Exposome, 2026, **6(1)**, osag010
https://doi.org/10.1093/exposome/osag010
Advance Access Publication Date: 13 March 2026
Research Article

Overview of mediation methods: from estimands to estimations

Benoît Lepage[1,2,*], Hélène Colineaux[3], Valérie Garès[4], Barbara Bodinier[3], Cyrille Delpierre[1,†], Marc Chadeau-Hyam[3,†]

[1]Equity team, CERPOP, UMR 1295, Université de Toulouse, Inserm, Toulouse, France
[2]Epidemiology Department, Toulouse University Hospital, Toulouse, France
[3]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, UK
[4]Université de Rennes, INRIA, IRMAR-CNRS 6625, IRSET-Inserm-1085, Rennes, France

*Corresponding author: Benoît Lepage, Université de Toulouse, Facuté de santé, 37, allées Jules Guesde, 31000, Toulouse, France (benoit.lepage@utoulouse.fr, benoit.lepage@univ-tlse3.fr)

[†]Joint senior authors

Abstract

The exposome paradigm aims to characterise the totality of environmental exposures shaping health across the life course, integrating chemical, physical, behavioural, and social domains. While exposome studies have been highly successful in describing complex exposure patterns and mixtures, they often rely on associative analytical frameworks, which can limit the interpretation of results in terms of causal mechanisms and potential intervention targets. Causal mediation analysis offers a natural framework to address these challenges by decomposing total exposure effects into pathway-specific components. However, the diversity of mediation estimands, assumptions, and analytical strategies developed in the causal inference literature may have limited their use in exposome research. This article provides a structured synthesis of modern causal mediation analysis approaches, with a focus on their conceptual foundations and relevance for exposome and life-course epidemiology. We review classical and contemporary mediation frameworks, including controlled, natural, and interventional direct and indirect effects, and discuss their identification assumptions under different causal structures. Particular attention is given to settings encountered in exposome research, such as time-varying exposures, exposure-induced confounding, high-dimensional mediators, and survival outcomes. By clarifying the conceptual landscape of causal mediation analysis and its applicability to exposome research, this work aims to support more interpretable, mechanism-oriented, and causally-informed investigations of how environmental exposures become biologically embodied across the life course.

Key words: mediation analyses, structural equation models, causality, exposome analytics, counterfactual.

Introduction

Why mediation analysis?

The exposome concept was originally proposed to complement the genomic paradigm by characterising the totality of environmental exposures across the life course likely to influence gene expression.[1] Since then, exposome research has developed into a interdisciplinary field, bringing together environmental sciences, epidemiology, toxicology, omics technologies and social sciences.[2] A central ambition of this paradigm is to move beyond isolated risk factors and to capture complex exposure profiles spanning chemical, physical, behavioural, and psychosocial domains, often using high-dimensional and data-driven analytical strategies.[3]

Alongside these methodological advances, exposome research has also faced persistent conceptual and interpretative challenges. One key tension concerns the integration of heterogeneous exposures within a coherent causal framework. In many exposome-wide association studies, psychosocial, behavioural, physical, and chemical exposures are modelled simultaneously as parallel predictors of health outcomes.[4] While this strategy is effective for identifying exposure signatures, it can blur distinctions between upstream and downstream determinants. In particular, social conditions are often treated as covariates or contextual modifiers, rather than as structuring forces that shape exposure distributions and biological responses over time.[5,6]

A related challenge is the predominance of associative analytical frameworks. Exposome studies excel at detecting correlations, clustering exposures, and characterising mixtures, but frequently stop short of articulating explicit causal questions.[7] As a result, estimated associations may be difficult to interpret in terms of mechanisms, intervention targets, or policy relevance. This limitation is especially salient when the scientific aim is not only to describe environmental complexity, but to understand how external environments become biologically embodied and translated into disease processes across the life course.[8,9]

Causal inference frameworks offer a natural extension to address these challenges. By making assumptions about temporal ordering, confounding, and causal pathways explicit, causal models enable researchers to distinguish between total effects, pathway-specific effects, and effects operating through intermediate variables.[10,11]

Received: February 23, 2026; **Revised:** March 3, 2026; **Accepted:** March 4, 2026

In the context of the exposome, such frameworks provide tools to reintroduce causal structure into complex exposure systems, while preserving the multidimensional perspective that motivated the exposome paradigm.[2]

Within this perspective, mediation analysis occupies a central position. Conceptually, mediation analysis aligns closely with the core objectives of exposome research, as it seeks to elucidate the processes through which external exposures influence internal biological systems and, ultimately, health outcomes.[12] Methodologically, mediation analysis allows the decomposition of total effects into components operating through specified intermediate mechanisms, while accounting for confounding, effect modification, and temporal ordering.[13-15] Over the past two decades, advances grounded in counterfactual reasoning and graphical causal models have substantially expanded the scope of mediation analysis beyond traditional regression-based approaches.[12,16]

These developments are particularly relevant for exposome research, where mediators may be time-varying, socially patterned, high-dimensional, and themselves affected by prior exposures. Biological intermediates such as biomarkers, multi-system scores, or omics-derived profiles are often influenced by complex exposure histories and may simultaneously act as confounders and mediators in longitudinal settings.[17] Under such conditions, naïve mediation approaches may yield biased or uninterpretable estimates, underscoring the importance of carefully defined causal estimands and identification assumptions. At the same time, the diversity of available mediation estimands, modelling strategies, and identification conditions can make their practical implementation challenging. Differences between controlled, natural, interventional, and stochastic effects, as well as the treatment of exposure-induced confounding and time-varying mediators, are not always transparent to applied researchers. This complexity may limit the uptake of causal mediation approaches in exposome studies, despite their conceptual relevance.

The objective of this article is to provide a structured synthesis of modern causal mediation analysis methods, with a focus on their conceptual foundations, underlying assumptions, and relevance for exposome research. Rather than proposing new methodology, we aim to clarify the relationships between classical and contemporary mediation frameworks, to highlight the causal questions they answer, and to discuss their applicability in complex exposure systems typical of exposome and life-course studies. By doing so, we seek to support the appropriate and transparent use of mediation analyses in exposome research and to contribute to more interpretable, mechanism-oriented, and policy-relevant investigations of environmental health.

Notations and examples

By convention, A denotes the exposure of interest (also referred to as "intervention" or "treatment") and Y denotes the outcome. The mediator of interest is represented by M. Temporal ordering assumptions are essential in mediation analyses (and in causal analyses in general), so we might use t to indicate the temporal ordering of a variable $V(t)$. As an illustrative example, we will use a possible research question inspired by work from the Expanse project on the urban exposome.[18,19] We might want to investigate whether the early exposure to physico-chemical pollution A (such as being exposed to high levels of $PM_{2.5}$) influences global death later in life (Y), and if so, whether this effect is mediated by an increase in the risk of type 2 diabetes (M) which would in turn increase the risk of death (Y) (Figure 1). The set of baseline confounders will be denoted $L(0)$. In the example, we could consider other components of the early environment (built, social, physical environment, …). As we will see later, it is also important to take into account possible confounders of the mediator-outcome relationship. In our example, we might think of overweight, chronic stress, inflammatory response, lifestyle habits, social position during adulthood, etc,

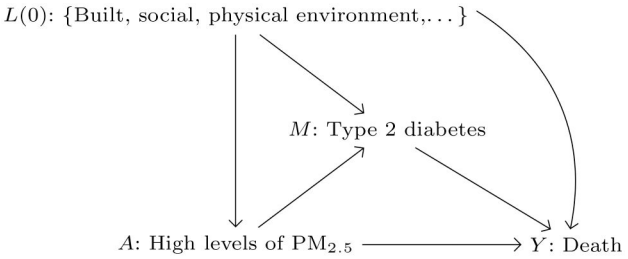


Figure 1. Directed acyclic graph summarising our example.

In order to answer the question, we want to decompose the total (causal) effect of being exposed to high levels of $PM_{2.5}$ on death into the sum of an indirect effect through type 2 diabetes ($A \rightarrow M \rightarrow Y$) and a direct effect ($A \rightarrow Y$). The causal model depicting the causal links between $L(0)$, A, M, and Y can be summarised in a directed acyclic graph (DAG, defined below) as illustrated in the model of Figure 1.

From Baron & Kenny to structural causal model

The founding methods of mediation analysis are the Baron and Kenny and the path analysis approaches.

Baron and Kenny approach

The Baron and Kenny approach is based on the sequential and step-wise estimation of linear regression models to explore simple causal structures.[20,21] This approach relies on the following steps:

1. Testing if A has an effect on Y. This *total effect* θ_A of A on Y can be tested using the following linear model: $\mathbb{E}(Y \mid A, L(0)) = \theta_0 + \theta_A A + \theta_L L(0)$. So, the effect θ_A is the linear regression coefficient of $A(0)$.
2. Testing if A has a significant effect β_A on the intermediate variable M, using the following linear regression: $\mathbb{E}(M \mid A, L(0)) = \beta_0 + \beta_A A + \beta_L L(0)$
3. Testing if the mediator M has a significant effect γ_M on the outcome Y, independently from A and $L(0)$, using the linear regression: $\mathbb{E}(Y \mid A, M, L(0)) = \gamma_0 + \gamma_A A + \gamma_M M + \gamma_L L(0)$.

Based on those three models, M would be considered a mediator of the $A-Y$ relationship if the coefficients β_A and γ_M are found to be statistically significant. In the third equation, γ_A is construed as the "direct effect" of A on Y, representing the effect that does not *pass through* M. Intuitively, if the null hypothesis $\{\mathcal{H}_0 : \gamma_A = 0\}$ is rejected and $\gamma_A < \theta_A$, the effect of A on Y could be considered as partially mediated by M. If the null cannot be rejected, the influence of A on Y is deemed to be entirely mediated by M.

Beyond null hypotheses testing, the "product method" or the "difference method" have been used to explicitly quantify the direct and indirect effects of A on Y.[21-24] From the models described above, the *total effect* of A on Y is estimated by the regression coefficient θ_A; the *direct effect* of A on Y is estimated by the regression coefficient γ_A; and the *indirect effect* of A on Y (corresponding to the path $A \rightarrow M \rightarrow Y$) is estimated by either (i) the "difference in coefficients" $\theta_A - \gamma_A$ using the first and third models, or (ii) the "product of coefficients" $\beta_A \times \gamma_M$ using the second and third models. If only linear least square regressions are

involved with quantitatives M and Y variables, the two methods give the same estimation of the indirect effect: $\theta_A - \gamma_A = \beta_A \times \gamma_M$.[23,24] In situations mixing categorical mediators M and quantitative outcomes Y, the "product method" can not be applied if the coefficients are estimated on different scales, however it is still possible to use the "difference method."

Because the indirect effect is derived from two different equations, there is no straightforward way to compute standard errors or 95% confidence intervals. Among other solutions, bootstrap approaches have shown good performance.[22,24]

Path analysis and structural equation modelling (SEM)

The development of path analysis started in the 1920s and has been predominantly applied within the domains of econometrics, social sciences, and psychology.[25-27] Path analysis is explicitly based on the integration of a graphical representation of causal structures, a set of linear models, and assumptions concerning the covariance structures of random residuals and latent variables. Concerning the graphical representation, the rules are akin to those used with Directed Acyclic Graphs (DAGs, see below); however, path diagrams may also encompass loops and additional nodes that represent variable transformations, useful in modelling non-linearity (eg, polynoms of $L(0)$ or interaction terms $(A * M)$, alongside the original nodes $L(0)$, A, and M).[23,28]

In our example, three endogenous variables, A, M and Y, are identified and a set of structural equations are defined and modelled using specific linear regressions setting their direct causes as explanatory variables, and we assume that the random residuals ε_A, ε_M and ε_Y are independent from one another:

$$\begin{aligned} A &= \lambda_0 + \lambda_L L(0) + \varepsilon_A \\ M &= \beta_0 + \beta_A A + \beta_L L(0) + \varepsilon_M \\ Y &= \gamma_0 + \gamma_A A + \gamma_M M + \gamma_L L(0) + \varepsilon_Y \end{aligned}$$

The coefficients $\{\lambda_L, \beta_A, \beta_L, \gamma_A, \gamma_M, \gamma_L\}$ are called *path coefficients* and measure the direct effect of a cause on its target covariate. For example, the *path coefficient* γ_M quantifies the direct effect of M on the outcome Y. Coefficients can be standardised: Γ_M is the standardised path coefficient (upper-case letter) of the unstandardised coefficient (lower-case letter) γ_M, defined by $\Gamma_M = \gamma_M \frac{\sigma_M}{\sigma_Y}$ (where σ_V is the variance of V).

According to Sewall Wright, the correlation between two variables is explained by the set of all paths which link these variables, ie, all the direct effects, indirect effects, and "joint effects," where joint effects correspond to confounding effects.[28] Assuming the underlying parametric hypotheses are true (uncorrelatedness of residuals, linearity and additivity), he proposed some graphical rules to decompose the correlation between two variables according to path coefficients and correlation between residuals. Such an analysis is called a *path analysis*.

In our example, the Pearson correlation between A and Y can be expressed as the sum of four paths connecting A and Y, each path being quantified by the standardised paths coefficients (for single arrows between A and Y) or by the product of standardised paths coefficients (for paths composed of a sequence of arrows):

- Path $A \rightarrow Y$: the "direct effect" of A on Y, quantified by Γ_A
- Path $A \rightarrow M \rightarrow Y$: the "indirect effect" of A on Y, quantified by $B_A \times \Gamma_M$
- Path $A \leftarrow L \rightarrow M \rightarrow Y$, quantified by $\Lambda_L \times B_L \times \Gamma_M$
- Path $A \leftarrow L \rightarrow Y$, quantified by $\Lambda_L \times \Gamma_L$

Path analysis approach in mediation analyses is therefore very similar to the "product of coefficients" approach described above, with the difference that unstandardised coefficients were used in the "product of coefficients" approach, eg, $\beta_A \times \gamma_M$. The correlation between A and Y, can be decomposed, as the sum of the first two paths corresponding to the total effect of interest of $A \rightarrow Y$ (the direct effect + the indirect effect) and the two other paths correspond to confounding effects:

$$\overbrace{\Gamma_A + B_A \Gamma_M}^{\text{Total effect}} + \overbrace{\Lambda_L B_L \Gamma_M + \Lambda_L \Gamma_L}^{\text{Confounding by } L}$$

Our structural assumptions and the principles of path analysis imply correlations that can be articulated as a combination of parameters to be estimated. These parameters are inferred by aligning the implied correlations with the observed correlations. In practice, maximum likelihood estimation and variants of generalised least squares are the predominant methods employed in structural equation modeling software.[29]

These methodologies can be applied to causal structures encompassing latent variables (ie, unobserved constructs), commonly referred to as Structural Equation Modelling (SEM). In this framework, observed measures are associated with latent constructs as in factor analyses, delineating measurement models. The integration of latent variables and measurement models constitutes the principal strength of this approach.[30]

Classical methods can be extended to accommodate binary or categorical mediator and/or outcome variables.[23,31,32] In scenarios involving interaction between the exposure A and the mediator M, specific procedures have been developed.[33-35] In cases of intra-individual interaction between A and M influencing Y, Kaufman and colleagues demonstrated that the classical "product of coefficients" or "difference in coefficients" methods may not be reliable to decompose the total effect into the sum of a direct and indirect effect.[36] More recently, those methods have been adapted to account for such interactions.[37-39]

A fundamental limitation in the estimation of path coefficients within both path analysis and SEM arises when the model is under-determined (or under-identified). A model is deemed under-identified if at least one parameter cannot be discerned from the observed correlations. Such under-identification may occur due to an insufficient number of indicators for one or more latent variables within the model, or to the presence of excessive reciprocal paths, feedback loops, or correlated residuals.[28] Furthermore, the practice of comparing alternative structural models using statistical tests or indicators is prevalent in SEM methodology to derive more parsimonious models.[28,29] However, determining the presence or absence of direct effects between two nodes based on statistical procedures may be inappropriate, as the results are often contingent on sample size and statistical power, while the absence of an arrow represents a strong assumption. Bollen and colleagues assert that the re-specification of an initial model is more aligned with an exploratory analysis approach and recommend prioritising expert knowledge before employing empirical statistical tests and fit measures.[29] Moreover, the conventional estimation method for SEMs, which involves estimating an extensive set of parameters through iterative maximisation of a fitness measure, may not be optimal for confirmatory analysis approaches.[40]

Certain authors argue that the primary utility of SEM or path analysis lies in the exploration of novel research hypotheses.[41] For confirmatory purposes, alternative methodologies for

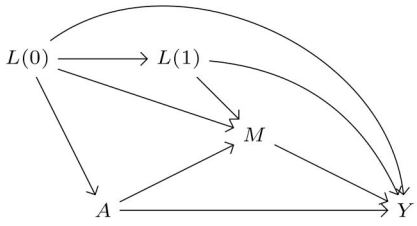
mediation analysis have been developed. These alternative approaches, grounded in the counterfactual framework and directed acyclic graphs (DAGs), distinctly separate statistical assumptions (that refer to the observed data distribution) and causal assumptions (that refer to knowledge external to the observed data, that might not be empirically testable), thereby facilitating the management of interactions, confounding, and sensitivity analyses.[10,41]

Non parametric structural causal models

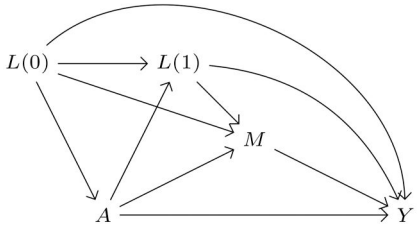
Another limitation of mediation analyses employing either the Baron and Kenny approach, path analysis or SEM, occurs with more complex systems. For example, it is necessary to consider confounders $L(1)$ of the mediator-outcome relationship to avoid biased estimations.[13,36,42] In longitudinal settings, it seems reasonable to assume that these "intermediate confounders" can be affected by the exposure A (as in Figure 2(b)). These variables are sometimes referred to as "time-varying covariates" or "recanting witness"[43]) In such causal systems, the "difference in coefficients" or "product of coefficients" approaches are inadequate, and the multiplication of paths between the exposure A and the outcome Y require a more precise formulation of the scientific question to better define the direct and indirect effects. Advancements in mediation analyses relied on concepts from the causal inference literature to express causal objectives more accurately and to develop estimation methods more focused on the targeted direct and indirect effects.[14]

Causal inference framework

Pearl integrated three complementary components to describe a structural causal model, combining "*features of the SEM [...], the potential outcome framework of Neyman and Rubin,*[44] *and the graphical models developed for probabilistic reasoning and causal analysis*" (ie, non-parametric structural equation models associated with DAGs).[14]



(a) DAG of model $\mathcal{M}_1$



(b) DAG of model $\mathcal{M}_2$

Figure 2. DAGs representing data-generating mechanisms for the distribution of $\{L(0), A, L(1), M, Y\}$.

Pearl's framework is based on counterfactual reasoning,[45] which seeks to address the hypothetical scenario of "what would have happened had the past been different." For instance, "what would the probability of death had been, had the whole population been exposed to low levels of $PM_{2.5}$?", or "what would the probability of death had been in a population exposed early to high levels of $PM_{2.5}$, but where the individual status of type 2 diabetes was changed to the status expected under low levels of $PM_{2.5}$?". Using counterfactuals enables inferences about scenarios not observed (or even unobservable) in the empirical world.

Donald Rubin and Judea Pearl proposed specific notations for conducting interventional and counterfactual causation analyses. Pearl employs a "do()" notation to signify hypothetical interventions: $\mathbb{P}(Y = y|do(A = 0))$ denotes the probability that the outcome Y would attain the value y in a hypothetical scenario where every participant is exposed to low levels of $PM_{2.5}$. Rubin's *potential outcome* notations correspond to random variables, defined as events that did not occur but could have. The notation $Y_{A=a}$ or Y_a represents the value the (potential) outcome Y would take had the exposure A been at level a. For example, the probability of death had the whole population been exposed to low levels of $PM_{2.5}$ is $\mathbb{P}(Y_{A=0} = 1)$. Conversely, the probability of death in a population fully exposed to high levels of $PM_{2.5}$ is denoted as $\mathbb{P}(Y_{A=1} = 1)$. For an individual i, the causal effect of a binary variable A on Y can be expressed using the contrast between the two potential outcomes, such as $Y_{A=1}(i) - Y_{A=0}(i)$. The notation of potential outcomes will be employed throughout the remainder of this manuscript. For simplicity, we will denote Y_a the counterfactual variable Y under the hypothetical scenario setting $A = a$ in the whole population, and Y_{am} under the scenario setting $\{A = a, M = m\}$ in the whole population.

Various types of counterfactual interventions or counterfactual scenarios can be defined, where the imaginary interventions can be static, dynamic or stochastic. Static interventions are characterised by setting the exposure of the entire population to a specific value. For example, $P(Y_{A=a} = y)$ is the probability the outcome Y would attain the value y, had the whole population been exposed to the value $A = a$. Dynamic interventions is usually used to describe dynamic regimes in which the imaginary intervention on $A(t)$ depends on the values of previous (time-varying) covariates $\{L(0), \ldots, L(t-1)\}$. As an example in Figure 2 (b), it is possible to define a joint exposure on $\{A, M\}$ setting the values of A and M as a function of the previous time-varying covariates $d_t\Big(L(0), \ldots, L(1)\Big)$. Different dynamic regimes can be defined based on alternative rules, which can be useful to define treatments according to monitoring variables, for example "change insulin therapy if blood glucose exceeds a given threshold." This approach has been generalised with "modified treatment policies" defined as hypothetical interventions where the post-intervention value of treatment can depend on the actual observed treatment level and the unit's history.[46] For *Stochastic interventions*, the hypothetical intervention corresponds to a random draw in a distribution specified by the analyst. For example, we can set the value of A as a random draw from a Bernoulli distribution of parameter π (setting $A \sim \mathcal{B}(\pi)$). In mediation analyses, some direct and indirect effects are defined based on hypothetical random draws of the mediator distribution, for example $M \sim \Gamma_{M_a} \mid L(0)$ corresponds to a random draw of the mediator from its distribution (within strata of $L(0)$) under the counterfactual intervention setting $A = a$.

Directed acyclic graph

Causal relationships, which are directed from a cause to an effect, can not be formulated with equations alone, which by nature can only describe symmetrical relationships and not directional ones. Wright[25] already suggested to combine graphs with parametric equations. These graphs are the "path diagrams," associated with the structural equations. Beyond the "path diagram" framework, Directed Acyclic Graphs (DAGs) can be used to represent nonparametric structural equations.[10,47] By definition, DAGs have the following principles:

- All the links are directed: every edge in a path is an arrow that points from one variable to the other.
- The graph is acyclic: a DAG does not contain loops.
- Every arrow $V \rightarrow W$ is interpreted as a "possible" effect of V on W, the absence of an arrow from a variable V to another W is a strong and explicit statement (based on prior knowledge) that there is no direct effect of V on W.
- Every common cause of two variables represented in a DAG should also appear in the DAG, even if the common cause is unmeasured in the observed data. In the literature, such unmeasured common causes are usually represented with U variables (for *unknown*), with dashed arrows or dashed double arrows (assuming a common unknown cause is present between the double arrows).
- Represented relationships should be stable over time and circumstances, like autonomous physical mechanisms.[10,48] This implies that changing one relationship without changing the others is conceivable (a relationship is unaffected by possible changes in the form of other functions).

Kinship terminology is generally used to describe the relationships between the variables in a DAG. For example in the DAG of Figure 1, Y is a *child* of M and M is a *parent* of Y. A and M are *ancestors* of Y, and M and Y are *descendents* of A.

DAGs can be used as a convenient way to formulate and visualise possible causal relationships and independence assumptions. For example in Figure 3, we can represent: a direct effect of A on Y ($A \rightarrow Y$ in Figure 3a and 3b); an indirect effect of A on Y through M (M is a mediator of the effect of A on Y, $A \rightarrow M \rightarrow Y$ in Figure 3a); a back-door path (ie, a confounding path) between A and Y ($A \leftarrow L \rightarrow Y$ in Figure 3b); and a collider C (ie, a common child) on the path between A and B (Figure 3c).[47]

DAGs can also help to deduce graphically what are the expected independences and conditional independences using the *d-separation* criterion.[49,50] For a DAG compatible with the data set under study, two variables V and W are said to be *d-separated* by a set of variables Z if all the paths between them are blocked conditional on Z. Such d-separation by Z implies that V and W are independent given Z. Graphically, a path connecting two variables V and W is "blocked" conditional on Z if: (1) there is a variable on the path which belongs to Z and which is not a collider, or (2) there is a *collider* on the path and the collider or any of its descendent does not belong to Z.[10]

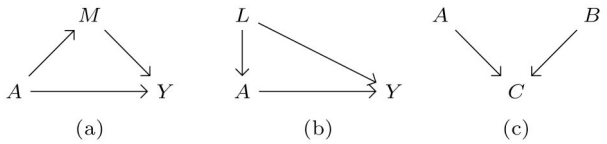


Figure 3. Direct effets, indirect effects, backdoor paths, and colliders.

It is possible to test the compatibility between a DAG and a data-set using the d-separation criterion (several DAGs can usually be compatible with a single data-set in terms of independence and conditional independence). For example, the R package DAGitty can be used to evaluate the testable implications associated with a given DAG and assess if the DAG is consistent with a data-set.[51]

Graphs can help to evaluate if a causal effect of interest is identifiable from observational data, under the assumptions depicted in the DAG. For example, the *backdoor criterion* can be used to select covariate adjustment sets required to identify causal effects. This criterion is formulated as: "Z is a sufficient adjustment set in order to test and estimate the effect of A on Y if : (i) no variable in the set Z is a descendant of A and (ii) each backdoor path between A and Y is blocked. A step-by-step method has been described to apply this criterion.[52]

Because DAGs are acyclic, dealing with bi-directional relationships requires distinguishing between two situations:

- A simple relationship $A \leftrightarrow Y$ is usually interpreted as confounding ($A \leftarrow L \rightarrow Y$), assuming the presence of common parents L between the two ends of the path.
- Or feedback loops, which should be disentangled and represented using temporal ordering notation $A(1) \rightarrow Y(1) \rightarrow A(2) \rightarrow Y(2) \rightarrow A(3)\ldots$

Non parametric structural equations—Markov factorisation

The DAG presented in Figure 1 can be formulated using the following set of non parametric structural equations:

$$\begin{aligned} L(0) &= f_{L(0)}\left(U_{L(0)}\right) \\ A &= f_A\left(L(0), U_A\right) \\ M &= f_M\left(A, L(0), U_M\right) \\ Y &= f_Y\left(M, A, L(0), U_Y\right) \end{aligned}$$

Where all residual terms U are assumed to be independent from each other. Note that these residuals can be considered as unmeasured exogenous variables affecting each of the endogenous variables. As we assumed their mutual independence here, it is not necessary to represent them in the DAG. The main difference between the set of structural equation in path analysis or SEMs and a set of non parametric structural equations is that the latter ones make no assumptions about the functional form of the equations.[14] Each of the f function determines the value of the output-variables from the value of the input-variables, and these can take any form. The joint probability of variables represented as nodes in the DAG can be expressed as a product of (conditional) probabilities:

$$\begin{aligned} P[l(0), a, m, y] = {} & P[l(0)] \\ & \times P[a \mid l(0)] \\ & \times P[m \mid l(0), a] \\ & \times P[y \mid l(0), a, m] \end{aligned}$$

Identification assumptions

Under Identification assumptions, it is possible to express the counterfactual (unobserved) causal quantities of interest as a parameter of the observed data. The following assumptions are necessary:[16]

Randomisation (or exchangeability) assumption

According to the DAGs in Figures 1 and 2, applying the backdoor criterion shows that adjusting for all the baseline confounders $L(0)$ is sufficient to identify the causal "total" effect of A on Y. In other words, conditional on $L(0)$, there is no unmeasured confounding between A and Y (denoted $Y_a \perp\!\!\!\perp A \mid L(0)$).

Positivity assumption

Also named *experimental treatment assignment*, the positivity assumption states that within each observed stratum of $L(0)$, each treatment level of interest $A = 1$ and $A = 0$ occurs with some positive probability:

$$\mathbb{P}\Big(A = a \mid L(0) = l(0)\Big) > 0 \quad \forall a \in \{0,1\} \text{ and } \mathbb{P}\Big(L(0) = l(0)\Big) \neq 0.$$

We can differentiate between "theoretical" positivity violations and "practical" positivity violations. For instance, if a treatment $A = 1$ is contraindicated for individuals over the age of 60, the probability of administering this treatment to an 80-year-old subject is assumed to be zero $\mathbb{P}(A = 1 | age = 80) = 0$, which constitutes a theoretical positivity violation. In such a scenario, a scientific objective comparing treatments $A = 1$ versus $A = 0$ in participants over 60 years of age would be irrelevant. Conversely, practical positivity violations may occur when participant profiles are characterised by a high-dimensional set of $L(0)$ variables, continuous $L(0)$ or continuous exposure A, leading to the possibility that some observed profiles are not exposed to one of the treatment levels of interest $A = a$ due to a limited sample size. In case of positivity violation, the estimation of some causal quantities of interest would not be supported by the data in the $l(0)$ strata.

In practice, positivity can be assessed by describing the distribution of the exposures A according to $L(0)$ and the distribution of M according to $\{L(0), A, L(1)\}$;[53] by describing the distribution of propensity scores $g_A = \mathbb{P}(A = 1 \mid L(0))$ and $g_M = \mathbb{P}(M = 1 \mid L(0), A, L(1))$; or by using some specific tools to diagnose positivity violation.[54]

Consistency assumption

This assumption states that "an individual's potential outcome under a hypothetical condition that happened to materialised is precisely the outcome experienced by that individual."[48] Under the consistency assumption, we can write: $\mathbb{P}[Y_a = y \mid A = a, L(0) = l(0)] = \mathbb{P}[Y = y \mid A = a, L(0) = l(0)]$. This statement is used to express an unobserved counterfactual concept (with Y_a on the left hand side of the equation) with a parameter of the observed data distribution (Y on the right hand side of the equation). It is linked to Rubin's "stable unit treatment value assumption" (ie, no hidden version of the treatment: "no matter how individual i received treatment $A = a$ the potential outcome that would be observed would be $Y_{A=a}(i)$").[55]

Its definition and position have been debated in the literature,[48,56-59] mainly around the notion of a "well defined intervention," which should be discussed transparently.

In our example, the consistency of the exposure to high or low levels of $PM_{2.5}$ and consistency of the type 2 diabetes status is questionnable, as these variables are not directly actionable: it is not possible to define unambiguous interventions that would enable an investigator to set the exposure to $PM_{2.5}$ or the type 2 diabetes status to a chosen value or a chosen distribution. Studying the effect of being exposed to air pollution regulation policies and lifestyle education interventions to prevent the occurrence of type 2 diabetes could be considered more reasonable regarding the consistency assumption (at the cost of changing the scientific question and provided the data are available). More generally in the context of exposome research, exposures are characterized by complex relationships between several components. For example, PM2.5 varies in chemical composition in time, by source and geography, and these compositional differences may lead to heterogeneous health effects.[60] Another example are environmental biomarkers which represent metabolic products of multiple parent compounds (for example phthalic acid arising from multiple and/or combinations of phthalates). These complex relationships could be represented on a DAG, but processing them would require their measurements to be available. The dimensionality of the exposure would be greatly increased, making statistical analyses more difficult to carry out.

Causal inference roadmap

Based on DAGs and new notations, the following steps have been suggested as a causal road map to investigate a causal question:[17,61,62]

1. *Define the causal question and causal estimand*. The aim is here to translate the scientific question into a causal quantity of interest using counterfactual notations (a *causal estimand*). Regarding mediation analysis, several quantities of interest have been defined in this way and are detailed below.
2. *Specify knowledge* about the data generating system to be studied using a causal model (using DAGs).
3. *Specify the observed data and their link to the causal model*. This step might help to clarify if some variables are unmeasured and how these missing variables might result in bias.
4. *Assess identifiability and define a statistical estimand*. Discuss the assumptions that make possible to represent the causal quantity of interest as a parameter of the observed data distribution (i.e. a statistical estimand). The assumptions include "no residual confounding assumptions," consistency and positivity assumptions. Software programs such as DAGitty for R can help to assess the exchangeability assumptions, as well as the compatibility between the data and the causal model (considered at steps 2 and 3).[51]
5. *State the statistical estimation problem and estimate*. From the estimand and the assumed statistical model, choose an estimator to approximate the causal quantity of interest. Several estimators have been described in the litterature and are detailed below.
6. *Interpret the results*. Results have to be interpreted by assessing the discrepancy between the data available for our analysis and the causal and statistical assumptions as well as the methodology employed. Sensitivity analyses can help discussing measurement error or "no residual confounding assumptions." This process enables the assessment of the causal gap ("the difference between the true values of the statistical and causal estimands").[17]

Causal quantities of interest in mediation analysis

Based on the concepts developed in the causal inference literature, several causal quantities of interest have been defined to explore the role of mediation variables. These quantities, (not exhaustively) listed in Table 1, correspond to 2-way, 3-way, or 4-way decompositions of a total effect of the exposure A on the outcome Y. The 3-way and 4-way decompositions are mainly

useful to separate the $(A*M)$ interaction effect from the direct and indirect effects of A on Y. The causal quantities of interest are expressed using potential outcome notations (corresponding to *causal estimands*). In this paper, causal effects are presented as contrasts on the additive scale, but they can also be defined on a multiplicative scale using relative risks or odds ratios.

A first essential step is to consider if the exposure A affects intermediate confounders $L(1)$ of the mediator-outcome relationship: do we assume that the causal model corresponds to the Figure 2(a) (model $\mathcal{M}_1$) or the Figure 2(b) (model $\mathcal{M}_2$)? In the latter case, some causal quantities presented below will not be identifiable. In our example, we should discuss if the exposure to high levels of $PM_{2.5}$ (A) can affect potential confounders $L(1)$ (such as overweight, chronic stress, inflammatory response, lifestyle habits, social position during adulthood, etc,) of the effect of type 2 diabetes (M) on death (Y).

Table 1. Synthesis of causal quantities of interest in mediation analyses.

Parameters	Definition
Total effects	
Average Total Effect (ATE)	$\mathbb{E}(Y_1)-\mathbb{E}(Y_0)$
Overall Effect (OE)	$\mathbb{E}\left(Y_{1,G_1\vert L(0)}\right)-\mathbb{E}\left(Y_{0,G_0\vert L(0)}\right)$
2-Way decomposition (1)	
Controlled Direct Effect (CDE_m)	$\mathbb{E}(Y_{1,m})-\mathbb{E}(Y_{0,m})$
Eliminated Effect (EE_m)	$[\mathbb{E}(Y_1)-\mathbb{E}(Y_0)]-[\mathbb{E}(Y_{1,m})-\mathbb{E}(Y_{0,m})]$
2-Way decomposition (2)	
Pure Natural Direct Effect (PNDE)	$\mathbb{E}(Y_{1,M_0})-\mathbb{E}(Y_{0,M_0})$
Total Natural Indirect Effect (TNIE)	$\mathbb{E}(Y_{1,M_1})-\mathbb{E}(Y_{1,M_0})$
2-Way decomposition (3)	
Total Natural Direct Effect (TNDE)	$\mathbb{E}(Y_{1,M_1})-\mathbb{E}(Y_{0,M_1})$
Pure Natural Indirect Effect (PNIE)	$\mathbb{E}(Y_{0,M_1})-\mathbb{E}(Y_{0,M_0})$
2-Way decomposition (4)†	
Marginal Randomised Direct Direct Effect (MRDE)	$\mathbb{E}\left(Y_{1,G_0\vert L(0)}\right)-\mathbb{E}\left(Y_{0,G_0\vert L(0)}\right)$
Marginal Randomised Indirect Effect (MRIE)	$\mathbb{E}\left(Y_{1,G_1\vert L(0)}\right)-\mathbb{E}\left(Y_{1,G_0\vert L(0)}\right)$
2-Way decomposition (5)	
Conditional Randomised Direct Effect (CRDE)	$\mathbb{E}\left(Y_{1,r_0\vert L(0),L(1)}\right)-\mathbb{E}\left(Y_{0,r_0\vert L(0),L(1)}\right)$
Conditional Randomised Indirect Effect (CRIE)	$\mathbb{E}\left(Y_{1,r_1\vert L(0),L(1)}\right)-\mathbb{E}\left(Y_{1,r_0\vert L(0),L(1)}\right)$
3-Way decomposition	
Pure Natural Direct Effect (PNDE)	$\mathbb{E}(Y_{1,M_0})-\mathbb{E}(Y_{0,M_0})$
Mediated Interactive Effect (MIE)	$\mathbb{E}[(Y_{1,1}-Y_{1,0}-Y_{0,1}+Y_{0,0})\times(M_1-M_0)]$
Pure Natural Indirect Effect (PNIE)	$\mathbb{E}(Y_{0,M_1})-\mathbb{E}(Y_{0,M_0})$
4-Way decomposition	
Controlled Direct Effect (CDE_0)	$\mathbb{E}(Y_{1,0})-\mathbb{E}(Y_{0,0})$
Mediated Interaction Effect (MIE)	$\mathbb{E}[(Y_{1,1}-Y_{1,0}-Y_{0,1}+Y_{0,0})\times(M_1-M_0)]$
Reference Interaction Effect (RIE)	$\mathbb{E}[(Y_{1,1}-Y_{1,0}-Y_{0,1}+Y_{0,0})\times M_0)]$
Pure Natural Indirect Effect (PNIE)	$\mathbb{E}(Y_{0,M_1})-\mathbb{E}(Y_{0,M_0})=\mathbb{E}[(Y_{0,1}-Y_{0,0})\times(M_1-M_0)]$

†*The sum is equal to the Overall Effect.*
(Table adapted from Refs.[15,63]).

Average total effect

The aim of mediation analyses is to decompose a total effect, so the first step is to define a total effect of interest. The most common total effect studied in causal analyses is the *average total effect* (ATE), defined as *the difference between the average outcome in the population had everyone been exposed to* $A=1$ (high levels of $PM_{2.5}$) *and the average outcome had everyone been exposed to* $A=0$ (low levels of $PM_{2.5}$). Using counterfactual notation, the ATE is defined as $ATE=\mathbb{E}(Y_{A=1})-\mathbb{E}(Y_{A=0})$ (see Table 1).

Under the identification assumption, the ATE can be expressed as a statistical estimand by the following *g-formula* (cf, Appendix 1 in Supplementary material):

$$\Psi^{ATE}=\sum_{l(0)}[\mathbb{E}(Y\mid A=1,l(0))-\mathbb{E}(Y\mid A=0,l(0))]\times P(L(0)=l(0))$$

Two-way decomposition of the total effect

Several approaches have been described in the literature to decompose a total effect into two components. Some of these quantities (Controlled Direct Effects, Marginal or Conditional Randomised Direct and Indirect effects) can be identified in both causal structures shown in Figures 2(a) and 2(b), but other quantities (Natural Direct and Indirect effects) can be identified only if confounders $L(1)$ of the $M\rightarrow Y$ relationship are not affected by the exposure A (as in Figure 2(a)).

Controlled direct effect

The *controlled direct effect* is defined as the effect of a joint hypothetical intervention that would change the exposure A from a reference value $A=0$ to the value $A=1$, while keeping the mediator constant to a given value $M=m$.[13,64] Using counterfactual notations, $CDE_m=\mathbb{E}(Y_{1,m})-\mathbb{E}(Y_{0,m})$.

Under the identification assumptions (consistency, positivity) and the following sequential randomisation assumptions:

- (A1) No unmeasured confounding between A and Y, given $L(0)$,
- (A2) No unmeasured confounding between M and Y, given $L(0)$, $L(1)$ and A

(cf, table S1 in the supplementary material), the CDE_m is identifiable and can be expressed by the g-formula indicated in table S2 (Supplementary material).

If the set of baseline and intermediate confounders $L(0)$ and $L(1)$ is sufficient, controlled direct effects is identifiable under both causal models represented in Figures 2(a) and (b).

In our example, we could consider estimating two CDEs: the effect of early exposure to $PM_{2.5}$ (contrasting $A=1$*versus*0) in a population where (i) no one had type 2 diabetes (setting $M=0$) or (ii) where everyone had type 2 diabetes (setting $M=1$) (note that this latter causal effect might not be of clinical interest). If there is an $(A*M)$ interaction effect on Y, $CDE_{M=0}$ will be different from $CDE_{M=1}$.

A controlled direct effect corresponds to an effect of A on Y that is not mediated by M. By analogy with path analyses, the CDE corresponds to the direct path $A\rightarrow Y$ in Figure 2(a) and to both paths $A\rightarrow Y$ and $A\rightarrow L(1)\rightarrow Y$ in Figure 2(b).

CDEs can be particularly useful if we aim to assess how intervening on the mediator can mitigate (or increase) a total effect. VanderWeele suggested to use the "eliminated effect" (EE_m) to express the part of the effect eliminated by a hypothetical

intervention setting $M = m$ in the whole population:[65] $EE_m = ATE - CDE_m$ and the "proportion eliminated" as the proportion of the average total effect eliminated by the hypothetical intervention $\frac{ATE - CDE_m}{ATE}$. However, the EE_m cannot be considered as a valid mediated effect: if there is no effect of A on M, the EE_m can still be non-null in the presence of an $(A * M)$ interaction effect on Y.[66]

Natural direct and indirect effect

Pure natural direct and total natural indirect effect

The *pure natural direct effect* (PNDE) was defined by Pearl[64] as the effect on Y that would be realised under the hypothetical intervention of changing the value of A from 0 to 1 (contrasting a population exposed to high *versus* low levels of $PM_{2.5}$), while the mediator was kept constant at the individual counterfactual values $M_{A=0}$ that would be (naturally) observed under the hypothetical intervention setting $A = 0$ (ie, setting the type 2 diabetes variable to the value expected (at the individual level) under exposure to low levels of $PM_{2.5}$):

$$PNDE = \mathbb{E}(Y_{1,M_0}) - \mathbb{E}(Y_{0,M_0})$$

In our example, a Natural Direct Effect can be interpreted as the effect of the exposure to high levels of $PM_{2.5}$ on death (Y) that is not mediated by the occurrence of type 2 diabetes (M).

Based on the following composition assumption: $Y_a = Y_{a,M_a}$ (ie, the potential outcome Y expected under the hypothetical intervention setting $A = a$ is equal to the potential outcome expected under the joint hypothetical intervention setting $A = a$ and M to the counterfactual value M_a expected had the exposure been $A = a$), it is possible to define the *total natural indirect effect* (TNIE) as the difference between the average total effect (ATE) and the pure natural direct effect :[64]

$$\begin{aligned} TNIE &= ATE - PNDE \\ &= [\mathbb{E}(Y_{1,M_1}) - \mathbb{E}(Y_{0,M_0})] - [\mathbb{E}(Y_{1,M_0}) - \mathbb{E}(Y_{0,M_0})] \\ TNIE &= \mathbb{E}(Y_{1,M_1}) - \mathbb{E}(Y_{1,M_0}) \end{aligned}$$

The total natural indirect effect (TNIE) is interpreted as the effect on Y that would be realised under the hypothetical intervention of changing the individual value of the mediator from the counterfactual value M_0 (individual values of type 2 diabetes had the population been exposed to low levels of $PM_{2.5}$) to the counterfactual value M_1 (had the population been exposed to high levels of $PM_{2.5}$), while the exposure to A was kept constant at $A = 1$ (had the population been constantly exposed to high levels of $PM_{2.5}$). These two definitions allow to decompose the ATE into the sum of a direct and an indirect effect: ATE = PNDE + TNIE.

Total natural direct and pure natural indirect effect

Alternatively, it is possible to define a *Total Natural Direct Effect* (TNDE) where the values of the mediator which are kept constant are the counterfactual values M_1 had A been set to $A = 1$ (instead of $A = 0$ as in the definition of PNDE described above): $TNDE = \mathbb{E}(Y_{1,M_1}) - \mathbb{E}(Y_{0,M_1})$. The *Pure Natural Indirect effect* (PNIE) can then be defined as: $PNIE = ATE - TNDE = \mathbb{E}(Y_{0,M_1}) - \mathbb{E}(Y_{0,M_0})$. The difference between TNDE/PNIE and PNDE/TNIE definitions of direct and indirect effects is that in the presence of an $(A * M)$ interaction affecting Y, the *mediated interactive effect* (see definition in 3-way and 4-way decompositions) appears in the "total" component of the direct or indirect effect.[67] Note that if there is no $(A * M)$ interaction affecting Y, the CDE, the TNDE and the PNDE will have the same value.

Under the identification assumptions (consistency, positivity), the sequential randomisation assumptions (A1) and (A2) described above, and the following Independence assumptions:

- (A3) No unmeasured confounding between A and M, given $L(0)$,
- (A4) A does not affect confounders $L(1)$ of the mediator-outcome relationship (corresponding to an Independence assumption between the counterfactuals $Y_{a,m}$ and M_{a^*})

(cf, table S1 in the supplementary material), the Natural Direct and Indirect Effects are identifiable and can be expressed by g-formulas indicated in table S2 (Supplementary material).

This means that if the set of confounders $L(0)$ and $L(1)$ is sufficient, Natural Direct and Indirect Effects are identifiable in causal models corresponding to Figure 2(a), but are not identifiable in causal models such as represented in Figure 2(b).

Natural direct and indirect effects may provide valuable insights into mediation mechanisms; however, they are defined through unobservable cross-world counterfactuals, rendering their intuitive interpretation challenging (Y_{1,M_0} refers to a hypothetical world where individuals are exposed to high levels of $PM_{2.5}$ and at the same time, to the occurrence of type 2 diabetes expected had they been exposed to low levels of $PM_{2.5}$, which combination cannot be observed in reality). While the causal quantity $\mathbb{E}(Y_{a,M_{a'}})$ is identifiable under the identification assumptions in Figure 2(a), it is not falsifiable by experimentation.

Interventional (or randomised) natural direct and indirect effect

Using the notion of stochastic counterfactual interventions, two other types of natural direct and indirect effects have been defined, identifiable despite the possible presence of intermediate confounders affected by the exposure (as in Figure 2(b)) and not requiring cross-world Independence assumptions (A4). These effects are referred to as "interventional" (or "randomised") direct and indirect effects. They are defined under hypothetical interventions on the mediator implying a random draw in the counterfactual distribution of the mediator M had the exposure been set to a given level, instead of setting the value of M to the individual potential values.[68,69] Moreover, because they involve distributions rather than unknown individual values, they can be considered more policy relevant.[70]

Marginal interventional natural direct and indirect effects

VanderWeele defined the marginal randomised (or interventional) natural effects.[71-76] The *marginal randomised natural direct effect* (MRDE) is the effect on Y that would be observed under the hypothetical intervention of changing the value of A from 0 to 1 (ie, from low to high levels of $PM_{2.5}$), while the mediator is set to a random draw for each subject from the (same) distribution of M_0 (the counterfactual distribution of type 2 diabetes had the exposure been set to low levels of $PM_{2.5}$), conditional on $L(0)$. Such counterfactual distribution of the mediator is denoted $G_0 \mid L(0)$:

$$MRDE = \mathbb{E}(Y_{1,G_0|L(0)}) - \mathbb{E}(Y_{0,G_0|L(0)})$$

The *marginal randomised natural indirect effect* (MRIE) is the effect on Y that would be observed under the hypothetical intervention of setting the value of the exposure to $A = 1$ (high levels of $PM_{2.5}$), while shifting the values of M (type 2 diabetes) from a random draw for each subject from the counterfactual distribution of the mediator (conditional on $L(0)$) had the exposure been

set to $A=0$ ($M \sim G_{0|L(0)}$), to a random draw from the counterfactual distribution of the mediator had the exposure been set to $A=1$ ($M \sim G_{1|L(0)}$):

$$\text{MRIE} = \mathbb{E}(Y_{1,G_{1|L(0)}}) - \mathbb{E}(Y_{1,G_{0|L(0)}}).$$

In case of intermediate confounder $L(1)$ of the $M-Y$ relationship affected by the exposure, as in Figure 2(b), the Marginal Randomised Indirect Effect (MRIE) corresponds to all the directed paths from A to Y going through the mediator M: $A \rightarrow M \rightarrow Y$ and $A \rightarrow L(1) \rightarrow M \rightarrow Y$; and the Marginal Randomised Direct Effect (MRDE) corresponds to all the directed paths from A to Y which do not go through the mediator M: $A \rightarrow Y$ and $A \rightarrow L(1) \rightarrow Y$.

Under the identification assumptions (consistency, positivity and the randomisation assumptions (A1), (A2) and (A3) described above and in supplementary table S1), the MRDE and MRIE are identified by g-formulas described in the supplementary table S3.

The sum of the MRDE and MRIE gives an *Overall Effect*: $\text{OE} = \mathbb{E}(Y_{a,G_{a|L(0)}}) - \mathbb{E}(Y_{a^*,G_{a^*|L(0)}})$. The Overall Effect can be interpreted as a total effect, however because it is defined using random draws from counterfactual distributions of the mediator (conditional on $L(0)$) rather than individual counterfactual values, the Overall Effect may differ from the Average Total Effect, especially in case of non-linear models and $L(0) * L(1)$ interaction effects affecting the mediator M.[70]

Marginal Randomized Direct and Indirect Effects were initially suggested as analogues of the PNDE and TNIE that could be used when intermediate confouders $L(1)$ are affected by the exposure (Figure 2(b)).[71] Indeed, if the causal model corresponds to Figure 2(a), the identifying g-formulas of MRDE and MRIE reduce to the g-formulas of the PNDE and TNIE (note that the definitions of MRDE and MRIE can easily be adapted to get analogues of TNDE and PNIE).[66] However, it has been shown that like the EE, the MRIE does not capture a true mediational effect: the MRIE does not satisfy the "sharp mediational null criteria" (even if the effect of A is not mechanistically mediated by M for each individual, the MRIE could still be non-null).[66] This can be verified, for example, in the presence of an $(A * L(1))$ interaction effect on M and a $(L(1) * M)$ interaction effect on Y.

Although MRDE and MRIE cannot be interpreted as "true mediational" direct and indirect effects, the can still be interpreted as contrasts between hypothetical scenarios implying stochastic interventions (and not relying on cross-world assumptions).[66] Another possible interpretation is that of natural direct effects given by Petersen et al.[77] as a weighted average of the controlled direct effects. For the causal model of Figure 2(a):

$$PNDE = \mathbb{E}_{L(0),L(1)} \sum_m \left[\mathbb{E}\left(Y_{1,m}|L(0),L(1)\right) - \mathbb{E}\left(Y_{0,m}|L(0),L(1)\right)\right] \times \mathbb{P}\left(M_0 = m|L(0),L(1)\right)$$

Similarly for the causal model of figure 2(b), the MRDE can be interpreted as an average of the CDE_m, weighted by $\mathbb{P}(M_0 = m \mid L(0))$.

Another limit described for MRDE and MRIE is that the counterfactual variables $Y_{a,G_{a'|L(0)}}$ might not be well-defined in survival settings where time-to-event outcomes can occur before the mediator: a participant still alive under $A=a$ would be allowed to draw the mediator value of a participant who has died under $A=a'$.[78]

Conditional interventional (or randomised) natural direct and indirect effects

Instead of hypothetical scenarios defined with random draws from the distribution of $M_{a'}$ conditional on $L(0)$, random draws can be defined conditional on both $L(0)$ and $L(1)_{a'}$.[16,78-80] The *conditional randomised natural direct effect* (CRDE) is defined as:

$$\text{CRDE} = \mathbb{E}(Y_{1,\Gamma_0|L(0),L(1)}) - \mathbb{E}(Y_{0,\Gamma_0|L(0),L(1)})$$

The CRIE corresponds the effect on Y that would be observed under the hypothetical intervention of changing the value of A from 0 to 1 (low to high levels of $PM_{2.5}$), while the mediator is set to a random draw for each subject from the distribution $\Gamma_{0|L(0),L(1)}$: the distribution of M_0 (the counterfactual distribution of type 2 diabetes had the population been exposed to low levels of $PM_{2.5}$), fully conditional on the past ($L(0)$ and $L(1)_{A=0}$).

The *conditional randomised natural indirect effect* (CRIE) is defined as:

$$\text{CRIE} = \mathbb{E}(Y_{1,\Gamma_1|L(0),L(1)}) - \mathbb{E}(Y_{1,\Gamma_0|L(0),L(1)})$$

The CRIE is the effect on Y that would be observed under the hypothetical intervention of setting the value of the exposure to $A=1$ (high levels of $PM_{2.5}$), while shifting the values of M from a random draw for each subject from $M \sim \Gamma_{0|L(0),L(1)}$ to $M \sim \Gamma_{1|L(0),L(1)}$: from the counterfactual distribution of type 2 diabetes (M), conditional on $L(0)$ and $L(1)$, had the population been exposed to low levels of $PM_{2.5}$, to its counterfactual distribution had the population been exposed to high levels of $PM_{2.5}$.

Under the consistency assumption, the positivity assumption and the randomisation assumptions (A1), (A2), (A3) and (A5) described in supplementary table S1, where

- (A5) No unmeasured confounding between A and $L(1)$, given $L(0)$,

The CRDE and CRIE are identified by g-formulas described in the supplementary table S3.

Identifiability assumptions for Conditional randomised natural (in)direct effects are stronger than for Marginal randomised natural (in)direct effects, but hold in both Figures 2(a) and 2(b). Like MRDE and MRIE, in causal structures such as described in Figure 2(a), the interpretation of CRDE and CRIE is analogous to the interpretation of the Pure Natural Direct Effect (PNDE) and the Total Natural Indirect Effect (TNIE), respectively (they are identified by the same g-formulas and give the same values). Note that the definitions of CRDE and CRIE can easily be adapted to get analogues of TNDE and PNIE. In case of intermediate confounder affected by the exposure, as in Figure 2(b), the CRIE can be interpreted as the path-specific effect of A to Y which goes only through the mediator M: $A \rightarrow M \rightarrow Y$. The Conditional Randomised Direct Effect (CRDE) corresponds to all the directed paths from A to Y, except the path going only through M: $A \rightarrow Y$, $A \rightarrow L(1) \rightarrow Y$, and $A \rightarrow L(1) \rightarrow M \rightarrow Y$. Because the CRDE includes one of the paths which goes through the mediator ($A \rightarrow L(1) \rightarrow M \rightarrow Y$), its interpretation might be less intuitive. In our example, assuming the correct causal model is depicted by Figure 2(b) and that the set $L(1)$ contains only overweight, the CRIE would be the effect of $PM_{2.5}$ mediated by type 2 diabetes, but excluding the mechanism of type 2 diabetes caused by overweight.

Interestingly, unlike marginal randomised (in)direct effects, the sum of the conditional randomised natural direct and

indirect effects is equal to the usual Average Total Effect (ATE). Moreover, the quantities are well-defined in survival settings.[78]

Three-way and four-way decomposition

In the presence of an interaction $A*M$ between the exposure and the mediator affecting the outcome Y, the effect of changing the exposure from $A=0$ to $A=1$ will depend on the value $M=m$ of the mediator. Consequently, the value of the CDE_m will depend on the value fixed for the mediator $M=m$, and the PNDE/TNIE will be different from the TNDE/PNIE. VanderWeele[15,81] defined several causal quantities to separate interaction effects from the direct and indirect effects, applying a 3-way or a 4-way decomposition.

Three-way decomposition

VanderWeele suggested a decomposition of the Average total effect into:[81]

- A *Pure Direct Effect*, equivalent to the PNDE described above. $PNDE=\mathbb{E}(Y_{1,M_0})-\mathbb{E}(Y_{0,M_0})$
- A *Pure Indirect Effect*, equivalent to the PNIE described above. $=\mathbb{E}(Y_{0,M_1})-\mathbb{E}(Y_{0,M_0})$
- And a *Mediated Interactive Effect* (MIE)

  $MIE=\mathbb{E}[(Y_{1,1}-Y_{1,0}-Y_{0,1}+Y_{0,0})\times(M_1-M_0)]$. The MIE is equal to the difference between the TNDE and the PNDE, as well as the difference between the TNIE and the PNIE. The MIE corresponds to an additive interaction which operates only if the exposure A has an effect on the mediator. The MIE is the average of the product between:

  - An additive interaction effect between the exposure A and the mediator M on the outcome Y, corresponding to the difference between the effect of a hypothetical joint modification of A and M from $(A=0,M=0)$ to $(A=1,M=1)$, contrasted with the sum of two individual changes in either A or M, while the other variable is set constant to the reference level $M=0$ or $A=0$:

$$[Y_{1,1}-Y_{0,0}]-[(Y_{1,0}-Y_{0,0})+(Y_{0,1}-Y_{0,0})] = (Y_{1,1}-Y_{1,0}-Y_{0,1}+Y_{0,0})$$

  - And the effect of the exposure A on the mediator M (denoting M_a the counterfactual value of M had the exposure been set to $A=a$): (M_1-M_0)

This decomposition enables us to isolate a mediated interactive effect (due to the $(A*M)$ interaction effect on Y) from the direct and indirect effects.

Four-way decomposition

VanderWeele further developped a 4-way decomposition of the Average total effect (ATE), into:[15]

- A "Controlled Direct Effect" (CDE_0) of A on Y, setting the level of the mediator to the reference value $M=0$: $CDE_0=\mathbb{E}(Y_{1,0})-\mathbb{E}(Y_{0,0})$. The CDE_0 corresponds to the standard CDE with the value of M fixed to 0, i.e. the effect of the exposure in the absence of the mediator (effect due neither to mediation nor to $(A*M)$ interaction);
- A "Reference Interaction Effect" (RIE) $RIE=\mathbb{E}[(Y_{1,1}-Y_{1,0}-Y_{0,1}+Y_{0,0})\times M_0]$. The RIE corresponds to the $(A*M)$ additive interaction effect on the outcome Y which operates only if the counterfactual mediator is present had the subject been unexposed to A (when $M_0=1$). This effect is due to the $(A*M)$ interaction only.
- A "mediated interaction" (MIE), similar to the MIE of the 3-way decomposition $MIE=\mathbb{E}[(Y_{1,1}-Y_{1,0}-Y_{0,1}+Y_{0,0})\times(M_1-M_0)]$. The MIE corresponds to the effect of A on Y due to both the mediation through the mediator and the interaction with the mediator.
- And a pure indirect effect equivalent to the PNIE previously described. $PNIE=\mathbb{E}(Y_{0,M_1})-\mathbb{E}(Y_{0,M_0})$. The PNIE corresponds to the effect of A on Y due to mediation only.

According to VanderWeele,[15] at least one of these 4 components must be non-null if the exposure A affects the outcome Y at the individual level.

Regarding the relationships between the 2-way, the 3-way and the 4-way decomposition, we have:[15]

$$\begin{aligned} ATE &= \text{TNDE}+\text{PNIE} \\ &\quad \text{where TNDE}=\text{PNDE}+\text{MIE} \\ ATE &= [\text{PNDE}+\text{MIE}]+\text{PNIE} \\ &\quad \text{and PNDE}=\text{CDE}_0+\text{RIE} \\ ATE &= [(\text{CDE}_0+\text{RIE})+\text{MIE}]+\text{PNIE} \end{aligned}$$

In the original articles describing the 3-way and the 4-way decompositions, the identification assumptions are the same as for the natural Direct and Indirect Effects (Supplementary table S1), so that the 3-way and 4-way decompositions were not identifiable with causal structure such as Figure 2(b).[15,81]

However, in causal structures such as Figure 2(b), analogues of 3-way and 4-way decompositions can be obtained (as in the CMAverse R package),[82] from the MRIE, the "Pure" Marginal Randomised Indirect Effect (PMRIE), the MRDE and the CDE_0:

$$\begin{aligned} MIE &= \text{MRIE}-\text{PMRIE} \\ &\quad \text{where PMRIE}=\mathbb{E}(Y_{0,G_1|L(0)})-\mathbb{E}(Y_{0,G_0|L(0)}) \\ RIE &= \text{MRDE}-\text{CDE}_0 \end{aligned}$$

Note that if there is no $(A*M)$ interaction effect on Y (a strong parametric assumption), then the MIE and RIE are null and $CDE_m=PNDE=TNDE$ and $EE_m=TNIE=PNIE$ for causal models corresponding to Figure 2(a). Similarly for causal models corresponding to Figure 2(b), $CDE_m=MRDE$ and $EE_m=MRIE$.

How to choose the relevant causal estimands?

Choosing among all the possible estimands can be guided by the scientific question and the identifiability assumptions:

1. The scientific question:

   If the aim is to study the potential benefit of intervening on the mediator to mitigate the total effect, the controlled direct effect (CDE_m) and the eliminated effect (EE_m) are the most relevant estimands.

   If the objective is to understand the mechanisms explaining the total effect, it is better to focus on natural direct and indirect effects (for example PNDE and TNIE) or their "interventional" analogues (MRDE and MRIE). The 3-way or 4-way decomposition also makes it possible to distinguish the effects of mediation from the effects of interaction between exposure and mediators if interaction issues are part of the scientific inquiry.
2. The identifiability assumptions:

   The randomization assumptions are easier to hold for controlled direct effects than for natural direct/indirect effect and

their interventional analogues. In case of intermediate confounder $L(1)$ affected by the exposure, natural direct and indirect effects (PNDE and TNIE) are no longer identifiable, which require a shift toward their interventional analogues (MRDE and MRIE).

Estimators

Several estimators of the causal quantities of interest have been developed. In this section, we present a summary of these estimators.

Traditional regression models

Using traditional regression models have been described for two-way decomposition (CDE and natural effects),[38,67,82] three-way decomposition[81] and four-way decomposition.[15] The approach is similar to the "product method" or the "difference method" previously described. When using traditional regression models, we have to assume that the models are correctly specified. If necessary, these models can accommodate $(A*M)$ interactions affecting the outcome Y.

These approaches can be applied in the absence of intermediate confounders affected by the exposure A (as in Figure 2(a)). With causal models corresponding to the Figure 2(b), Natural Effects are not identifiable and the use of traditional regression models adjusted for the mediator results in biased estimations of direct or indirect effects due to a collider stratification bias.[13,42] This bias can be large in case of strong effects of $A(0)$ on M combined to strong effects of $L(1)$ on M.[83] If intermediate confounders $L(1)$ are affected by the exposure A (as the DAG in Figure 2(b)), other estimators are required: G-computation, IPTW or TMLE, described below.

G-computation

A simple example of estimation by g-computation is given below for the estimation of the Average Total Effect (ATE). G-computation can be described as a "simple substitution estimator," based on the g-formula defining Ψ^{ATE} (Supplementary table S2):.[84,85]

1. Firstly, fit a regression of Y on the exposure A and baseline confounders $L(0)$ (using a logistic regression for binary outcomes for example): $\bar{Q}(A,L(0)) = \mathbb{P}(Y=1 \mid A,L(0))$.
2. Secondly, for each individual, predict the expected values of $\mathbb{E}(Y_a \mid L(0))$ using this model, had the whole population been exposed to $A=1$ (denoted $\widehat{\bar{Q}}(A=1,L(0))$, and had the whole population been exposed to $A=0$ ($\widehat{\bar{Q}}(A=0,L(0))$.
3. The predicted values are then plugged in the g-formula (for a sample of size n).

$$\widehat{\Psi}\text{gcompATE} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\bar{Q}}(A=1,L(0)) - \widehat{\bar{Q}}(A=0,L(0))\right]$$

4. Confidence intervals can be obtained using bootstrapping.

Parametric G-computation

G-formula estimands can be obtained using Monte Carlo simulations of the $L(1)_{a'}$, $M_{a'}$, $Y_{a',m}$, $Y_{a,M_{a'}}$, $Y_{a,G_{a'|L(0)}}$ or $Y_{a,\Gamma_{a'|L(0),L(1)}}$ variables under the counterfactual scenarios considered to define the causal quantities of interest.[86,87]

Estimations using parametric g-computation have been described to estimate Controlled Direct Effects;[88] Natural Direct and Indirect Effects,[88] where an additional step to estimate the density function of the mediator is necessary, in order to simulate individual values of the mediator $M_{a'}$ under the counterfactual scenario setting $A=a'$; and Marginal Randomised Direct and Indirect Effects,[89] where an additional step to estimate the density function of the mediator is also necessary, in order to simulate and randomly permute individual values of the mediator $M_{a'}$ under the counterfactual scenario setting $A=a'$. The approach described for MRDE and MRIE can be adapted to estimate Conditional Randomised Direct and Indirect Effects.

G-computation by iterative conditional expectation

A limit of parametric G-computation is the difficulty to estimate density functions of $L(1)$ variables. Moreover, it is necessary to fit a model for each variable in the set $L(1)$. An alternative approach is g-computation by iterative conditional expectation (ICE), which have been described to analyse counterfactual scenarios relevant for Controlled direct effects, Marginal or Randomised natural direct and indirect effects.[78,90,91] This approach relies on a smaller number of models to fit (especially if several variables are included in the set $L(1)$). As an example for Controlled direct effects, the estimand can be reformulated by iterative conditional expectation :[91]

$$\begin{aligned}\Psi^{\text{CDE}} = {} & \mathbb{E}(\mathbb{E}_{L(1)}[\mathbb{E}_Y(Y \mid L(1),L(0),A,M=m) \mid L(0),A=1]) \\ & - \mathbb{E}(\mathbb{E}_{L(1)}[\mathbb{E}_Y(Y \mid L(1),L(0),A,M=m) \mid L(0),A=0])\end{aligned}$$

Statistical properties of g-computation estimators

Estimations of direct and indirect effects by g-computation are expected to be unbiased if the models fitted during the procedures ($\bar{Q}$ regressions) are correctly specified.[84,85] Estimates are unaffected by deviation from the positivity assumption, so that the procedure is able to extrapolate beyond the observed data. Considering the positivity assumption is all the more important to avoid conclusions that are only weakly supported by the available data.[84] Moreover, G-computation is not an asymptotically linear estimator, so that its efficiency properties are not optimal.[85]

Marginal structural models (MSM)

Marginal structural models are models of the expected value of a counterfactual outcome under study. They are used to summarise the causal relationship between the expectation of the counterfactual outcome and the exposures of interest.[92,93] In the context of mediation analyses, exposures of interest are the initial exposure A and the mediator M.

MSM for controlled direct effects

The following MSM can be considered to estimate Controlled direct effects:[94,95] $\mathbb{E}(Y_{a,m}) = \alpha_0 + \alpha_A a + \alpha_M m$. If we suspect the presence of $A*M$ interaction affecting the outcome, it is possible to add an interaction term:

$\mathbb{E}(Y_{a,m}) = \alpha_0 + \alpha_A a + \alpha_M m + \alpha_{AM}(a \times m)$. The controlled direct effects CDE_m can then be expressed using the coefficients of the MSM:

$$\begin{aligned}\Psi^{\text{CDE}_m} &= \mathbb{E}(Y_{a,m}) - \mathbb{E}(Y_{a^*,m}) \\ \Psi^{\text{CDE}_m}_{\text{MSM}} &= (\alpha_0 + \alpha_A a + \alpha_M m + \alpha_{AM}(a \times m)) \\ &\quad - (\alpha_0 + \alpha_A a^* + \alpha_M m + \alpha_{AM}(a^* \times m)) \\ \Psi^{\text{CDE}_m}_{\text{MSM}} &= \alpha_A(a-a^*) + \alpha_{AM}(a-a^*) \times m\end{aligned}$$

MSM for natural direct and indirect effects

For Pure Natural Direct Effects and Total Natural Indirect Effects, VanderWeele suggested using two MSMs:[94] a model of $Y_{a,m}$ and a

model of M_a, conditional on the baseline confounders $L(0)$, where h and k are the link functions chosen by the analyst.

$$\mathbb{E}(Y_{a,m} \mid l(0)) = h^{-1}(a, m, l(0))$$
$$\mathbb{E}(M_a \mid l(0)) = k^{-1}(a, l(0)).$$

If h^{-1} is linear in m (meaning that interactions between M and other variables are possible, but not polynomial functions of M or other transformations of M such as log(M), $\sqrt{M}$, etc,), and if A does not affect confounders $L(1)$ of the $M \rightarrow Y$ relationship, then $\mathbb{E}(Y_{a,M_{a^*}} \mid l(0)) = h^{-1}[a, k^{-1}[a^*, l(0)], l(0)]$.

Then the PNDE and TNIE can be reformulated using the MSM functions:

$$\begin{aligned}\Psi_{MSM}^{PNDE} &= \mathbb{E}(Y_{a,M_{a*}}) - \mathbb{E}(Y_{a*,M_{a*}}) \\ &= \sum_{l(0)} [h^{-1}\left(a, k^{-1}[a^*, l(0)], l(0)\right) \\ &\quad - h^{-1}\left(a^*, k^{-1}[a^*, l(0)], l(0)\right)] \times \mathbb{P}\left(L(0) = l(0)\right) \\ \Psi_{MSM}^{TNIE} &= \mathbb{E}(Y_{a,M_a}) - \mathbb{E}(Y_{a,M_{a*}}) \\ &= \sum_{l(0)} [h^{-1}\left(a, k^{-1}[a, l(0)], l(0)\right) \\ &\quad - h^{-1}\left(a, k^{-1}[a^*, l(0)], l(0)\right)] \times \mathbb{P}\left(L(0) = l(0)\right)\end{aligned}$$

Alternatively, Lange et al.[96] introduced another MSM enabling a "unified" approach for estimating natural direct and indirect effects.

MSM for marginal randomised natural direct and indirect effects

As for Natural direct and indirect effects, VanderWeele suggested to use two marginal structural models:[72]

$\mathbb{E}(Y_{a,m}) = h^{-1}(a, m)$ and $\mathbb{P}(M_a = m) = k^{-1}(a)$. Those two MSMs can then be combined in order to define the causal quantities necessary for Marginal Randomised direct and indirect effects:

$$\mathbb{E}(Y_{a,G_{a^*|L(0)}}) = \sum_m \mathbb{E}(Y_{a,m}) \times P(M_{a^*} = m).$$

Estimation of the MSM parameters

Because MSM are models of unobserved counterfactual variables, estimators of the MSM parameters are necessary. Several methods have been described to estimate MSM parameters, based on g-computation, Inverse Probability of Treatment Weighting or double robust methods.[72,84,92-94,97] Most often, MSM parameters are estimated using Inverse Probability of Treatment Weighting.

Inverse probability of treatment weighting (IPTW)

Intuitively, estimators based on Inverse probability of treatment weighting (IPTW) operates by assigning a weight to each individual so that baseline and intermediate confounders are balanced relative to the exposure A and the mediator M in the new pseudo-population, so that there is no confounding between A (or M) and $Y_{a,m}$.[98]

Estimating ATE and CDE by IPTW

For the *Average Total Effect*, it is possible to estimate $\mathbb{E}(Y_{a'})$ applying the following weight to the outcome of each individual (where $g(A_i \mid L(0)_i)$ is the probability of receiving his observed exposure A_i, given $L(0)_i$): $w_{ATE} = \frac{I(A_i = a')}{g(A_i \mid L(0)_i)}$ so that the ATE can be estimated by the following Horvitz and Thompson estimator:[98,99]

$$\begin{aligned}\widehat{\Psi}_{IPTW}^{ATE} &= \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{\hat{g}(A_i = a \mid L(0)_i)} Y_i \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a^*)}{\hat{g}(A_i = a^* \mid L(0)_i)} Y_i\end{aligned}$$

In case of positivity violation (if $g(A_i \mid L(0)_i) = 0$ in some strata $L(0)$), weights cannot be computed. Near positivity violation (if $g(A_i \mid L(0)_i) \approx 0$ in some strata $L(0)$) will result in extreme weights, increasing the variance of the IPTW estimator. In order to reduce variability resulting from near positivity violation, common approaches are: to truncate the weights (for example at the 1st and 99th percentiles, or applying a data-adaptive selection of the truncation level); or to trim the weights (drop units with propensity scores outside a given interval), but this will also result in a biased IPTW estimator.[83,100-105] Alternatively, a "stabilised" IPTW estimator can be applied using a modified Horvitz-Thomson estimator:[98]

$$\widehat{\Psi}_{sIPTW}^{ATE} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a) g^*(A_i = a)}{\hat{g}(A_i = a \mid L(0)_i)} Y_i}{\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a) g^*(A_i = a)}{\hat{g}(A_i = a \mid L(0)_i)}}$$

$$- \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a^*) g^*(A_i = a^*)}{\hat{g}(A_i = a^* \mid L(0)_i)} Y_i}{\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a^*) g^*(A_i = a^*)}{\hat{g}(A_i = a^* \mid L(0)_i)}}$$

where $g^*(A_i = a)$ is a non-null function of A (for example, $g^*(A = a) = P(A = a)$). Using stabilised IPTW estimator enables to get a bounded estimator and a weaker positivity assumption (the denominator can be zero if the numerator is zero).[100]

Similarly, an Horvitz-Thomson IPTW estimator and a "stabilised" IPTW estimator can be used to estimate *Controlled direct effects*. They will depend on propensity scores (treatment mechanisms) for the exposure A and for the mediator M, conditional on the past: $\hat{g}(A_i = a \mid L(0)_i)$ and $\hat{g}(M_i = m \mid L(1)_i, A_i, L(0)_i)$

For conditional randomised direct and indirect effects, Zheng[78] described the following IPTW estimator:

$$\begin{aligned}\widehat{\mathbb{E}}\left(Y_{a,\Gamma_{a'}|L(0),L(1)}\right) &= \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a) \times Y_i}{\hat{g}(A_i = a_i \mid L(0)_i)} \\ &\quad \times \frac{\hat{g}(M_i = m_i \mid a', L(1)_i, L(0)_i)}{\hat{g}(M_i = m_i \mid a, L(1)_i, L(0)_i)}\end{aligned}$$

Estimation of MSM parameters

In order to estimate the parameters of the various Marginal Structural Models described previously, IPTW methods are frequently applied. The general approach is to fit weighted generalised linear models corresponding to the MSM models, where the weights are defined using the propensity scores of the exposure A and the mediator M ($\hat{g}(A_i = a \mid L(0)_i)$ and $\hat{g}(M_i = m \mid L(1)_i, A_i, L(0)_i)$). A recommended approach is to use "stabilised" weights for weaker positivity assumptions. This approach has been described to estimate the parameters of the MSMs for controlled direct effects,[94] for natural direct and indirect effects,[94] for marginal randomised direct and indirect effects,[72,97] and for the unified MSMs of Lange et al.[96]

Statistical properties of IPTW estimators

IPTW estimators are expected to be unbiased if the models fitted to construct the weights ($g(A \mid L(0))$ and $g(M \mid L(0), A, L(1))$ are consistent.[84,85] As indicated earlier, IPTW estimators can be strongly affected by positivity violation, which is expected with data sparsity (with large sets of $L(0)$ and $L(1)$ confounders including continuous variables, or if the exposures A and M are high-dimensional variables). Positivity violation will result in IPTW estimators with increased variance. Using stabilised weights can partially mitigate this variability.[83,85,92,100] Other approaches to reduce variability of IPTW estimators are to truncate the weights or to trim the weights. However, weight truncation will also result in increased bias (estimators of $g(A \mid L(0))$ and $g(M \mid L(0), A, L(1))$ are no longer consistent after weight truncation).[83,100] Several authors suggested improved procedures to choose the truncation levels, using data-adaptive selection of optimal truncation levels.[101,104,105] Regarding weights trimming (dropping units with propensity scores outside a given interval), several estimators have been suggested to optimise the strategy and the standard error of the estimations.[103,106,107]

Doubly robust efficient methods

Double robust methods can be used to mitigate the influence of misspecification of models applied in g-computation or IPTW estimators. For example, Targeted Maximum Likelihood Estimation (TMLE) or Augmented Inverse Probability of Treatment Weighted (A-IPTW) have been described as doubly-robust estimators. They rely on both models of the outcomes ($\bar{Q}$ used in iterative g-computation) and propensity score models (g used in IPTW). If either $\bar{Q}$ or g models are consistently estimated, double robust methods will be consistent. Moreover, they are efficient if both $\bar{Q}$ and g models are consistently estimated: they can achieve the Cramer-Rao lower bound for the variance of unbiased estimators (the smallest asymptotic variance).[85,108,109] Both approach rely on the estimation of the efficient influence curve. Compared to TMLE, A-IPTW is described as less robust to positivity violation, and it might produce estimates outside of the statistical model space.[85] Data-adaptive (machine learning) algorithm can be applied in order to obtain consistent estimates of $\bar{Q}$ and g functions and optimise the statistical properties of the estimators. TMLE and A-IPTW procedures have been described, with statistical packages, for the estimation of Average treatment effects (ATE).[85,110-112] A TMLE procedure for repeated exposures has been developed and can be applied to estimate controlled direct effects (CDE).[91,113,114] TMLE and an alternative double robust "one-step" estimator have been described for marginal randomised direct and indirect effects,.[73,76,115,116] as well as for conditional randomised direct and indirect effects.[78,80]

More general longitudinal structures and time-to-event outcomes

In survival contexts and causal models without mediator-outcome confounders affected by the exposure, Lange and Hansen suggested using the Aalen additive hazard model to estimate Natural direct and indirect effects, under the usual "no unmeasured confounding" assumptions (A1), (A2),(A3) and (A4).[117]

In the same context, VanderWeele described alternative approaches using accelerated failure time models and proportional hazard models, which can be applied to estimate PNDE and TNIE on the mean survival time scale, and proportional hazard models, which can be applied to estimate natural direct and indirect effects on the log hazard ratio difference scale, provided the outcome is rare. In case of death-truncation of the mediator, Tai et al. redefined Natural direct and indirect effects.[118] Their proposal is akin to the conditional randomised effects (CRDE and CRIE) described earlier, focusing on the path-specific effect going only through the mediator M for the indirect effect.[78,79]

More general longitudinal structures implying repeated exposures $A(t)$ and mediators $M(t)$, with time-varying covariates $L(t)$ and $R(t)$ have been described (Figure S2 in supplementary document). In those complex causal models, it is possible to estimate:

- Controlled direct effects, which can be considered as effects of repeated exposures with time-varying covariates;[86]
- Marginal randomised direct and indirect effects.[71,72,89,116] However, as indicated earlier, MRDE and MRIE are not well defined in survival settings if participants can die before mediator occurrences;[78]
- Conditional randomised direct and indirect effects.[78,80] Considering that $Y(t)$ is a time dependant outcome included in the $L(t)$ set, CRDE and CRIE are well defined when conditioning on the participant's time-varying history. It's also possible to model informative censoring mechanisms, considering an indicator of remaining uncensored at time t in the $A(t)$ set of exposure variables. An indicator of being monitored at time t can also be added in the $A(t)$ set, in order to take into account missing values at time t for participants who missed a visit at time t but who are still alive and uncensored.

Dealing with high dimensionality

Implementing causal mediation analyses involves formulating a scientific question focusing on decomposing the effect of an exposure of interest on a outcome, through one or more mediators of interest. As described previously, the exposure and the mediators may be repeated over time, and it will be necessary to identify the baseline and intermediate confounders to be considered for the analysis. When analysing the human exposome, difficulties linked to the high dimensionality of the data can quickly arise, for a number of reasons: specification of the causal model, dealing with multiple mediators and dealing with high-dimensional variables.

The first difficulty is to specify a hypothesised causal model in which the variables can be unambiguously divided among the relevant sets of variables according to their causal sequence (baseline confounders, exposure(s) of interest, intermediate confounders, mediator(s) of interest and outcome).[17,119] A purely agnostic mediation analysis is not possible: using DAGitty, we can show that a dataset compatible with the causal model represented in Figure 2(b) would also be compatible with the same causal model where the $L(1)$ and M variables are switched, resulting in different direct and indirect effects.[51] This difficulty is inherent to causal analyses, which are positioned in a confirmatory rather than exploratory framework.

Multiple mediators

The number of ways of decomposing a total effect into a sum of direct and indirect effects increases exponentially with the number of mediators.[120] In causal structures without mediator-outcome confounders affected by the exposure (as in Figure 2(a)), VanderWeele et al.[97] described a situation with multiple mediators where $M = (M^{(1)}, M^{(2)}, \ldots, M^{(k)})$ is a vector including all the mediators of interest. In this context, they suggested using traditional regression approaches in order to estimate Natural direct

and indirect effects or controlled direct effets: (i) Assess all the mediators of interest as a single mediator M, defined as the entire vector of mediators. (ii) or assess mediators sequentially: first $M^{(1)}$, then $(M^{(1)},M^{(2)})$ jointly, then $(M^{(1)},M^{(2)},M^{(3)})$ jointly, etc, The first approach does not require knowing the ordering of the mediators. If ordering of the mediator is known, the second approach can correctly give more details on the portion of the total effect mediated through $M^{(1)}$, the portion of the total effect mediated through both $(M^{(1)},M^{(2)})$ and deduce the additional contribution of $M^{(2)}$ beyond $M^{(1)}$. This additional contribution would correspond to the additional effect mediated only through $M^{(2)}$, which is added to the possible paths going through both $M^{(1)}$ and $M^{(2)}$ (if $M^{(1)}$ affects $M^{(2)}$). The sequential analysis can then be continued to assess the effect mediated through $(M^{(1)},M^{(2)},M^{(3)})$, etc,

Steen et al.[121] extended this sequential method in a similar context (no mediator-outcome confounder affected by the exposure) and gave an example implying 2 mediators. They applied the "unified" MSM approach in order to obtain a 3-way decomposition with:[96] a Natural direct effect (not mediated by $M^{(1)}$, nor $M^{(2)}$), corresponding to the path $A \rightarrow Y$; a Natural indirect effect with respect to $M^{(1)}$, implying two paths: $A \rightarrow M^{(1)} \rightarrow Y$ and $A \rightarrow M^{(1)} \rightarrow M^{(2)} \rightarrow Y$; a partial indirect effect with respect to $M^{(2)}$, implying the remaining path $A \rightarrow M^{(2)} \rightarrow Y$. They described 6 possible decompositions, according to the way interaction terms are considered in definitions of direct and indirect effects.[121]

More generally, Marginal and Conditional Randomised Direct and Indirect effects can be applied when dealing with a set of intermediate variables (with known ordering) in which some are considered as mediators of interest and the others as time-varying confounders.[71,72,78] Tai and Lin developed a general approach for a number of ordered mediator and intermediate confounders, enabling to estimate "interventional path specific effects" through the mediators of interest.[122] In a causal structure implying 2 mediators $M^{(1)}$ and $M^{(2)}$, Vansteelandt and Daniel[70] applied the principles of interventional effects (drawing mediators in counterfactual distributions under $A = a$ or $A = a^*$) to decompose a total effect into: a direct effect of $A \rightarrow Y$; an indirect effect via the first mediator; an indirect effect via the second mediator, including the path via the first and second mediator if the first mediator affects the second; and an indirect effects corresponding to the effect of dependence between mediators on the outcome. Interestingly, this approach can be applied if the structural dependence between the mediators is unknown (regarding the direction of the causal effects from one mediator to the other, or the presence of an unmeasured common cause).[70] With numerous mediators of interest, Loh et al. defined slightly different Interventional Direct and Indirect effects (where mediators are separately set to random draws from their counterfactual distributions) enabling to estimate indirect effects for each mediator, without having to make assumptions about the causal sequence between mediators.[123]

High-dimensional variables

Even if the positivity assumption holds theoretically, some causal quantities can quickly have no support in a finite sample, including when the number of subjects is far greater than the number of variables.[91] Deviation from the positivity assumption will increase (i) with the dimensionality of the exposure or the mediators (causal models involving repeated exposures and mediators, and whenever they are multicategorical, continuous or a mixture of several exposures) and (ii) with the dimensionalilty of the baseline and intermediate confounders (especially if they are numerous, multicategorical or continuous).

When the exposure and the mediator are binary or categorical with few levels, classical dimension reduction methods can be applied to functions conditional on high-dimensional baseline or intermediate confounders (variable selection, regularisation, pruning, etc,). For example, TMLE estimators are usually coupled with data-adaptive algorithms that include various dimension reduction techniques.[124]

When the exposure or the mediator are continuous variables (because they are measured on a concentration scale, or after being defined as an exposure mixture by weighted quantile sum regression,[125] etc,), it can be judicious to use the concept of stochastic counterfactual interventions to define causal quantities of interest. As a general advice, Nguyen et al. recommend a flexible approach to define the relevant causal estimands for our mediation analyses, beyond the list of effects described in Table 1.[16]

Using of stochastic interventions allows us to simulate more realistic interventions and to weaken the positivity assumption. For example, Diàz and van der Laan described how to assess the potential effect of policies enforcing pollution levels below a certain cutoff point, by contrasting stochastic counterfactual distributions of a continuous exposure.[126] Kennedy proposed using "incremental propensity score intervention," a stochastic dynamic intervention which replaces the observational exposure process with a shifted version. This approach enables identification and estimation of causal effects without any positivity or parametric assumptions.[127] Within the framework of mediation analyses, Hejazi et al. showed how interventional direct and indirect effect can be defined using stochastic interventions applied to both the exposure and mediators, whether they are categorical or continuous.[128]

Methods which have been developed to deal with continuous exposures or continuous mediators can probably be generalized in order to deal with mixtures or multiple exposures, which are typical of exposome research.

Discussion

In summary

Due to the complex and multidimensional data that characterizes exposome research, the ability to interpret the results of complex statistical analyses in a causal manner is a key issue, particularly if we want to be able to identify levers for action and make public health recommendations.[129]

For the last twenty years, classical methods in mediation analyses (difference in coefficients, product of coefficients, path analyses and structural equation modelling) have been supplemented by concepts and methods from the causal inference literature: Non parametric causal models and graphical approaches (DAGs), to describe hypotheses on the causal structure of the data generating system; Counterfactual expressions and notations of causal quantities of interest, allowing more precise definitions of direct and indirect effects dealing with interactions and intermediate confounding affected by the initial exposure; Specification of assumptions needed to identify and estimate the causal quantities of interest (consistency, sequential randomisation assumption, positivity); And several estimators (g-computation, IPTW, double robust estimators) which can be implemented, with statistical properties varying from one family of estimators to another. Integrating these approaches may help exposome research move from the description of complex

exposure patterns toward a more explicit understanding of the mechanisms linking environmental conditions to health across the life course.

Addressing potential biases: measurement errors, unmeasured confounding, selection bias

Measurement errors and misclassification regarding the exposure, mediator, outcome or confounders can lead to bias in the estimation of the causal quantities of interest.

In mediation analyses, various methods have been described to take into account measurement errors on binary or continuous mediators,[130-134] on binary exposures,[135,136] or on the outcome.[137] Most of those methods and results were discussed in the causal framework of Figure 2(a) without mediator-outcome confounding affected by the exposure. Methods to correct bias or apply sensitivity analyses included regression calibration, EM algorithms, SIMEX, and method of moments.[133-138]

Estimation of causal quantities of interest in mediation analyses rely strongly on the sequential randomisation assumptions or other forms of "no unmeasured confounding." We can consider that Controlled Direct Effects are easier to identify (relying on 2 randomisation assumptions) than Natural Direct and Indirect Effects (relying on 4 randomisation assumptions), while Randomised Natural Direct and Indirect effects are between both (3 randomisation assumptions). These assumptions can be assessed by sensitivity analysis (or bias analysis). Sensitivity analyses aim to assess "the combination of bias parameters that could wholly explain the observed association if no effect truly existed."[139,140] In mediation analyses, several sensitivity analysis approaches have been developed, mostly to assess unmeasured confounding between the mediator and the outcome, for the estimation of CDE,[141] and Natural (in)direct effects.[142-147] Sensitivity analyses have been less developed in the context of multiple and time-varying mediators.[119,120,148]

In the case of several intermediate variables (mediators and intermediate confounders) between the exposure A and the outcome Y, specifying incorrectly the order of the variables might result in some bias: mis-specifying a parent $L(1)$ of the mediator as a child of the mediator would result in some residual confounding. However, within a set of intermediate confounders $L(t)$ or within a set of mediators of interest $M(t)$ considered at the same time t in the analysis, temporal ordering can be chosen arbitrarily without causing bias. Sensitivity analyses could help to check the consequences of including or not a variable in the set of intermediate confounders $L(1)$ between the mediator of interest and the outcome. As mentioned earlier, some methods have also been developed to deal with unknown sequence of mediators in order to estimate an indirect effect for each mediator, at the cost of a slightly different definition of the direct and indirect effects.[123]

It is also possible to define more general approaches to sensitivity analyses. For example, Rijnhart et al. described how to apply a "multiverse" analysis, where the robustness of the results are assessed regarding the arbitrary analytical decisions that are made in mediation analyses.[149] Applying another approach, Díaz and van der Laan suggested to integrate sensitivity parameters directly into the target quantities in order to assess violation of randomisation assumptions or bias due to measurement errors.[150] Their approach does not rely on additional models and can be implemented with asymptotically linear estimators such as TMLE.

Valeri and Coull discussed selection bias arising from missing data and its consequences on the estimation of direct and indirect effects. They suggest using nonparametric sensitivity analyses.[151] More generally, the usual approaches described for dealing with missing data can be applied for mediation analysis.[152]

Current and future prospects

The causal inference approaches are framed in confirmatory analyses where the structural hypotheses are assumed to be correct for the causal model. To our knowledge, we lack some tools and guidelines to conduct mediation analyses with more exploratory objectives. For example, beginning with a general scientific objective of exploring the intermediate mechanisms between an initial exposure and the outcome, using a large set of variables:

- How can we state relevant structural causal models (DAGs) combining theoretical knowledge and observed data? The recent development of causal discovery methods could enable us to make further progress in this area.[153-157]
- Dealing with a large set of intermediate variables, methods aiming at estimating indirect effects with arbitrary ordered mediators could help to explore and to identify the larger or most interesting indirect effects, to understand mechanisms or suggest possible interventions.[123,158]
- Recently Correia et al. introduced a flexible workflow in the context of ecological research, moving toward an exploratory causal discovery approach or toward a more confirmatory causal inference approach, depending on the extent of pre-existing knowledge. The authors point out that both approaches rely on untestable assumptions of causal sufficiency (no unmeasured confounding), causal Markov condition and faithfulness (required to infer the absence of causation from independence).[159]

Among the limitations of the mediation analysis methods presented in this review, we can mention in particular:

- Marginal Randomised Indirect Effects (MRIE) cannot be stricly interpreted as a "true mediational" indirect effects, as it does not satisfy the "sharp null criteria." Moreover, it is not well defined in survival settings where the mediator can be truncated by death.
- Conditional Randomised Indirect Effects (CRIE) can be defined in survival settings where the mediator can be truncated by death, however it does not capture a "full" indirect effect through the mediator of interest, as the path $A \rightarrow L(1) \rightarrow M \rightarrow Y$ is part of the direct rather than the indirect effect.

Two recent developments offer promising solutions to these difficulties:

Robins and Richardson described an "interventionist" approach to mediation analyses, that can be applied in survival settings, in which hypothetical scenarios are defined on a conceptual decomposition of the exposure A into two separable components: an A_Y component with a direct effect on the outcome $Y(t)$ and an A_M component with an indirect effect on the outcome through the mediator $M(t)$ (cf, Figure S3 in supplementary material).[160,161] In this decomposition, the counterfactual intervention applies only to the exposure A and no longer to the mediator. This conceptual decomposition allows us to obtain mediated effects that are well defined for survival analyses, the sum of the direct and indirect effects is equal to the average total effect (ATE), and the sharp null criteria is respected. Importantly,

their identification will rely on additional assumptions, such as isolation conditions and dismissible component conditions.[162] Until now, this concept has been used mainly to clarify mediation analyses in survival settings and to take into account the occurrence of competing or truncation events.[163-165]

Díaz proposed an intervention which alters the information propagated through the edges of the graph rather the usual counterfactual interventions used to define (in)direct effects in causal analysis. This approach allowed to define path-specific decompositions of the total effect, which are identified with intermediate confounding affected by the exposure and satisfies the "sharp null criteria."[166] Vo et al. recently combined this approach with the separable effects concepts developed by Robins and Richardson.[167]

Conclusion

Causal methods are relevant approaches for addressing some of the challenges in exposome research, particularly those related to the interpretability of observed associations and the consideration of causal structures within the data. As such, these methods could contribute to more interpretable, mechanism-driven research that is more relevant to help developing environmental health interventions and public policies. Recent perspectives and developments in causal methods should help researchers to conduct mediation analyses in more complex contexts typically encountered in the field of exposomics, thereby enabling a better understanding of the actual mechanisms at work.

Author contributions

Benoit Lepage(Conceptualization [Equal], Investigation [Equal], Methodology [Equal], Validation [Equal], Writing—original draft [Equal], Writing—review & editing [Equal]), Helene Colineaux (Conceptualization [Equal], Data curation [Equal], Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Writing—original draft [Equal], Writing—review & editing [Equal]), Valerie Gares(Conceptualization [Equal], Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Writing—original draft [Equal]), Barbara Bodinier(Conceptualization [Equal], Methodology [Equal], Writing—original draft [Equal]), and Cyrille Delpierre (Conceptualization [Equal], Investigation [Equal], Methodology [Equal], Writing—original draft [Equal], Writing—review & editing [Equal]), Marc Chadeau-Hyam(Conceptualization [Equal], Investigation [Equal], Methodology [Equal], Project administration [Equal], Resources [Equal], Writing—original draft [Equal], Writing—review & editing [Equal])

Supplementary material

Supplementary material is available at *Exposome* online.

Funding

This work was supported by the EXPANSE project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874627.

Conflict of interest

Marc Chadeau-Hyam holds shares of the O-SMOSE company. Consulting activities of the company are independent of the present work. All other authors declare no competing interests. Marc Chadeau-Hyam holds the position of Associate Editor for Exposome and has not peer reviewed or made any editorial decisions for this paper.

Data availability

No new data were generated or analyzed in support of this research.

References

1. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prevent*. 2005;14(8):1847–1850.
2. Vineis P, Robinson O, Chadeau-Hyam M, Dehghan A, Mudway I, Dagnino S, What is new in the exposome? *Environ Int*. 2020; 143:105887.
3. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: where chemistry meets biology. *Science*. 2020;367(6476):392–396.
4. Giroux É, Fayet Y, Serviant-Fine T. L'Exposome – Tensions Entre Holisme et Ŕeductionnisme. *Med Sci (Paris)*. 2021;37 (8–9):774–778.
5. Neufcourt L, Castagné R, Mabile L, Khalatbari-Soltani S, Delpierre C, Kelly-Irving M. Assessing how social exposures are integrated in exposome research: a scoping review. *Environ Health Perspect*. 2022;130(11):116001.
6. Vineis P. Invited perspective: the mysterious case of social determinants of health. *Environ Health Perspect*. 2022;130 (11):111303.
7. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS ONE*. 2010;5(5): e10746.
8. Krieger N. Embodiment: a conceptual glossary for epidemiology. *J Epidemiol Community Health*. 2005;59(5):350–355.
9. Glass TA, McAtee MJ. Behavioral science at the crossroads in public health: extending horizons, envisioning the future. *Soc Sci Med*. 2006;62(7):1650–1671.
10. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press; 2009.
11. Hernan M, Robins J. *Causal Inference: What If*. Chapman & Hall/CRC Press; 2025.
12. VanderWeele TJ. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press; 2015.
13. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143–155.
14. Pearl J. An introduction to causal inference. *Int J Biostat*. 2010;6 (2):Article 7.
15. VanderWeele TJ. A unification of mediation and interaction: a four-way decomposition. *Epidemiology*. 2014;25(5):749–761.
16. Nguyen TQ, Schmid I, Ogburn EL, Stuart EA. Clarifying causal mediation analysis: effect identification via three assumptions and five potential outcomes. *J Causal Inference*. 2022;10 (1):246–279.
17. Dang LE, Gruber S, Lee H, *et al.* A causal roadmap for generating high-quality real-world evidence. *J Clin Transl Sci*. 2023;7 (1):e212.
18. Münzel T, Sørensen M, Hahad O, Nieuwenhuijsen M, Daiber A. The contribution of the exposome to the burden of cardiovascular disease. *Nat Rev Cardiol*. 2023;20(10):651–669.

19. Beulens JWJ, Pinho MGM, Abreu TC, *et al*. Environmental risk factors of type 2 diabetes-an exposome approach. *Diabetologia* 2022;65(2):263–274.
20. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51 (6):1173–1182.
21. Valeri L. *Statistical Methods for Causal Mediation Analysis*. Ph.d. thesis. Havard University; 2013.
22. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods*. 2002;7(1):83–104.
23. Iacobucci D. *Mediation Analysis. Quantitative Applications in the Social Sciences*. SAGE Publications, Inc.; 2008.
24. MacKinnon DP. *Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates; 2008.
25. Wright S. Correlation and causation. *J Agric Res*. 1921;20 (7):557–585.
26. Wright S. Path coefficients and path regressions: alternative or complementary concepts? *Biometrics* 1960;16(2):189–202.
27. Tarka P. An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Qual Quant*. 2018;52(1):313–354.
28. Loehlin J, Beaujean A. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*. 5th ed. Routledge; 2017.
29. Bollen KA. *Structural Equation Models with Observed Variables*. John Wiley & Sons, Ltd; 1989.
30. Gunzler D, Morris N, Tu XM. Causal mediation analysis using structure equation models. In: He H, Wu P, Chen DGD, eds. *Statistical Causal Inferences and Their Applications in Public Health Research*. Springer International Publishing; 2016:5–314.
31. Kupek E. Log-linear transformation of binary variables: a suitable input for SEM. *Struct Equ Modeling*. 2005;12(1):28–40.
32. Kupek E. Beyond logistic regression: structural equations modelling for binary variables and its application to investigating unobserved confounders. *BMC Med Res Methodol*. 2006;6 (1):13.
33. Kenny DA, Judd CM. Estimating the nonlinear and interactive effects of latent variables. *Psychol Bull*. 1984;96(1):201–210.
34. Bollen KA, Paxton P. Interactions of latent variables in structural equation models. *Struct Equ Modeling*. 1998;5(3):267–293.
35. Jöreskog KG, Yang F. Nonlinear structural equation models: the Kenny-Judd model with interaction effects. In: Marcoulides G, Schumacker R, eds. *Advanced Structural Equation Modeling: Issues and Techniques*. 1st ed. Psychology Press; 1996:7–88.
36. Kaufman JS, Maclehose RF, Kaufman S. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiol Perspect Innov*. 2004;1(1):4.
37. Preacher KJ, Rucker DD, Hayes AF. Addressing moderated mediation hypotheses: theory, methods, and prescriptions. *Multivariate Behav Res*. 2007;42(1):185–227.
38. Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods*. 2013;18(2):137–150.
39. Hayes AF. An index and test of linear moderated mediation. *Multivariate Behav Res*. 2015;50(1):1–22.
40. Pearl J. Causality and structural models in social science and economics. In: Pearl J, ed. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press; 2009:5–194.
41. VanderWeele TJ. Invited commentary: structural equation models and epidemiologic analysis. *Am J Epidemiol*. 2012;176 (7):608–612.
42. Cole SR, Hernán MA. Fallibility in estimating direct effects. *Int J Epidemiol*. 2002;31(1):163–165.
43. Tchetgen Tchetgen E, Vanderweele T. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*. 2014;25(2):282–291.
44. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66 (5):688–701.
45. Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. 1st ed. Basic Books, Inc.; 2018.
46. D'Iaz I, Williams N, Hoffman KL, Schenck EJ. Nonparametric causal effects based on longitudinal modified treatment policies. *J Am Stat Assoc* 2023;118(542):846–857.
47. Tennant PWG, Murray EJ, Arnold KF, *et al*. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*. 2021;50 (2):620–632.
48. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*. 2010;21 (6):872–875.
49. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco*. Morgan Kaufmann; 1988.
50. Pearl J. Causal diagrams for empirical research. *Biometrika* 1995;82(4):669–688.
51. Textor J, Zander B, Gilthorpe M, Liskiewicz M, Ellison G. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology* 2016;45 (6):1887–1894.
52. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37–48.
53. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010;171(6):674– 63.
54. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012;21(1):31–54.
55. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc*. 2005;100(469):322–331.
56. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945–960.
57. Hernán MA. Does water kill? A call for less casual causal inferences. *Ann Epidemiol*. 2016;26(10):674–680.
58. Rehkopf DH, Glymour MM, Osypuk TL. The consistency assumption for causal inference in social epidemiology: when a rose is not a rose. *Curr Epidemiol Rep*. 2016;3(1):63–71.
59. Schwartz S, Prins SJ, Campbell UB, Gatto NM. Is the "well-defined intervention assumption" politically conservative? *Soc Sci Med*. 2016;166:254–257.
60. Park M, Joo HS, Lee K, *et al*. Differential toxicities of fine particulate matters from various sources. *Sci Rep*. 2018;8(1):17007.
61. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology*. 2014;25(3):418–426.
62. Goetghebeur E, Cessie S, De Stavola B, Moodie EE, Waernbaum I. Formulating causal questions and principled statistical answers. *Stat Med*. 2020;39(30):4922–4948.
63. Wang A, Arah OA. G-computation demonstration in causal mediation analysis. *Eur J Epidemiol*. 2015;30(10):1119–1127.
64. Pearl J. Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial*

IntelligenceUAI'01.Morgan Kaufmann Publishers Inc.; 2001:411–420.
65. VanderWeele TJ. Policy-relevant proportions for direct effects. *Epidemiology*. 2013;24(1):175–176.
66. Miles CH. On the causal interpretation of randomised interventional indirect effects. *J R Stat Soc Series B Stat Methodol*. 2023;85(4):1154–1172.
67. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*. 2009;2(4):457–468.
68. Didelez V, Dawid AP, Geneletti S. Direct and indirect effects of sequential treatments. In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence UAI'06*. AUAI Press; 2006:138–146.
69. Nguyen TQ, Schmid I, Stuart EA. Clarifying causal mediation analysis for the applied researcher: defining effects based on what we want to learn. *Psychol Methods*. 2021;26(2):255–271.
70. Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology*. 2017;28 (2):258–265.
71. Vanderweele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*. 2014;25(2):300–306.
72. VanderWeele TJ, Tchetgen Tchetgen EJ. Mediation analysis with time varying exposures and mediators. *J R Stat Soc Series B Stat Methodol*. 2017;79(3):917–938.
73. Rudolph KE, Sofrygin O, Zheng W, van der Laan MJ. Robust and flexible estimation of stochastic mediation effects: a proposed method and example in a randomized trial setting. *Epidemiol Methods*. 2018;7(1):20170007.
74. Rudolph KE, Sofrygin O, Schmidt NM, *et al.* Mediation of neighborhood effects on adolescent substance use by the school and peer environments. *Epidemiology*. 2018;29(4):590–598.
75. Rudolph KE, Goin DE, Paksarian D, Crowder R, Merikangas KR, Stuart EA. Causal mediation analysis with observational data: considerations and illustration examining mechanisms linking neighborhood poverty to adolescent substance use. *Am J Epidemiol*. 2019;188(3):598–608.
76. Díaz I, Hejazi NS, Rudolph KE, van der Laan MJ. Nonparametric efficient causal mediation with intermediate confounders. *Biometrika*. 2021;108(3):627–641.
77. Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology*. 2006;17(3):276–284.
78. Zheng W, Van Der Laan M. Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *J Causal Inference*. 2017;5(2):20160006.
79. Miles CH, Shpitser I, Kanki P, Meloni S, Tchetgen Tchetgen EJ. On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika* 2020; 107(1):159–172.
80. Wang Z, van der Laan L, Petersen M, Gerds T, Kvist K, van der Laan M. Targeted maximum likelihood based estimation for longitudinal mediation analysis. *J Causal Inference*. 2025;13 (1):20230013.
81. VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*. 2013; 24(2):224–232.
82. Shi B, Choirat C, Coull BA, VanderWeele TJ, Valeri L. CMAverse: a suite of functions for reproducible causal mediation analyses. *Epidemiology*. 2021;32(5):e20–e22.
83. Lepage B, Dedieu D, Savy N, Lang T. Estimating controlled direct effects in the presence of intermediate confounding of the mediator–outcome relationship: comparison of five different methods. *Stat Methods Med Res*. 2016;25(2):553–570.
84. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173(7):731–738.
85. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. 1st ed. Springer Series in Statistics. Springer; 2011.
86. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice GM, ed. *Longitudinal Data Analysis*. Chapman & Hall/CRC Press; 2009:3–599.
87. Daniel R, Cousens S, De Stavola B, Kenward MG, Sterne JAC. Methods for dealing with time-dependent confounding. *Stat Med*. 2013;32(9):1584–1618.
88. Daniel RM, De Stavola BL, Cousens SN. Gformula: estimating causal effects in the presence of time-varying confounding or mediation using the G-Computation formula. *Stata J*. 2011;11 (4):479–517.
89. Lin SH, Young J, Logan R, Tchetgen Tchetgen EJ, VanderWeele TJ. Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators and confounders. *Epidemiology*. 2017;28(2):266–274.
90. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005;61(4):962–973.
91. Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *J Causal Inference*. 2014;2(2):147–185.
92. Robins JM, HernáN MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11 (5):550–560.
93. Neugebauer R, van der Laan M. Nonparametric causal effects based on marginal structural models. *J Stat Plan Inference*. 2007; 137(2):419–434.
94. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*. 2009;20 (1):18–26.
95. Vansteelandt S. Estimating direct effects in cohort and case-control studies. *Epidemiology*. 2009;20(6):851–860.
96. Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. *Am J Epidemiol*. 2012;176(3):190–195.
97. VanderWeele T, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Methods*. 2014;2(1):95–115.
98. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60 (7):578–586.
99. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47 (260):663–685.
100. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168 (6):656–664.
101. Bembom O, van der Laan M. Data-adaptive selection of the adjustment set in variable importance estimation. *UC Berkeley Division of Biostatistics Working Paper Series* 2008; Paper 231.
102. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009;96(1):187–199.
103. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol*. 2010;172(7):843–854.

104. Xiao Y, Moodie EE, Abrahamowicz M. Comparison of approaches to weight truncation for marginal structural cox models. *Epidemiol Methods*. 2013;2(1):1–20.
105. Ju C, Schwab J, van der Laan MJ. On adaptive propensity score truncation in causal inference. *Stat Methods Med Res*. 2019;28 (6):1741–1760.
106. Yang S, Ding P. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*. 2018;105(2):487–493.
107. Garès V, Chauvet G, Hajage D. Variance estimators for weighted and stratified linear dose–response function estimators using generalized propensity score. *Biom J*. 2022;64 (1):33–56.
108. Porter KE, Gruber S, van der Laan MJ, Sekhon JS. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int J Biostat*. 2011;7(1):31.
109. Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: a tutorial. *Stat Med*. 2018;37(16):2530–2546.
110. van der Laan M, Rubin D. Targeted maximum likelihood learning. *The International Journal of Biostatistics*. 2006;2(1)
111. Gruber S, van der Laan M. tmle: An R package for targeted maximum likelihood estimation. *J Stat Soft*. 2012;51(13):1–35.
112. Zhong Y, Kennedy EH, Bodnar LM, Naimi AI. AIPW: an R package for augmented inverse probability–weighted estimation of average causal effects. *Am J Epidemiol*. 2021;190(12):2690–2699.
113. van der Laan MJ, Gruber S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int J Biostat*. 2012;8(1):Article 8.
114. Lendle SD, Schwab J, Petersen ML, van der Laan MJ. ltmle: An R package implementing targeted minimum loss-based estimation for longitudinal data. *J Stat Soft*. 2017;81(1):1–21.
115. Hejazi NS, Rudolph KE, Díaz I. medoutcon: Nonparametric efficient causal mediation analysis with machine learning in 'R'. *Joss*. 2022;7(69):3979.
116. Díaz I, Williams N, Rudolph KE. Efficient and flexible mediation analysis with time-varying mediators, treatments, and confounders. *J. Causal Inference*. 2023;11(1):20220077.
117. Lange T, Hansen JV. Direct and indirect effects in a survival context. *Epidemiology*. 2011;22(4):575–581.
118. Tai AS, Tsai CA, Lin SH. Survival mediation analysis with the death-truncated mediator: the completeness of the survival mediation parameter. *Stat Med*. 2021;40(17):3953–3974.
119. Schuler M, Coffman D, Stuart E, Nguyen T, Vegetabile B, Mccaffrey D. Practical challenges in mediation analysis: a guide for applied researchers. *Health Serv Outcomes Res Methodol*. 2025;25(1):57–84.
120. Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics* 2015;71 (1):1–14.
121. Steen J, Loeys T, Moerkerke B, Vansteelandt S. Flexible mediation analysis with multiple mediators. *Am J Epidemiol*. 2017;186 (2):184–193.
122. Tai A, Lin S. Complete effect decomposition for an arbitrary number of multiple ordered mediators with time-varying confounders: a method for generalized causal multi-mediation analysis. *Stat Methods Med Res*. 2023;32(1):100–117.
123. Loh WW, Moerkerke B, Loeys T, Vansteelandt S. Nonlinear mediation analysis with high-dimensional mediators whose causal structure is unknown. *Biometrics* 2022;78(1):46–59.
124. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1):Article25.
125. Keil AP, Buckley JP, O'Brien KM, Ferguson KK, Zhao S, White AJ. A quantile-based g-computation approach to addressing the effects of exposure mixtures. *Environ Health Perspect*. 2020;128 (4):47004.
126. Díaz I, van der Laan MJ. Assessing the causal effect of policies: an example using stochastic interventions. *Int J Biostat*. 2013;9 (2):161–174.
127. Kennedy EH. Nonparametric causal effects based on incremental propensity score interventions. *J Am Stat Assoc*. 2019; 114(526):645–656.
128. Hejazi NS, Rudolph KE, van der Laan MJ, D'Iaz I. Nonparametric causal mediation analysis for stochastic interventional (in)direct effects. *Biostatistics*. 2023;24(3):686–707.
129. Ponzano M, Rotem RS, Bellavia A. Complex methods for complex data: key considerations for interpretable and actionable results in exposome research. *Eur J Epidemiol*. 2025;40 (12):1399–1403.
130. Ogburn EL, VanderWeele TJ. analytic results on the bias due to nondifferential misclassification of a binary mediator. *Am J Epidemiol*. 2012;176(6):555–561.
131. VanderWeele TJ, Valeri L, Ogburn EL. The role of measurement error and misclassification in mediation analysis: mediation and measurement error. *Epidemiology*. 2012;23(4):561–564.
132. Blakely T, McKenzie S, Carter K. Misclassification of the mediator matters when estimating indirect effects. *J Epidemiol Community Health*. 2013;67(5):458–466.
133. Valeri L, Vanderweele TJ. The estimation of direct and indirect causal effects in the presence of misclassified binary mediator. *Biostatistics*. 2014;15(3):498–512.
134. Valeri L, Lin X, VanderWeele TJ. Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. *Stat Med*. 2014;33 (28):4875–4890.
135. Valeri L, Reese SL, Zhao S, *et al*. Misclassified exposure in epigenetic mediation analyses. Does DNA methylation mediate effects of smoking on birthweight? *Epigenomics* 2017;9 (3):253–265.
136. Jiang Z, VanderWeele T. causal mediation analysis in the presence of a misclassified binary exposure. *Epidemiol Methods*. 2019;8(1):20160006.
137. Jiang Z, VanderWeele TJ. Causal mediation analysis in the presence of a mismeasured outcome. *Epidemiology*. 2015;26 (1):e8–e9.
138. Le Cessie S, Debeij J, Rosendaal FR, Cannegieter SC, Vandenbroucke JP. Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology*. 2012;23(4):551–560.
139. Lash TL, Fink AK, Fox MP. Unmeasured and unknown confounders. In: Lash TL, Fox MP, Fink AK, eds. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Statistics for biology and health. Springer New York; 2009:9–78.
140. Imbens GW, Rubin DB. Sensitivity analysis and bounds. In: Imbens GW, Rubin DB, eds. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*.Cambridge University Press; 2015:496–510.
141. Vanderweele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*. 2010;21(4):540–551.
142. Tchetgen Tchetgen E, Shpitser I. Semiparametric Theory for Causal Mediation Analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Ann Stat* 2012;40(3):1816–1845.
143. VanderWeele T, Chiba Y. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-

outcome confounders. *Epidemiol Biostat Public Health*. 2014;11 (2):e9027.
144. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology*. 2016;27(3):368–377.
145. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the e-value. *Ann Intern Med*. 2017;167 (4):268–274.
146. Smith LH, VanderWeele TJ. Mediational e-values: approximate sensitivity analysis for unmeasured mediator-outcome confounding. *Epidemiology*. 2019;30(6):835–837.
147. McCandless LC, Somers JM. Bayesian sensitivity analysis for unmeasured confounding in causal mediation analysis. *Stat Methods Med Res*. 2019;28(2):515–531.
148. Wickramarachchi D, Lim L, Sun B. Mediation analysis with multiple mediators under unmeasured mediator-outcome confounding. *Stat Med*. 2023;42(4):422–432.
149. Rijnhart JJM, Twisk JWR, Deeg DJH, Heymans MW. Assessing the robustness of mediation analysis results using multiverse analysis. *Prev Sci*. 2022;23(5):821–831.
150. Díaz I, van der Laan MJ. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *Int J Biostat*. 2013;9(2):149–160.
151. Valeri L, Coull BA. Estimating causal contrasts involving intermediate variables in the presence of selection bias. *Stat Med*. 2016;35(26):4779–4793.
152. Carpenter JR, Smuk M. Missing data: a statistical framework for practice. *Biom J*. 2021;63(5):915–947.
153. Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. *Appl Inform (Berl)*. 2016;3:3.
154. Malinsky D, Danks D. Causal discovery algorithms: a practical guide. *Philos Compass*. 2018;13(1):e12470. https://doi.org/10.1111/phc3.12470
155. Vowels MJ, Camgoz NC, Bowden R. D'ya Like DAGs? A survey on structure learning and causal discovery. *ACM Comput Surv*. 2023;55(4):1–36.
156. Squires C, Uhler C. Causal structure learning: a combinatorial perspective. *Found Comut Math*. 2022;1:1–35.
157. Zanga A, Ozkirimli E, Stella F. A survey on causal discovery: theory and practice. *Int J Approxim Reason*. 2022;151:101–129.
158. Hou L, Yu Y, Sun X, *et al.* Causal mediation analysis with multiple causally non-ordered and ordered mediators based on summarized genetic data. *Stat Methods Med Res*. 2022;31 (7):1263–1279.
159. Correia HE, Dee LE, Byrnes JEK, *et al.* Best practices for moving from correlation to causation in ecological research. *Nat Commun*. 2026;17(1):1981.
160. Robins JM, Richardson TS. Alternative graphical causal models and the identification of direct effects. In: *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. Oxford University Press; 2011.
161. Robins JM, Richardson TS, Shpitser I. An interventionist approach to mediation analysis. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. Association for Computing Machinery; 2022:713–764.
162. Stensrud MJ, Hernán MA, Tchetgen Tchetgen EJ, Robins JM, Didelez V, Young JG. A generalized theory of separable effects in competing event settings. *Lifetime Data Anal*. 2021;27 (4):588–631.
163. Didelez V. Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Anal*. 2019;25 (4):593–610.
164. Stensrud MJ, Young JG, Didelez V, Robins JM, Hernán MA. Separable effects for causal inference in the presence of competing events. *J Am Stat Assoc*. 2022;117(537):175–183.
165. Stensrud MJ, Robins JM, Sarvet A, Tchetgen Tchetgen EJ, Young JG. Conditional separable effects. *J Am Stat Assoc*. 2023; 118(544):2671–2683.
166. Díaz I. Non-agency interventions for causal mediation in the presence of intermediate confounding. *J R Stat Soc Series B Stat Methodol*. 2024;86(2):435–460.
167. Vo TT, Williams N, Liu R, Rudolph KE, Dıaz I. Recanting twins: addressing intermediate confounding in mediation analysis. *Stat Med*. 2024;45(3–5):e70432.

Exposome, 2026, 6, 1–20
https://doi.org/10.1093/exposome/osag010
Research Article