Dawid, Philip; Humphreys, Macartan; Musio, Monica

Article — Published Version

Bounding Causes of Effects With Mediators

Sociological Methods & Research

Article

Sociological Methods & Research
2024, Vol. 53(1) 28–56

DOI: 10.1177/00491241211036161
journals.sagepub.com/home/smr
Sage

Bounding Causes of Effects With Mediators

Philip Dawid[1], Macartan Humphreys[2,3] and Monica Musio[4]

Abstract

Suppose *X* and *Y* are binary exposure and outcome variables, and we have full knowledge of the distribution of *Y*, given application of *X*. We are interested in assessing whether an outcome in some case is due to the exposure. This "probability of causation" is of interest in comparative historical analysis where scholars use process tracing approaches to learn about causes of outcomes for single units by observing events along a causal path. The probability of causation is typically not identified, but bounds can be placed on it. Here, we provide a full characterization of the bounds that can be achieved in the ideal case that *X* and *Y* are connected by a causal chain of complete mediators, and we know the probabilistic structure of the full chain. Our results are largely negative. We show that, even in these very favorable conditions, the gains from positive evidence on mediators is modest.

Keywords

causal pathway, causes of effects, interval bounds, mediator, probability of causation, process tracing, qualitative methods

[1] University of Cambridge, United Kingdom
[2] Columbia University, New York, NY, United States
[3] WZB Berlin, Germany
[4] Università degli Studi di Cagliari, Italy

Corresponding Author:
Macartan Humphreys, WZB Berlin, Berlin 10785, Germany.
Email: macartan.humphreys@wzb.eu

Introduction

Even the best possible evidence regarding the effects of a treatment on an outcome in a population is generally not enough to identify the probability that a positive outcome in an individual treated case was in fact caused by the treatment.

For instance, researchers conducting randomized controlled trials may determine that providing a medicine to school children increases the overall probability of good health from one third to two thirds. This information, no matter how precise, is not enough to answer the following question: Is Ann healthy because she took the medicine? It is not even enough to answer the question probabilistically. The reason is that, consistent with these results, it may be that the medicine makes a positive change for two out of three children, but a negative change for the remainder: In that case, the medicine certainly helped Ann. But it might alternatively be that the medicine makes a positive change for one in three children but no change for the others. In that case, the chances it helped Ann are just one in two. For, of the children taking the medicine, two thirds are healthy. Half of these are healthy because of the medicine, whereas the other half would have been healthy anyway.

Put differently, the experimental data identifies the "effects of causes," but we are interested in the reverse problem, of quantifying "causes of effects." The causes of effects task of defining and assessing the *probability of causation* (Robins and Greenland 1989) in an individual case have been considered by Tian and Pearl (2000); Dawid (2011); Yamamoto (2012); Pearl (2015); Dawid, Musio and Fienberg (2016); and Murtas, Dawid, and Musio (2017).[1] Note that this is distinct from the "reverse causal question" of Gelman and Imbens (2013), which is a collection of effects of causes questions aimed at ascertaining *which* causes have an effect on an outcome—the difference being that the estimand in this formulation does not condition on observed values of treatments and outcomes. The question is of interest for historical analyses that seek to *explain* outcomes, for judicial determinations of innocence or guilt, and policy analysis seeking to assign responsibility for outcomes to interventions. For these outcomes, bounds are useful when they are narrow—in which case they can be treated like point estimates despite the lack of identification. But even less narrow bounds can sometimes be useful and support claims of the form: *For any possible priors you might hold* you should conclude that Y was more likely than not due to X. Finally, knowing that bounds are *not* narrow is useful since it clarifies that claims about causal attribution reflect prior beliefs about causal processes and not beliefs justified by data. For all these cases, we highlight that determining that X caused

Y does not in any way mean that *X* is the only cause of *Y* or the most important cause of *Y*. For this reason, the attribution question can be addressed without needing to take account of other possible causes—although, as we will show, taking account of these may sometimes sharpen conclusions.

A common approach to learning about causes of effects is to seek additional evidence along causal pathways. Observation of such ancillary evidence can then act like a test, leading to updating on overall causal relations. Using the language in Van Evera (1997), a “smoking gun test” searches for evidence that, though unlikely to be found, would give great confidence in a claim if it were to be found; a “hoop” test is a search for evidence that we expect to find, but which, if found to be absent, would provide compelling evidence against a proposition (as if the proposition were asked to jump through a hoop).

Though these tests do not *require* that causal process observations lie along a simple chain—what Weller and Barnes (2016) call scenario 1 chains and we call a chain with complete mediation—in many applications, researchers presume that they do. In the account provided in Mahoney (2012), Skocpol (1979) produced a hoop test by identifying a mediator *M* (local events) such that *X* (community solidarity) was necessary for *M* and *M* was sufficient for *Y* (peasant revolution). As described also by Mahoney (2012), researchers might use chains to justify smoking gun tests, seeking “chains of necessary conditions.” A common practice among researchers evaluating development programs is to specify “theories of change” and seek evidence for intermediate outcomes along a pathway linking treatment to outcomes (Ghate 2018): Was the treatment received? Was the medicine ingested? Knight and Winship (2013) review a long history in sociology of “mechanism-focused scholarship,” including in Max Weber, Karl Marx, and Paul Lazerfeld. Gross (2018) describes the many different classes of causal chains used in sociological research, many of which involve complete mediation (or linearity, to use his term).

This strategy of looking at values of a mediating variable is often extended by examining multiple points on a chain. Seeing supportive evidence at many points along such a causal chain would appear to give confidence that the final outcome is indeed *due* to the conjectured cause. This is a common idea in process tracing (Collier 2011) as well as of mixed methods research as used in development evaluation (White 2009). As described by Mahoney (2012), “[a]lthough a hypothesis that passes any one straw in the wind test may not be well supported, a hypothesis that passes several straw in the wind tests may generate a good deal of confidence in its validity.” In the most optimistic accounts of observation of causal chains, it is reasoned that, as one gets close enough to a process, by observing more and more links in a chain, the link between any two steps becomes less questionable—intuitively obvious—and eventually the causal process reveals itself (Mahoney 2012:581).

We here provide a comprehensive treatment of the scope for inferences of this form. Our analyses employs causal models for justifying mechanistic accounts as advocated by Knight and Winship (2013). The analysis builds on logic found in Mahoney (2012) by quantifying the learning that can be made from cases involving necessity and sufficiency as well as probabilistic relations. Whereas existing results (Dawid, Murtas, and Musio 2016) have considered the case of a single unobserved mediator, we generalize by considering situations with chains of arbitrary length and we calculate bounds for general data, that is, for situations in which the values of none, some, or all the mediators are observed. We obtain a general formula for calculating bounds on the probability of causation, derive implications of this formula, and calculate the largest and smallest upper and lower bounds achievable from any causal chain consistent with known relation between X and Y.

We emphasize that we focus on what might appear to be ideal conditions: those in which we believe causal processes follow a simple causal chain and in which researchers have complete evidence about the probabilistic relationship between any two consecutive nodes in the chain. Thus, we exclude more complex situations in which there are both direct and indirect effects connecting nodes. We explore still more optimistic conditions in which the chain is arbitrarily long, in which the causal effect of each intermediate variable on its successor climbs to 1, and in which researchers observe outcomes consistent with positive effects at every point on the chain.

Insofar as these are best case settings, the negative results we provide are, we believe, all the more striking. Our key results imply that our ability to raise lower bounds is often modest. Consistent observations along a causal chain, for instance, do increase confidence that an outcome can be attributed to a cause; moreover, for "homogeneous" chains (chains for which causal processes look the same at every step)-the longer the chain the better. However, even under these ideal conditions, the narrowing of bounds is often small. In the example of attributing Ann's health to good medicine, a smooth process with arbitrarily many positive intermediate steps observed would only tighten the bounds from $[0.5, 1]$ to $[0.58, 1]$. Other processes can tighten the bounds more. For example, suppose Ann was prescribed the medicine and recovered. If we know that being prescribed the medicine is the only way in which Ann could have obtained and taken the medicine, and that taking the medicine helps anyone who would otherwise be sick, then with positive evidence on a single intermediate point on the causal chain—that Ann did indeed take the medicine—we can identify the probability that prescribing the medicine caused Ann's recovery at two thirds. A process like this, in which we observe a "necessary condition for a sufficient condition," provides the largest possible lower bound on the probability of causation available from any observations on any chain. At this point, we have done the best possible and more data along the chain will not help. No data pattern supports an inference closer to 1.

Although achieving identification of the probability of causation at 1 is generally elusive, even on long chains, negative data can yield identification at 0, even when observed at single node. In this sense, information on mediators can support "hoop" tests but not "smoking gun" tests.

The intuition for why identification at 0 is possible is the following. If we know that $A = 1$ is necessary for $B = 1$, then we know that A cannot induce a negative effect on B. But then if we observe $A = 1, B = 0$ we can infer that A did not have any effect—positive or negative—on B, and so the causal chain is broken. The intuition for why positive evidence is not so informative for updating towards 1 is that positive evidence is always consistent with both $A = 1$ causing $B = 1$ and $B = 1$ arising regardless of A. The only time in which we do not face this ambiguity at all is when we know that $B = 1$ does not arise regardless of A in which case we would not learn anything new from observation of A. The intuition for why longer chains of positive evidence have modest effects on bounds is that while a decomposition of a process with many steps means greater confidence of causal effects at each step, each additional step also creates another point at which a causal chain might be broken. As a numerical example, in Ann's one-step process, we had a lower bound of the probability that X caused Y of .5. If we had five steps and transition matrices, identical at each step and consistent with the known distribution of Y given X, then we would have to have quite strong average effects at each step—around 0.8 rather than one third (since $0.8^5 \approx 1/3$); these in turn induce a lower bound that each outcome was caused by its predecessor of around 0.89. While 0.89 for a single step appears promising, the implied lower bound for the entire chain is then just $0.89^5 \approx 0.56$, which is only a modest increase in what we had before: in short, the parsing into steps gives more scope to find positive evidence but is accompanied by an accumulation of points at which a chain might be broken.

Our results have implications for qualitative and quantitative scholars. Most immediately they can be used to assess what inferences can be drawn from observations along a causal path and thus inform decisions about whether to gather data of this form. They can also help clarify the background knowledge about causal processes needed to make these inferences.

The result can also be used to help determine *which* observations to examine in settings where researchers have a choice. Yet the negative results also carry a caution: Argumentation for attribution built on evidence along causal chains can rarely support positive claims for causal effects.

We proceed as follows. The next section introduces the setup and gives general formulae for bounding the probability of causation for a simple one-step process. In the third section, we provide new results for cases in which all mediators are unobserved, all are observed, or just some are observed. Theorem 2 provides a general formula applicable to all cases. Then, theorem 3 details the maximum and minimum upper and lower bounds for all possible processes. In all cases, these can be achieved by processes of at most two steps. In the fourth section, we compare the extrema with the bounds obtained from smooth (homogeneous) processes, with bounds achievable when processes are known to be monotonic, and bounds obtainable from knowledge of covariates, which can be much tighter. We summarize our results, and consider some implications, in the fifth section. Various technical details for the proofs in the paper are elaborated in Online Appendices (which can be found at Supplementary material for this article, available online).

Preliminaries

Consider a binary treatment or exposure variable X, and binary outcome variable Y. We let $\mathbf{Y} = (Y(0), Y(1))$ denote a pair of *potential outcomes*, for Y where we conceive of $Y(x)$ as the value Y would take, if X were set to the value x by external intervention. We regard both $Y(0)$ and $Y(1)$ as existing simultaneously, even prior to setting the value of X, and as having a bivariate probability distribution.

Throughout, we invoke two assumptions:

Consistency: Even when X is not set by intervention, the outcome Y will be $Y(X)$.

No confounding: This is expressed as independence of $\mathbf{Y}$ and X.

Consistency is generally uncontroversial, but no confounding is a strong assumption. Under these assumptions,

$$\Pr(Y = y|X = x) = \Pr(Y(x) = y). \tag{1}$$

We suppose we have access to extensive data supplying exact values for expression (1), for $x, y \in \{0, 1\}$.

Define

$$\tau := \Pr(Y(1) = 1) - \Pr(Y(0) = 1),$$

$$\rho := \Pr(Y(1) = 1) - \Pr(Y(0) = 0).$$

Then, τ is the *average causal effect* of X on Y, while ρ is an indicator of how common $Y = 1$ is (as seen more immediately when we rearrange to write $\rho = \Pr(Y(1) = 1) + \Pr(Y(0) = 1) - 1$). We note that both τ and ρ can be calculated from the available data.

The transition matrix P from X to Y (where the row and column labels of any such matrix are implicitly 0 and 1 in that order) has as entries expression (1) for $x, y = 0, 1$. It is helpful to express it in terms of τ and ρ:

$$P = P(\tau, \rho) := \begin{pmatrix} \frac{1}{2}(1 + \tau - \rho) & \frac{1}{2}(1 - \tau + \rho) \\ \frac{1}{2}(1 - \tau - \rho) & \frac{1}{2}(1 + \tau + \rho) \end{pmatrix}. \tag{2}$$

All entries of P must be nonnegative. This holds if and only if

$$|\rho| + |\tau| \leq 1. \tag{3}$$

We have equality in inequality (3) if and only if one of the entries of matrix (2) is 1, in which case we term P *degenerate*. For $\tau \geq 0$, this will happen if either $\rho = 1 - \tau$, in which case $\Pr(Y = 1|X = 1) = 1$ and $X = 1$ can be thought of as a sufficient condition for $Y = 1$; or $\rho = \tau - 1$, in which case $\Pr(Y = 1|X = 0) = 0$, and $X = 1$ can be thought of as a necessary condition for $Y = 1$. Define

$$\sigma := \begin{cases} \frac{\rho}{1 - \tau} & (\tau \in [0, 1)) \\ 1 & (\tau = 1) \end{cases}. \tag{4}$$

Then, $\sigma \in [-1, 1]$ is a measure the *relative sufficiency* of $X = 1$ for $Y = 1$. Intuitively σ captures the distribution of weight between the lower left and upper-right cells of the matrix (2) with $\tau \in [0, 1)$. In this case, the entries in these cells sum to $1 - \tau$ with share $(1 - \sigma)/2$ in the lower-left cell and share $(1 + \sigma)/2$ in the upper-right cell.

Causes of Effects

While knowledge of the transition matrix P, and in particular the "average causal effect" τ, is directly relevant for studying "effects of causes," it is not enough for analyzing "causes of effects."

Using the notation $\bar{x}$ to denote $1 - x$, we can now define the following events in terms of $\mathbf{Y}$:

General causation: $C^{(X,Y)} :=$ "$Y(1) \neq Y(0)$".

That is, changing the value of X will result in a change to the value of Y. We can also describe this as "X affects Y."

When the relevant variables (here X and Y) are clear from the context, we will simplify the notation to C.

Specific causation: $C_{xy}^{(X,Y)} :=$ "$Y(x) = y, Y(\bar{x}) = \bar{y}$" (for $x, y = 0$ or 1).

That is, changing the value of X from x to $\bar{x}$ would change the value of Y from y to $\bar{y}$. We can also describe this as "$X = x$ causes $Y = y$." When the relevant variables X and Y are clear from the context, we will simplify the notation to C_{xy}.

We note that $C_{xy} = C_{\overline{xy}}$.

Probability of Causation

In cases of interest, we will have observed $X = x, Y = y$, and want to know *the probability that X caused Y*, given this information. We denote this quantity by $\mathrm{PC}_{xy}^{(X,Y)}$, or PC_{xy} when the relevant variables X and Y are clear from the context. Thus,

$$\mathrm{PC}_{xy} = \Pr(C | X = x, Y = y) = \Pr(C_{xy} | Y(x) = y), \tag{5}$$

by consistency and no confounding.

Note that, unlike for the definition of the average causal effect, the probability of causation conditions on a value for the outcome. Our PC_{11} is what Pearl (1999) terms the "probability of necessity," PN, while our PC_{00} is his "probability of sufficiency," PS.

Simple Bounds

The joint distribution for $\mathbf{Y}$, while constrained by knowledge of the transition matrix P, is in general not fully determined by it. Rather, we can only deduce

Table 1. Joint distribution of $Y(0)$ and $Y(1)$.

	$Y(1)=0$	$Y(1)=1$	
$Y(0)=0$	$\frac{1}{2}(1-\rho-\xi)$	$\frac{1}{2}(\xi+\tau)$	$\frac{1}{2}(1+\tau-\rho)$
$Y(0)=1$	$\frac{1}{2}(\xi-\tau)$	$\frac{1}{2}(1+\rho-\xi)$	$\frac{1}{2}(1-\tau+\rho)$
	$\frac{1}{2}(1-\tau-\rho)$	$\frac{1}{2}(1+\tau+\rho)$	1

that it has the form of Table 1, where the marginal probabilities agree with the entries of matrix (2).

However, the internal entries of Table 1 are not determined by P but have one degree of freedom, expressed by the "slack" quantity $\xi = \xi(P)$. We see that

$$\xi = \Pr(Y(0)=0, Y(1)=1) + \Pr(Y(0)=1, Y(1)=0) = \Pr(C), \tag{6}$$

the probability of general causation.

The only constraints on ξ are that all internal entries of Table 1 must be nonnegative, which holds if and only if

$$|\tau| \leq \xi \leq 1 - |\rho|. \tag{7}$$

In particular ξ, and thus the bivariate distribution of $(Y(0), Y(1))$ in Table 1, is uniquely determined by P if and only P is degenerate. More generally from equation (7), we see the distinct roles played by τ and ρ. The larger is τ in absolute magnitude, the greater the lower bound on ξ. The larger is ρ in absolute magnitude, the lower is the upper bound on ξ: If $Y=1$ is either very common or very uncommon then one or other off-diagonal cell in equation (2) is small, thus limiting the share of cases with $Y(0) \neq Y(1)$.

We further note

$$\Pr(C_{00}) = \Pr(C_{11}) = \frac{1}{2}(\xi+\tau), \tag{8}$$

$$\Pr(C_{01}) = \Pr(C_{10}) = \frac{1}{2}(\xi-\tau), \tag{9}$$

whence, by inequality (7),

$$\max\{0, \tau\} \leq \Pr(C_{00}) = \Pr(C_{11}) \leq \frac{1}{2}(1+\tau-|\rho|), \tag{10}$$

$$\max\{0, -\tau\} \leq \Pr(C_{01}) = \Pr(C_{10}) \leq \frac{1}{2}(1 - \tau - |\rho|). \tag{11}$$

Since $C_{xy} \Rightarrow Y(x) = y$,

$$PC_{xy} = \frac{\Pr(C_{xy})}{\Pr(Y(x) = y)}$$

which is thus subject to the interval bounds, given by equation (10) or (11), as appropriate, divided by the known entry $\Pr(Y(x) = y)$ of the transition matrix P.

This analysis delivers the following lower and upper bounds (superscript "s" for "simple"):

$$L^s_{00} := \frac{\max\{0, \tau\}}{\Pr(Y(0) = 0)} \leq \mathrm{PC}_{00} \leq \frac{\frac{1}{2}(\tau + 1 - |\rho|)}{\Pr(Y(0) = 0)} =: U^s_{00}, \tag{12}$$

$$L^s_{10} := \frac{\max\{0, -\tau\}}{\Pr(Y(1) = 0)} \leq \mathrm{PC}_{10} \leq \frac{\frac{1}{2}(1 - |\rho| - \tau)}{\Pr(Y(1) = 0)} =: U^s_{10}, \tag{13}$$

$$L^s_{01} := \frac{\max\{0, -\tau\}}{\Pr(Y(0) = 1)} \leq \mathrm{PC}_{01} \leq \frac{\frac{1}{2}(1 - |\rho| - \tau)}{\Pr(Y(0) = 1)} =: U^s_{01}, \tag{14}$$

$$L^s_{11} := \frac{\max\{0, \tau\}}{\Pr(Y(1) = 1)} \leq \mathrm{PC}_{11} \leq \frac{\frac{1}{2}(\tau + 1 - |\rho|)}{\Pr(Y(1) = 1)} =: U^s_{11}. \tag{15}$$

In the absence of additional information, the above bounds constitute the best available inference regarding the probability of causation.

Specifically, when $\tau \geq 0$, on defining

$$\gamma := \frac{1 - \tau - |\rho|}{1 - \tau + |\rho|} = \frac{1 - |\sigma|}{1 + |\sigma|}, \tag{16}$$

$$\delta := \frac{1 + \tau - |\rho|}{1 + \tau + |\rho|} \tag{17}$$

we have the upper bounds given in Table 2.

A particular interest is in cases where $\tau > 0$ (so the overall effect of X and Y is positive), and we observe positive outcomes, $X = 1$, $Y = 1$. In this case, we omit the subscript 11. We have

Table 2. U^s_{xy} Denotes the upper bound on the probability that $X = x$ caused $Y = y$ in a one-step process.

	$\rho \geq 0$	$\rho < 0$
U^s_{00}	1	δ
U^s_{01}	γ	1
U^s_{10}	1	γ
U^s_{11}	δ	1

$$PC = \frac{\xi + \tau}{2\Pr(Y(1) = 1)}, \tag{18}$$

and interval bounds given by

$$L^s = \frac{2\tau}{1 + \tau + \rho} \leq \text{PC} \leq U^s = \begin{cases} \delta & (\rho \geq 0) \\ 1 & (\rho < 0). \end{cases} \tag{19}$$

This result agrees with Tian and Pearl (2000) and Dawid (2011).

PC is identified (i.e., the interval inequality (19) reduces to a single point) if and only if $|\rho| = 1 - \tau$, which holds when P is degenerate with either the lower-left or upper-right element of P being 0. In the former case $PC = \tau$, while in the latter case $PC = 1$.

More generally, we have $L^s = \tau/\Pr(Y(1) = 1) \geq \tau$, and so $PC \geq \tau$.

Bounds From Mediation

We now suppose that, in addition to X and Y, we can gather data on one or more binary mediator variables $M_1, \ldots, M_{n-1}$. We also define $M_0 \equiv X$ and $M_n \equiv Y$. We are interested in assessing the probability that $X = x$ caused $Y = y$ for a new case where we have information on the values of some or all of the mediators $M_1, \ldots, M_{n-1}$.

Assumptions

We confine attention to the case of a *complete mediation sequence*, where for every $i \in \{0, \ldots, n-1\}$, M_{i+1} depends on M_i but not on $M_j, j < i$. Formally, we introduce, for $i \geq 1$, bivariate variables

$$\mathbf{M}_i := (M_i(0), M_i(1)),$$

where $M_i(m)$ denotes the potential value for M_i when we intervene to set M_j to m_j, $j < i - 1$, and M_{i-1} to m. As the notation expresses,[2] this value is supposed not to depend on the values set for M_j's prior to the immediate predecessor.

We assume:

Consistency: Even when some or all of the previous M's are not set by intervention, the value of (M_i) will be $M_i(M_{i-1})$.

No confounding: We have mutual independence between X, $\mathbf{M}_1, \ldots, \mathbf{M}_n$.

Then,

$$\Pr(M_{i+1} = m_{i+1} | M_j = m_j, j = 0, \ldots, i) = \Pr(M_{i+1}(m_i) = m_{i+1}).$$

Thus, the sequence $(X \equiv M_0, \ldots, M_n \equiv Y)$ forms a (generally nonstationary) Markov chain. This is an empirically testable consequence of our assumptions. Our assumptions would therefore be falsified if the Markov property is found to fail, for instance if we found that X were correlated with Y conditional on $M_1 = 1$. We note that the converse does not hold: These assumptions are not guaranteed to be valid when the Markov property is not found to fail.

Finally, we assume that we have access to data sufficient to accurately determine the one-step transition probabilities

$$\Pr(M_{i+1}(m_i) = m_{i+1}) = \Pr(M_{i+1} = m_{i+1} | M_i = m_i), \quad (i = 0, \ldots, n-1). \tag{20}$$

Inferences on Chains

In this section, we establish that the probability that X caused Y is given by the probabilities that each step in the chain from X to Y was caused by its predecessor.

Let the transition matrix from M_{i-1} to M_i be $P_i = P(\tau_i, \rho_i)$, and the overall transition matrix from X to Y be $P = P(\tau, \rho)$. We shall write,

$$P = P_1 | P_2 \ldots | P_n \tag{21}$$

to indicate that we are assuming the above mediation sequence, and refer to equation (21) as a *decomposition* of the matrix P. In particular, we then have $P = P^{(n)} := \prod_{i=1}^{n} P_i$.

We can readily show by induction that

$$\tau = \tau^{(n)} := \prod_{i=1}^{n} \tau_i, \tag{22}$$

$$\rho = \rho^{(n)} := \sum_{i=1}^{n} \left(\rho_i \prod_{j=i+1}^{n} \tau_j \right). \tag{23}$$

In particular, for the case $n = 2$, equation (23) becomes

$$\rho = \rho_1 \tau_2 + \rho_2. \tag{24}$$

On account of equation (22), we have the following result:

Lemma 1: The average causal effect of X on Y is the product of the successive average causal effects of each variable in the sequence on the following one.

Lemma 2: $C^{(X,Y)} = \cap_{i=0}^{n-1} C^{(M_i, M_{i+1})}$. That is to say, $M_0 \equiv X$ affects $M_n \equiv Y$ if and only if each M_i affects the next.

Proof. Suppose first that each variable affects the next. Then, changing the value of X will change that of M_1, which in turn will change that of M_2, and so on until the value of Y is changed, so showing that X affects Y. Conversely, if, for some $j < n$, M_j does not affect M_{j+1}, then, whether or not M_j has been changed, the value of M_{j+1} will be unchanged, whence so too will that of M_{j+2}, and so on until the value of Y is unchanged, whence X does not affect Y. □

We have as a corollary that for any decomposition, the probability that X affects Y is the product of the probabilities that each variable in the sequence from X to Y affects the next in the sequence.

Corollary 1.

i. $\Pr(C^{(X,Y)}) = \prod_{i=1}^{n} \Pr(C^{(M_{i-1}, M_i)})$,

ii. $\xi(P) = \prod_{i=1}^{n} \xi(P_i)$,

iii. Given knowledge of the decomposition (21), the constraints on $\xi = \xi(P)$ are now:

$$|\tau| \leq \xi \leq \prod_{i=1}^{n} (1 - |\rho_i|). \tag{25}$$

Proof.

i. By the assumed mutual independence of the $(\mathbf{M}_i)$.
ii. By equation (6).
iii. By (ii), inequality (7) for each P_i , and equation (22). □

On comparing inequality (25) with inequality (7), we see that detailed knowledge of the mediation process has not changed the lower bound for ξ. However, the upper bound is typically reduced:

Theorem 1. The upper bound that results from knowledge of the decomposition of P is no greater than the upper bound that results from P alone. It will be strictly less if for some $i > 1$, P_i is nondegenerate and $\rho_{i-1} \neq 0$.

Proof. We compare the upper bound of inequality (25) with that of inequality (7).

Consider first the case $n = 2$. Then,

$$|\rho| = |\rho_1\tau_2 + \rho_2|, \quad \text{by equation (24)},$$

$$\leq |\rho_1||\tau_2| + |\rho_2|, \tag{26}$$

$$\leq |\rho_1|(1 - |\rho_2|) + |\rho_2| \quad \text{by inequality (3)}. \tag{27}$$

It follows that

$$(1 - |\rho_1|)(1 - |\rho_2|) \leq 1 - |\rho|. \tag{28}$$

Moreover, we shall have strict inequality in (27), and hence also in (28), if P_2 is nondegenerate and $\rho_1 \neq 0$, since these together imply $|\rho_1|(1 - |\rho_2|) < |\rho_1||\tau_2|$.

Noting that if $(1 - |\rho_1|)(1 - |\rho_2|) = 1 - |\rho|$, then $\rho_2 \neq 0$ implies $\rho \neq 0$, the result for general n follows by induction. □

We note that the above condition for strict inequality (28), while sufficient, is not necessary. For example, in the case $n = 2$, it will also hold if $\rho_1\tau_2$ and ρ_2 have different signs, since then we would have strict inequality in (26).

It follows from inequalities (25) and (28) that collapsing two mediators into a single one (for instance, by removing M_i and replacing P_i, P_{i+1} with $Q = P_iP_{i+1}$) can only increase the upper bound for ξ.

Corollary 2. Consider two decompositions $P = P_1|P_2| \ldots |P_n$ and $P = P_1| \ldots |P_i|Q|P_{i+2}| \ldots |P_n$, where $Q = P_iP_{i+1}$. Then the upper bound for ξ for the former does not exceed that for the latter.

Unobserved Mediators

Suppose first that, for the new case, we have observed $X = x$, $Y = y$, but the values of the mediators are not observed. That is, although we have data supplying the transition probabilities in equation (20) as before, we do not know the values of the mediators for the case in question. Even in this case, as was shown for the two-term decomposition in Dawid et al. (2016), knowledge of the decomposition (equation [21]) of P can alter the bounds for PC.

Indeed, in this case, equation (5) still applies, where $\Pr(C_{xy})$ is given by equation (8) or (9) as appropriate, but now with ξ subject to the revised bounds of equation (25). In each case, the lower bound is unaffected, but, by theorem 1, the upper bound is reduced.

This analysis delivers the following revised bounds (superscript "$\emptyset$" for "not observed"):

$$L_{00}^{\emptyset} := L_{00}^{s} = \frac{\max\{0, \tau\}}{\Pr(Y(0) = 0)} \leq \mathrm{PC}_{00} \leq \frac{\tau + \prod_{i=1}^{n}(1 - |\rho_i|)}{2\Pr(Y(0) = 0)} =: U_{00}^{\emptyset}, \tag{29}$$

$$L_{10}^{\emptyset} := L_{10}^{s} = \frac{\max\{0, -\tau\}}{\Pr(Y(1) = 0)} \leq \mathrm{PC}_{10} \leq \frac{\prod_{i=1}^{n}(1 - |\rho_i|) - \tau}{2\Pr(Y(1) = 0)} =: U_{10}^{\emptyset}, \tag{30}$$

$$L_{01}^{\emptyset} := L_{01}^{s} = \frac{\max\{0, -\tau\}}{\Pr(Y(0) = 1)} \leq \mathrm{PC}_{01} \leq \frac{\prod_{i=1}^{n}(1 - |\rho_i|) - \tau}{2\Pr(Y(0) = 1)} =: U_{01}^{\emptyset}, \tag{31}$$

$$L_{11}^{\emptyset} := L_{11}^{s} = \frac{\max\{0, \tau\}}{\Pr(Y(1) = 1)} \leq \mathrm{PC}_{11} \leq \frac{\tau + \prod_{i=1}^{n}(1 - |\rho_i|)}{2\Pr(Y(1) = 1)} =: U_{11}^{\emptyset}. \tag{32}$$

Note, in particular, for the case $\tau > 0$, where we observe $X = 1$, $Y = 1$ (but the values of mediators are not observed), we have revised bounds

$$L^{\emptyset} := \frac{2\tau}{1 + \tau + \rho} \leq \mathrm{PC} \leq \frac{\tau + \prod_{i=1}^{n}(1 - |\rho_i|)}{1 + \tau + \rho} =: U^{\emptyset}. \tag{33}$$

For $n = 2$, this agrees with the analysis of Dawid et al. (2016).

Bounds When Some or All Mediators are Observed

Now suppose that, in addition to $X = x$, $Y = y$, we also observe data on k mediators ($0 \leq k \leq n-1$) for the new case. In particular, we observe $M_{i_r} = m_{i_r}$, for $0 < i_1 < \ldots i_r \ldots < i_k < n$. For notational simplicity, we write $\widetilde{M}_r$ for M_{i_r}, $\widetilde{m}_r$ for m_{i_r}. We also identify $\widetilde{M}_0 \equiv X$ and $\widetilde{M}_{k+1} \equiv Y$ (so $\widetilde{m}_0 = x$, $\widetilde{m}_{k+1} = y$).

The relevant probability of causation is now

$$\widetilde{\mathrm{PC}}_{xy} := \Pr(C | \widetilde{M}_r = \widetilde{m}_r, r = 0, \ldots, k+1).$$

Note that in contrast to the difference between equations (29)–(32), on the one hand, and equations (12)–(15), on the other hand, which relate to the same quantity PC_{xy} but express different conclusions about it (since based on different external evidence), $\widetilde{\mathrm{PC}}_{xy}$ is a genuinely different quantity from PC_{xy}, as it conditions on different information about the case in question. For this reason, it is possible that the upper bound on the probability of causation for a particular case when M is observed is *higher* than the upper bound on the probability of causation for a particular case given M is not observed.

Theorem 2. Given observations on $X, \widetilde{M}_1, \ldots, \widetilde{M}_k, Y$, the probability that X caused Y is given by the product of the probabilities that each observed term in the sequence caused the next observed term:

$$\widetilde{\mathrm{PC}}_{xy} = \prod_{r=0}^{k} PC_{\widetilde{m}_r \widetilde{m}_{r+1}}^{(\widetilde{M}_r, \widetilde{M}_{r+1})}.$$

Proof. From lemma 2, we have

$$C = \bigcap_{r=0}^{k} C^{(\widetilde{M}_r, \widetilde{M}_{r+1})},$$

whence, using the "no-confounding" independence properties,

$$\widetilde{\mathrm{PC}}_{xy} = \prod_{r=0}^{k} \Pr(C^{(\widetilde{M}_r, \widetilde{M}_{r+1})} | \widetilde{M}_r = \widetilde{m}_r, \widetilde{M}_{r+1} = \widetilde{m}_{r+1}),$$

$$= \prod_{r=0}^{k} PC_{\widetilde{m}_r \widetilde{m}_{r+1}}^{(\widetilde{M}_r, \widetilde{M}_{r+1})}. \tag{34}$$

□

Now since we have the decomposition information about the mediators (if any) occurring between $\widetilde{M}_r \equiv M_{i_r}$ and $\widetilde{M}_{r+1} \equiv M_{i_{r+1}}$, but not their values for the new case, the bounds on any factor in equation (34) will, *mutatis mutandis*, have the form of the relevant expressions for $L_{xy}^{\varnothing}$ and $U_{xy}^{\varnothing}$, as displayed in inequalities (29)–(32). Then, the overall lower [resp., upper] bound on $\widetilde{\mathrm{PC}}_{xy}$ will be the product of these lower [resp., upper] bounds, across all terms. This procedure supplies a complete recipe for determining the appropriate bounds on $\widetilde{\mathrm{PC}}_{xy}$ in the knowledge of the full decomposition of P and the values of the observed mediators for the new case.

Again consider the special case with $\tau > 0$, $X = Y = 1$. On account of equation (22), we can, after possibly switching the labels 0 and 1 for some of the M_i's, take $\tau_i > 0$, all i. We assume henceforth that this is the case. The above procedure then delivers lower bound 0 unless $\widetilde{m}_i = \widetilde{m}_{i-1}$, all i, so that $m_i = 1$, all i. In that case, we obtain lower bound (with superscript + for "positive mediators"):

$$\begin{aligned} L^{+} &:= \frac{\tau}{\prod_{r=0}^{k} \Pr(\widetilde{M}_{r+1} = 1 | \widetilde{M}_r = 1)} \\ &= \frac{\tau}{\Pr(Y = 1, \widetilde{M}_r = \widetilde{m}_r, r = 2, \ldots, k | X = 1)}. \end{aligned} \tag{35}$$

It is easy to see that this lower bound can only increase if we introduce further observed mediators. It follows that the smallest lower bound occurs when there are no observed mediators, when it reduces to $L^{\varnothing} = L^s$ as in inequalities (33) and (19); while the largest lower bound occurs when all mediators are observed (all taking value 1)—that is to say, there is positive evidence for every link in the mediation chain.

In the remainder of this article, we shall give special attention to this case and write simply $\widetilde{\mathrm{PC}}$ for $\widetilde{\mathrm{PC}}_{11}$, and so on. The bounds for $\widetilde{\mathrm{PC}}$ are then:

$$L^{+} := \prod_{i=1}^{n} \left(\frac{2\tau_i}{1 + \tau_i + \rho_i} \right) \leq \widetilde{\mathrm{PC}} \leq \prod_{i=1}^{n} \left(\frac{1 + \tau_i - |\rho_i|}{1 + \tau_i + \rho_i} \right) =: U^{+}. \tag{36}$$

The following result follows directly from the above considerations:

Lemma 3. The lower bound L^{+} of inequality (36) is at least as large as the lower bound L^s of inequality (19).

Table 3. Largest and smallest achievable upper and lower bounds from decompositions of any length, given no mediators observed ($L^{\varnothing}$, $U^{\varnothing}$), positive evidence observed for all mediators (L^{+}, U^{+}), or when some negative evidence is observed (L^{-},U^{-}).

		No Evidence	Positive Evidence	Some Negative Evidence
Largest	Upper	$\overline{U^{\varnothing}} = \frac{1+\tau-\lvert\rho\rvert}{1+\tau+\rho}$	$\overline{U^{+}} = \min\{1, 1-\rho\}$	$\overline{U^{-}} = 1$
	Lower	$\overline{L^{\varnothing}} = \frac{2\tau}{1+\tau+\rho}$	$\overline{L^{+}} = \frac{1+\tau-\rho}{2}$	$\overline{L^{-}} = 0$
Smallest	Upper	$\underline{U^{\varnothing}} = \frac{2\tau}{1+\tau+\rho}$*	$\underline{U^{+}} = \frac{2\tau}{1+\tau+\rho}$*	$\underline{U^{-}} = 0$*
	Lower	$\underline{L^{\varnothing}} = \frac{2\tau}{1+\tau+\rho}$	$\underline{L^{+}} = \frac{2\tau}{1+\tau+\rho}$	$\underline{L^{-}} = 0$

*PC can be identified.

However, it will follow from theorem 3 below that U^{+} can be smaller or larger than U^{s}.

Largest and Smallest Upper and Lower Bounds

Equation (34) provides a general formula for calculating bounds on the probability of causation for any pattern of data observed on mediating variables (including no data). We now use this result to assess the largest and smallest possible upper bounds from observation of possible values on mediating variables.

Consider an arbitrary decomposition of P:

$$P = P_1|P_2| \ldots |P_n, \tag{37}$$

with $P = P(\tau, \rho)$, $P_i = P(\tau_i, \rho_i)$. We restrict attention to the case $\tau > 0$ and assume that variables are labeled so that each $\tau_i > 0$.

We investigate the smallest and largest achievable values for $L^{\varnothing}, U^{\varnothing}, L^{+}, U^{+}, L^{-}$, U^{-} (superscript $-$ for some negative evidence) and show that in each case, these are achievable by decompositions involving at most one mediator.

Theorem 3. Consider transition matrix $P = P(\tau, \rho)$ from X to Y with $\tau > 0$ and $|\rho| < 1 - \tau$. The largest and smallest upper and lower bounds on the probability that $X = 1$ caused $Y = 1$, from any complete mediation process for (a) the case with mediators unobserved (b) the case with positive outcomes on all mediators observed and (c) cases that include some negative evidence on the mediators, are as given in Table 3. These can all be achieved by decompositions of length 1 or 2.

Proof. See Online Appendix A (which can be found at Supplementary material for this article, available online). □

The largest upper bound with mediators unobserved, $\overline{U^{\varnothing}}$, can be achieved without any mediators. Since unobserved mediators do not alter the lower bound, we have $\overline{L^{\varnothing}} = \underline{L^{\varnothing}} = L^s$. In addition, we have $\underline{U^{\varnothing}} = L^s$, which is achievable, for example, from the following decomposition:

$$P = \begin{pmatrix} \frac{2\tau}{1+\tau+\rho} & \frac{1-\tau+\rho}{1+\tau+\rho} \\ 0 & 1 \end{pmatrix} \Bigg| \begin{pmatrix} 1 & 0 \\ \frac{1-\tau-\rho}{2} & \frac{1+\tau+\rho}{2} \end{pmatrix}. \tag{38}$$

Note that, with this decomposition, PC is identified *via* two degenerate transition matrices: $X = 1$ is a sufficient condition for $M = 1$, while $M = 1$ is a necessary condition for $Y = 1$.

The smallest upper and lower bounds available when mediators are observed agree with the simple lower bound. Positive evidence cannot reduce the lower bound, but it can reduce the upper bound to the lower bound, at which point $\widetilde{\mathrm{PC}}$ is identified. This can be achieved by the same decomposition given in equation (38).

The largest upper bound with positive evidence on mediators, $\overline{U^{+}}$, can exceed the simple upper bound when $\rho > 0$. It results from the following two-term decomposition, involving a single mediator:

$$P = \begin{pmatrix} \frac{1-\rho+\tau}{2(1-\rho)} & \frac{1-\rho-\tau}{2(1-\rho)} \\ \frac{1-\rho-\tau}{2(1-\rho)} & \frac{1-\rho+\tau}{2(1-\rho)} \end{pmatrix} \Bigg| \begin{pmatrix} 1-\rho & \rho \\ 0 & 1 \end{pmatrix}. \tag{39}$$

The lower bound can be raised with positive information on mediators and takes its largest value with the following degenerate two-term decomposition $P = P_1|P_2$, involving a single mediator:

$$P = \begin{pmatrix} 1 & 0 \\ \frac{1-\tau-\rho}{1+\tau-\rho} & \frac{2\tau}{1+\tau-\rho} \end{pmatrix} \Bigg| \begin{pmatrix} \frac{1+\tau-\rho}{2} & \frac{1-\tau+\rho}{2} \\ 0 & 1 \end{pmatrix}. \tag{40}$$

With this decomposition $\widetilde{\text{PC}}$ is identified *via* two degenerate transition matrices: In this case, $X = 1$ is a necessary condition for $M = 1$, while $M = 1$ is a sufficient condition for $Y = 1$. The largest lower bound with positive evidence from this decomposition is $\frac{1+\tau-\rho}{2}$ which can fall far short of 1, implying that in general mediators cannot provide "smoking gun" evidence that $X = 1$ caused $Y = 1$. A benchmark of 50 percent—a balance of probabilities—is sometimes used (e.g., in civil legal proceedings) as the standard of proof. This result shows that this standard cannot be met by any information on mediators if $\tau < \rho$, or equivalently, if $\Pr(Y = 1|X = 0) > 0.5$.

For the case with some negative evidence on the mediators, the lower bound is always 0. The smallest upper bound is also 0, which can be achieved by the decomposition of equation (40) above, with the single mediator observed at 0 (the key feature of this decomposition is that $Y = 1$ cannot be caused by $M = 0$). In this case, $\widetilde{\text{PC}}$ is identified at 0, showing that it is possible for negative data on mediators to provide "hoop" evidence that $X = 1$ did not cause $Y = 1$. The highest upper bound when there is some negative evidence, $U^- = 1$, can be achieved by a two-step decomposition, $P(\tau, \rho) = P(\tau_1, \rho_1)|P(\tau_2, \rho_2)$, with the mediator taking value 0. For $\rho \leq 0$, this occurs with the decomposition with parameters:

$$\tau_1 = \frac{2\tau}{1+\tau+\rho} \qquad \rho_1 = 0 \qquad \tau_2 = \frac{1+\tau+\rho}{2} \qquad \rho_2 = \rho. \tag{41}$$

For $\rho \geq 0$, it occurs with decomposition parametrized by:

$$\tau_1 = \frac{\tau(1+\rho+\tau)}{2(\tau+\rho)} \qquad \rho_1 = \frac{\rho(1+\rho+\tau)}{2(\tau+\rho)} \qquad \tau_2 = \frac{2(\tau+\rho)}{1+\tau+\rho} \qquad \rho_2 = 0. \tag{42}$$

Comparisons

Although knowledge of mediators can narrow bounds, the scope for learning from knowledge of mediation processes—and the specific values taken on by mediators—is often small. In particular, although negative evidence can yield low upper bounds, providing confidence that an outcome was *not* due to a putative cause, positive evidence generally does not raise lower bounds substantially.

To put these claims in context, we compare the extrema on bounds in theorem 3 with bounds that can be achieved from "homogeneous" processes, from knowledge of monotonicity, and from covariate information.

Table 4. Upper and lower bounds from homogeneous decompositions of length $n \to \infty$, given no mediators observed, positive evidence observed for all mediators, and alternating evidence.

	No Evidence	Positive Evidence	Alternating Evidence
Upper	$U_\infty^\varnothing = \frac{\tau+\tau^{\lvert\sigma\rvert}}{1+\tau+\rho}$	$U_\infty^+ = \min\{1, \tau^\sigma\}$	$U_\infty^- = \begin{cases} 0 & \text{if} \rho \neq 0 \\ 1 & \text{if} \rho = 0 \end{cases}$
Lower	$L_\infty^\varnothing = \frac{2\tau}{1+\tau+\rho}$	$L_\infty^+ = \tau^{\frac{1}{2}(1+\sigma)}$	$L_\infty^- = 0$

Homogeneous Processes

First, we consider bounds for a special case: long homogeneous processes—that is, cases in which we have a potentially unlimited sequence of variables directly mediating between X and Y, with one-step transition matrices that are identical at each step (and having positive average causal effect). For such processes, $\Pr(M_{i+1}(m) = m') = \Pr(M_i(m) = m')$.

Intuitively, a lot of data at many points in a chain should lead to stronger inferences. This intuition is however not in line with our finding that the extrema on the bounds given in theorem 3 are generally achieved through two-step processes in which transition matrix P_1 is different from transition matrix P_2. The bounds from long processes can be no better than those described in theorem 3, but how different are they?

Table 4 shows the upper and lower bounds achievable with homogeneous processes of unbounded length, for three cases: cases in which there are no data on the values of the mediators, cases in which all mediators are observed and positive ($M_t = 1$ for all t), and cases in which values alternate between 1 and 0. For further details, see Online Appendix B (which can be found at Supplementary material for this article is available online).

We see that, for $\rho \neq 0$, with alternating evidence, identification can be achieved in the limit, at 0. In other cases, however, identification is not achieved. In particular, the lower bound with positive evidence can fall far short of the highest possible lower bound, especially when $|\rho|$ and τ are small. For example, if $\rho = 0$, then $\overline{L^+} - L_\infty^+ = \frac{1}{2}(1-\sqrt{\tau})^2$.

Monotonicity

Suppose that we knew that there are no cases for which the exposure would prevent the outcome, that is, such that $Y(0) = 1, Y(1) = 0$. We note that since monotonicity is an attribute of the typically unidentifiable joint distribution of $(Y(0), Y(1))$, it is not easy to justify without additional knowledge.

One case where this is possible is when we know of the existence of a mediation process with decomposition as in equation (38).

From Table 1, we have that monotonicity implies $\xi = \tau$, that is, ξ is identified at its lower limit. In turn, this implies that PC, given by equation (18), is identified at its lower limit, $L^s = 2\tau/(1 + \tau + \rho)$. In this case, knowledge of the value of mediators does nothing to raise the lower bound.

Observed Covariate

Suppose that, in addition to X and Y, we can observe a binary covariate W, pretreatment to X, which can affect the dependence of Y on X. Let $\pi := \Pr(W = 1)$, and let P_i be the transition matrix from X to Y, conditional on $W = i$; for consistency with the known $P = P(\tau, \rho)$ we must have $P = \pi P_1 + (1 - \pi)P_0$.

In particular, it could then be the case that $\pi = (1 + \tau - \rho)/2$, and

$$P_1 = \begin{pmatrix} 1 & 0 \\ \frac{1-\tau-\rho}{2} & \frac{1+\tau+\rho}{2} \end{pmatrix} \qquad P_0 = \begin{pmatrix} 0 & 1 \\ \frac{1-\tau-\rho}{2} & \frac{1+\tau+\rho}{2} \end{pmatrix}.$$

In this case, knowledge that an individual with $X = Y = 1$ also has $W = 1$ is enough to identify PC at 1. We emphasize that we use an extreme decomposition here not to argue that such a decomposition is likely but rather to highlight that there is always a possibility for full identification at 1 with observed covariates whereas identification at 1 with mediators is generally not obtainable.

Unobserved Covariate

As shown in Dawid (2011), knowledge of covariates can improve bounds, even if their values are not observed for the case at hand. In particular, this can let us identify PC at the upper bound, $U^s = \min\{1, \frac{1+\tau-\rho}{1+\tau+\rho}\}$. For this to be possible, however, the average treatment effect must be negative for some value of W. Thus, suppose again the W is pretreatment to X and $\pi := \Pr(W = 1)$. Suppose then that $\pi = \frac{1+\tau+\rho}{2}$, and the conditional transition matrices are the following:

For $\rho < 0$,

$$P_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad P_0 = \begin{pmatrix} \frac{-2\rho}{1-\tau-\rho} & \frac{1+\tau-\rho}{1-\tau-\rho} \\ 1 & 0 \end{pmatrix}.$$

For $\rho \geq 0$,

$$P_1 = \begin{pmatrix} \frac{1+\tau-\rho}{1+\tau+\rho} & \frac{2\rho}{1+\tau+\rho} \\ 0 & 1 \end{pmatrix} \qquad P_0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

In either case, knowledge that $X = Y = 1$ is sufficient to infer that $W = 1$. This identifies the probability of causation: $PC = 1$ for $\rho < 0$, $PC = \frac{1+\tau-\rho}{1+\tau+\rho}$ for $\rho \geq 0$. In both cases we hit the upper bound.

Figure 1 compares the bounds obtained, under various assumptions, for a range of values of τ and ρ. It illustrates how, in general, lower bounds rise with τ and fall with ρ. For homogeneous processes, the lower bounds improve on the simple bounds, although the gain from unlimited steps is not a striking improvement on that for just two steps. The gains from nonhomogeneous decompositions can be substantial. The best lower bounds achievable from knowledge of covariates are higher than lower bounds achieved from any knowledge of mediators.

Conclusion

We provide a general formula for calculating bounds on the probability of causation for complete mediation processes involving binary variables of arbitrary length and with arbitrary data patterns. In addition, we characterize the largest and smallest achievable bounds obtainable from any data. Knowledge of these bounds is useful for assessing when there can be gains from learning about processes in a population and gains from learning about the values of mediators for cases.

Our analysis focuses on ideal cases in which there is a very simple known causal structure in which nodes are connected in a simple causal chain—excluding situations such as one in which X has a direct effect on Y as well as an indirect effect through M. We show, however, that even in these ideal conditions, access to even unlimited data on mediators has only a modest, and asymmetric, impact on inferences. Knowledge of mediation processes, and of positive values for some mediators in a particular case, can raise the

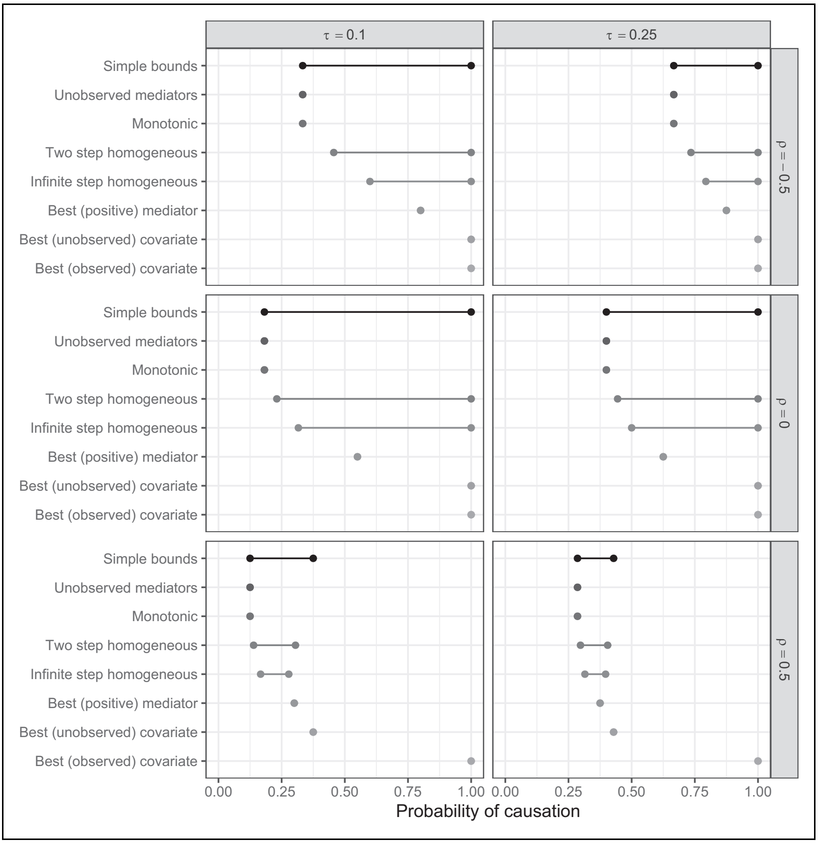

Figure 1. Comparison of bounds on PC. Simple bounds are derived from the distribution of Y given X and are given by inequality (19). Tightest bounds from unobserved mediators are given by the decomposition in (38). Monotonicity implies the same bounds. Bounds from a homogeneous two-step decomposition and positive evidence can be calculated from theorem 3. Infinite-step bounds, assuming positive evidence observed at every step from a homogeneous process, are given in Table 4. Best two-step bounds show the highest lower bound achievable from information on mediation shown in Table 3 and can be achieved with positive evidence for the decomposition of equation (40). Greatest lower bounds given information on an unobserved and observed binary covariate are as described in the Comparisons section.

lower bound on the probability of causation, thus providing some evidence against a skeptic who doubts that the outcome in the case can be attributed to the putative cause. Moreover, this information can be enough to achieve identification. However, the gains are generally modest and may not be sufficient to convince a skeptic. For instance, if most outcomes are positive for untreated units, then it follows from our results that there is no evidence on mediators for a treated unit with positive outcomes that can raise the lower bound on the probability that the outcome was due to the treatment above 50 percent. More generally, identification at 1 is not possible. In contrast, for some processes, observing negative evidence on a single mediator can effectively convince a skeptic that the outcome is *not* due to the exposure.

These general results have implications for when gathering further intermediate data on particular cases can be useful. We see, for instance, starkly contrasting implications for a process in which X is a necessary condition for a sufficient condition for Y and a process in which X is a sufficient condition for a necessary condition for Y. In the first case, consistent with arguments in Mahoney (2012), negative evidence on mediators implies no causal effect—we have a hoop test. In addition, we show, positive evidence on mediators yields the largest possible upper bound and identifies the probability of causation. For example, if it is known that the effect of delivering a deworming medicine passes uniquely through ingestion, and ingestion is sufficient for effective deworming, then evidence of ingestion raises the lower bound and identifies the probability of causation. These features, we note, depend on the chain structure we specify: were there a possible direct effect from X to Y, then necessity followed by sufficiency would not imply a hoop test because knowledge that X did not cause M is not sufficient to conclude that X did not cause Y.

In contrast for a process in which X is a sufficient condition for a necessary condition for Y, we already enjoy identification and there is no gain from gathering data on the mediator. For instance, if ingesting medicine is a sufficient condition for good health, and good health is a necessary condition for good school performance, then observing ingestion and good school performance is sufficient to achieve identification. There are no additional gains from measuring health, since good health is already implied by good performance. A similar logic holds for any chain of necessary relations, suggesting that these do not in fact aggregate to form a smoking gun test since if $M = 1$ is necessary for $Y = 1$, then the value of M is already known from observing $Y = 1$.

The main result can also be used to guide choice of *which* causal process observations to examine. For instance, consider a homogeneous process with n steps (n even) and suppose that researchers can observe the value of just one mediator M_i. In this case, we can show that the lower bound on the probability of causation, following observation of positive data, is maximized if the central mediator in the sequence is observed. For intuition, there is more ex ante certainty about the values of mediators close to the edges; ex ante uncertainty increases, and the scope for learning increases accordingly far from the edges. See Appendix C for details.

Finally, these results also have implications for the potential gains from research agendas that seek to learn about mediation processes (as, e.g., in the designs described in Imai, Keele, and Tingley 2010) compared to the potential gains from learning about effect heterogeneity (as, e.g., is done in factorial designs; Fisher 1926). The scope for gains from knowledge of mediation processes is typically weaker than potential gains from knowledge of conditions under which interventions are more or less effective. While of course the actual gains from knowledge of mediators and covariates depends on underlying causal relations, by providing extrema on bounds, the results we provide can inform the choice of experimental design.

Authors' Note

Monica Musio's research supported by the project GESTA of the Fondazione di Sardegna and Regione Autonoma di Sardegna, Sardegna, Italy.

Acknowledgment

The authors thank Steffen Huck, Alan Jacobs, Lily Medina, and Michael Zürn for their generous comments.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Philip Dawid https://orcid.org/0000-0002-7410-6882
Macartan Humphreys https://orcid.org/0000-0001-7029-2326

Supplemental Material

The supplemental material is available in the online version of the article.

Notes

1. General procedures for deriving bounds on causal queries are given in Sachs et al. (2021) though unfortunately these cannot be used for the problem considered here as our causal query is in general not linear, or, in their formulation, not a linear function of joint probabilities of response function variables.
2. But note that, although we are identifying M_n with Y, we will distinguish between $M_n(a)$, the potential value of $M_n = Y$ when setting M_{n-1} to a, and $Y(a)$, the potential value of Y when setting X to a $(a = 0, 1)$.

References

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science & Politics* 44(4):823-30.

Dawid, Alexander Philip. 2011. "The Role of Scientific and Statistical Evidence in Assessing Causality." Pp. 133-47 in *Perspectives on Causation*, edited by Richard Goldberg. Oxford, England: Hart.

Dawid, Alexander Philip, Rossella Murtas, and Monica Musio. 2016. "Bounding the Probability of Causation in Mediation Analysis." Pp. 75-84 in *Topics on Methodological and Applied Statistical Inference*, edited by Tonio Di Battista, Elías Moreno, and Walter Racugno. Cham, Switzerland: Springer.

Dawid, Alexander Philip, Monica Musio, and Stephen E. Fienberg. 2016. "From Statistical Evidence to Evidence of Causality." *Bayesian Analysis* 11:725-52.

Fisher, Ronald A. 1926. "The Arrangement of Field Experiments." *Journal of the Ministry of Agriculture of Great Britain* 33:503-13.

Gelman, Andrew and Guido Imbens. 2013. Why Ask Why? Forward Causal Inference and Reverse Causal Questions. Working Paper 19614 National Bureau of Economic Research. Retreived on January 26, 2022. DOI 10.3386/w19614.

Ghate, Deborah. 2018. "Developing Theories of Change for Social Programmes: Co-producing Evidence-supported Quality Improvement." *Palgrave Communications* 4(1):1-13.

Gross, Neil. 2018. "The Structure of Causal Chains." *Sociological Theory* 36(4): 343-67.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15(4):309.

Knight, Carly R and Christopher Winship. 2013. "The Causal Implications of Mechanistic Thinking: Identification Using Directed Acyclic Graphs (DAGs)." Pp. 275-99 in *Handbook of Causal Analysis for Social Research*, edited by Stephen L Morgan. Cham, Switzerland: Springer.

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods & Research* 41(4):570-97.

Murtas, Rossella, Alexander Philip Dawid, and Monica Musio. 2017. "New Bounds for the Probability of Causation in Mediation Analysis." arXiv:1706.04857. math. ST.

Pearl, Judea. 1999. "Probabilities of Causation: Three Counterfactual Interpretations and Their Identification." *Synthese* 121:93-149.

Pearl, Judea. 2015. "Causes of Effects and Effects of Causes." *Sociological Methods & Research* 44(1):149-64.

Robins, James and Sander Greenland. 1989. "The Probability of Causation under a Stochastic Model for Individual Risk." *Biometrics* 45:1125-38.

Sachs, Michael C, Gustav Jonzon, Erin E. Gabriel, and Arvid Sjölander. 2021. "A General Method for Deriving Tight Symbolic Bounds on Causal Effects." arXiv: 2003.10702. stat.ME.

Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge, England: Cambridge University Press.

Tian, Jin and Judea Pearl. 2000. "Probabilities of Causation: Bounds and Identification." *Annals of Mathematics and Artificial Intelligence* 28:287-313.

Van and Stephen Evera. 1997. *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Weller, Nicholas and Jeb Barnes. 2016. "Pathway Analysis and the Search for Causal Mechanisms." *Sociological Methods & Research* 45(3):424-57.

White, Howard. 2009. "Theory-based Impact Evaluation: Principles and Practice." *Journal of Development Effectiveness* 1(3):271-84.

Yamamoto, Teppei. 2012. "Understanding the Past: Statistical Analysis of Causal Attribution." *American Journal of Political Science* 56(1):237-56.

Author Biographies

Philip Dawid is Emeritus Professor of Statistics of the University of Cambridge and a Fellow of Darwin College, Cambridge. His research focuses on foundations of probability, statistics and causal inference, and forensic statistic and legal reasoning. Recent publications include "Extended Conditional Independence and Applications in Causal Inference," *Annals of Statis*tics 45(6):2618-53 with P. Constantinou; "The Probability of Causation," *Law, Probability and Risk* 16(4):163-79 with M. Musio and R. Murtas; and "On Individual Risk," *Synthese* 194(9):3445-74.

Macartan Humphreys is professor of political science at Columbia University and director of the Institutions and Political Inequality Unit at WZB Berlin. His research interests include the political economy of development and causal inference. His recent publications include "The Aggregation Challenge," *World Development* with A Scacco (2020); "Information Technology and Political Engagement: Mixed Evidence from Uganda," *Journal of Politics* with G. Grossman and G. Sacramone Lutz

(2020); and "Declaring and Diagnosing Research Designs," *American Political Science Review* with G. Blair, J. Cooper, and A. Coppock (2019).

Monica Musio is professor of statistics at Università di Cagliari. Recent work focuses on causal inference and time-series analysis. Recent publications include "The Hyvärinen Scoring Rule in Gaussian Linear Time Series Models," *Journal of Statistical Planning and Inference* (2021) with Columbu, V. Mameli, and A. P. Dawid; "What Can Group Level Data Tell Us About Individual Causality?" in *Stephen E. Fienberg—A Public Statistician*, edited by Alicia Carriquiry, Bill Eddy, and Judith Tanur, Springer, with A. P. Dawid; and "Robust Inference for Nonlinear Regression Models from the Tsallis Score: Application to Covid-19 Contagion in Italy," *Stat* (2020) with P. Girardi, L. Greco, V. Mameli, W.Racugno, E. Ruli, and L. Ventura.