Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2021, Vol. 30(6) 1395–1412

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280221998409
journals.sagepub.com/home/smm
SAGE

Mediation effects that emulate a target randomised trial: Simulation-based evaluation of ill-defined interventions on multiple mediators

Margarita Moreno-Betancur[1,2], Paul Moran[3], Denise Becker[2], George C Patton[1,2] and John B Carlin[1,2]

Abstract

Many epidemiological questions concern potential interventions to alter the pathways presumed to mediate an association. For example, we consider a study that investigates the benefit of interventions in young adulthood for ameliorating the poorer mid-life psychosocial outcomes of adolescent self-harmers relative to their healthy peers. Two methodological challenges arise. First, mediation methods have hitherto mostly focused on the elusive task of discovering pathways, rather than on the evaluation of mediator interventions. Second, the complexity of such questions is invariably such that there are no well-defined mediator interventions (i.e. actual treatments, programs, etc.) for which data exist on the relevant populations, outcomes and time-spans of interest. Instead, researchers must rely on exposure (non-intervention) data, that is, on mediator measures such as depression symptoms for which the actual interventions that one might implement to alter them are not well defined. We propose a novel framework that addresses these challenges by defining mediation effects that map to a target trial of hypothetical interventions targeting multiple mediators for which we simulate the effects. Specifically, we specify a target trial addressing three policy-relevant questions, regarding the impacts of hypothetical interventions that would shift the mediators' distributions (separately under various interdependence assumptions, jointly or sequentially) to user-specified distributions that can be emulated with the observed data. We then define novel interventional effects that map to this trial, simulating shifts by setting mediators to random draws from those distributions. We show that estimation using a g-computation method is possible under an expanded set of causal assumptions relative to inference with well-defined interventions, which reflects the lower level of evidence that is expected with ill-defined interventions. Application to the self-harm example in the Victorian Adolescent Health Cohort Study illustrates the value of our proposal for informing the design and evaluation of actual interventions in the future.

Keywords

Mediation, ill-defined interventions, interventional effects, natural effects, target trial, multiple mediators, randomised controlled trial, causal inference

[1] Department of Paediatrics, University of Melbourne, Melbourne, Australia
[2] Murdoch Children's Research Institute, Melbourne, Australia
[3] Centre for Academic Mental Health, School of Social & Community Medicine, University of Bristol, Bristol, UK

Corresponding author:
Margarita Moreno-Betancur, Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Royal Children's Hospital, 50 Flemington Road, Parkville, Victoria 3052, Australia.
Email: margarita.moreno@mcri.edu.au

1 Introduction

In areas such as life course and social epidemiology, questions arise around potential interventions to alter pathways presumed to mediate an association, such as between an early-life marker of vulnerability and later outcomes. Our motivating example investigated potential interventions to counter the poorer psychosocial outcomes in adulthood of adolescents who self-harm relative to their healthy peers, such as targeting substance use and mental health problems in young adulthood. Addressing such questions raises two key methodological challenges, which this paper aims to tackle.

The first challenge relates to the focus of the mediation literature on the discovery of mechanistic pathways.[1] The prevailing logic is to assume a pre-existing (axiomatic) notion of mediation and then to define "indirect" effects so as to detect and quantify this, with the modern definitions in the potential outcomes framework referred to as "natural" effects.[2–5] These effects are not defined in a way that makes them empirically measurable, even hypothetically, in a randomised experiment,[1,6] and alternative methods that would explicitly address the issue of mediator intervention evaluation have been lacking. This is striking given that the implied appeal of discovering pathways is often to reveal potential intervention points. It also contrasts with current thinking in the broader epidemiological literature, where the elusive nature of the notion of "causation"[7,8] (of which "mediation" is an extension), tied to aspirations for an epidemiology of consequence,[9] has brought a move away from the quest for the discovery of causes. Instead, emphasis is given to the more tangible goal of assessing effects of causes conceptualised as interventions,[7,10–12] with analyses designed to emulate a "target trial",[13,14] defined as the ideal randomised trial that one would hypothetically perform to evaluate the intervention in question.

The second challenge is that the endeavour of intervention evaluation presupposes the existence of well-defined interventions. However, the complexity of the questions being asked in many areas, such as the self-harm example, is often such that there are no well-defined interventions for which data have been or could be collected to directly assess impact for the populations, outcomes and time-spans of interest. Instead, to address their questions, researchers have to rely on observational exposure (non-intervention) data, for example from long-term longitudinal cohort studies, and use mediator measures such as depression symptoms for which the actual interventions that one might implement to alter them are not well defined. There has been much criticism of such "exposure epidemiology" for causal inference, yet producing some evidence, even if imperfect, is arguably a key first step to future intervention development and evaluation.[15] This explains a recent push[10,15–17] for addressing, rather than shunning, the methodological challenge of ill-defined interventions, and it has been suggested that simulation-based evaluation of hypothetical interventions might be needed.[10]

In this work, we reverse the logic that has driven the mediation literature: rather than assuming a pre-existing notion of mediation, we propose to start with specific policy-relevant questions relating to mediator interventions and then define effects to address these in explicit correspondence to a target trial. We show that, within this logic, mediation effects are not required if the question and available data pertain to well-defined mediator interventions, but mediation regains its relevance in the context of ill-defined interventions, in the form of so-called "interventional effects" (a.k.a. "interventional randomised analogues").[18–21]

Specifically, recent work shows that interventional mediation effects implicitly emulate effects in target trials that evaluate the impacts of distributional shifts in the mediators.[22] We propose that conceptualising such distributional shifts as arising from hypothetical interventions provides a useful framework for simulating potential effects and thus tackle the issue of ill-defined mediator interventions, in particular as this acknowledges the composite nature of the exposures under consideration.[16] However, given their unintentional (implicit) nature, the target trials emulated by previously proposed interventional effects for the setting with multiple mediators and a time-fixed–exposure[19,23] are not necessarily relevant for informing policy (see section "Summary and comparison with previous effects"). Therefore, we define novel interventional effects explicitly in terms of a target trial that addresses three specific policy-relevant questions, regarding the impacts of intervening to shift mediators separately (under various interdependence assumptions), jointly or sequentially.

The paper is structured as follows. First, we introduce the self-harm example. Second, we introduce the issue of ill-defined mediator interventions and propose a novel conceptual framework under a set of principles for tackling it via simulation of hypothetical interventions. Third, we describe the target trial integrating these principles and derive novel definitions of interventional effects that map to that trial, with a description of how these compare with previous proposals. Fourth, we determine identification assumptions and describe a g-computation estimation method, providing example R code. Finally, we illustrate the value of the proposed approach in the self-harm example and conclude with a discussion.

2 Self-harm example

Adolescent self-harm is on the rise[24–26] and is associated with substantial disease burden[27] through immediate effects on health and mortality,[28] as well as through persisting associations with poor health and social functioning in later life, including higher rates of substance use,[29,30] depression[29] and financial hardship.[31] A question of considerable public health interest is whether policies targeting young adulthood processes may have benefit in reducing these impacts. We focus on the financial hardship outcome, and consider four young adulthood mediators: depression or anxiety, cannabis use, lack of higher education and unemployment.[31] We draw data from the Victorian Adolescent Health Cohort Study, a 10-wave longitudinal population-based cohort study of health across adolescence to the fourth decade of life in the state of Victoria, Australia (1992–2014). Data collection protocols for this study were approved by the Ethics in Human Research Committee of the Royal Children's Hospital, Melbourne. Informed parental consent was obtained before inclusion in the study. In the adult phase, all participants were informed of the study in writing and gave verbal consent before being interviewed. The Supplementary Materials provide more details on study design, with the key measures of relevance for our illustrative analysis summarised next.

The main exposure, denoted A, was adolescent self-reported self-harm across waves 3–6 (age 15–18 years), with $A = 1$ if self-harm was present at any wave during adolescence and $A = 0$ otherwise, including when all wave-specific measures were either negative or missing. The outcome (Y) was self-reported financial hardship at wave 10 (median age 35 years), with $Y = 1$ if financial hardship was present and $Y = 0$ otherwise. The mediators, measured at wave 8 (median age 24 years), were depression or anxiety (M_1), weekly or more frequent cannabis use over the past year (M_2), not having completed a university degree (M_3), and not being in paid work (M_4). We define $M_k = 1$ if the mediator was present and $M_k = 0$ if it was absent ($k = 1, \ldots, 4$).

Pre-exposure confounders $(\boldsymbol{C})$ of the exposure-mediator, mediator-outcome and exposure-outcome associations were selected on an a priori basis: participant sex, parental completion of high school (as a marker of socio-economic position), parental divorce or separation up to and including wave 6, and adolescent antecedents of the mediators where present, specifically participant completion of high school, adolescent depression or anxiety, and cannabis use (weekly or more frequent). The latter two were summarised across waves 3–6 in the same way as the exposure. Figure 1 shows the assumed causal structure for the observed data following prior evidence.[25,29–31] Although the mediators are assumed to be correlated, the causal diagram is agnostic to their causal ordering.

3 Proposed framework for tackling ill-defined interventions

We consider the general case of K mediators and, initially, the question of assessing the impact in the exposed ($A = 1$) of K hypothetical interventions, each targeting a single mediator (in the next section we consider other possibilities). Let $B_k = 1$ if the intervention targeting M_k is received and $B_k = 0$ if not ($k = 1, \ldots, K$). If these interventions were well-defined and existed, for instance, in the form of specific programs for mental health care, substance use reduction, and career development targeted at self-harmers in the example, and we had relevant data,

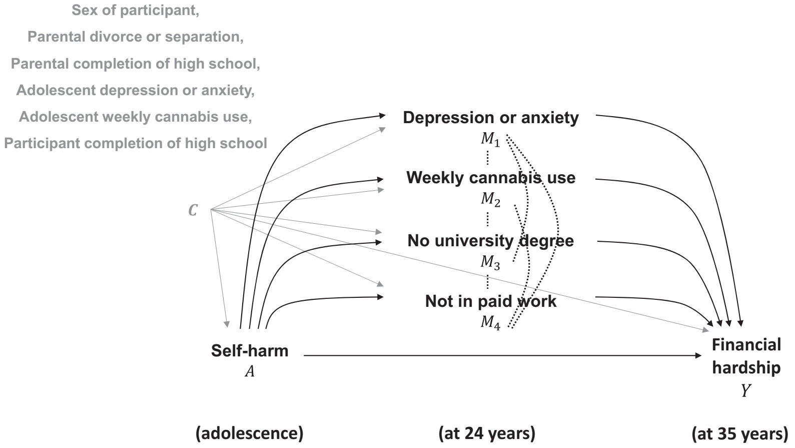


Figure 1. Directed acyclic graph portraying the assumed causal structure for the observed data, conceptualising the pathways from adolescent self-harm to financial hardship, via the four mediators of interest. Dotted undirected arrows indicate where we are agnostic about the directionality of causal influences. Pre-exposure confounders and arrows from these are shown in grey to improve clarity.

we could address the questions of interest by separately assessing the effect of each intervention in the exposed. That is, letting $Y_{B_k=b_k}$ denote the potential outcome when setting $B_k = b_k$, we would compute and compare their causal effects in the exposed, which in the difference scale and framed in terms of the reduction achieved by the intervention, are given by $E\big(Y_{B_k=0}|A=1\big) - E\big(Y_{B_k=1}|A=1\big)$, $k = 1, \ldots, K$. Here the unexposed group and mediation effects are not relevant.

However, with no well-defined interventions for which data are available, the most common approach is to simply estimate the contrasts $E\big(Y_{M_k=1}|A=1\big) - E\big(Y_{M_k=0}|A=1\big)$, $k = 1, \ldots, K$, but this raises the following issues. The potential outcomes $Y_{M_k=m_k}$ are ill-defined: for example, considering the first mediator in the example, there are many potential interventions for improving the mental health of individuals (i.e. achiving $\mathrm{M}_1 = 0$) that could lead to very different conclusions regarding causal effects.[32,33] Furthermore, any intervention is unlikely to result in complete elimination of depression and anxiety in the self-harm group, which is the scenario that $E\big(Y_{\mathrm{M}_1=0}|A=1\big)$ corresponds to, given that these conditions remain present at a certain level in the unexposed. An additional issue, also related to the fact that we are dealing with constructs rather than well-defined interventions, is that we do not know the order of the mediators, which would be needed for confounding control in the simple approach.

We propose the following principles to tackle these issues:

- Explicitly acknowledge that evidence for actual interventions in this context is not possible. Instead, one can address a more modest goal: that of informing "intervention targets", that is, the constructs that future hypothetical interventions might target, which are what is captured in available data. Although such evidence should be regarded as of lower level than causal inference about well-defined interventions, it might be the only available in the field.
- Define effects that map to a target trial assessing the impact of the distributional mediator shifts that those hypothetical interventions might achieve; these shifts can be individualised, i.e. conditional on covariates. Similar to effects studied by VanderWeele and Hernan,[33] this amounts to setting mediators to random draws from distributions specified to reflect realistic, user-specified benchmarks, to simulate the potential impacts of hypothetical interventions. The unexposed population (and thus the concept of mediation) regain relevance in specifying these "estimand assumptions". In addition to these, "identification assumptions" are required to ensure that the estimand can be estimated from available data. An expanded set of assumptions is required for causal inference with ill-defined vs. well-defined interventions, as should be expected.[10]
- In specifying relevant distributional shifts, consider the joint distribution of the mediators. This enables the mediator interrelatedness to be accounted for even without making causal ordering assumptions. The price to pay for this is a need to make unverifiable assumptions regarding the correlations amongst the mediators (at a population, distributional level) under the hypothetical interventions, as these correlations cannot be expected to remain as in the observed data, i.e. without intervention. For example, mental health in a subpopulation offered widespread provision of psychotherapy might be more or less correlated (on average) with substance use than in one offered widespread provision of antidepressants.

Figure 2 provides a conceptual overview of the proposed framework for evaluating hypothetical interventions. The approach can be seen as an intermediate step between traditional causal inference, which relies predominantly on data, and simulation-based approaches like agent-based modelling, which depend less on data and more on theory and modelling, i.e. assumptions. As Hernán has noted,[34] such approaches to causal inference are needed in disciplines that ask more complex questions, like in our example.

4 Target trial

We now describe the target trial that integrates these principles, with focus on three specific policy-relevant questions.

Question 1: If targeting only one mediator ("one-policy premise"), which of these separate interventions would provide the "biggest bang for the buck", in terms of reducing disparities between exposure groups?

This question is of relevance under resource (e.g. financial) constraints implying that the policy maker would implement only one of the K hypothetical interventions $B_1, \ldots,\ B_K$, in the exposed population.

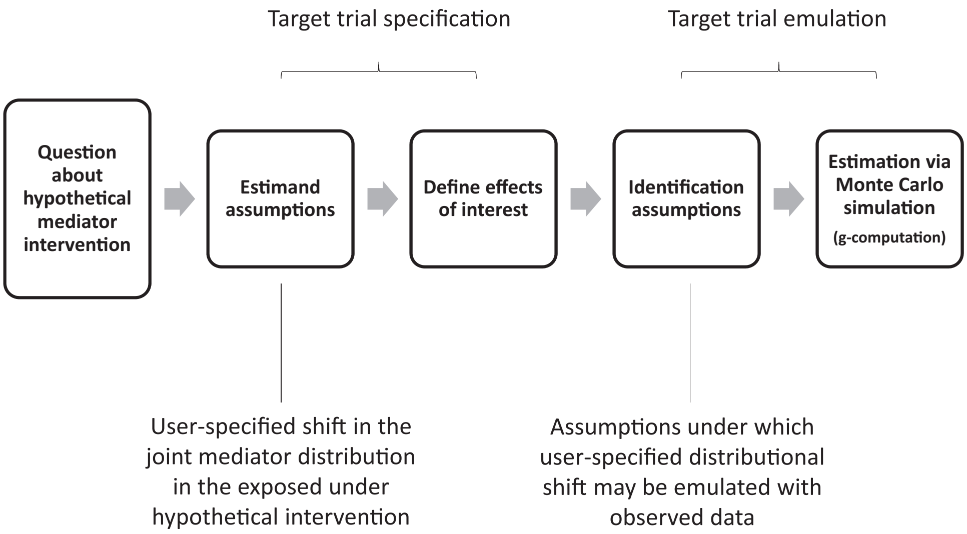

Target trial specification
Target trial emulation
Question about hypothetical mediator intervention
Estimand assumptions
Define effects of interest
Identification assumptions
Estimation via Monte Carlo simulation
(g-computation)
User-specified shift in the joint mediator distribution in the exposed under hypothetical intervention
Assumptions under which user-specified distributional shift may be emulated with observed data

Figure 2. Conceptual overview of the proposed approach for tackling the issue of ill-defined interventions via simulation of hypothetical interventions

(a) Approach under minimal estimand assumptions

We first consider the following reduced set of estimand assumptions (E1–E3), which allows for less assumption-laden and thus clearer comparisons and is likely to be widely applicable as a starting point:

E1. Intervention B_k would be applied independently of the other mediators, for $k = 1, \dots, K$;
E2. Intervention B_k would shift the distribution of mediator M_k to what it would be in the unexposed given $\boldsymbol{C}$, for $k = 1, \dots, K$. This is equivalent to setting M_k to a random draw from the distribution it would have under no exposure given $\boldsymbol{C}$; and
E3. Intervention B_k would sever the dependence on average between M_k and the other mediators, so that the joint distribution of the other mediators is held at what it would be under exposure given $\boldsymbol{C}$, for $k = 1, \dots, K$.

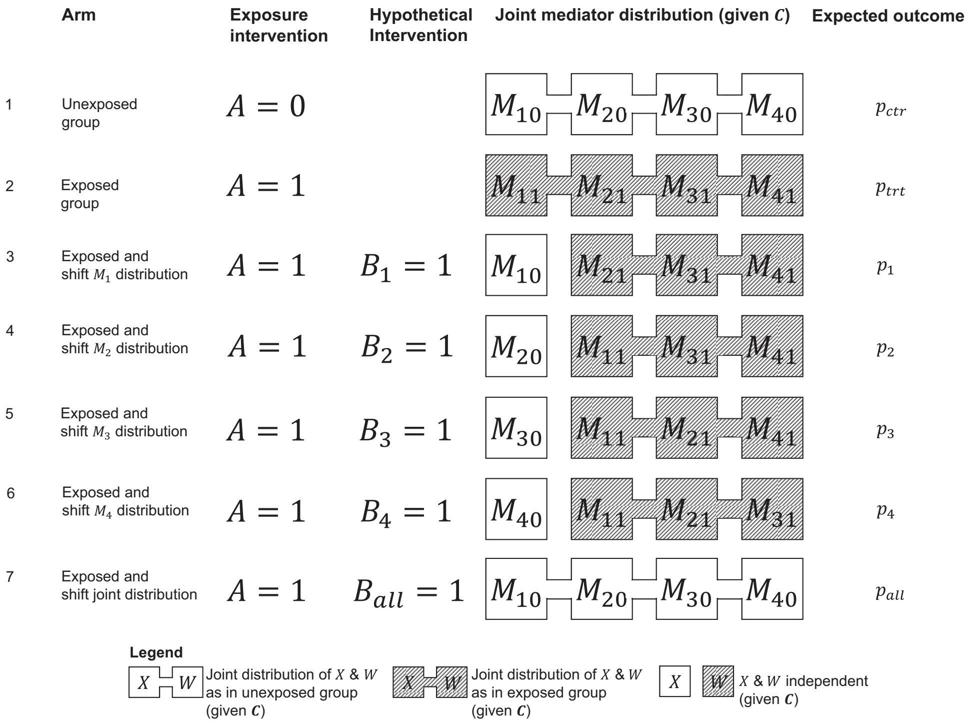
Formally, we represent E1–E3 as the assumption that the hypothetical intervention B_k would set the mediators to a random draw from the following joint distribution

$$P(M_{k0} = m_k | \boldsymbol{C}) \times P\big(\boldsymbol{M}_{(-k)1} = \boldsymbol{m}_{(-k)} | \boldsymbol{C}\big)$$

where M_{ka} denotes the status of M_k when setting A to a; $\boldsymbol{M}_{\cdot a}$ denotes the vector $(M_{1a}, \dots, M_{Ka})$; and $\boldsymbol{M}_{(-k)a}$ denotes $\boldsymbol{M}_{\cdot a}$ without the kth component.

Assumption E1 could be modified if the policy maker intended to personalise treatments conditional on other mediators. However, this would require an expanded set of estimand assumptions, e.g. delineating which mediators, etc. Assumption E2 is justified on the basis that realistically we cannot expect effects beyond bringing levels to those in the unexposed, which can be estimated from the data. Furthermore, this benchmark is meaningful in that it addresses the question of how disparities in the outcome between exposure groups reduce when disparities in the mediator are eliminated. Other benchmarks could be specified by the user if they make sense in the specific context, but again this may require additional unverifiable assumptions. Assumption E3 can be considered a worst case scenario in the sense that it precludes any effects of the hypothetical intervention flowing onto other mediators that may be causal descendants. This seems appropriate for the purpose of comparing potential intervention targets, but it can be relaxed to allow for correlations between M_k and the other mediators under the hypothetical intervention. Although this would be more realistic, it requires further unverifiable assumptions, regarding the extent of such correlations. Indeed, as in the aforementioned psychotherapy versus antidepressant example, the correlation between mediators would not be as in the observed data and rather would depend on the hypothetical intervention. Under (b) below we consider an approach relaxing E3, showing the additional estimand assumptions required for this.

A target trial for the self-harm example under assumptions E1–E3 is depicted in Figure 3. Arms 1 and 2, referred to as the unexposed and exposed groups, correspond to those in a classic two-arm parallel trial design: the intervention is only to set the exposure to $A = 0$ and $A = 1$, respectively, leading to a naturally arising joint distribution



	Arm	Exposure intervention	Hypothetical Intervention	Joint mediator distribution (given $\boldsymbol{C}$)	Expected outcome
1	Unexposed group	$A = 0$		M_{10} M_{20} M_{30} M_{40}	p_{ctr}
2	Exposed group	$A = 1$		M_{11} M_{21} M_{31} M_{41}	p_{trt}
3	Exposed and shift M_1 distribution	$A = 1$	$B_1 = 1$	M_{10} M_{21} M_{31} M_{41}	p_1
4	Exposed and shift M_2 distribution	$A = 1$	$B_2 = 1$	M_{20} M_{11} M_{31} M_{41}	p_2
5	Exposed and shift M_3 distribution	$A = 1$	$B_3 = 1$	M_{30} M_{11} M_{21} M_{41}	p_3
6	Exposed and shift M_4 distribution	$A = 1$	$B_4 = 1$	M_{40} M_{11} M_{21} M_{31}	p_4
7	Exposed and shift joint distribution	$A = 1$	$B_{all} = 1$	M_{10} M_{20} M_{30} M_{40}	p_{all}

Legend
X W Joint distribution of X & W as in unexposed group (given C)
X W Joint distribution of X & W as in exposed group (given C)
X W X & W independent (given C)

Figure 3. Graphical depiction of arms in the "target trial" designed to examine the effects of hypothetical interventions resulting in individualised shifts in the distributions of four interdependent mediators. This figure shows the arms required to evaluate effects addressing Question 1 (one-policy premise) under approach (a), and Question 2 (remaining disparities).

of the mediators in each arm. For each of arms 3–6, A is set to 1 and in addition one of the hypothetical interventions is applied, shifting the joint distribution of the mediators (given $\boldsymbol{C}$) in some way. For example, in arm 4, intervention B_2 is set to 1 so that the distribution of M_2 is shifted to be as it is in the unexposed group (following E2) given baseline characteristics but independently of the other mediators (E1), while the joint distribution of M_1, M_3 and M_4 remains as it naturally arises in the exposed group (E3). That is, in arm 4 the intervention regime is to set (A, B_2) to (1,1) and has the effect of setting A to 1, which results in M_1, M_3 and M_4 being set to a random draw from their joint distribution under exposure given $\boldsymbol{C}$; and setting M_2 to a random draw from the distribution that it would have had under no exposure given $\boldsymbol{C}$, and this independently from other mediators.

(b) Approach under causal ordering and mediator interdependence estimand assumptions

To relax E3, we need to make additional assumptions about the order of the mediators as well as the correlations between the mediators after the hypothetical intervention. For instance, we can consider alternative interventions B'_k for $k = 1, \dots, K$ under the following extended set of assumptions:

E3′. Assume that:

(i) The order of the mediators is $M_1, \dots, M_K$

(ii) Under intervention B'_k, the joint distribution of causally antecedent mediators of M_k is unaffected, remaining at what it would be under exposure.

(iii) Under intervention B'_k, the *conditional* joint distribution of the causally descendent mediators of M_k given $M_1, \dots, M_k$ and $\boldsymbol{C}$ is what it would be under exposure.

Formally, E1, E2 and E3′ can be expressed as the assumption that the hypothetical intervention B'_k would set the mediators to a random draw from the following joint distribution

$$P\left(\boldsymbol{M}_{\overline{k-1}1} = \boldsymbol{m}_{\overline{k-1}}|\boldsymbol{C}\right) \times P(M_{k0} = m_k|\boldsymbol{C}) \times P\left(\boldsymbol{M}_{\underline{k+1}1} = \boldsymbol{m}_{\underline{k+1}}|\boldsymbol{C}, \boldsymbol{M}_{\overline{k-1}1} = \boldsymbol{m}_{\overline{k-1}}, M_{k1} = m_k\right)$$

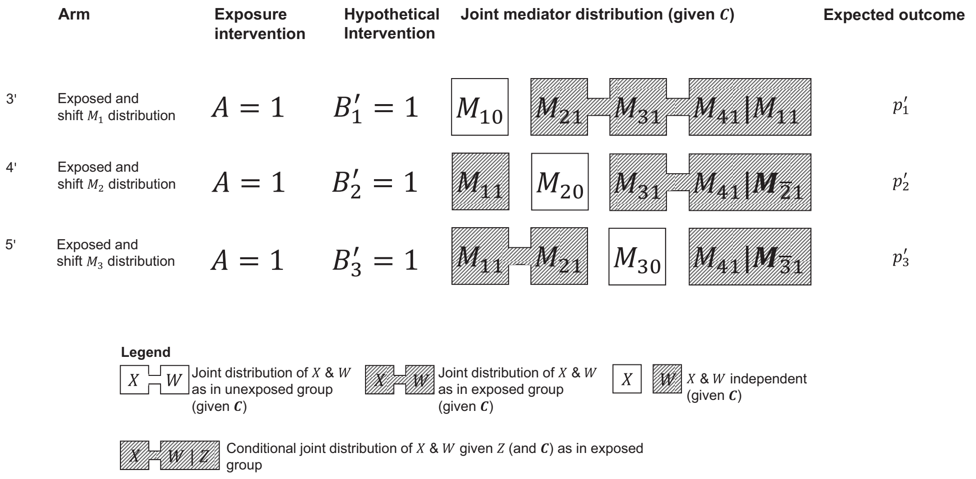

Arm
Exposure intervention
Hypothetical Intervention
Joint mediator distribution (given $\boldsymbol{C}$)
Expected outcome
3'
Exposed and shift M_1 distribution
$A = 1$
$B_1' = 1$
M_{10}
M_{21} M_{31} $M_{41}|M_{11}$
p_1'
4'
Exposed and shift M_2 distribution
$A = 1$
$B_2' = 1$
M_{11}
M_{20}
M_{31} $M_{41}|\boldsymbol{M}_{\overline{2}1}$
p_2'
5'
Exposed and shift M_3 distribution
$A = 1$
$B_3' = 1$
M_{11} M_{21}
M_{30}
$M_{41}|\boldsymbol{M}_{\overline{3}1}$
p_3'
Legend
X W Joint distribution of X & W as in unexposed group (given $\boldsymbol{C}$)
X W Joint distribution of X & W as in exposed group (given $\boldsymbol{C}$)
X W X & W independent (given $\boldsymbol{C}$)
X $W|Z$ Conditional joint distribution of X & W given Z (and $\boldsymbol{C}$) as in exposed group

Figure 4. Extension of target trial of Figure 3, including arms required to evaluate effects addressing Question 1 (one-policy premise) under approach (b).

where $\boldsymbol{M}_{\overline{k-1}a}$ denotes the vector $(M_{1a}, \ldots, M_{k-1a})$ and $\boldsymbol{M}_{\underline{k+1}a}$ denotes the vector $(M_{k+1a}, \ldots, M_{Ka})$. This assumption would be suitable in situations where we have some knowledge of ordering and can assume that the hypothetical intervention would have no impact on the interdependence between the descendent mediators, given previous ones, with these associations remaining as they would be under exposure at a population level (given confounders). Figure 4 depicts the additional target trial arms that could be added to examine these hypothetical interventions, under alternative assumptions, in the self-harm example. Only three arms are added as the assumptions under E3 relating to B_4 are equivalent to those under E3$'$ relating to B_4'. That is, $B_4 = B_4'$.

Question 2: What would be the remaining disparities between exposure groups if it were possible to jointly target all the mediators?

We can address this question by considering a hypothetical intervention B_{all}, targeting all the mediators. We make the following estimand assumption:

E4. The hypothetical intervention B_{all} shifts the joint distribution of the mediators to be as in the unexposed given $\boldsymbol{C}$.

Formally, assumption E4 states that the hypothetical intervention B_{all} sets the mediators to a random draw from the joint distribution $P(\boldsymbol{M}_{\cdot 0} = \boldsymbol{m}|\boldsymbol{C})$.

Arm 7 in Figure 3 shows what this translates to in the target trial: in this arm, A is set to 1 and B_{all} to 1, which shifts the joint distribution of the mediators to what it is in the unexposed group, including mediator interdependencies. This thus represents an intervention where the exposed group is set to be exactly like the unexposed in terms of the joint mediator distribution. Large remaining disparities after this intervention would suggest that there is a need to investigate additional intermediate processes. A different benchmark could be used for E4, but this would require additional assumptions if there are no data from which to estimate it.

Question 3: What would be the benefit of sequential policies, applying the separate mediator interventions under Question 1 approach (a) sequentially?

Let $B_{\{k\}}$ denote an intervention applying all interventions in the sequence $B_1, \ldots, B_K$ up to B_k $(k = 1, \ldots, K)$, so that setting $B_{\{k\}}$ to 1 means that each of $B_1, \ldots, B_k$ is set to 1. We consider assumptions E1–E3 applied to $B_{\{k\}}$, but their interpretation is extended to mean that, however this is done (e.g. simultaneously), it shifts the distribution of each mediator $1, \ldots, k$ to what it would be in the unexposed given $\boldsymbol{C}$, independently of other mediators and severing the dependence on average from the subsequent ones in the sequence.

Formally, E1–E3 in this setting can be expressed as the assumption that $B_{\{k\}}$ is a hypothetical intervention that sets the mediators to a random draw from the joint distribution

$$P(M_{10} = m_1|\boldsymbol{C}) \times \cdots \times P(M_{k0} = m_k|\boldsymbol{C}) \times P\big(\boldsymbol{M}_{\underline{k+1}1} = \boldsymbol{m}_{\underline{k+1}}|\boldsymbol{C}\big)$$

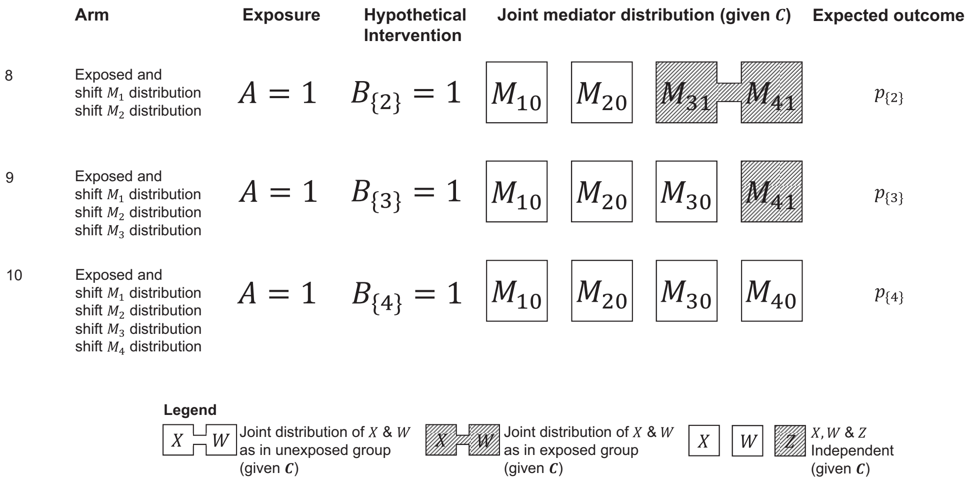

Arm
Exposure
Hypothetical Intervention
Joint mediator distribution (given C)
Expected outcome
8
Exposed and shift M_1 distribution shift M_2 distribution
$A = 1$
$B_{\{2\}} = 1$
M_{10}
M_{20}
M_{31}
M_{41}
$p_{\{2\}}$
9
Exposed and shift M_1 distribution shift M_2 distribution shift M_3 distribution
$A = 1$
$B_{\{3\}} = 1$
M_{10}
M_{20}
M_{30}
M_{41}
$p_{\{3\}}$
10
Exposed and shift M_1 distribution shift M_2 distribution shift M_3 distribution shift M_4 distribution
$A = 1$
$B_{\{4\}} = 1$
M_{10}
M_{20}
M_{30}
M_{40}
$p_{\{4\}}$
Legend
X W Joint distribution of X & W as in unexposed group (given C)
X W Joint distribution of X & W as in exposed group (given C)
X W Z X, W & Z Independent (given C)

Figure 5. Extension of target trial of Figures 3 and 4, including arms required to evaluate effects addressing Question 3 (sequential policies).

with the last factor omitted for $k = K$.

To evaluate the impact of the sequential interventions, we can add more arms to the trial, as depicted in Figure 5 for the case of four mediators. Only three arms are added as $B_{\{1\}} = B_1$. For each of arms 8–10, A is set to 1 and $B_{\{k\}}$ is set to 1, $k = 2{,}3, 4$. The order of the sequence, here assumed to be $B_1, \ldots, B_K$ (E5), should be determined by the research question: which order is of interest from a policy perspective? If a different order were of interest, then the new trial arms, and resulting effects (next section) would be different.

The target trial in Figures 3 to 5 extends in the natural way to the case of K mediators.

5 Mediation effect definitions

We define interventional effects addressing each question by contrasting the outcome expectation between relevant trial arms. Following the notation in the last column of Figures 3 to 5, but considering the general case of K mediators, let p_{ctr} and p_{trt} denote the outcome expectation in the unexposed ("control") and exposed ("treated") groups, respectively; p_k, for $k = 1, \ldots, K$, denote the outcome expectation in the arm where the distribution of M_k is shifted under B_k; p'_k, for $k = 1, \ldots, K$, denote the outcome expectation in the arm where the distribution of M_k is shifted under B'_k (noting $p_K = p'_K$); and p_{all} the outcome expectation in the arm shifting the joint distribution of all mediators. Further, let $p_{\{0\}} = p_{trt}$ and $p_{\{1\}} = p_1$, and let $p_{\{k\}}$ for $k > 1$ denote the outcome expectation in the arm in which the interventions B_1 to B_k have been applied sequentially (Figure 5).

The total causal effect (TCE) in the difference scale is given by: TCE $= p_{trt} - p_{ctr}$.

5.1 Effects for Question 1: one-policy premise

(a) Effects under minimal estimand assumptions

We define a type of interventional indirect effect via the kth mediator, IIE$_k$ ($k = 1, \ldots, K$), as the contrast between the outcome expectation in the exposed group and the arm in which the M_k distribution is shifted by B_k

$$\mathrm{IIE}_k = p_{trt} - p_k$$

This quantifies the impact of an intervention targeting M_k, while the joint distribution of the other mediators remains as it would be under exposure. In the example, for M_2 (weekly cannabis use), the corresponding effect IIE$_2$ is the reduction in risk of financial hardship in self-harmers that would be achieved by reducing their rates of weekly cannabis use to those in the non-self-harmers, while the joint distribution of all other mediators remains unaffected (given covariates).

In a previous section, we mentioned that in the context of well-defined interventions, i.e. if we had data on an intervention B_k for a given k, then the effect in the exposed group, defined as $E\big(Y_{B_k=0}|A=1\big) - E\big(Y_{B_k=1}|A=1\big)$, would be of interest. Within the identifiability proofs in the Supplementary Materials we show that under certain assumptions, $\mathrm{IIE}_k = E_{\boldsymbol{C}}\left[E\left(Y_{B_k=0}|A=1,\boldsymbol{C}\right) - E\left(Y_{B_k=1}|A=1,\boldsymbol{C}\right)\right]$, that is, the proposed effect is a whole-population standardised version of that effect. Specifically, it is the average of the exposed-group-specific effects within confounder strata, where the average is taken with respect to the confounder distribution in the whole population (exposed and unexposed).

(b) Effects under causal ordering and mediator interdependence estimand assumptions

Similarly, we define an alternative type of interventional indirect effect via the kth mediator, $\mathrm{IIE}'_k (k = 1, \ldots, K)$, as the contrast between the outcome expectation in the exposed group and the arm in which the M_k distribution is shifted by B'_k

$$\mathrm{IIE}'_k = p_{trt} - p'_k$$

This quantifies the impact of an intervention targeting M_k, while the joint distribution of the antecedent mediators remains as it would be under exposure and the conditional joint distribution of the causally descendent mediators given all past mediators remains as under exposure. In the example, assuming the causal ordering M_1, M_2, M_3, M_4, for M_2 (weekly cannabis use), the corresponding effect IIE'_2 is the reduction in risk of financial hardship in self-harmers that would be achieved by reducing their rates of weekly cannabis use to those in the non-self-harmers, and allowing this shift to flow on to causally descedent mediators through the interdependence between M_2 and (M_3, M_4), the strength of which is assumed to be as it would have been under exposure (given covariates). As for effects under approach (a), it can be proved that IIE'_k is a whole-population standardised version of the exposed-group-specific effect of B'_k.

Effects under both (a) and (b) differ from those proposed by Vansteelandt and Daniel,[19] which implicitly emulate other distributional shifts (see section "Summary and comparison with previous effects" below).[22]

5.2 Effects for Question 2: Remaining disparities

We consider the following interventional direct effect not via any mediator (IDE)

$$\mathrm{IDE} = p_{all} - p_{ctr}$$

The IDE quantifies disparities between exposed and unexposed that would remain even if it were possible to intervene simultaneously on all the mediators to shift their joint distribution (mean levels and interdependence) to be as in the unexposed group (given covariates). While the IDE answers the question regarding *remaining disparities*, it might also be interesting to consider the effect of the joint intervention, defined in terms of the reduction achieved as follows

$$\mathrm{IIE}_{all} = p_{\mathrm{trt}} - p_{all}$$

5.3 Effects for Question 3: Sequential policies

We define the interventional indirect effect of the kth intervention in the sequence, $\mathrm{IIE}_{\{k\}}$ $(k = 1, \ldots, K)$, as

$$\mathrm{IIE}_{\{k\}} = p_{\{k-1\}} - p_{\{k\}}$$

The sum of these effects provides an interventional indirect effect quantifying the overall impact of the sequential intervention ($\mathrm{IIE}_{\{seq\}}$) and is equal to

$$\mathrm{IIE}_{\{seq\}} = p_{trt} - p_{\{K\}}$$

5.4 Decompositions of the TCE and other interesting effects

There are many possible decompositions of the TCE but it is important to focus on component effects that address relevant questions. For example, the TCE may be decomposed as: $\text{TCE} = \text{IDE} + \text{IIE}_1 + \cdots + \text{IIE}_\text{K} + \text{IIE}_{int}$, where the last term is a type of interventional indirect effect via the mediators' interdependence, contrasting the benefit of the aforementioned joint intervention with the sum of the benefits of individual interventions: $\text{IIE}_{int} = \text{IIE}_{all} - (\text{IIE}_1 + \cdots + \text{IIE}_\text{K})$. This effect does not have a policy-relevant interpretation so it is not of much interest. Similarly, $\text{TCE} = \text{IDE} + \text{IIE}'_1 + \cdots + \text{IIE}'_K + \text{IIE}'_{int}$ where $\text{IIE}'_{int} = \text{IIE}_{all} - \left(\text{IIE}'_1 + \cdots + \text{IIE}'_K\right)$, is a type of interventional effect with a difficult and not very useful interpretation.

The decomposition that focusses on sequential policies is: $\text{TCE} = \text{IDE} + \text{IIE}_{\{seq\}} + \text{IIE}_{\{int\}}$. Here $\text{IIE}_{\{int\}}$ contrasts the benefit of the joint intervention B_{all} with the benefit of sequentially applying $B_1, \ldots, B_K$: $\text{IIE}_{\{int\}} = (p_{trt} - p_{all}) - \text{IIE}_{\{seq\}} = p_{\{K\}} - p_{all}$. The expression after the second equality shows that this effect captures what one would intuitively conceive as the effect via the mediators' interdependence: by contrasting the expected outcome under a shift in the joint mediator distribution with that when a sequence of independent shifts is made across the mediators, this effect quantifies the effect via mediator correlations under no exposure as they are in the data.

Other contrasts that could be of interest are $\text{IDE}_k = p_k - p_{ctr} = \text{TCE} - \text{IIE}_k$ and $\text{IDE}'_k = p'_k - p_{ctr} = \text{TCE} - \text{IIE}'_k$, for $k = 1, \ldots, K$, with IDE_k and IDE'_k quantifying the disparities remaining after intervening on M_k alone via B_k or B'_k, respectively. Each effect can be expressed as a proportion of the TCE to gauge relative size.

5.5 Summary and comparison with previous effects

Table 1 summarises the proposed effects in terms of the assumed mediator distribution shifts under hypothetical interventions and the contrasting (pre-intervention) state. Next to each intervention effect, defined as the contrast between pre- and post-intervention states, we show in brackets the estimand expressing the remaining between-exposure-group differences after the intervention, i.e. the difference remaining between unexposed and exposed after the intervention. Whether it is the intervention effect or the remaining difference that is of most interest depends on the question, e.g. Question 1 is focused on intervention effects while Question 2 is framed around remaining differences. The table also shows the estimand assumptions underlying other effects that have been proposed in the literature for the setting with multiple mediators and a time-fixed–exposure,[19,23] viewing them through the lens of our proposed framework for evaluating the effects of hypothetical interventions on mediator distributions (Figure 2).

It is seen that previous effects are different from the proposed effects and their interpretability in answering policy-relevant questions about hypothetical interventions requires consideration. For example, the mediator-specific effects of Vansteelandt and Daniel,[19] denoted in Table 1 by VD-IIE_k, emulate an intervention that shifts the kth mediator to levels in the unexposed independently of previous mediators, with the joint distribution of the subsequent mediators (as they have been numbered, since these authors do not assume a causal order) assumed to reduce to levels in the unexposed independently of the kth mediator. This could be of interest if we assume that the numbering reflects a causal order and that the hypothetical intervention impacts subsequent mediators very strongly. The pre-intervention mediator distribution could be difficult to interpret as it does not correspond to that naturally arising in the exposed, which would be the natural benchmark for policy-makers, but one where the joint distribution of the subsequent mediators is also at the unexposed levels. The pre- and post-intervention mediator distributions for the effects of Lin and VanderWeele[23] prove very difficult to interpret through this hypothetical intervention lens.

Of note, other previously proposed estimands that can be considered to fall under the "interventional effects" umbrella either focus on the setting where only a single mediator is of substantive interest[18,20,21,35–37] or when the exposure is time-varying[38,39] so are not directly comparable with our proposal. Other related effects are those that correspond to shifting confounder[40–42] or exposure[43,44] distributions (see section 8).

6 Identification and estimation

To identify and emulate these effects, it suffices to consider the identifiability and estimation of the outcome expectation in a given target trial arm subject to a mediator distribution shift (arms 3–10 and 3′–5′). Let B indicate receipt of the corresponding hypothetical intervention (e.g. B stands for B_1 in arm 3, B'_1 in arm 3′, B_{all} in arm 7 and $B_{\{2\}}$ in arm 8). Further, for $a, b = 0,1$ and $k = 1, \ldots, K$, let Y_{ab} denote the outcome when A is set to a and B to b. Recall that M_{ka} denotes the status of M_k when setting A to a; $\boldsymbol{M}_{\cdot a}$ denotes the vector $(M_{1a}, \ldots, M_{Ka})$; and

Table 1. Comparison of proposed and previously published interventional effects for multiple mediators with a point exposure, in terms of their interpretation as effects of hypothetical interventions shifting mediator distributions.

Effects of hypothetical interventions	Mediator distribution under intervention (post-intervention)	Mediator distribution for contrast (pre-intervention)
Effects defined in this paper		
IIE_k (remaining difference[a]: IDE_k)	$P(M_{k0} = m_k \vert \boldsymbol{C}) \times P\big(\boldsymbol{M}_{(-k)1} = \boldsymbol{m}_{(-k)} \vert \boldsymbol{C}\big)$	$P(\boldsymbol{M}_{\cdot 1} = \boldsymbol{m} \vert \boldsymbol{C})$
IIE'_k (remaining difference: IDE'_k)	$P\big(\boldsymbol{M}_{\overline{k-1}1} = \boldsymbol{m}_{\overline{k-1}} \vert \boldsymbol{C}\big) \times P(M_{k0} = m_k \vert \boldsymbol{C})$ $\times P\big(\boldsymbol{M}_{\underline{k+1}1} = \underline{m_{k+1}} \vert \boldsymbol{C}, \boldsymbol{M}_{\overline{k-1}1} = \boldsymbol{m}_{\overline{k-1}}, M_{k1} = m_k\big)$	$P(\boldsymbol{M}_{\cdot 1} = \boldsymbol{m} \vert \boldsymbol{C})$
$\text{IIE}_{\{k\}}$	$P(M_{10} = m_1 \vert \boldsymbol{C}) \times \cdots \times P(M_{k0} = m_k \vert \boldsymbol{C}) \times P\big(\boldsymbol{M}_{\underline{k+1}1} = \boldsymbol{m}_{\underline{k+1}} \vert \boldsymbol{C}\big)$	$P(\boldsymbol{M}_{\cdot 1} = \boldsymbol{m} \vert \boldsymbol{C})$
$\text{IIE}_{\{seq\}}$ (equal to sum of the $\text{IIE}_{\{k\}}$ effects)	$P(M_{10} = m_1 \vert \boldsymbol{C}) \times \cdots \times P(M_{K0} = m_k \vert \boldsymbol{C})$	$P(\boldsymbol{M}_{\cdot 1} = \boldsymbol{m} \vert \boldsymbol{C})$
IIE_{all} (remaining difference: IDE [b])	$P(\boldsymbol{M}_{\cdot 0} = \boldsymbol{m} \vert \boldsymbol{C})$	$P(\boldsymbol{M}_{\cdot 1} = \boldsymbol{m} \vert \boldsymbol{C})$
IIE_{int}, IIE'_{int} and $\text{IIE}_{\{int\}}$	Contrast IIE_{all} with sums of the IIE_k IIE'_k and $\text{IIE}_{\{k\}}$ effects, respectively	
Effects of Vansteelandt & Daniel[19]		
VD-IIE_k	$P\big(\boldsymbol{M}_{\overline{k-1}1} = \boldsymbol{m}_{\overline{k-1}} \vert \boldsymbol{C}\big) \times P(M_{k0} = m_k \vert \boldsymbol{C}) \times P\big(\boldsymbol{M}_{\underline{k+1}0} = \boldsymbol{m}_{\underline{k+1}} \vert \boldsymbol{C}\big)$	$P\big(\boldsymbol{M}_{\overline{k-1}1} = \boldsymbol{m}_{\overline{k-1}} \vert \boldsymbol{C}\big) \times P(M_{k1} = m_k \vert \boldsymbol{C}) \times P\big(\boldsymbol{M}_{\underline{k+1}0} = \boldsymbol{m}_{\underline{k+1}} \vert \boldsymbol{C}\big)$
VD-IIE_{int}	Contrast IIE_{all} with sum of the $\text{VD}_{\text{IIE}}k$ effects	
Effects of Lin & VanderWeele[23] [c]		
$\text{LV-IIE}_{A \to M_1 \to Y}$	$P(M_{10} = m_1 \vert \boldsymbol{C}) \times \Big[\sum_{m_1'} P\big(M_{20m'_1} = m_2 \vert \boldsymbol{C}\big) \times P\big(M_{10} = m_1' \vert \boldsymbol{C}\big)\Big]$	$P(M_{11} = m_1 \vert \boldsymbol{C}) \times \Big[\sum_{m'_1} P\big(M_{20m'_1} = m_2 \vert \boldsymbol{C}\big) \times P\big(M_{10} = m'_1 \vert \boldsymbol{C}\big)\Big]$
$\text{LV-IIE}_{A \to M_2 \to Y}$	$P(M_{11} = m_1 \vert \boldsymbol{C}) \times \Big[\sum_{m'_1} P\big(M_{20m'_1} = m_2 \vert \boldsymbol{C}\big) \times P\big(M_{10} = m'_1 \vert \boldsymbol{C}\big)\Big]$	$P(M_{11} = m_1 \vert \boldsymbol{C}) \times \Big[\sum_{m'_1} P\big(M_{21m'_1} = m_2 \vert \boldsymbol{C}\big) \times P\big(M_{10} = m'_1 \vert \boldsymbol{C}\big)\Big]$
$\text{LV-IIE}_{A \to M_1 \to M_2 \to Y}$	$P(M_{11} = m_1 \vert \boldsymbol{C}) \times \Big[\sum_{m'_1} P\big(M_{21m'_1} = m_2 \vert \boldsymbol{C}\big) \times P\big(M_{10} = m'_1 \vert \boldsymbol{C}\big)\Big]$	$P(M_{11} = m_1 \vert \boldsymbol{C}) \times \Big[\sum_{m'_1} P\big(M_{21m'_1} = m_2 \vert \boldsymbol{C}\big) \times P\big(M_{11} = m'_1 \vert \boldsymbol{C}\big)\Big]$

[a] In addition to the effect of the intervention, defined as a contrast between pre- and post-intervention states shown in this table, we also defined effects in terms of the difference that remains post-intervention between the exposed and unexposed

[b] IDE is equal to an effect also defined by Vansteelandt and Daniel,[19] though they do not explicitly define IIE_{all}.

[c] Lin and VanderWeele[23] define effects for the setting $\boldsymbol{K} = 2$ and do not actually condition on $\boldsymbol{C}$ in their definitions but we include that conditioning here for comparability. For their estimands we need to extend the notation, so that M_{2am_1} is the value of mediator M_2 when setting $A = a$ and $M_1 = m_1$

$\boldsymbol{M}_{(-k)a}$ denotes $\boldsymbol{M}_{\cdot a}$ without the kth component; $\boldsymbol{M}_{\overline{k}a}$ denotes the vector $(M_{1a}, \dots, M_{ka})$; and $\boldsymbol{M}_{\underline{k}}a$ denotes the vector $(M_{ka}, \dots, M_{Ka})$. The observed counterparts are denoted by removing the subscript a, e.g. $\boldsymbol{M}_{\cdot}$ for the observed joint distribution of the mediators.

In addition to standard positivity assumptions,[45] we make the following identification assumptions:

A1. There is no causal effect of B on the outcome other than through mediator distributional shifts, that is, other than through setting the mediators to a random draw from the specified distribution;
A2. The following conditional independence assumptions hold

$$(i)\ \ Y_{ab} \perp (A, B) | \boldsymbol{C}$$

$$(ii)\ \ (M_{1a}, \dots, M_{Ka}) \perp A | \boldsymbol{C}$$

A3. $Y_{ab} = Y$ when $A = a$ and $B = b$; $M_{ka} = M_k$ when $A = a$ for $k = 1, \dots, K$

A1–A3 are similar to those considered by VanderWeele and Hernán.[33] With the intervention B being hypothetical, it is not possible to assess whether these assumptions are plausible, except for assumptions not pertaining to B, which are similar to assumptions in Vansteelandt and Daniel.[19] Further, A3 relies partly on the possibility of identifying the exposure A with a well-defined intervention. This can be assessed but, with the main goal being to evaluate mediator interventions, it can be argued that application of the proposed method remains meaningful even with no well-defined exposure intervention, as others have proposed in related settings.[35,42,46]

Under A1–A3, the outcome expectation in the given arm can be emulated using observational data. Complete identification formulae and proofs are given in the Supplementary Materials. For illustration, consider the arm where intervention B_k is applied to shift mediator k under the one-policy premise. From A2(i) and A3, it follows that the outcome expectation p_k can be expressed as: $p_k = E(Y_{11}) = E_{\boldsymbol{C}}[E(Y|A = 1, B_k = 1, \boldsymbol{C})]$. By A1, setting $B_k = 1$ is equivalent to setting the mediators to a random draw from the joint distribution $P(M_{k0} = m_k|\boldsymbol{C}) \times P\big(\boldsymbol{M}_{(-k)1} = \boldsymbol{m}_{(-k)}|\boldsymbol{C}\big)$, which from A2(ii) and A3 is equal to $P(M_k = m_k|A = 0, \boldsymbol{C}) \times$ $P\big(\boldsymbol{M}_{(-k)} = \boldsymbol{m}_{(-k)}|A = 1, \boldsymbol{C}\big)$. This leads to the following identification formula

$$p_k = E_{\boldsymbol{C}}\left[\sum_{\boldsymbol{m}=(m_1,\dots,m_K)} E(Y|A = 1, \boldsymbol{M} = \boldsymbol{m},\ \boldsymbol{C})\ \times P(M_k = m_k|A = 0, \boldsymbol{C}) \times P\big(\boldsymbol{M}_{(-k)} = \boldsymbol{m}_{(-k)}|A = 1, \boldsymbol{C}\big)\right]$$

Estimation can be performed using the Monte Carlo simulation-based g-computation approach described by Vansteelandt and Daniel[19] (see Supplementary Materials). To reduce the risk of misspecification bias, it is recommended to use rich parametric models, including various interaction terms and higher-order terms (for continuous variables).[46] Example code in R[47] for implementing the method, including a function and a worked example on simulated data, can be accessed at the first author's GitHub repository (https://github.com/moreno-betancur/medRCT).

7 Results for self-harm example

Table 2 shows descriptive statistics based on the 1786 participants (out of 1943 in the cohort study) with the adolescent self-harm exposure available. As all other analysis variables had missing data, subsequent analyses were based on multiple imputation using 40 imputations (details in Supplementary Materials). Table 3 shows preliminary estimates of unadjusted and regression-adjusted exposure-outcome, exposure-mediator and mediator-outcome associations, which were obtained using main-effects multivariable logistic regression models. These provide an idea of the strength of some of the hypothesised pathways in Figure 1.

We estimated the proposed effects using the g-computation method with multivariable logistic regressions including all two-way interactions (see Supplementary Materials); see Table 4 for results. Adolescent self-harmers had an increased risk of financial hardship in adulthood compared to non-self-harmers in our study: TCE $=7.2\%$ (95% CI: -1.7 to 16.1%). Under the one-policy premise and minimal estimand assumptions, we estimated that the highest impact would be achieved by an intervention that would improve the rates of university completion in adolescent self-harmers (IIE$_3$=0.9%; -1.3 to 3.2%). This corresponds to a 13% reduction in the between-group difference, with the remaining difference being IDE$_3$=TCE $-$ IIE$_3$=6.3%. Other intervention targets have lower impact. Under causal ordering and mediator interdependence assumptions, results were very

Table 2. Descriptive statistics by exposure group in the self-harm example.

	Adolescent self-harm[b]		
	No	Yes	Missing %[c]
Number[a]	1638	148	
Pre-exposure confounders			
Sex of participant: Female (%)	846 (51.6)	95 (64.2)	0.0
Parental divorce or separation (%)	339 (20.7)	45 (30.4)	0.0
Neither parent completed secondary school (%)	515 (32.7)	46 (33.3)	4.1
Adolescent depression or anxiety (%)	495 (30.2)	111 (75.0)	0.0
Adolescent weekly cannabis use (%)	155 (9.5)	41 (27.9)	0.5
Participant did not complete secondary school (%)	232 (14.8)	32 (23.2)	4.6
Mediators (at age 24 years)			
Depression or anxiety (%)	263 (20.0)	32 (26.0)	19.6
Weekly cannabis use (%)	143 (10.9)	25 (20.3)	19.7
No university degree (%)	805 (61.3)	96 (78.0)	19.5
Not in paid work (%)	140 (10.6)	22 (17.9)	19.5
Outcome (at age 35 years)			
Financial hardship	258 (21.9)	41 (38.3)	28.0
Any analysis variable missing (%)	546 (33.3)	47 (31.8)	0

[a] The total number of participants in each exposure group.
[b] Descriptive statistics for each characteristic are based on the records with available data for that variable in the given exposure group.
[c] Proportion of missing data across both exposure groups for that variable.

Table 3. Associations amongst exposure, outcome and mediators estimated using multivariable logistic regression models and multiple imputation (40 imputations).

Associations	Crude OR	95% CI	Adjusted OR[a]	95% CI
Exposure (adolescence) – Outcome (35 years)				
Self-harm – Financial hardship	2.20	(1.49; 3.25)	1.56	(1.01; 2.42)
Exposure (adolescence) – Mediators (24 years)				
Self-harm – Depression or anxiety	1.46	(0.96; 2.22)	0.93	(0.59; 1.45)
Self-harm – Weekly cannabis use	2.06	(1.31; 3.23)	1.29	(0.76; 2.19)
Self-harm – No university degree	2.07	(1.34; 3.20)	1.56	(0.95; 2.53)
Self-harm – Not in paid work	1.89	(1.16; 3.08)	1.42	(0.84; 2.40)
Mediators (24 years) – Outcome (35 years)				
Depression or anxiety – Financial hardship	1.64	(1.17; 2.30)	1.37	(0.96; 1.95)
Weekly cannabis use – Financial hardship	1.47	(1.00; 2.16)	1.34	(0.87; 2.08)
No university degree – Financial hardship	2.97	(2.16; 4.08)	2.53	(1.78; 3.59)
Not in paid work – Financial hardship	2.23	(1.53; 3.26)	1.77	(1.18; 2.64)

OR: Odds ratio; CI: Confidence Interval.
[a] Adjusted for pre-exposure confounders and, for outcome-mediator associations, the exposure.

similar, with a slightly higher reduction of the difference (14%) for interventions shifting university completion. A hypothetical intervention shifting the joint distribution of the mediators in the self-harm group to be as under no self-harm, given covariates, would lead to a 23% reduction ($\text{IIE}_{all} = 1.6\%$; -1.6% to 4.9%), therefore still leaving 77% of the difference between the two groups remaining: IDE $= 5.6\%$ (-3.1% to 14.3%).

The overall sequential policy could, in principle, achieve a reduction of 27% of the total effect ($\text{IIE}_{\{\text{seq}\}} = 1.9\%$; -1.4% to 5.2%). This is decomposed into the effects of applying each policy on top of the previous ones in the sequence. Each of the effects from M_2 onwards is of slightly lower magnitude than under the one-policy premise. The effect via the interdependence $\text{IIE}_{\{\text{int}\}}$ is negative, indicating that the sequential intervention would achieve a larger reduction in risk than the joint intervention. This is explained by the severing of the dependence amongst the mediators under assumption E3, which, as mentioned, is a pragmatic assumption to avoid making further unverifiable assumptions. The direction of this effect indicates that we might estimate a smaller effect for the

Table 4. Estimates of proposed interventional mediation effects to address each policy-relevant question, obtained using the Monte Carlo simulation-based g-computation approach (200 replications), along with the bootstrap (1000 runs) and multiple imputation (40 imputations).

Effect		Estimate	95% CI	Proportion of TCE(%)
TCE		0.072	(−0.016; 0.161)	100
IDE	(remaining after joint intervention, cf IIE$_{\text{all}}$ below)	0.056	(−0.031; 0.142)	77
Question 1: Effects under one-policy premise				
(a) Under minimal assumptions				
IIE$_1$	(depression or anxiety)	0.002	(−0.016; 0.019)	2
IIE$_2$	(weekly cannabis use)	0.005	(−0.011; 0.020)	6
IIE$_3$	(no university degree)	0.009	(−0.013; 0.032)	13
IIE$_4$	(not in paid work)	0.006	(−0.010; 0.023)	9
IIE$_{int}$	(mediators' interdependence)	−0.006	(−0.021; 0.009)	−8
(b) Under causal ordering and interdependence assumptions				
IIE$'_1$	(depression or anxiety)	−0.002	(−0.016; 0.013)	−2
IIE$'_2$	(weekly cannabis use)	0.005	(−0.009; 0.020)	7
IIE$'_3$	(no university degree)	0.010	(−0.011; 0.031)	14
IIE$'_4$	(not in paid work)	0.006	(−0.010; 0.023)	9
IIE$'_{int}$	(mediators' interdependence)	−0.003	(−0.013; 0.007)	−5
Question 2: Effect under joint mediator intervention				
IIE$_{\text{all}}$	(joint intervention on all)	0.016	(−0.016; 0.049)	23
Question 3: Effects under sequential policies				
IIE$_{\{seq\}}$	(full sequence)	0.019	(−0.013; 0.052)	27
IIE$_{\{1\}}$	(depression or anxiety)	0.002	(−0.016; 0.019)	2
IIE$_{\{2\}}$	(weekly cannabis use)	0.004	(−0.010; 0.018)	5
IIE$_{\{3\}}$	(no university degree)	0.009	(−0.012; 0.029)	12
IIE$_{\{4\}}$	(not in paid work)	0.005	(−0.010; 0.020)	7
IIE$_{\{int\}}$	(mediators' interdependence)	−0.003	(−0.008; 0.002)	−4

TCE: Total Causal Effect; IDE: Interventional Direct Effect; IIE: Interventional Indirect Effect; CI: Confidence Interval.

sequential intervention under the additional assumption that mediator correlations after the interventions are similar to what they are under no exposure (given confounders), as estimated from the observed data.

8 Discussion

While avoiding previous "axiomatic" definitions of mediation, this paper proposed a novel framework that uses interventional mediation effects for tackling the issue of ill-defined interventions that abounds in various areas of epidemiology.[10,15–17] Building on previous work,[22] novel interventional effects are defined that explicitly emulate target trials of hypothetical interventions that result in individualised (covariate-specific) mediator distributional shifts. Simulating the effects of hypothetical interventions in this way addresses the realistic if relatively modest goal of informing intervention targets and requires an expanded set of assumptions both to define the estimand and to identify it with observational data. This is commensurate with the lower-level evidence and increased subtlety in interpretation that is to be expected with ill-defined interventions, towards the left-hand end of the Galea-Hernán causal spectrum, for which one must simulate "in silico hypothetical experiments".[10,34] Although uncertainty of estimation precludes any strong conclusions being drawn, the self-harm example illustrated the value of our proposal for addressing policy-relevant questions.

We retained mediation terminology ("direct", "indirect", etc.) for the proposed effects, consistent with the view that there is no clear definition of these notions beyond these and so-called "separable" effects (see below). Although we suggest that it is more realistic to focus on the benchmark of our proposed direct effects, which is the distribution in the unexposed given covariates, it is straightforward to apply the same methodology to evaluate hypothetical interventions that set the mediators to another user-specified distribution, even a degenerate (constant-valued) distribution; for example, one could even assess the extreme case where mediators are eliminated. In this sense, interventional direct effects generalise "controlled direct effects", which can be seen as setting the mediator to a draw from a degenerate distribution. Others have also considered more realistic benchmarks in the definition of direct effects.[37,48] More broadly, although the estimand assumptions outlined here are likely to be of relevance in a range of settings, alternative assumptions might well be warranted in other contexts. In particular, further work could consider estimand assumptions that individualise mediator shifts by conditioning on a set of baseline covariates that may overlap with but is not necessarily equal to the minimal confounding adjustment set $\boldsymbol{C}$. A further refinement would be to specify mediator shifts that completely differ across specific subgroups of the population or that depend on causally antecedent mediators if the order is known, to emulate further individualised hypothetical interventions.

The identification assumptions that concern hypothetical interventions are not assessable without considering a concrete intervention. As has been noted,[10,32,33] confounder selection is complex in this context: considering common causes of the intervention and its target is difficult with no concrete intervention in mind. Nonetheless, the mapping to a target trial makes it clear that all identification assumptions underlying interventional effects would be assessable in randomised experiments of the hypothetical interventions. This contrasts with natural effects, which require "cross-world independence" assumptions that are not empirically verifiable, even in hypothetical experiments,[1,6] as well as further untestable assumptions in the context of multiple mediators.[18,22,49,50] This difference is due to interventional effects being population-level quantities, like the total causal effect, whilst natural effects are individual-level effects.[22] An exception for natural effects is when the exposure is separable into components acting through distinct pathways,[6,51,52] with the resulting separable effects emulating hypothetical trials of intervention regimes on the exposure components.

Assumptions about the causal ordering of the mediators are not needed for defining and identifying the proposed effects except those under approach (b) to Question 1. This is facilitated by the fact that estimand assumptions pertain to the joint distribution and, for sequential policies, the choice of question for the policy-maker (e.g. which sequence of policies is of interest?). As previously mentioned, the price to pay for considering the joint distribution in the estimand assumptions, even under approach (b) to Question 1, is the need for unverifiable assumptions about the dependence between the mediators under the hypothetical interventions, which, as the shifts themselves, would not be identifiable from the data. It was interesting to note, however, that results under approaches (a) and (b) to Question 1 were very similar, which is consistent with the expectation that effects via interdependence are small, following Vansteelandt and Daniel.[19] It may therefore be that assumptions about mediator interdependencies do not have much impact on estimates.

It is important to make a connection with the literature on estimation of causal effects under distributional interventions on an exposure[43,44] – called "population intervention effects of stochastic interventions". In the context of one mediator, interventional mediation effects are equivalent to population intervention effects of stochastic interventions in the mediator within the exposed group. With multiple mediators, as mentioned previously, our approach focusing on stochastic interventions on the joint mediator distribution provides a way of identifying those effects in the context of unknown causal ordering of the mediators, which would otherwise be needed for appropriate confounding control. This connection will be important when extending this approach to continuous mediators because the scenario of continuous exposures has been considered in depth in the literature on population intervention effects of stochastic interventions. Indeed, careful thought would be needed regarding sensible estimand assumptions for, say, a two-parameter distribution, e.g. to specify how the hypothetical intervention affects the mean and the variance of the target mediator. As such, this would be best investigated in the context of a real example, as we did here.

A (non-causal) ordering needs to be chosen for estimating the joint mediator distribution if a sequential regression approach is used. We implemented g-computation using highly flexible regression models, but parametric misspecification bias is still a possibility. Development of doubly or multiply robust methods for estimation with machine learning, building on recent work,[53] would be desirable to counter parametric misspecification bias.

Importantly, our goal in this work was to define the contrasts of interest in the context of questions regarding ill-defined mediator interventions, acknowledging that this is only one step of a full "target trial approach," which must also consider further protocol components of the target trial.[14] Further applications and future extensions of our proposal, e.g. to time-varying mediators and dynamic policies, should consider the broader set of target trial principles. Nonetheless, our proposal opens new avenues for causal inference about policy-relevant effects with ill-defined interventions.

Acknowledgements

We thank the reviewers for invaluable feedback and Carolyn Coffey for sharing her knowledge of the Victorian Adolescent Health Cohort Study.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a Discovery Early Career Researcher Award fellowship to MMB from the Australian Research Council [DE190101326]. This work was also supported by The University of Melbourne (MMB) and the NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol, England (PM). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care. The Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

Availability of data and code for replication

Example R code for implementing the method, including a worked example on simulated data, can be accessed at the first author′s GitHub repository (https://github.com/moreno-betancur/medRCT). Data from the Victorian Adolescent Health Cohort Study are not publicly available but those interested in replicating these findings are welcome to contact the study team.

ORCID iD

Margarita Moreno-Betancur https://orcid.org/0000-0002-8818-3125

Supplemental material

Supplemental material for this article is available online.

References

1. Naimi AI, Kaufman JS and MacLehose RF. Mediation misgivings: Ambiguous clinical and public health interpretations of natural direct and indirect effects. *Int J Epidemiol* 2014; **43**: 1656–1661.
2. Robins JM and Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; **3**: 143–155.
3. Pearl J. Direct and indirect effects. In: Proceedings of the seventeenth conference on uncertainty and artificial intelligence, San Francisco, USA, August 2001. San Francisco, CA: Morgan Kaufmann, 2001, pp.411–420.
4. VanderWeele TJ. *Explanation in causal inference: methods for mediation and interaction*. New York, NY: Oxford University Press, 2015.
5. Imai K, Keele L and Tingley D. Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci* 2010; **25**: 51–71.
6. Robins JM and Richardson TS. Alternative graphical causal models and the identification of direct effects. In: Shrout P, Keyes K, Ornstein K (eds) Causality and psychopathology: finding the determinants of disorders and their cures. Oxford, UK: Oxford University Press, 2011, pp.103–158.
7. Hernan MA. Does water kill? A call for less casual causal inferences. *Ann Epidemiol* 2017; **26**: 674–680.
8. Hernan MA. Do you believe in causes? The distinction between causality and causal inference. In: European causal inference meetings – EuroCIM, Bremen, Germany, March 2019.
9. Galea S. An argument for a consequentialist epidemiology. *Am J Epidemiol* 2013; **178**: 1185–1191.
10. Galea S and Hernán MA. Win-win: reconciling social epidemiology and causal inference. *Am J Epidemiol* 2020; **189**: 167–170.
11. Gelman A and Imbens G. Why ask why? Forward causal inference and reverse causal questions. NBER Working Paper Series, Working Paper 19614, 2013.
12. Holland PW. *Causation and race*. *ETS Res Rep Ser* 2003; **2003**: i–21.
13. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008; **19**: 766–779.
14. Hernán MA and Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016; **183**: 758–764.
15. Jackson JW and Arah OA. Making causal inference more social and (social) epidemiology more causal. *Am J Epidemiol* 2020; **189**: 179–182.
16. Vanderweele TJ. Counterfactuals in social epidemiology: thinking outside of "The Box." *Am J Epidemiol* 2020; **189**: 175–178.
17. Robinson WR and Bailey Z. What social epidemiology brings to the table: reconciling social epidemiology and causal inference. *Am J Epidemiol* 2020; **189**: 171–174.

18. VanderWeele TJ, Vansteelandt S and Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* 2014; **25**: 300–306.
19. Vansteelandt S and Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology* 2017; **28**: 258–265.
20. Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *J R Stat Soc Ser B* 2007; **69**: 199–215.
21. Didelez V, Dawid AP and Geneletti S. Direct and indirect effects of sequential treatments. In: Dechter R and Richardson T (eds) Proceedings of the 22nd annual conference on uncertainty in artificial intelligence. Arlington, VA: AUAI Press, 2006, pp.138-164.
22. Moreno-Betancur M and Carlin JB. Understanding interventional effects: a more natural approach to mediation analysis? *Epidemiology* 2018; **29**: 614–617.
23. Lin S-H and VanderWeele T. Interventional approach for path-specific effects. *J Causal Inference* 2017; **5**.
24. Hawton K, Saunders KE and O'Connor RC. Self-harm and suicide in adolescents. *Lancet* 2012; **379**: 2373–2382.
25. Moran P, Coffey C, Romaniuk H, et al. The natural history of self-harm from adolescence to young adulthood: a population-based cohort study. *Lancet* 2012; **379**: 236–243.
26. Morgan C, Webb RT, Carr MJ, et al. Incidence, clinical management, and mortality risk following self-harm among children and adolescents: cohort study in primary care. *BMJ* 2017; **359**: j4351.
27. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators, Abajobir AA, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 2017; **390**: 1211–1259.
28. Bergen H, Hawton K, Waters K, et al. Premature death after self-harm: a multicentre cohort study. *Lancet* 2012; **380**: 1568–1574.
29. Mars B, Heron J, Crane C, et al. Clinical and social outcomes of adolescent self-harm: population based birth cohort study. *BMJ* 2014; **349**: g5954.
30. Moran P, Coffey C, Romaniuk H, et al. Substance use in adulthood following adolescent self-harm: a population-based cohort study. *Acta Psychiatr Scand* 2015; **131**: 61–68.
31. Borschmann R, Becker D, Coffey C, et al. 20-year outcomes in adolescents who self-harm: a population-based cohort study. *Lancet Child Adolesc Heal* 2017; **1**: 195–202.
32. Hernan MA and Vanderweele TJ. Compound treatments and transportability of causal inference. *Epidemiology* 2011; **22**: 368–377.
33. VanderWeele TJ and Hernan MA. Causal inference under multiple versions of treatment. *J Causal Inference* 2013; **1**: 1–20.
34. Hernán MA. Invited commentary: agent-based models for causal inference-reweighting data and theory in epidemiology. *Am J Epidemiol* 2015; **181**: 103–105.
35. VanderWeele TJ and Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* 2014; **25**: 473–484.
36. Rudolph KE, Sofrygin O, Zheng W, et al. Robust and flexible estimation of stochastic mediation effects: a proposed method and example in a randomized trial setting. *Epidemiol Method* 2018; **7**: 1–26.
37. Popham F. Controlled mediation as a generalization of interventional mediation. *Epidemiology* 2019; **30**: e21–e22.
38. VanderWeele TJ and Tchetgen Tchetgen EJ. Mediation analysis with time varying exposures and mediators. *J R Stat Soc Ser B Stat Methodol* 2017; **79**: 917–938.
39. Lin S-H, Young J, Logan R, et al. Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators, and confounders. *Epidemiology* 2017; **28**: 266–274.
40. Moreno-Betancur M, Koplin JJ, Anne-Louise P, et al. Measuring the impact of differences in risk factor distributions on cross-population differences in disease occurrence: a causal approach. *Int J Epidemiol* 2018; **47**: 217–225.
41. Jackson JW and VanderWeele TJ. Intersectional decomposition analysis with differential exposure, effects, and construct. *Soc Sci Med* 2019; **226**: 254–259.
42. Jackson JW and VanderWeele TJ. Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology* 2018; **29**: 825–835.
43. Díaz I and van der Laan M. Population intervention causal effects based on stochastic interventions. *Biometrics* 2012; **68**: 541–549.
44. Díaz I and Hejazi NS. Causal mediation analysis for stochastic interventions. *J R Stat Soc Ser B (Statistical Methodol)* 2020; **82**: 661–683.
45. Hernan MA and Robins J. *Causal inference: what if*. Boca Raton, FL: Chapman & Hall/CRC, 2020.
46. Micali N, Daniel RM, Ploubidis GB, et al. Maternal prepregnancy weight status and adolescent eating disorder behaviors a longitudinal study of risk pathways. *Epidemiology* 2018; **29**: 579–589.
47. R Core Team. *R: a language and environment for statistical computing*, 2013, http://www.r-project.org
48. Naimi AI, Moodie EEM, Auger N, et al. Stochastic mediation contrasts in epidemiologic research: interpregnancy interval and the educational disparity in preterm delivery. *Am J Epidemiol* 2014; **180**: 436–445.

49. Daniel RM, De Stavola BL, Cousens SN, et al. Causal mediation analysis with multiple mediators. *Biometrics* 2015; **71**: 1–14.
50. Avin C, Shpitser I and Pearl J. Identifiability of path-specific effects. In: Proceedings of international joint conference on artificial intelligence, Edinburgh, Scotland, 2005, pp.357-363.
51. Didelez V. Defining causal meditation with a longitudinal mediator and a survival outcome. *Lifetime Data Anal* 2019; **25**: 593–610.
52. Aalen OO, Stensrud MJ, Didelez V, et al. Time-dependent mediators in survival analysis: Modeling direct and indirect effects with the additive hazards model. *Biometrical J* 2020; **62**: 532–549.
53. Benkeser D. *Nonparametric inference for interventional effects with multiple mediators*, 2020, http://arxiv.org/abs/2001.06027 (accessed 17 March 2020).