden jene Bedingungen, die für die Identifikation nötig sind, wenn sowohl M als auch Z_2 stetig sind, kurz erläutert. Erstens müssen die Instrumente (Z_1, Z_2) unabhängig von den unbeobachtbaren Variablen (U, V, W) konditional auf X sein und die Ausschlussrestriktionen erfüllen, wie sie im nichtparametrischen Modell oben postuliert wurden. Zweitens muss Z_1 unabhängig von Z_2 konditional auf X sein. Beide Annahmen sind bei einer getrennten (d. h. unabhängigen) Randomisierung der Instrumente erfüllt. Darüber hinaus und in Analogie zu Imbens und Angrist (1994) muss gelten, dass Z_1 für niemanden einen negativen Effekt auf D aufweist (eine Bedingung, die als Monotonie bekannt ist) und für zumindest eine Subpopulation einen positiven Effekt, was die Existenz von Interventionsübereinstimmern („compliers“) impliziert. Eine weitere Bedingung ist die strikte Monotonie des Mediators im unbeobachteten und gemäß Annahme stetig verteilten Term V. Außerdem muss auch eine Annahme bezüglich des gemeinsamen Trägers erfüllt sein, nämlich, dass $\Pr(Z_1 = z_1|M, V, X, D(1) = 1, D(0) = 0)$ für $z_1 \in \{1, 0\}$ größer als null ist.

Unter diesen Bedingungen können die mittleren hypothetischen Ergebnisse für die Gruppe mit $D(1) = 1$, $D(0) = 0$ identifiziert werden, z. B.

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

$\Omega = \frac{E[(D-1)\cdot\{Z_1 - \Pr(Z_1=1)\}|M,C]}{E[D\cdot\{Z_1 - \Pr(Z_1=1)\}|M,C]}$ stellt eine Gewichtungsfunktion dar und $\pi(X) = \Pr(Z_1 = 1|X)$ ist der sogenannte Instrumenten-Propensity-Score. C bezeichnet eine nichtparametrische Kontrollfunktion, siehe z. B. Imbens und Newey (2009), welche die Verteilung von V identifiziert und damit für die Endogenität des Mediators

kontrolliert. Dieser Ansatz ist im „causalweight"-Paket von Bodory und Huber (2018) für die Software R implementiert.

Zu guter Letzt analysiert Miquel (2002) ein nichtparametrisches Modell mit binären Instrumenten Z_1, Z_2 und endogenen Variablen D, M. Sie zeigt die Identifikation von kontrollierten direkten Effekten (und dynamischen Interventionseffekten im Allgemeinen) für bestimmte Subpopulationen, die anhand dessen definiert sind, wie die Intervention und der Mediator auf das jeweilige Instrument reagieren.

4.4 Ansätze mit einem Instrument

Mehrere in der Mediationsliteratur vorgeschlagene Identifikationsstrategien stützen sich auf nur ein Instrument. Sie behandeln daher weniger allgemeine Modelle oder Methoden als der in Abschn. (4.3) erörterte Doppel-IV-Ansatz, könnten aber aus praktischer Sicht relevant sein, da (mehrere) glaubwürdige Instrumente in empirischen Daten in der Regel schwer zu finden sind. Ein Teil dieser Literatur geht davon aus, dass die Intervention exogen ist, so dass sie nicht instrumentiert werden muss, wie dies bei einer randomisierten Intervention der Fall ist, und konzentriert sich ausschließlich auf die Endogenität des Mediators. In diesem Kontext nehmen Robins und Greenland (1992) ein „perfektes" Instrument Z_2 für M an, das stark genug ist, den Mediator auf jeden möglichen Wert zu setzen, ganz so, als ob man M direkt und exogen manipulieren könnte. Solche perfekten Instrumente sind in empirischen Anwendungen aber wohl die große Ausnahme.

Imai et al. (2013) untersuchen Experimente mit einer randomisierten Intervention und einem diskreten Instrument Z_2 für einen binären Mediator M. Z_2 hat einen Effekt auf M in der Subpopulation der „mediator compliers" (also der „Mediatorübereinstimmer" in Bezug auf das Instrument, während in Abschn. 2.3 Übereinstimmung als Funktion der Intervention definiert wurde). Trotz der Randomisierung von Z_2 und D können natürliche direkte und indirekte Effekte (im Gegensatz zum kontrollierten direkten Effekt) im Allgemeinen nicht identifiziert werden. Während die Randomisierung der Intervention die Verteilungen von $M(d)$ und $Y(d, M(d))$ in der Gesamtpopulation wiedergibt, identifiziert die Anwendung von Z_2 zwecks exogener Variation von M die Verteilung von $Y(d, m)$ für die Mediatorübereinstimmer in den Gruppen mit und ohne Intervention. Die Verteilung von $Y(d, M(1-d))$ wird jedoch in keiner Population identifiziert, und zwar aus denselben Gründen wie für das parallele experimentelle Design in Abschn. 4.2: Wir können für die Gruppe mit $D = d$ nicht beobachten, für welche Individuen $M(1-d) = m$ gilt, so dass die Information über $Y(d, m)$ für die Mediatorübereinstimmer es nicht erlaubt, auf die Verteilung von $Y(d, M(1-d))$ zu schließen.

Aufgrund der Unmöglichkeit einer punktgenauen Identifikation der kausalen Parameter leiten Imai et al. (2013) und Mattei und Mealli (2011) obere und untere Schranken für ein Intervall an möglichen natürlichen bzw. direkten Stratumspezifischen Effekten her, unter der Annahme eines randomisierten D und eines diskreten Instruments Z_2 für den binären Mediator M. Ein weiterer Ansatz besteht in der Unterstellung bestimmter funktionaler Restriktionen für das Mediations-

modell, z. B. das Fehlen von Interaktionen zwischen der Intervention und dem Mediator in (33) oder die Annahme eines stetigen und starken Instruments Z_2 wie in Frölich und Huber (2017), siehe Abschn. 4.3. Ein Beispiel für ein parametrisches IV-Mediationsmodell findet sich in Chen et al. (2017). Die Autoren untersuchen den direkten Effekt des (als randomisiert angenommenen) Geschlechts des zweiten Kindes in einer Familie (D) auf die Bildung des ersten Kindes (Y) sowie den indirekten Effekt über die Familiengröße (M), für die Zwillingsgeburten (Z_2) als Instrument verwendet werden. Ten Have et al. (2007) verwenden (starke) parametrische Modellannahmen direkt als Instrumente für den Mediator. So dienen Interaktionen zwischen der Intervention und den Kovariaten als Instrumente für M, während die Abwesenheit von Interaktionen zwischen der Intervention und dem Mediator, dem Mediator und den Kovariaten sowie der Intervention und den Kovariaten im Modell für die Ergebnisvariable vorausgesetzt wird. Siehe Dunn und Bentall (2007) und Albert (2008) für verwandte Ansätze sowie Small (2012), der im Gegensatz zu den vorherigen Studien bestimmte Formen der Effektheterogenität zulässt.

Einen nichtparametrischen Ansatz stellt hingegen das experimentelle Kreuz-Design („cross-over design“) von Imai et al. (2013) dar, das bereits in Abschn. 4.2 erörtert wurde. Im ersten Experiment werden die Intervention randomisiert und die resultierenden Mediator- und Ergebniswerte gemessen. Im zweiten Experiment erhalten alle Probanden den Interventionsstatus, der ihrem Interventionsstatus im ersten Experiment entgegengesetzt ist, und M wird durch das Instrument Z_2 „angeregt“, dem beobachteten Mediatorenwert im ersten Experiment zu entsprechen. Unter bestimmten IV-Annahmen und dem Ausschluss von Verhaltenseffekten aus dem ersten Experiment, welche die hypothetischen Ergebnisse im zweiten Experiment kontaminieren, identifiziert dieser Ansatz die natürlichen Effekte für die Mediatorübereinstimmer.

Andere Studien gehen davon aus, dass sowohl D als auch M endogen sind, verwenden aber nur ein Instrument, um beide Probleme zu lösen. Yamamoto (2013) befasst sich mit der nichtparametrischen Identifikation, wenn die Intervention endogen ist und anhand von Z_1 instrumentiert wird, wie es im konventionellen LATE-Modell von Imbens und Angrist (1994) der Fall ist. Um in Abwesenheit eines zweiten Instruments für die Endogenität des Mediators zu kontrollieren, wird in Bezug auf M eine Annahme der „latenten Ignorierbarkeit“ („latent ignorability“) getroffen, siehe Frangakis und Rubin (1999). Diese unterstellt, dass der Mediator innerhalb der Subpopulation von „Interventionsübereinstimmern“ („treatment compliers“) mit $D(1) = 1, D(0) = 0$ und konditional auf X exogen ist. Das bedeutet, dass Übereinstimmung („compliance“) gemeinsam mit den Kovariaten ausreichend informativ ist, um für unbeobachtete Störfaktoren des Mediators zu kontrollieren. Auch Brunello et al. (2016) nehmen ein Instrument Z_1 für die Intervention D an, während sie beobachtbare Kovariaten mit bestimmten funktionalen Restriktionen kombinieren, um für die Endogenität des Mediators zu kontrollieren.

Joffe et al. (2008) gehen von einem einzigen Instrument aus, welches die Intervention und den Mediator gemeinsam beeinflusst und sind aus diesem Grunde auf relativ strikte parametrische Annahmen für die Identifikation der Effekte

angewiesen. Auch Dippel et al. (2017) identifizieren direkte und indirekte Effekte durch ein einziges Instrument. Sie treffen die Annahme, dass die unbeobachteten Störterme der Intervention und des Mediator und jene des Mediators und der Ergebnisvariablen voneinander unabhängig sind. Die Autoren untersuchen anhand dieses Ansatzes, ob der Effekt von Importen nach Deutschland (welche in Konkurrenz mit der lokalen Produktion stehen könnten) auf das Wahlverhalten über von den Importen induzierte Arbeitsmarktanpassungen wirkt.

4.5 Mediation auf der Grundlage von Differenz-von-Differenzen

Im Gegensatz zur konventionellen Politikevaluation, die sich mit Gesamteffekten wie dem ATE befasst, fanden Ansätze, die auf sogenannten natürlichen Experimenten basieren, in der kausalen Mediationsanalyse bisher selten Anwendung. Eine Ausnahme ist die Studie von Deuchert et al. (2019), die eine Differenz-in-Differenzen-Strategie („difference-in-differences“) verwenden, um Stratum-spezifische direkte und indirekte Effekte zu identifizieren, siehe Abschn. 2.3. Dazu unterstellen die Autoren eine randomisierte Intervention, die Monotonie des (binären) Mediators in der Intervention und spezifische Annahmen zu gemeinsamen Trends („common trends“) in bestimmten mittleren hypothetischen Ergebnissen über verschiedene Strata (oder Subpopulationen) hinweg. Diese Trendannahmen besagen, dass sich die mittleren hypothetischen Ergebnisse („mean potential outcomes“) unter bestimmten Interventions- und Mediatorzuständen für bestimmte, unterschiedliche Subpopulationen über die Zeit hinweg gleich verändern. Ein Beispiel ist die Annahme, dass die mittleren hypothetischen Ergebnisse für $d = 0$ und $m = 0$ in den Strata (i) der „niemals Mediierten“ („never mediated“) und (ii) der „Übereinstimmer“ („mediator compliers“), deren Mediator mit der Intervention übereinstimmt, über die Zeit einem gemeinsamen Trend folgen, also die gleiche Veränderung aufweisen (während sich die Niveaus der Ergebnisse zwischen den Strata aber unterscheiden können):

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

wobei $t = 0$ einen Zeitpunkt vor der Intervention bezeichnet, d. h. vor der Messung von D und M, und $t = 1$ einen späteren Zeitpunkt, in dem die Effekte auf das Ergebnis zu evaluieren sind. Unter dieser Annahme ergibt sich der direkte Stratum-spezifische Effekt auf die „niemals Mediierten“ ($M(1) = M(0) = 0$) anhand von

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

Die Unterstellung weiterer gemeinsamer Trend- und Effekthomogenitätsannahmen ermöglicht die Identifikation von direkten und indirekten Effekten auf die Übereinstimmer ($M(1) = 1, M(0) = 0$) und, sofern die Annahmen die Identifikation in allen Strata erlauben, auch auf die Gesamtpopulation, siehe Deuchert et al. (2019). Siehe auch Huber et al. (2019) und Sawada (2019), welche den mit der Differenz-in-Differenzen-Strategie verwandten „changes-in-changes"-Ansatz von Athey und Imbens (2006) für die Evaluation direkter und indirekter Effekte in Subpopulationen anwenden.

5 Erweiterungen

Die nachfolgende Diskussion stellt einige Erweiterungen des im Abschn. (3.1) behandelten Standardmodells mit sequentieller bedingter Unabhängigkeit vor, um die Analyse an von der Gesamtpopulation abweichende Zielgruppen, an Funktionen oder Transformationen der Ergebnisvariablen oder an mehrfache (statt binäre) Interventionen anzupassen. Messfehlerprobleme bei der Evaluation des Mediators sowie eine endogene Stichprobenselektion aufgrund teilweise unbeobachtbarer Ergebnisse werden ebenfalls behandelt.

5.1 Unterschiedliche Populationen und Ergebnisfunktionen

Während die meisten Mediationsstudien kausale Effekte auf die Gesamtbevölkerung evaluieren, können auch alternative Zielgruppen, wie z. B. die Subpopulation, welche die Intervention erhalten hat („treated population"), von Interesse sein. In Analogie zu den gewichteten Interventionseffekten in Hirano et al. (2003) können wir direkte und indirekte Effekte für eine bestimmte Zielgruppe unter den Annahmen 1 bis 3 messen, indem wir Beobachtungen entsprechend der Verteilung der Kovariaten X in der interessierenden Zielgruppe gewichten. Bezeichnen wir deshalb mit $\omega(X)$ eine Gewichtungsfunktion, die von X abhängt. Dann identifiziert die Verwendung von $\frac{\omega(X)}{E[\omega(X)]}$ in die Erwartungswerte von (17) die direkten und indirekten Effekte auf die interessierende Zielgruppe:

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

Das Subskript $\omega(X)$ verdeutlicht, dass sich die Effekte auf eine bestimmte Zielgruppe beziehen, die durch eine Funktion von X charakterisiert wird. Wichtige Beispiele dafür sind $\omega(X) = \Pr(D = 1|X)$ und $\omega(X) = 1 - \Pr(D = 1|X)$, wodurch sich die direkten und indirekten Effekte auf die Subpopulationen mit und ohne Intervention ergeben. Die identifizierenden Annahmen können für die Analyse dieser Gruppen etwas abgeschwächt werden, siehe die Diskussion in Vansteelandt und VanderWeele (2012).

Darüber hinaus kann die Mediationsanalyse auch auf Funktionen oder Transformationen des Ergebnisses anstatt auf Y selbst angewandt werden. Ersetzen wir beispielsweise Y in (13), (14) oder (15) durch eine Indikatorfunktion, die den Wert 1 annimmt, falls Y nicht größer als ein bestimmter Wert a ist, und ansonsten den Wert 0 annimmt, oder formal: $I\{Y \leq a\}$. Dadurch erhalten wir die kumulative Verteilungsfunktion der hypothetischen Ergebnisse (anstatt deren Mittelwert). Das Invertieren der Verteilungsfunktion ermöglicht die Identifikation direkter und indirekter Quantilseffekte der Intervention für bestimmte Quantile (oder Perzentile) der (unbedingten) Verteilung der hypothetischen Ergebnisse. Siehe auch Schmidpeter (2018), der direkte und indirekte Quantilseffekte auf die Subpopulation mit Intervention unter Anwendung der sogenannten (und für die Quantilsregression vorgeschlagenen) Check-Funktion von Koenker und Bassett (1978) identifiziert. Weitere Studien analysieren direkte und indirekte Effekte für bedingte Quantile des Ergebnisses und/oder des Mediators gegeben X (statt unkonditonale Quantile über die Verteilung von X), siehe Dominici et al. (2006), Imai et al. (2010a), Shen et al. (2014), Bind et al. (2017) und Geraci und Mattei (2017). Bedingte direkte und indirekte Quantilseffekte addieren sich jedoch nicht notwendigerweise zum Gesamteffekt an einem unkonditionalen Quantil der hypothetischen Ergebnisse in der Gesamtpopulation auf.

5.2 Mehrfache und stetige Interventionen

Ein großer Teil der Studien im Bereich der kausalen Mediationsanalyse untersucht eine binäre Intervention. In empirischen Anwendungen können Interventionen jedoch auch mehrere diskrete Werte annehmen, wie z. B. alternative Arbeitsmarktprogramme (z. B. keine Ausbildung, Bewerbungstraining, Sprachkurs, Computerkurs) oder gar stetig verteilt sein, z. B. die Zeit in einer Weiterbildung. Hayes und Preacher (2014) diskutieren die Analyse von mehrfachen Interventionen in linearen Modellen. In nichtparametrischen Modellen können wir die Evaluation von mehrfachen diskreten (z. B. geordneten oder multinomial verteilten) Interventionen auf der Grundlage der Erkenntnisse von Abschn. 3.2 implementieren. Für jedes mögliche Wertepaar $d \neq d'$ der Intervention D sind die Ausdrücke in (13), (14) und (15) für die hypothetischen Ergebnisse unmittelbar gültig, sofern Annahmen 1 bis 3 in Bezug auf die nichtbinären Werte d und d' erfüllt sind. Direkte und indirekte Effekte lassen sich dann analog zum binären Fall durch paarweise Vergleiche von hypothetischen Ergebnissen unter unterschiedlichen Interventionen berechnen (z. B. $d = 1, d' = 0$ oder $d = 2, d' = 0$ oder $d = 2, d' = 1$, wenn $d, d' \in \{0, 1, 2\}$).

Falls D stetig verteilt ist, bleibt die Identifikation im Fall eines linearen Modells, wie es durch die Gleichungen (9) bis (12) definiert wird, im Vergleich zu einer binären Intervention unverändert. Unter der Annahme eines nichtparametrischen Modells müssen die Ergebnisse aus Abschn. (3.2) jedoch modifiziert werden, um der Tatsache Rechnung zu tragen, dass stetig verteilte Interventionen (im Gegensatz zu diskreten) keine Punktmasse haben. Hsu et al. (2018) passen die gewichtungsbasierte Identifikation hypothetischer Ergebnisse, wie sie in (14) gegeben ist, an eine stetig verteilte Intervention an, unter der Voraussetzung, dass die Annahmen 1 bis 3 auch für den stetigen Fall gültig sind. Insbesondere werden alle Indikatorfunktionen für Interventionswerte durch sogenannte Kern-Funktionen und die Propensity Scores für die Intervention durch bedingte Dichtefunktionen ersetzt. Letztere sind auch als verallgemeinerte Propensity Scores („generalized propensity scores“) bekannt, siehe z. B. Hirano und Imbens (2004) und Imai und van Dyk (2004). Daraus folgt, dass das mittlere hypothetische Ergebnis unter Interventionswerten $d \neq d'$ wie folgt identifiziert wird:

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

Die Gewichtungsfunktion $\omega(D; d, h) = K\left((D - d)/h\right)/h$, wobei K eine symmetrische Kern-Funktion zweiter Ordnung ist, die Beobachtungen höher gewichtet, die näher an d liegen, und h die Bandweite darstellt. Wenn h gegen null geht, d. h. $\lim_{h \to 0}$, dann entsprechen die bedingten Erwartungswerte $E[\omega(D; d', h)|X]$ und $E[\omega(D; d', h)|M, X]$ den verallgemeinerten Propensity Scores $f(D = d|X)$ bzw. $f(D = d|M, X)$.

Hsu et al. (2018) diskutieren die Schätzung von direkten und indirekten Effekten durch Gewichtung basierend auf parametrischen oder nichtparametrischen verallgemeinerten Propensity Scores. Die parametrische Variante ist auch im „causalweight“-Paket von Bodory und Huber (2018) verfügbar. Alternativ zur Gewichtung kann die imputationsbasierte Schätzung hypothetischer Ergebnisse, wie von Vansteelandt et al. (2012) vorgeschlagen, sowohl auf mehrfache diskrete als auch auf stetige Interventionen angewendet werden. Dieser Ansatz ist im „medflex“-Paket von Steen et al. (2017b) implementiert. Auch das regressionsbasierte „mediation“-Paket von Tingley et al. (2014) erlaubt die Analyse von nichtbinären Interventionen sowohl in linearen als auch nichtlinearen Modellen.

5.3 Messfehler in Mediatoren und fehlende Ergebnisse

VanderWeele (2012b) weist auf ein subtiles, aber wichtiges Problem im Falle einer zu „groben“ Messung des Mediators hin, wie sie wahrscheinlich in vielen empirischen Fragestellungen auftritt. Nehmen wir zur Veranschaulichung an, dass der Mediator als Beschäftigung definiert ist. Nehmen wir weiter an, dass in unseren Daten ein binärer Indikator für die Erwerbstätigkeit verfügbar ist, d. h. ob eine Person keine oder eine positive Anzahl an Arbeitsstunden leistet, aber keine

Information über die Anzahl der von den Erwerbstätigen geleisteten Arbeitsstunden. Dies stellt ein Problem dar, wenn der Effekt der Intervention auf den Mediator nicht ausschließlich die binäre Entscheidung „Erwerbstätigkeit vs. keine Erwerbstätigkeit" beeinflusst, sondern auch manche Erwerbstätige dazu veranlasst, ihr mit und ohne Intervention stets positives Arbeitspensum zu verändern. Der letztere Effekt wird nämlich im Fall einer binären Definition des Mediators nicht dem indirekten Effekt zugeordnet, sondern dem direkten. Dieser unintuitiven Interpretation der Effekte aufgrund eines (zu) „groben" Maßes für den Mediator sollte man sich bewusst sein. Generell bringt jede Form von Messfehlern im Mediator Probleme für die Evaluation und Interpretation direkter und indirekter Effekte mit sich. Heckman und Pinto (2015) bieten unter bestimmten Annahmen ein Verfahren zur Korrektur von Messfehlern im Mediator an, welches beispielsweise voraussetzt, dass das Mediationsmodell parametrisch ist und der Messfehler unabhängig von den hypothetischen Mediatoren ist.

Eine weitere Komplikation für die Mediationsanalyse ergibt sich, wenn die Ergebnisse aufgrund von endogener Stichprobenselektion oder Datenschwund nur für eine Teilpopulation beobachtet werden. Solche Probleme treten häufig in empirischen Anwendungen auf, z. B. in Dekompositionen von geschlechtsspezifischen Lohnunterschieden, wo Löhne ausschließlich für Erwerbstätige beobachtet werden, oder wenn Personen an einer Folgeerhebung einer Studie nicht mehr teilnehmen, in der die Ergebnisvariable gemessen wird. Huber und Solovyeva (2018) diskutieren die Identifikation von natürlichen Effekten sowie des direkten kontrollierten Effekts, wenn man die Annahme der sequentiellen bedingten Unabhängigkeit für die Intervention und den Mediator mit bestimmten Annahmen bezüglich der teilweisen Beobachtbarkeit des Ergebnisses kombiniert. So kann man z. B. unterstellen, dass diese Beobachtbarkeit konditional auf die Intervention, den Mediator und die Kovariaten exogen (also so gut wie zufällig) ist oder dass ein Instrument zur Verfügung steht, welches die Beobachtungswahrscheinlichkeit des Ergebnisses, aber nicht das Ergebnis selbst, beeinflusst (z. B. finanzielle Anreize, an einer Folgeerhebung teilzunehmen). Die Autoren schlagen für diese Szenarien gewichtungsbasierte Schätzer vor, die auf Propensity Scores für die Intervention, den Mediator und/oder die Beobachtbarkeit des Ergebnisses basieren.

6 Zusammenfassung

Dieses Kapitel gab einen Überblick über methodische Weiterentwicklungen in der kausalen Mediationsanalyse, mit besonderem Schwerpunkt auf Anwendungen in den Wirtschaftswissenschaften. Nach der Definition der interessierenden direkten und indirekten Effekte wurde die Identifikation und Schätzung basierend auf Annahmen der bedingten Unabhängigkeit der Intervention und des Mediators erörtert. Die Diskussion umfasste sowohl statische als auch dynamische Szenarien, d. h. wenn zumindest manche Kovariaten, die den Mediator und das Ergebnis beeinflussen, selbst eine Funktion der Intervention sind. Das Kapitel befasste sich auch mit weiteren Evaluationsstrategien, die auf partieller Identifikation, Randomisierung der

Intervention und des Mediators, Instrumentenvariablen für die Intervention und/oder den Mediator und Differenz-von-Differenzen-Ansätzen basieren. Zu guter Letzt wurden mehrere Erweiterungen des Standardmodells skizziert, wie mehrfache statt binäre Interventionen, sich von der Gesamtpopulation unterscheidende Zielpopulationen, Messfehler in Mediatoren und endogene Stichprobenselektion bzw. die partielle Beobachtbarkeit des Ergebnisses.

Literatur

Albert JM (2008) Mediation analysis via potential outcomes models. Stat Med 27:1282–1304
Albert JM, Nelson S (2011) Generalized causal mediation analysis. Biometrics 67:1028–1038
Angrist J, Imbens G, Rubin D (1996) Identification of causal effects using instrumental variables. J Am Stat Assoc 91:444–472 (with discussion)
Athey S, Imbens GW (2006) Identification and inference in nonlinear difference-in-differences models. Econometrica 74:431–497
Avin C, Shpitser I, Pearl J (2005) Identifiability of path-specific effects. In: IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, S 357–363
Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51:1173–1182
Bellani L, Bia M (2019) The long-run effect of childhood poverty and the mediating role of education. J R Stat Soc Ser A (Stat Soc) 182:37–68
Bijwaard GE, Jones AM (2018) An IPW estimator for mediation effects in hazard models: with an application to schooling, cognitive ability and mortality. Empir Econ 57:1–47
Bind M-A, VanderWeele TJ, Schwartz JD, Coull BA (2017) Quantile causal mediation analysis allowing longitudinal data. Stat Med 36:4182–4195
Bodory H, Huber M (2018) The causalweight package for causal inference in R. SES Working Paper 493, University of Fribourg
Brunello G, Fort M, Schneeweis N, Winter-Ebmer R (2016) The causal effect of education on health: what is the role of health behaviors? Health Econ 25:314–336
Burgess S, Daniel RM, Butterworth AS, Thompson SG (2015) Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. Int J Epidemiol 44:484–495
Cai Z, Kuroki M, Pearl J, Tian J (2008) Bounds on direct effects in the presence of confounded intermediate variables. Biometrics 64:695–701
Chan K, Imai K, Yam S, Zhang Z (2016) Efficient nonparametric estimation of causal mediation effects. Working paper arXiv:1601.03501
Chen SH, Chen YC, Liu JT (2017) The impact of family composition on educational achievement. J Hum Res 1:122–170
Chen Y-T, Hsu Y-C, Wang H-J (2018) A stochastic frontier model with endogenous treatment status and mediator. J Bus Econ Stat 38(2):243–256
Cochran WG (1957) Analysis of covariance: its nature and uses. Biometrics 13:261–281
Conti G, Heckman JJ, Pinto R (2016) The effects of two influential early childhood interventions on health and healthy behaviour. Econ J 126:F28–F65
De Stavola BL, Daniel RM, Ploubidis GB, Micali N (2015) Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. Am J Epidemiol 181:64–80
Deuchert E, Huber M, Schelker M (2019) Direct and indirect effects based on difference-in-differences with an application to political preferences following the Vietnam draft lottery. J Bus Econ Stat 37:710–720

Dippel C, Gold R, Heblich S, Pinto R (2017) Instrumental variables and causal mechanisms: unpacking the effect of trade on workers and voters. National Bureau of Economic Research (No. w23209)
Dominici F, Zeger SL, Parmigiani G, Katz J, Christian P (2006) Estimating percentile-specific treatment effects in counterfactual models: a case-study of micronutrient supplementation, birth weight and infant mortality. J R Stat Soc Ser C 55:261–280
Dunn G, Bentall R (2007) Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). Stat Med 26:4719–4745
Farbmacher H, Huber M, Lafférs L, Langen H, Spindler M (2022) Causal mediation analysis with double machine learning. Econ J 25:277–300
Flores CA, Flores-Lagunes A (2009) Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. IZA DP No. 4237
Flores CA, Flores-Lagunes A (2010) Nonparametric partial identification of causal net and mechanism average treatment effects. mimeo, University of Florida
Frangakis C, Rubin D (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, Rubin D (2002) Principal stratification in causal inference. Biometrics 58:21–29
Frölich M, Huber M (2017) Direct and indirect treatment effects – causal chains and mediation analysis with instrumental variables. J R Stat Soc Ser B 79(5):1645–1666
Gelman A, Imbens GW (2013) Why ask why? Forward causal inference and reverse causal questions. NBER Working Paper No. 19614
Geraci M, Mattei A (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 AF, Preacher KJ (2014) Statistical mediation analysis with a multicategorical independent variable. Br J Math Stat Psychol 67:451–470
Heckman J, Pinto R, Savelyev P (2013) Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. Am Econ Rev 103:2052–2086
Heckman JJ, Pinto R (2015) Econometric mediation analyses: identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. Econ Rev 34:6–31
Hernandez-Diaz S, Schisterman EF, Hernan MA (2006) The birth weight „paradox“ uncovered? Am J Epidemiol 164:1115–1120
Hirano K, Imbens GW (2004) The propensity score with continuous treatments. In: Gelman A, Meng XL (Hrsg) Applied Bayesian modeling and causal inference from incomplete-data perspectives. Wiley, New York, S 73–84
Hirano K, Imbens GW, Ridder G (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. American Statistical Association, Alexandria, S 2401–2415
Hong G (2015) Causality in a Social World: Moderation, Meditation and Spill-Over. Wiley, West Sussex
Hong G, Deutsch J, Hill HD (2015) Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. J Educ Behav Stat 40:307–340
Hong G, Qin X, Yang F (2018) Weighting-based sensitivity analysis in causal mediation studies. J Educ Behav Stat 43:32–56
Hsu Y, Huber M, Lee Y, Pipoz L (2018) Direct and indirect effects of continuous treatments based on generalized propensity score weighting. SES Working Paper 495, University of Fribourg
Hsu Y, Huber M, Lai T (2019) Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting. J Econ Methods 8:621–654
Huber M (2014) Identifying causal mechanisms (primarily) based on inverse probability weighting. J Appl Econ 29:920–943

Huber M (2015) Causal pitfalls in the decomposition of wage gaps. J Bus Econ Stat 33:179–191
Huber M, Solovyeva A (2018) Direct and indirect effects under sample selection and outcome attrition. SES Working Paper 496, University of Fribourg
Huber M, Lechner M, Mellace G (2016) The finite sample performance of estimators for mediation analysis under sequential conditional independence. J Bus Econ Stat 34:139–160
Huber M, Lechner M, Mellace G (2017) Why do tougher caseworkers increase employment? The role of program assignment as a causal mechanism. Rev Econ Stat 99:180–183
Huber M, Schelker M, Strittmatter A (2019) Direct and indirect effects based on changes-in-changes, arXiv preprint, arXiv:1909.04981
Imai K, van Dyk DA (2004) Causal inference with general treatment regimes. J Am Stat Assoc 99:854–866
Imai K, Yamamoto T (2013) Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. Polit Anal 21:141–171
Imai K, Keele L, Tingley D (2010a) A general approach to causal mediation analysis. Psychol Methods 15:309–334
Imai K, Keele L, Yamamoto T (2010b) Identification, inference and sensitivity analysis for causal mediation effects. Stat Sci 25:51–71
Imai K, Tingley D, Yamamoto T (2013) Experimental designs for identifying causal mechanisms. J R Stat Soc Ser A 176:5–51
Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86:4–29
Imbens GW, Angrist J (1994) Identification and estimation of local average treatment effects. Econometrica 62:467–475
Imbens GW, Newey WK (2009) Identification and estimation of triangular simultaneous equations models without additivity. Econometrica 77:1481–1512
Joffe MM, Small D, Have TT, Brunelli S, Feldman HI (2008) Extended instrumental variables estimation for overall effects. Int J Biostat 4:Article 4
Judd CM, Kenny DA (1981) Process analysis: estimating mediation in treatment evaluations. Eval Rev 5:602–619
Kaufman JS, MacLehose RF, Kaufman S (2004) A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. Epidemiol Perspect Innov 1:4
Kaufman S, Kaufman J, MacLenose R, Greenland S, Poole C (2005) Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. Stat Med 24:1683–1702
Keele L, Tingley D, Yamamoto T (2015) Identifying mechanisms behind policy interventions via causal mediation analysis. J Pol Anal Manag 34:937–963
Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46(1):33–50
Lange T, Vansteelandt S, Bekaert M (2012) A simple unified approach for estimating natural direct and indirect effects. Am J Epidemiol 176:190–195
Lange T, Rasmussen M, Thygesen LC (2014) Assessing natural direct and indirect effects through multiple pathways. Am J Epidemiol 179(4):513–518
Lechner M (2009) Sequential causal models for the evaluation of labor market programs. J Bus Econ Stat 27:71–83
Lechner M, Miquel R (2010) Identification of the effects of dynamic treatments by sequential conditional independence assumptions. Empir Econ 39:111–137
Mattei A, Mealli F (2011) Augmented designs to assess principal strata direct effects. J R Stat Soc Ser B 73:729–752
Miquel R (2002) Identification of dynamic treatment effects by instrumental variables. University of St. Gallen Economics Discussion Paper Series, 2002–2011
Pearl J (2001) Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufman, San Francisco, S 411–420
Petersen ML, Sinisi SE, van der Laan MJ (2006) Estimation of direct causal effects. Epidemiology 17:276–284

Pirlott AG, MacKinnon DP (2016) Design approaches to experimental mediation. J Exp Soc Psychol 66:29–38
Powdthavee N, Lekfuangfu WN, Wooden M (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 KJ, Rucker DD, Hayes AF (2007) Addressing moderated mediation hypotheses: theory, methods, and prescriptions. Multivar Behav Res 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. Math Model 7:1393–1512
Robins JM (2003) Semantics of causal DAG models and the identification of direct and indirect effects. In: Green P, Hjort N, Richardson S (Hrsg) In highly structured stochastic systems. Oxford University Press, Oxford, S 70–81
Robins JM, Greenland S (1992) Identifiability and exchangeability for direct and indirect effects. Epidemiology 3:143–155
Robins JM, Richardson T (2010) Alternative graphical causal models and the identification of direct effects. In: Shrout P, Keyes K, Omstein K (Hrsg) Causality and psychopathology: finding the determinants of disorders and their cures. Oxford University Press, Oxford
Robins JM, Hernan MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701
Rubin DB (2004) Direct and indirect causal effects via potential outcomes. Scand J Stat 31:161–170
Sawada M (2019) Non-compliance in randomized control trials without exclusion restrictions, arXiv preprint, arXiv:1910.03204
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, Chou C-P, Pentz MA, Berhane K (2014) Quantile mediation models: a comparison of methods for assessing mediation across the outcome distribution, multivariate behavioral research. Multivar Behav Res 49:471–485
Simonsen M, Skipper L (2006) The costs of motherhood: an analysis using matching estimators. J Appl Econ 21:919–934
Sjölander A (2009) Bounds on natural direct effects in the presence of confounded intermediate variables. Stat Med 28:558–571
Small DS (2012) Mediation analysis without sequential ignorability: using baseline covariates interacted with random assignment as instrumental variables. J Stat Res 46:91–103
Steen J, Loeys T, Moerkerke B, Vansteelandt S (2017a) Flexible mediation analysis with multiple mediators. Am J Epidemiol 186:184–193
Steen J, Loeys T, Moerkerke B, Vansteelandt S (2017b) Medflex: an R package for flexible mediation analysis using natural effect models. J Stat Softw 76:1–46
Tchetgen Tchetgen EJ (2013) Inverse odds ratio-weighted estimation for causal mediation analysis. Stat Med 32:4567–4580
Tchetgen Tchetgen EJ, Shpitser I (2012) Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. Ann Stat 40:1816–1845
Tchetgen Tchetgen EJ, VanderWeele TJ (2014) On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. Epidemiology 25:282–291
Ten Have TR, Joffe MM, Lynch KG, Brown GK, Maisto SA, Beck AT (2007) Causal mediation analyses with rank preserving models. Biometrics 63:926–934
Tingley D, Yamamoto T, Hirose K, Imai K, Keele L (2014) Mediation: R package for causal mediation analysis. J Stat Softw 59:1–38
Van der Laan MJ, Petersen ML (2008) Direct effect models. Int J Biostat 4:1–27
VanderWeele TJ (2008) Simple relations between principal stratification and direct and indirect effects. Stat Prob Lett 78:2957–2962

VanderWeele TJ (2009) Marginal structural models for the estimation of direct and indirect effects. Epidemiology 20:18–26
VanderWeele TJ (2010) Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology 21:540–551
VanderWeele TJ (2012a) Comments: should principal stratification be used to study mediational processes? J Res Educ Effect 5(3):245–249
VanderWeele TJ (2012b) Mediation analysis with multiple versions of the mediator. Epidemiology 23:454–463
VanderWeele TJ (2016) Mediation analysis: a practitioner's guide. Annu Rev Public Health 37(1):17–32
VanderWeele TJ, Chiba Y (2014) Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. Epidemiol Biostat Public Health 11:e9027
VanderWeele TJ, Vansteelandt S (2009) Conceptual issues concerning mediation, interventions and composition. Stat Interface 2:457–468
VanderWeele TJ, Vansteelandt S (2010) Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol 172:1339–1348
VanderWeele TJ, Vansteelandt S (2014) Mediation analysis with multiple mediators. Epidemiol Methods 2:95–115
Vansteelandt S (2009) Estimating direct effects in cohort and case-control studies. Epidemiology 20(6):851–860
Vansteelandt S, Bekaert M, Lange T (2012) Imputation strategies for the estimation of natural direct and indirect effects. Epidemiol Methods 1:129–158
Vansteelandt S, VanderWeele TJ (2012) Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. Biometrics 68:1019–1027
Wilcox AJ (2001) On the importance-and the unimportance-of birthweight. Int J Epidemiol 30:1233–1241
Wunsch C, Strobl R (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, van der Laan MJ (2012) Targeted maximum likelihood estimation of natural direct effects. Int J Biostat 8:1–40