Mediation Analysis for Probabilities of Causation

Yuta Kawakami, Jin Tian

Mohamed bin Zayed University of Artificial Intelligence, UAE
{Yuta.Kawakami, Jin.Tian}@mbzuai.ac.ae

Abstract

Probabilities of causation (PoC) offer valuable insights for informed decision-making. This paper introduces novel variants of PoC-controlled direct, natural direct, and natural indirect probability of necessity and sufficiency (PNS). These metrics quantify the necessity and sufficiency of a treatment for producing an outcome, accounting for different causal pathways. We develop identification theorems for these new PoC measures, allowing for their estimation from observational data. We demonstrate the practical application of our results through an analysis of a real-world psychology dataset.

Introduction

Pearl (1999) introduced three types of *probabilities of causation* (PoC), that is, the probability of necessity and sufficiency (PNS), the probability of necessity (PN), and the probability of sufficiency (PS). PoC quantify whether one event was the real cause of another in a given scenario (Robins and Greenland 1989; Tian and Pearl 2000; Pearl 2009; Kuroki and Cai 2011; Dawid, Murtas, and Musio 2014; Murtas, Dawid, and Musio 2017; Shingaki and Kuroki 2021; Kawakami, Shingaki, and Kuroki 2023). PoC are valuable for decision-making (Hannart and Naveau 2018; Li and Pearl 2019, 2022) and for explaining AI-based decision-making systems (Galhotra, Pradhan, and Salimi 2021; Watson et al. 2021).

Various variants of PoC have been studied, including for multi-valued discrete variables (Li and Pearl 2024a,b) and for continuous and vector variables (Kawakami, Kuroki, and Tian 2024). Rubinstein, Cuellar, and Malinsky (2024) introduced direct and indirect mediated PoC to decompose total PoC when there exists a mediator between the treatment and outcome.

Causal mediation analysis is a key method for uncovering the influence of different pathways between the treatment and outcome through mediators (Wright 1921, 1934; Baron and Kenny 1986; Robins and Greenland 1992; Imai, Keele, and Tingley 2010; Imai, Keele, and Yamamoto 2010; Tchetgen and Shpitser 2012). Notably, Pearl (2001) formally defined direct and indirect effects for general nonlinear models. Causal mediation analysis is also a valuable technique for explainable artificial intelligence (XAI) (Shin 2021). In this paper, we aim to provide causal mediation analysis for PoC, to reveal the necessity and sufficiency of the treatment through different pathways. Once a treatment is revealed to be necessary and sufficient to induce a particular event via PNS, other causal questions would arise:

(**Q1**). *Would the treatment still be necessary and sufficient had the value of the mediator been fixed to a certain value?*

(**Q2**). *Would the treatment still be necessary and sufficient had there been no influence via the mediator?*

(**Q3**). *Would the treatment still be necessary and sufficient had the influence only existed via the mediator?*

We introduce new variants of PoC - controlled direct, natural direct, and natural indirect PNS (CD-PNS, ND-PNS, and NI-PNS) to answer these questions. We further define direct and indirect PoC with evidence to capture more sophisticated counterfactual information useful for decision-making. These quantities can retrospectively answer questions (Q1), (Q2), and (Q3) for a specific subpopulation. We provide identification results for each type of PoC we introduce. Finally, we apply our results to a real-world psychology dataset.

Notations and Background

We represent a single or vector variable with a capital letter (X) and its realized value with a small letter (x). Let $\mathbb{I}(\cdot)$ be an indicator function that takes 1 if the statement in $(\cdot)$ is true and 0 otherwise. Denote Ω_Y be the domain of variable Y, $\mathbb{E}[Y]$ be the expectation of Y, $\mathbb{P}(Y < y)$ be the cumulative distribution function (CDF) of continuous variable Y, and $\mathfrak{p}(Y = y)$ be the probability density function (PDF) of continuous variable Y. We use $X \perp\!\!\!\perp Y|C$ to denote that X and Y are conditionally independent given C. We use $\preceq$ to denote a total order. A formal definition of total order is given in Appendix A in (Kawakami and Tian 2024).

Structural causal models (SCM). We use the language of SCMs as our basic framework and follow the standard definition in the following (Pearl 2009). An SCM $\mathcal{M}$ is a tuple $\langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, \mathbb{P}_{\boldsymbol{U}} \rangle$, where $\boldsymbol{U}$ is a set of exogenous (unobserved) variables following a distribution $\mathbb{P}_{\boldsymbol{U}}$, and $\boldsymbol{V}$ is a set of endogenous (observable) variables whose values are

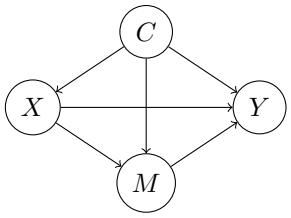


Figure 1: A causal graph representing SCM $\mathcal{M}$.

determined by structural functions $\mathcal{F} = \{f_{V_i}\}_{V_i \in \boldsymbol{V}}$ such that $v_i := f_{V_i}(\mathbf{pa}_{V_i}, \boldsymbol{u}_{V_i})$ where $\mathbf{PA}_{V_i} \subseteq \boldsymbol{V}$ and $\boldsymbol{U}_{V_i} \subseteq \boldsymbol{U}$. Each SCM $\mathcal{M}$ induces an observational distribution $\mathbb{P}_{\boldsymbol{V}}$ over $\boldsymbol{V}$, and a causal graph $G(\mathcal{M})$ in which there exists a directed edge from every variable in $\mathbf{PA}_{V_i}$ and $\boldsymbol{U}_{V_i}$ to V_i. An intervention of setting a set of endogenous variables $\boldsymbol{X}$ to constants $\boldsymbol{x}$, denoted by $do(\boldsymbol{x})$, replaces the original equations of $\boldsymbol{X}$ by the constants $\boldsymbol{x}$ and induces a *sub-model* $\mathcal{M}_{\boldsymbol{x}}$. We denote the potential outcome Y under intervention $do(\boldsymbol{x})$ by $Y_{\boldsymbol{x}}(\boldsymbol{u})$, which is the solution of Y in the sub-model $\mathcal{M}_{\boldsymbol{x}}$ given $\boldsymbol{U} = \boldsymbol{u}$.

Probabilities of causation (PoC). Kawakami, Kuroki, and Tian (2024) defined the (multivariate conditional) PoC for vectors of continuous or discrete variables as follows:

Definition 1 (PoC)**.** *(Kawakami, Kuroki, and Tian 2024) The (multivariate conditional) PoC are defined by*

$$\text{PNS}(y; x', x, c) \triangleq \mathbb{P}(Y_{x'} \prec y \preceq Y_x | C = c), \quad (1)$$

$$\text{PN}(y; x', x, c) \triangleq \mathbb{P}(Y_{x'} \prec y | y \preceq Y, X = x, C = c), \quad (2)$$

$$\text{PS}(y; x', x, c) \triangleq \mathbb{P}(y \preceq Y_x | Y \prec y, X = x', C = c). \quad (3)$$

$\text{PNS}(y; x', x, c)$ provides a measure of the necessity and sufficiency of x w.r.t. x' to produce $Y \succeq y$ given $C = c$. $\text{PN}(y; x', x, c)$ and $\text{PS}(y; x', x, c)$ provide a measure of the necessity and sufficiency, respectively, of x w.r.t. x' to produce $Y \succeq y$ given $C = c$.

We will often call PNS *total PNS (T-PNS)* and denote it by $\text{T-PNS}(y; x', x, c)$ for convenience. When treatment X and outcome Y are binary, PNS, PS, and PS become (setting $y = 1$) $\text{PNS}(c) = \mathbb{P}(Y_0 = 0, Y_1 = 1 | C = c)$, $\text{PN}(c) = \mathbb{P}(Y_0 = 0 | Y = 1, X = 1, C = c)$, and $\text{PS}(c) = \mathbb{P}(Y_1 = 1 | Y = 0, X = 0, C = c)$ for any $c \in \Omega_C$, which reduce to Pearl's (1999) original definition when $C = \emptyset$.

Causal mediation analysis. Causal mediation analysis reveals the strength of different pathways between treatment and outcome through a mediator. Researchers often consider the following SCM $\mathcal{M}$:

$$\begin{aligned} Y := f_Y(X, M, C, U_Y), M := f_M(X, C, U_M), \\ X := f_X(C, U_X), C := f_C(U_C), \end{aligned} \quad (4)$$

where all variables can be vectors, and U_X, U_C, U_Y, and U_M are latent exogenous variables. Assume that the domains Ω_Y and $\Omega_{U_Y} \times \Omega_{U_M}$ are totally ordered sets with $\preceq$. Figure 1 shows the causal graph of SCM $\mathcal{M}$ (with latent variables dropped).

One widely used model in the mediation analysis is a linear SCM $\mathcal{M}^L$ (Baron and Kenny 1986) consisting of $Y := \beta_0 + \beta_1 X + \beta_2 M + \beta_3 C + U_Y$ and $M := \alpha_0 + \alpha_1 X + \alpha_2 C + U_M$, where $U_Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and $U_M \sim \mathcal{N}(\mu_M, \sigma_M^2)$ are independent normal distribution. Under SCM $\mathcal{M}^L$, the total effect of X on Y is $\beta_1 + \alpha_1\beta_2$, the indirect effect is $\alpha_1\beta_2$, and the direct effect is β_1.

Pearl (2001) defined the total, controlled direct, natural direct, and natural indirect effects for general (nonlinear and nonparametric) SCM $\mathcal{M}$.

Definition 2 (TE, CDE, NDE, and NIE)**.** *(Pearl 2001) The total, controlled direct, natural direct, and natural indirect effects are defined by:*

1. *Total Effect (TE):* $\text{TE}(y; x', x) \triangleq \mathbb{E}[Y_x] - \mathbb{E}[Y_{x'}]$
2. *Controlled Direct Effect (CDE):* $\text{CDE}(y; x', x, m) \triangleq \mathbb{E}[Y_{x,m}] - \mathbb{E}[Y_{x',m}]$
3. *Natural Direct Effect (NDE):* $\text{NDE}(y; x', x) \triangleq \mathbb{E}[Y_{x,M_{x'}}] - \mathbb{E}[Y_{x'}]$
4. *Natural Indirect Effect (NIE):* $\text{NIE}(y; x', x) \triangleq \mathbb{E}[Y_{x',M_x}] - \mathbb{E}[Y_{x'}]$

CDE represents the causal effect of changing the treatment from x' to x had the value of the mediator been fixed at a certain value. NDE represents the causal effect of changing the treatment from x' to x had the value of the mediator been kept to the same value $M_{x'}$ that M attains under x'. NIE represents the causal effect of changing the mediator from $M_{x'}$ to M_x had the value of the treatment been fixed to x'. TE can be decomposed into NDE and NIE by $\text{TE}(y; x', x) = \text{NDE}(y; x', x) - \text{NIE}(y; x, x') = \text{NIE}(y; x', x) - \text{NDE}(y; x, x')$.

These direct and indirect effects may be identified from observational distributions under various settings (Pearl 2001; Avin, Shpitser, and Pearl 2005; Shpitser and Pearl 2008; Shpitser 2013; Malinsky, Shpitser, and Richardson 2019). A widely used assumption for identifying causal mediation effects is the following sequential ignorability assumption (Imai, Keele, and Tingley 2010):

Assumption 1 (Sequential ignorability)**.** *The following two conditional independence statements hold:*

$$(1)\ \{Y_{x,m}, M_x\} \perp\!\!\!\perp X | C = c \text{ and } (2)\ M_x \perp\!\!\!\perp Y_{x,m} | C = c$$

for any $m \in \Omega_M$ and $x \in \Omega_X$, where $\mathfrak{p}(X = x | C = c) > 0$ and $\mathfrak{p}(M = m | X = x, C = c) > 0$ for any $m \in \Omega_M$, $x \in \Omega_X$, and $c \in \Omega_C$.

We have

Proposition 1 (Identification of $\mathbb{P}(Y_{x',M_x} \prec y | C = c)$)**.** *(Imai, Keele, and Tingley 2010; VanderWeele and Knol 2014) Under SCM $\mathcal{M}$ and Assumption 1, the counterfactual $\mathbb{P}(Y_{x',M_x} \prec y | C = c)$ is identifiable by*

$$\begin{aligned} &\mathbb{P}(Y_{x',M_x} \prec y | C = c) \\ &= \int_{\Omega_M} \mathbb{P}(Y \prec y | X = x', M = m, C = c) \\ &\qquad \times \mathfrak{p}(M = m | X = x, C = c) dm \end{aligned} \quad (5)$$

for any $x', x \in \Omega_X$, $y \in \Omega_Y$, and $c \in \Omega_C$.

Direct and Indirect PNS

In this section, we introduce new concepts of direct and indirect PNS and provide corresponding identification results. We will focus our attention on PNS, and show in the next section that direct and indirect PN and PS can be derived as special cases of direct and indirect PNS with evidence.

Definitions of CD-PNS, ND-PNS, and NI-PNS

We define controlled direct, natural direct, and natural indirect probabilities of necessity and sufficiency.

Definition 3 (CD-PNS, ND-PNS, and NI-PNS)**.** *The controlled direct, natural direct, and natural indirect PNS (CD-PNS, ND-PNS, and NI-PNS) are defined by*

$$\text{CD-PNS}(y; x', x, m, c) \triangleq \mathbb{P}(Y_{x',m} \prec y \preceq Y_{x,m} | C = c), \tag{6}$$

$$\text{ND-PNS}(y; x', x, c) \triangleq \mathbb{P}(Y_{x'} \prec y \preceq Y_x, Y_{x',M_x} \prec y | C = c), \tag{7}$$

$$\text{NI-PNS}(y; x', x, c) \triangleq \mathbb{P}(Y_{x'} \prec y \preceq Y_x, y \preceq Y_{x',M_x} | C = c). \tag{8}$$

First, the controlled direct PNS (CD-PNS) provides a measure of the necessity and sufficiency of x w.r.t. x' to produce $Y \succeq y$ given $C = c$ when the mediator is fixed to a value $M = m$. CD-PNS can be used to answer the causal question (Q1). CD-PNS consists of two counterfactual conditions:

(**A1**). *"had the treatment and the mediator been* (x', m)*, the outcome would be* $Y \prec y$*"* $(Y_{x',m} \prec y)$; and

(**A2**). *"had the treatment and the mediator been* (x, m)*, the outcome would be* $y \preceq Y$*"* $(y \preceq Y_{x,m})$.

Conditions (A1) and (A2) have different values of treatment and the same values of mediator. The relative values of the potential outcomes $Y_{x,m}$ are shown in Figure 2 (b). For comparison, Figure 2 (a) shows the situation for T-PNS.

Second, ND-PNS has three counterfactual conditions:

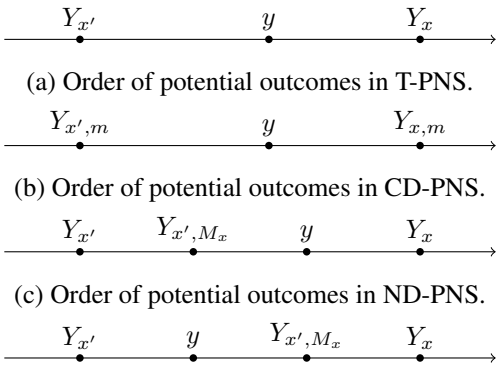
(**B1**). *"had the treatment been* x'*, the outcome would be* $Y \prec y$*"* $(Y_{x'} = Y_{x',M_{x'}} \prec y)$,

(**B2**). *"had the treatment been* x' *but the mediator was kept at the same value* M_x *when the treatment is* x*, the outcome would be* $Y \prec y$*"* $(Y_{x',M_x} \prec y)$, and

(**B3**). *"had the treatment been* x*, the outcome would be* $y \preceq Y$*"* $(y \preceq Y_x = Y_{x,M_x})$,

The relative values of the potential outcomes are shown in Figure 2 (c). Conditions (B1) and (B3) mean $Y_{x'} \prec y \preceq Y_x$, which is the same condition in T-PNS and represents that the treatment x is necessary and sufficient w.r.t. x' to provoke the event $y \preceq Y$ given $C = c$. Conditions (B2) and (B3) mean $Y_{x',M_x} \prec y \preceq Y_{x,M_x}$, which represents the necessity and sufficiency of x w.r.t. x' to produce $Y \succeq y$ given $C = c$ when keeping the values of the mediator by the same as M_x. In other words, they mean that the treatment would be necessary and sufficient even if there were no influences via the mediator. Therefore, ND-PNS can answer the causal question (Q2).

Third, NI-PNS has three counterfactual conditions:



(a) Order of potential outcomes in T-PNS.

(b) Order of potential outcomes in CD-PNS.

(c) Order of potential outcomes in ND-PNS.

(d) Order of potential outcomes in NI-PNS.

Figure 2: Order of potential outcomes in each PNS.

(**C1**). *"had the treatment been* x'*, the outcome would be* $Y \prec y$*"* $(Y_{x'} = Y_{x',M_{x'}} \prec y)$,

(**C2**). *"had the treatment been* x' *but the mediator was kept at the same value* M_x *when the treatment is* x*, the outcome would be* $y \preceq Y$*"* $(y \preceq Y_{x',M_x})$, and

(**C3**). *"had the treatment been* x*, the outcome would be* $y \preceq Y$*"* $(y \preceq Y_x = Y_{x,M_x})$,

The relative values of the potential outcomes are shown in Figure 2 (d). Conditions (C1) and (C3) mean $Y_{x'} \prec y \preceq Y_x$, which is the same condition in T-PNS and states that the treatment x is necessary and sufficient w.r.t. x' to provoke the event $y \preceq Y$ given $C = c$. Conditions (C1) and (C2) mean $Y_{x',M_{x'}} \prec y \preceq Y_{x',M_x}$, which represents the necessity and sufficiency of M_x w.r.t. $M_{x'}$ to produce $Y \succeq y$ given $C = c$ when setting the treatment to x'. In other words, they mean that the treatment would be necessary and sufficient if the influence is only via the mediator. Therefore, NI-PNS can answer the causal question (Q3).

Then, the following proposition holds.

Proposition 2. *We have*

$$\begin{aligned} &\text{T-PNS}(y; x', x, c) \\ &= \text{ND-PNS}(y; x', x, c) + \text{NI-PNS}(y; x', x, c). \end{aligned} \tag{9}$$

Eq. (9) states that the total PNS can be decomposed into a summation of the natural direct and natural indirect PNS, a desired property of causal mediation analysis.

Remark 1. Researchers have considered the proportion of direct or indirect influence in the total influence, which captures how important each pathway is in explaining the total influence (VanderWeele 2013). However, the proportions of direct and indirect effects in the total effects under linear SCM $\mathcal{M}^L$ or the proportions of NDE and NIE in TE may not always make sense since these quantities may take negative values. In contrast, the proportions of ND-PNS and NI-PNS in T-PNS are given by $\text{ND-PNS}(y; x', x, c)/\text{T-PNS}(y; x', x, c) = \mathbb{P}(Y_{x',M_x} \prec y | Y_{x'} \prec y \preceq Y_x, C = c)$ and $\text{NI-PNS}(y; x', x, c)/\text{T-PNS}(y; x', x, c) = \mathbb{P}(y \preceq$

$Y_{x',M_x}|Y_{x'} \prec y \preceq Y_x, C = c)$, respectively, which do not take negative values. Additionally, the sum of the proportions of ND-PNS and NI-PNS is always equal to 1.

Remark 2. Rubinstein, Cuellar, and Malinsky (2024) defined, for binary treatment, outcome, and mediator, the total mediated PoC by $\delta(c) \triangleq \mathbb{P}(Y_0 = 0|Y_1 = 1, M_1 = 1, C = c)$, the direct mediated PoC by $\psi(c) \triangleq \mathbb{P}(Y_{1,M_0} = 0, Y_{0,M_0} = 0|Y_{1,M_1} = 1, M_1 = 1, C = c)$, and the indirect mediated PoC by $\zeta(c) \triangleq \mathbb{P}(Y_{1,M_0} = 1, Y_{0,M_0} = 0|Y_{1,M_1} = 1, M_1 = 1, C = c)$. While we focus on the necessity and sufficiency of the treatment to provoke an event, their definitions of mediated PoC differ from ours and are aimed at answering different questions. For example, their total mediated PoC is motivated by the question: "Given that subjects would experience events $Y = 1$ and $M = 1$ had they taken a treatment $X = 1$, what is the probability that they would not have experienced the event $Y = 1$ in the absence of the treatment?". We note that their mediated PoC satisfy the property $\delta(c) = \psi(c) + \zeta(c)$. We provide a detailed comparison in Appendix E in (Kawakami and Tian 2024).

Identification of CD-PNS, ND-PNS, and NI-PNS

Next, we provide identification theorems for the direct and indirect PNSs we have introduced.

Assumptions The identification of PoC relies on monotonicity assumptions in the literature (Tian and Pearl 2000). We will make similar assumptions, specifically similar to those in (Kawakami, Kuroki, and Tian 2024).

Assumption 2. *Potential outcome $Y_{x,m}$ has conditional PDF $\mathfrak{p}_{Y_{x,m}|C=c}$ for each $x \in \Omega_X$, $m \in \Omega_M$, and $c \in \Omega_C$, and its support $\{y \in \Omega_Y : \mathfrak{p}_{Y_{x,m}|C=c}(y) \neq 0\}$ is the same for each $x \in \Omega_X$, $m \in \Omega_M$, and $c \in \Omega_C$.*

Assumption 3. *Potential outcome Y_{x',M_x} has conditional PDF $\mathfrak{p}_{Y_{x',M_x}|C=c}$ for each $x', x \in \Omega_X$ and $c \in \Omega_C$, and its support $\{y \in \Omega_Y : \mathfrak{p}_{Y_{x',M_x}|C=c}(y) \neq 0\}$ is the same for each $x', x \in \Omega_X$ and $c \in \Omega_C$.*

Assumptions 2 and 3 are reasonable for continuous variables. For example, potential outcomes $Y_{x,m}$, Y_{x',M_x} often has $[-\infty, \infty]$ support, such as in linear SCM $\mathcal{M}^L$.

We assume the following monotonicity condition for identifying CD-PNS:

Assumption 4 (Monotonicity over f_Y)**.** *The function $f_Y(x, m, c, U_Y)$ is either monotonic increasing on U_Y for all $x \in \Omega_X$, $m \in \Omega_M$, and $c \in \Omega_C$, or monotonic decreasing on U_Y for all $x \in \Omega_X$, $m \in \Omega_M$, and $c \in \Omega_C$, almost surely w.r.t. $\mathbb{P}_{U_Y}$.*

Alternatively, one may assume monotonicity over potential outcomes:

Assumption 4' (Conditional monotonicity over $Y_{x,m}$) *The potential outcomes $Y_{x,m}$ satisfy: for any $x', x \in \Omega_X$, $m \in \Omega_M$, $y \in \Omega_Y$, and $c \in \Omega_C$, either $\mathbb{P}(Y_{x',m} \prec y \preceq Y_{x,m}|C = c) = 0$ or $\mathbb{P}(Y_{x,m} \prec y \preceq Y_{x',m}|C = c) = 0$.*

Assumptions 4 and 4' are equivalent under Assumption 2 (a straightforward extension of Theorem 4.1 in (Kawakami, Kuroki, and Tian 2024)). We note that the widely used linear SCM $\mathcal{M}^L$ satisfies Assumption 4. Furthermore, another popular model, a nonlinear SCM with normal distribution $\mathcal{M}^N$, consisting of $Y := f_Y(X, M, C) + U_Y$ and $M := f_M(X, C) + U_M$, where $U_Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and $U_M \sim \mathcal{N}(\mu_M, \sigma_M^2)$, also satisfies Assumptions 2-4.

Let the compound function $f_Y \circ f_M$ represent $(f_Y \circ f_M)(x', x, c, \tilde{U}) = f_Y(x', f_M(x, c, U_M), c, U_Y)$ for all $x', x \in \Omega_X$ and $c \in \Omega_C$, where $\tilde{U} = (U_Y, U_M)$. We assume the following for identifying ND-PNS and NI-PNS:

Assumption 5 (Monotonicity over $f_Y \circ f_M$)**.** *The function $(f_Y \circ f_M)(x', x, c, \tilde{U})$ is either monotonic increasing on $\tilde{U}$ for all $x', x \in \Omega_X$ and $c \in \Omega_C$, or monotonic decreasing on $\tilde{U}$ for all $x', x \in \Omega_X$ and $c \in \Omega_C$, almost surely w.r.t. $\mathbb{P}_{\tilde{U}}$.*

Or, alternatively,

Assumption 5' (Conditional monotonicity over Y_{x',M_x}) *The potential outcomes Y_{x',M_x} satisfy: for any $x, x', x'', x''' \in \Omega_X$, $y \in \Omega_Y$, and $c \in \Omega_C$, either $\mathbb{P}(Y_{x',M_x} \prec y \preceq Y_{x''',M_{x''}}|C = c) = 0$ or $\mathbb{P}(Y_{x''',M_{x''}} \prec y \preceq Y_{x',M_x}|C = c) = 0$.*

Similarly, Assumptions 5 and 5' are equivalent under Assumption 3. We note that both the linear SCM $\mathcal{M}^L$ and the nonlinear SCM with normal distribution $\mathcal{M}^N$ satisfy Assumption 5 with $\tilde{U} = U_Y + U_M$.

Lemmas. Then, we obtain the following results.

Lemma 1. *Under SCM $\mathcal{M}$, and Assumptions 2 and 4,*

$$\text{CD-PNS}(y; x', x, m, c) = \max\Big\{\mathbb{P}(Y_{x',m} \prec y|C = c) - \mathbb{P}(Y_{x,m} \prec y|C = c), 0\Big\}. \tag{10}$$

Lemma 2. *Under SCM $\mathcal{M}$, and Assumptions 3 and 5,*

$$\text{ND-PNS}(y; x', x, c) = \max\Big\{\min\{\mathbb{P}(Y_{x'} \prec y|C = c), \mathbb{P}(Y_{x',M_x} \prec y|C = c)\} - \mathbb{P}(Y_x \prec y|C = c), 0\Big\}, \tag{11}$$

$$\text{NI-PNS}(y; x', x, c) = \max\Big\{\mathbb{P}(Y_{x'} \prec y|C = c) - \max\{\mathbb{P}(Y_x \prec y|C = c), \mathbb{P}(Y_{x',M_x} \prec y|C = c)\}, 0\Big\}. \tag{12}$$

The lemmas mean that, under monotonicity, CD-PNS, ND-PNS, and NI-PNS can be computed from the CDF of certain counterfactual outcomes.

Identification theorems. The CDF of the counterfactual outcomes $\mathbb{P}(Y_{x,M_{x'}} \prec y|C = c)$ is identifiable under the sequential ignorability Assumption 1 by Proposition 1 as $\mathbb{P}(Y_{x,M_{x'}} \prec y|C = c) = \rho(y; x', x, c)$, where we denote

$$\rho(y; x', x, c) \triangleq \int_{\Omega_M} \mathbb{P}(Y \prec y|X = x', M = m, C = c) \times \mathfrak{p}(M = m|X = x, C = c)dm. \tag{13}$$

Then, we obtain the following identification theorems by combining Lemmas 1 and 2 and Proposition 1:

Theorem 1 (Identification of CD-PNS)**.** *Under SCM $\mathcal{M}$, and Assumptions 1, 2, and 4, CD-PNS is identifiable by*

$$\begin{aligned}&\text{CD-PNS}(y; x', x, m, c) = \\ &\quad \min\Big\{\mathbb{P}(Y \prec y | X = x', M = m, C = c) \\ &\qquad - \mathbb{P}(Y \prec y | X = x, M = m, C = c), 0\Big\}.\end{aligned} \tag{14}$$

Theorem 2 (Identification of ND-PNS and NI-PNS)**.** *Under SCM $\mathcal{M}$, and Assumptions 1, 3, and 5, ND-PNS and NI-PNS are identifiable by*

$$\begin{aligned}&\text{ND-PNS}(y; x', x, c) \\ &= \max\Big\{\min\{\mathbb{P}(Y \prec y | X = x', C = c), \\ &\qquad \rho(y; x', x, c)\} - \mathbb{P}(Y \prec y | X = x, C = c), 0\Big\},\end{aligned} \tag{15}$$

$$\begin{aligned}&\text{NI-PNS}(y; x', x, c) = \max\Big\{\mathbb{P}(Y \prec y | X = x', C = c) \\ &\quad - \max\{\mathbb{P}(Y \prec y | X = x, C = c), \rho(y; x', x, c)\}, 0\Big\}.\end{aligned} \tag{16}$$

As a consequence, under SCM $\mathcal{M}$ and Assumptions 1, 3, and 5, the proportions of ND-PNS and NI-PNS in T-PNS are also identifiable.

Direct and Indirect PNS with Evidence

In this section, we define CD-PNS, ND-PNS, and NI-PNS with evidence and provide corresponding identification theorems. Specifically, we consider two types of evidence:

$$\mathcal{E} \triangleq (X = x^*, M = m^*, Y \in \mathcal{I}_Y), \tag{17}$$

$$\mathcal{E}' \triangleq (X = x^*, Y \in \mathcal{I}_Y), \tag{18}$$

where $\mathcal{I}_Y$ is a half-open interval $[y^l, y^u)$ or a closed interval $[y^l, y^u]$ w.r.t. $\prec$. PNS with evidence allows us to examine PNS for a specific subpopulation characterized by the evidence.

The main distinction between the evidence $\mathcal{E}$ or $\mathcal{E}'$ and the subject's covariates C in the definition of CD-PNS, ND-PNS, and NI-PNS (Def. 3) is that C in the SCM $\mathcal{M}$ are pre-treatment variables but $\mathcal{E}$ are post-treatment variables. Conditioning on post-treatment variables differs from traditional conditioning on pre-treatment variables and has been discussed in the context of PN or PS (Pearl 1999) and the posterior causal effects (Lu et al. 2022; Li et al. 2023). They have applications in various fields, such as attribution of risk factors in public health and epidemiology, medical diagnosis of diseases, root-cause diagnosis in equipment and production processes, and reference measures for penalties in law.

Definitions of CD-PNS, ND-PNS, and NI-PNS with Evidence

First, we define CD-PNS with evidence $\mathcal{E}$, and T-PNS, ND-PNS, and NI-PNS with evidence $\mathcal{E}'$ [1].

[1]Kawakami, Kuroki, and Tian (2024) have studied T-PNS with evidence $(X = x^*, Y = y^*)$, which is a special case of $\mathcal{E}'$.

Definition 4 (CD-PNS, T-PNS, ND-PNS, and NI-PNS with evidence)**.** *CD-PNS with evidence $\mathcal{E}$, and T-PNS, ND-PNS, and NI-PNS with evidence $\mathcal{E}'$ are defined by*

$$\begin{aligned}&\text{CD-PNS}(y; x', x, m, \mathcal{E}, c) \\ &\quad \triangleq \mathbb{P}(Y_{x',m} \prec y \preceq Y_{x,m} | \mathcal{E}, C = c),\end{aligned} \tag{19}$$

$$\text{T-PNS}(y; x', x, \mathcal{E}', c) \triangleq \mathbb{P}(Y_{x'} \prec y \preceq Y_x | \mathcal{E}', C = c), \tag{20}$$

$$\begin{aligned}&\text{ND-PNS}(y; x', x, \mathcal{E}', c) \\ &\quad \triangleq \mathbb{P}(Y_{x'} \prec y \preceq Y_x, Y_{x', M_x} \prec y | \mathcal{E}', C = c),\end{aligned} \tag{21}$$

$$\begin{aligned}&\text{NI-PNS}(y; x', x, \mathcal{E}', c) \\ &\quad \triangleq \mathbb{P}(Y_{x'} \prec y \preceq Y_x, y \preceq Y_{x', M_x} | \mathcal{E}', C = c).\end{aligned} \tag{22}$$

CD-PNS with evidence can answer questions: "What is the probability that the situation in the question (Q1) holds for the subjects, when, in reality, their treatment is x^*, their mediator is m^*, their outcome is in $\mathcal{I}_Y$, and their covariates is c?" ND-PNS and NI-PNS with evidence can answer questions: "What is the probability that the situation in the questions (Q2) and (Q3) hold, when, in reality, their treatment is x^*, their outcome is in $\mathcal{I}_Y$, and their covariates is c?" CD-PNS, ND-PNS, and NI-PNS with evidence can retrospectively answer questions for the specific subpopulation characterized by the evidence.

The following desired decomposition property holds:

Proposition 3.

$$\begin{aligned}&\text{T-PNS}(y; x', x, \mathcal{E}', c) \\ &= \text{ND-PNS}(y; x', x, \mathcal{E}', c) + \text{NI-PNS}(y; x', x, \mathcal{E}', c).\end{aligned} \tag{23}$$

Remark 3. We do not use mediator information in evidence for T-PNS, ND-PNS, and NI-PNS because a more strict assumption is required for identification to exploit mediator information. In Appendix C in (Kawakami and Tian 2024), we provide an identification theorem (Theorem 4') of T-PNS, ND-PNS, and NI-PNS with evidence $\mathcal{E}' = (X = x^*, M \in \mathcal{I}_M, Y \in \mathcal{I}_Y)$ with an additional assumption, where $\mathcal{I}_M$ is a half-open interval $[m^l, m^u)$ w.r.t. the total order on Ω_M.

ND-PN, NI-PN, ND-PS, and NI-PS. So far, we have focused our attention on PNS in the PoC family. It turns out that PN and PS, the other two members of the PoC family defined in Def. 1, can be computed as special cases of T-PNS with evidence. Specifically, PN is equivalent to T-PNS with the evidence $\mathcal{E}' = (y \preceq Y, X = x)$, and PS is equivalent to T-PNS with the evidence $\mathcal{E}' = (Y \prec y, X = x')$ as follows.

Proposition 4. *We have the following:*

$$\text{PN}(y; x', x, c) = \mathbb{P}(Y_{x'} \prec y \preceq Y_x | y \preceq Y, X = x, C = c), \tag{24}$$

$$\text{PS}(y; x', x, c) = \mathbb{P}(Y_{x'} \prec y \preceq Y_x | Y \prec y, X = x', C = c). \tag{25}$$

Then, direct and indirect PN and PS can be naturally defined by extending the definitions of ND-PNS and NI-PNS with evidence in Def. 4.

Definition 5 (ND-PN, NI-PN, ND-PS, and NI-PS)**.** *The natural direct PN (ND-PN), natural indirect PN (NI-PN), natural direct PS (ND-PS), and natural indirect PS (NI-PS) are defined by*

$$\begin{aligned}&\text{ND-PN}(y; x', x, c)\\ &\triangleq \mathbb{P}(Y_{x'} \prec y, Y_{x', M_x} \prec y | y \preceq Y, X = x, C = c),\end{aligned} \tag{26}$$

$$\begin{aligned}&\text{NI-PN}(y; x', x, c)\\ &\triangleq \mathbb{P}(Y_{x'} \prec y, y \preceq Y_{x', M_x} | y \preceq Y, X = x, C = c),\end{aligned} \tag{27}$$

$$\begin{aligned}&\text{ND-PS}(y; x', x, c)\\ &\triangleq \mathbb{P}(y \preceq Y_x, Y_{x', M_x} \prec y | Y \prec y, X = x', C = c),\end{aligned} \tag{28}$$

$$\begin{aligned}&\text{NI-PS}(y; x', x, c)\\ &\triangleq \mathbb{P}(y \preceq Y_x, y \preceq Y_{x', M_x} | Y \prec y, X = x', C = c).\end{aligned} \tag{29}$$

ND-PN, NI-PN, ND-PS, and NI-PS provide a measure of the necessity or the sufficiency of the treatment for the outcome through direct or indirect pathways. We have the desirable decomposition property that $\text{PN}(y; x', x, c) = \text{ND-PN}(y; x', x, c) + \text{NI-PN}(y; x', x, c)$ and $\text{PS}(y; x', x, c) = \text{ND-PS}(y; x', x, c) + \text{NI-PS}(y; x', x, c)$.

Identification of CD-PNS, T-PNS, ND-PNS, and NI-PNS with Evidence

We obtain the following two identification theorems under the same assumptions for Theorems 1 or 2.

Theorem 3 (Identification of CD-PNS with evidence $\mathcal{E}$)**.** *Let $\mathcal{I}_Y$ be a half-open interval $[y^l, y^u)$ in evidence $\mathcal{E}$. Under SCM $\mathcal{M}$, and Assumptions 1, 2, and 4, for each $x', x \in \Omega_X$, $m \in \Omega_M$, $y \in \Omega_Y$, and $c \in \Omega_C$, we have*

(A). *If $\mathbb{P}(Y \prec y^u | X = x^*, M = m^*, C = c) \neq \mathbb{P}(Y \prec y^l | X = x^*, M = m^*, C = c)$, then*

$$\text{CD-PNS}(y; x', x, m, \mathcal{E}, c) = \max\{\alpha/\beta, 0\}, \tag{30}$$

where

$$\begin{aligned}\alpha = \min\Big\{&\mathbb{P}(Y \prec y | X = x', M = m, C = c),\\ &\mathbb{P}(Y \prec y^u | X = x^*, M = m^*, C = c)\Big\}\\ -\max\Big\{&\mathbb{P}(Y \prec y | X = x, M = m, C = c),\\ &\mathbb{P}(Y \prec y^l | X = x^*, M = m^*, C = c)\Big\},\end{aligned} \tag{31}$$

$$\begin{aligned}\beta = &\mathbb{P}(Y \prec y^u | X = x^*, M = m^*, C = c)\\ &- \mathbb{P}(Y \prec y^l | X = x^*, M = m^*, C = c).\end{aligned} \tag{32}$$

(B). *If $\mathbb{P}(Y \prec y^u | X = x^*, M = m^*, C = c) = \mathbb{P}(Y \prec y^l | X = x^*, M = m^*, C = c)$, then*

$$\begin{aligned}&\text{CD-PNS}(y; x', x, m, \mathcal{E}, c) = \mathbb{I}\Big(\mathbb{P}(Y \prec y | X = x', C = c)\\ &\leq \mathbb{P}(Y \prec y^l | X = x^*, M = m^*, C = c)\\ &< \mathbb{P}(Y_x \prec y | C = c)\Big).\end{aligned} \tag{33}$$

Theorem 4 (Identification of T-PNS, ND-PNS, and NI-PNS with evidence $\mathcal{E}'$)**.** *Let $\mathcal{I}_Y$ be a half-open interval $[y^l, y^u)$ in evidence $\mathcal{E}'$. Under SCM $\mathcal{M}$, and Assumptions 1, 3, and 5, for each $x', x \in \Omega_X$, $y \in \Omega_Y$, and $c \in \Omega_C$, we have*

(A). *If $\mathbb{P}(Y \prec y^u | X = x^*, C = c) \neq \mathbb{P}(Y \prec y^l | X = x^*, C = c)$, then*

$$\text{T-PNS}(y; x', x, \mathcal{E}', c) = \max\left\{\gamma^T/\delta, 0\right\}, \tag{34}$$

$$\text{ND-PNS}(y; x', x, \mathcal{E}', c) = \max\left\{\gamma^D/\delta, 0\right\}, \tag{35}$$

$$\text{NI-PNS}(y; x', x, \mathcal{E}', c) = \max\left\{\gamma^I/\delta, 0\right\}, \tag{36}$$

where

$$\begin{aligned}\gamma^T = \min\Big\{&\mathbb{P}(Y \prec y | X = x', C = c),\\ &\mathbb{P}(Y \prec y^u | X = x^*, C = c)\}\\ -\max\{&\mathbb{P}(Y \prec y | X = x, C = c),\\ &\mathbb{P}(Y \prec y^l | X = x^*, C = c)\Big\},\end{aligned} \tag{37}$$

$$\begin{aligned}\gamma^D = \min\Big\{&\mathbb{P}(Y \prec y | X = x', C = c),\\ &\mathbb{P}(Y \prec y^u | X = x^*, C = c), \rho(y; x', x, c)\}\\ -\max\{&\mathbb{P}(Y \prec y | X = x, C = c),\\ &\mathbb{P}(Y \prec y^l | X = x^*, C = c)\Big\},\end{aligned} \tag{38}$$

$$\begin{aligned}\gamma^I = \min\Big\{&\mathbb{P}(Y \prec y | X = x', C = c),\\ &\mathbb{P}(Y \prec y^u | X = x^*, C = c)\}\\ -\max\{&\mathbb{P}(Y \prec y | X = x, C = c),\\ &\mathbb{P}(Y \prec y^l | X = x^*, C = c), \rho(y; x', x, c)\Big\},\end{aligned} \tag{39}$$

$$\begin{aligned}\delta = &\mathbb{P}(Y \prec y^u | X = x^*, C = c)\\ &- \mathbb{P}(Y \prec y^l | X = x^*, C = c).\end{aligned} \tag{40}$$

(B). *If $\mathbb{P}(Y \prec y^u | X = x^*, C = c) = \mathbb{P}(Y \prec y^l | X = x^*, C = c)$, then*

$$\begin{aligned}&\text{T-PNS}(y; x', x, \mathcal{E}', c) = \mathbb{I}\Big(\mathbb{P}(Y \prec y | X = x', C = c) \leq\\ &\mathbb{P}(Y \prec y^l | X = x^*, C = c) < \mathbb{P}(Y \prec y | X = x, C = c)\Big),\end{aligned} \tag{41}$$

$$\begin{aligned}&\text{ND-PNS}(y; x', x, \mathcal{E}', c) = \mathbb{I}\Big(\mathbb{P}(Y \prec y | X = x', C = c) \leq\\ &\mathbb{P}(Y \prec y^l | X = x^*, C = c) < \mathbb{P}(Y \prec y | X = x, C = c),\\ &\rho(y; x', x, c) \leq \mathbb{P}(Y \prec y^l | X = x^*, C = c)\Big),\end{aligned} \tag{42}$$

$$\begin{aligned}&\text{NI-PNS}(y; x', x, \mathcal{E}', c) = \mathbb{I}\Big(\mathbb{P}(Y \prec y | X = x', C = c) \leq\\ &\mathbb{P}(Y \prec y^l | X = x^*, C = c) < \mathbb{P}(Y \prec y | X = x, C = c),\\ &\mathbb{P}(Y \prec y^l | X = x^*, C = c) < \rho(y; x', x, c)\Big).\end{aligned} \tag{43}$$

Remark 4. When $\mathcal{I}_Y$ is a closed intervel $[y^l, y^u]$ in evidence $\mathcal{E}$ or $\mathcal{E}'$, the identification results are obtained by changing "$Y \prec y^u$" to "$Y \preceq y^u$" in Theorems 3 and 4.

Remark 5. When $\mathcal{I}_Y$ is a point $y^l = y^u$, the identification of T-PNS with evidence $(X = x^*, Y = y^l)$ in Theorem 4 reduces to Theorem 5.1 in (Kawakami, Kuroki, and Tian 2024). Thus, T-PNS identification in Theorem 4 is an extension of Theorem 5.1 in (Kawakami, Kuroki, and Tian 2024).

Simulated Experiments

Estimation from Finite Sample Size

We perform numerical experiments to illustrate the properties of the estimators from finite sample size. Theoretically, the estimators in this paper are consistent and it is expected that the estimates are reliable when the sample size is large.

Estimation methods. All identification theorems in the paper compute all quantities through conditional CDFs. Using dataset $\{x_i, m_i, y_i\}_{i=1}^N$, we estimate the conditional CDFs by the empirical conditional CDFs, i.e., $\hat{\mathbb{P}}(Y \prec y | X = x, M = m) = \sum_{i=1}^N \mathbb{I}(y_i \prec y, x_i = x, m_i = m) / \sum_{i=1}^N \mathbb{I}(x_i = x, m_i = m)$, $\hat{\mathbb{P}}(M = m | X = x) = \sum_{i=1}^N \mathbb{I}(m_i = m, x_i = x) / \sum_{i=1}^N \mathbb{I}(x_i = x)$, and, in addition, $\hat{\rho}(y; x', x) = \sum_{m \in \Omega_M} \hat{\mathbb{P}}(Y \prec y | X = x', M = m) \hat{\mathbb{P}}(M = m | X = x)$. We conduct the bootstrapping (Efron 1979) to reveal the distribution of the estimators, and provide the means and 95% confidential intervals (CI) for each estimator.

Setting. We assume the following SCM:

$$X := \text{Bern}(0.5), M := \text{Bern}(\pi(X)), \\ Y := \text{Bern}(\pi(X + M)), \tag{44}$$

where $\pi(x) = \exp(1 + 0.5x)/(1 + \exp(1 + 0.5x))$. $\text{Bern}(z)$ represents a Bernoulli distribution with probability z. X, M, and Y are all binary variables. We simulate 1000 times with the sample size $N = 100, 1000, 10000$, respectively, and assess the means and 95% confidential intervals (CIs) of the estimators.

Results. The ground truths of T-PNS, ND-PNS, and NI-PNS are 0.074, 0.066, and 0.008. When $N = 100$, the estimates are

T-PNS: 0.083 (CI: $[0.000, 0.228]$),
ND-PNS: 0.074 (CI: $[0.000, 0.220]$),
NI-PNS: 0.009 (CI: $[0.000, 0.046]$).

When $N = 1000$, the estimates are

T-PNS: 0.075 (CI: $[0.029, 0.125]$),
ND-PNS: 0.068 (CI: $[0.021, 0.116]$),
NI-PNS: 0.007 (CI: $[0.000, 0.017]$).

When $N = 10000$, the estimates are

T-PNS: 0.074 (CI: $[0.060, 0.088]$),
ND-PNS: 0.067 (CI: $[0.052, 0.082]$),
NI-PNS: 0.008 (CI: $[0.005, 0.011]$).

When the sample size is small ($N = 100$), the estimators have relatively wide 95% CIs. When the sample size is large enough ($N = 1000$ or $N = 10000$), the estimators are close to the ground truths and have relatively narrow 95% CIs. We perform additional experiments for T-PN, ND-PN, NI-PN, T-PS, ND-PS, and NI-PS and the results are presented in Appendix F in (Kawakami and Tian 2024).

Illustration of the Proposed Measures

To illustrate the behavior of the proposed direct and indirect PoC measures, we simulate data from an SCM and plot the measures against the covariate. The results are discussed in Appendix F in (Kawakami and Tian 2024).

Application to a Real-world Dataset

We show an application to a real-world psychology dataset.

Dataset. We take up a dataset from the Job Search Intervention Study (JOBS II) (Vinokur and Schul 1997). This dataset is open through the R package "mediation" (https://cran.r-project.org/web/packages/mediation/index.html). JOBS II was a randomized job training intervention for unemployed subjects aiming at increasing the prospect of reemployment and improving their mental health. In the experiment, the unemployed workers were randomly assigned to treatment and control groups. Those in the treatment group participated in job-skills workshops, and they learned job-search skills and coping strategies for dealing with setbacks in the job-search process. Those in the control group received a booklet of job-search tips. In follow-up interviews, a measure of depressive symptoms based on the Hopkins Symptom Checklist was assessed. The sample size is 899 with no missing values.

Variables. Let the randomly assigned interventions be treatment variable (X) (`treat`), which takes 0 for the control group and 1 for the treatment group. We choose the measure of depressive symptoms based on the Hopkins Symptom Checklist (`depress2`) as the outcome (Y). We consider job-search self-efficacy (M) (`job_seek`) as a discrete mediating variable. We set $C = \emptyset$. We let the threshold of the depression be $y = 3$ in all the definitions of PoC variants, and let $x' = 0$ and $x = 1$. We assume Assumptions 1-3. These are reasonable because the interventions are randomly assigned, and the linear model used in the previous study (Vinokur and Schul 1997) satisfies these assumptions. On this dataset, it is reasonable that $X = 0$ increases the depression compared to $X = 1$ and we assume 4' and 5' for monotonic increasing. Assumption 4' for monotonic increasing represents $\mathbb{P}(Y_{1,m} \succ y \succeq Y_{0,m} | C = c) = 0$, which means that there do not exist subjects whose potential depression score when setting the value of job-search self-efficiency by m and receiving no intervention is under the given threshold y, and whose potential depression score when setting the value of job-search self-efficiency by

m and receiving an intervention is over the given threshold y. This seems reasonable. Assumption 5' for monotonic increasing represents $\mathbb{P}(Y_{1,M_1} \succ y \succeq Y_{1,M_0}|C=c)=0$, $\mathbb{P}(Y_{1,M_1} \succ y \succeq Y_{0,M_1}|C=c)=0$, $\mathbb{P}(Y_{1,M_1} \succ y \succeq Y_{0,M_0}|C=c)=0$, $\mathbb{P}(Y_{1,M_0} \succ y \succeq Y_{0,M_0}|C=c)=0$, and $\mathbb{P}(Y_{0,M_1} \succ y \succeq Y_{0,M_0}|C=c)=0$. For example, $\mathbb{P}(Y_{1,M_1} \succ y \succeq Y_{1,M_0}|C=c)=0$ means that there do not exist subjects whose potential depression score when receiving an intervention and keeping the value of job-search self-efficiency by M_0 is under the given threshold y, and whose potential depression score when receiving an intervention and keeping the value of job-search self-efficiency by M_1 is over the given threshold y. This also seems reasonable.

Results. The estimated T-PNS is 23.840% (CI: [19.021%,29.254%]). Then, we consider the following three questions:

(**Q1'**). *Would the intervention be necessary and sufficient to cure the depression had the job-search self-efficacy been fixed to a value* $(m=5)$*?*

(**Q2'**). *Would the intervention still be necessary and sufficient to cure the depression had there been no influence via the job-search self-efficacy?*

(**Q3'**). *Would the intervention still be necessary and sufficient to cure the depression had the influence only existed via the job-search self-efficacy?*

We evaluate CD-PNS $(m=5)$, ND-PNS, and NI-PNS specified in Def. 3 and obtain the following results:

CD-PNS: 7.484% (CI: [0.000%,41.676%]),
ND-PNS: 0.000% (CI: [0.000%,0.000%]),
NI-PNS: 23.840% (CI: [19.021%,29.254%]).

CD-PNS, ND-PNS, and NI-PNS answer the questions (Q1'), (Q2'), and (Q3'), respectively. CD-PNS is less than T-PNS. T-PNS is equal to NI-PNS, and this means that the necessity and sufficiency of the treatment is entirely due to the indirect influence via the mediator. The proportions of ND-PNS and NI-PNS in T-PNS are 0 and 1.

While Vinokur and Schul (1997) reported both direct and indirect effects as statistically significant, our results decompose the total influence entirely into the indirect. However, this does not contradict the observation that treatment has a direct effect on the outcome. Our results imply that the treatment would be necessary and sufficient at the same level of T-PNS had the influence only existed via the mediator, and the treatment would not be necessary and sufficient had there been no influence via the mediator. Our results do not imply that the treatment has no direct effect on outcome.

Next, we study PoC for a specific subpopulation described by evidence. We evaluate T-PNS, CD-PNS $(m=5)$, ND-PNS, and NI-PNS with evidence specified in Def. 4. We consider the evidence of $x^*=0, \mathcal{I}_Y=[y^l,y^u)$ where $y^l=1.5$ and $y^u=2.5$ for ND-PNS and NI-PNS, and additionally $m^*=5$ for CD-PNS. We obtain:

T-PNS with evidence: 57.899%(CI: [39.130%,76.190%]),
CD-PNS with evidence: 0.000%(CI: [0.000%,0.000%])[2],
ND-PNS with evidence: 0.000%(CI: [0.000%,0.000%]),
NI-PNS with evidence: 57.899%(CI: [39.130%,76.190%]).

CD-PNS, ND-PNS, and NI-PNS can answer the questions (Q1'), (Q2'), and (Q3'), respectively, for the subpopulation specified by the evidence. CD-PNS is 0, and the proportions of ND-PNS and NI-PNS in T-PNS are 0 and 1. T-PNS and NI-PNS for this subpopulation are larger than those of the whole population.

Conclusion

We consider mediation analysis for PoC and introduce new direct and indirect variants of PoC to represent the necessity and sufficiency of the treatment to produce an outcome event directly or through a mediator. We provide identification theorems for each type of PoC we introduce. The results expand the family of PoC and provide tools for researchers to answer more sophisticated causal questions. In addition, we show in Appendix D in (Kawakami and Tian 2024) how these direct and indirect variants of PoC look like for binary treatment, outcome, and mediator variables. In settings where the identification assumptions (sequential ignorability, monotonicity) do not hold, bounding (Tian and Pearl 2000; Dawid, Musio, and Murtas 2017; Dawid, Humphreys, and Musio 2024; Li and Pearl 2024a) or sensitivity analysis (Imai, Keele, and Tingley 2010; Imai, Keele, and Yamamoto 2010; VanderWeele 2016) is desired. Also, researchers are often interested in path-specific effects, of which direct and indirect effects are special instances (Daniel et al. 2015; Xia and Chan 2022; Zhou and Yamamoto 2023). Extending our results to these cases will be interesting future work.

Acknowledgements

The authors thank the anonymous reviewers for their time and thoughtful comments.

References

Avin, C.; Shpitser, I.; and Pearl, J. 2005. Identifiability of path-specific effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, 357–363. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Baron, R. M.; and Kenny, D. A. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6): 1173.

Daniel, R. M.; De Stavola, B. L.; Cousens, S. N.; and Vansteelandt, S. 2015. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1): 1–14.

[2]The result of bootstrap CI width 0 is due to the "max" function in the estimators. In Eq. 30, CD-PNS with evidence are identified using the "max" function, i.e., $\max\{\cdot,0\}$. If the upper bound of the bootstrap CI of the inside value of the "max" function is negative (i.e., 95% chance the inside value is in a range all negative), the estimated CD-PNS is 0% with bootstrap CI width 0. The interpretation is that CD-PNS is 0% with 95% confidence.

Dawid, A. P.; Murtas, R.; and Musio, M. 2014. Bounding the Probability of Causation in Mediation Analysis. *ArXiv*, abs/1411.2636.

Dawid, A. P.; Musio, M.; and Murtas, R. 2017. The probability of causation1. *Law, Probability and Risk*, 16(4): 163–179.

Dawid, P.; Humphreys, M.; and Musio, M. 2024. Bounding Causes of Effects With Mediators. *Sociological Methods & Research*, 53(1): 28–56.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1 – 26.

Galhotra, S.; Pradhan, R.; and Salimi, B. 2021. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, 577–590. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383431.

Hannart, A.; and Naveau, P. 2018. Probabilities of Causation of Climate Changes. *Journal of Climate*, 31(14): 5507–5524.

Imai, K.; Keele, L.; and Tingley, D. 2010. A general approach to causal mediation analysis. *Psychol Methods*, 15(4): 309–334.

Imai, K.; Keele, L.; and Yamamoto, T. 2010. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1): 51 – 71.

Kawakami, Y.; Kuroki, M.; and Tian, J. 2024. Probabilities of Causation for Continuous and Vector Variables. *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI-2024)*.

Kawakami, Y.; Shingaki, R.; and Kuroki, M. 2023. Identification and Estimation of the Probabilities of Potential Outcome Types Using Covariate Information in Studies with Non-compliance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10): 12234–12242.

Kawakami, Y.; and Tian, J. 2024. Mediation Analysis for Probabilities of Causation. arXiv:2412.14491.

Kuroki, M.; and Cai, Z. 2011. Statistical Analysis of 'Probabilities of Causation' Using Co-variate Information. *Scandinavian Journal of Statistics*, 38(3): 564–577.

Li, A.; and Pearl, J. 2019. Unit Selection Based on Counterfactual Logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 1793–1799. International Joint Conferences on Artificial Intelligence Organization.

Li, A.; and Pearl, J. 2022. Unit Selection with Causal Diagram. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5): 5765–5772.

Li, A.; and Pearl, J. 2024a. Probabilities of Causation with Nonbinary Treatment and Effect. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-2024)*.

Li, A.; and Pearl, J. 2024b. Unit Selection with Nonbinary Treatment and Effect. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-2024)*.

Li, W.; Lu, Z.; Jia, J.; Xie, M.; and Geng, Z. 2023. Retrospective causal inference with multiple effect variables. *Biometrika*, 111(2): 573–589.

Lu, Z.; Geng, Z.; Li, W.; Zhu, S.; and Jia, J. 2022. Evaluating causes of effects by posterior effects of causes. *Biometrika*, 110(2): 449–465.

Malinsky, D.; Shpitser, I.; and Richardson, T. 2019. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 3080–3088. PMLR.

Murtas, R.; Dawid, A. P.; and Musio, M. 2017. New bounds for the Probability of Causation in Mediation Analysis. *arXiv: Statistics Theory*.

Pearl, J. 1999. Probabilities Of Causation: Three Counterfactual Interpretations And Their Identification. *Synthese*, 121(1): 93–149.

Pearl, J. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, 411–420. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Robins, J.; and Greenland, S. 1989. The Probability of Causation under a Stochastic Model for Individual Risk. *Biometrics*, 45(4): 1125–1138.

Robins, J. M.; and Greenland, S. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2): 143–155.

Rubinstein, M.; Cuellar, M.; and Malinsky, D. 2024. Mediated probabilities of causation. *arXiv preprint arXiv:2404.07397*.

Shin, D. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146: 102551.

Shingaki, R.; and Kuroki, M. 2021. Identification and Estimation of Joint Probabilities of Potential Outcomes in Observational Studies with Covariate Information. In *Advances in Neural Information Processing Systems*, volume 34, 26475–26486. Curran Associates, Inc.

Shpitser, I. 2013. Counterfactual Graphical Models for Longitudinal Mediation Analysis With Unobserved Confounding. *Cognitive Science*, 37(6): 1011–1035.

Shpitser, I.; and Pearl, J. 2008. Complete Identification Methods for the Causal Hierarchy. *J. Mach. Learn. Res.*, 9: 1941–1979.

Tchetgen, E. J. T.; and Shpitser, I. 2012. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *The Annals of Statistics*, 40(3): 1816 – 1845.

Tian, J.; and Pearl, J. 2000. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1): 287–313.

VanderWeele, T. J. 2013. Policy-relevant proportions for direct effects. *Epidemiology*, 24(1): 175–176.

VanderWeele, T. J. 2016. Mediation analysis: a practitioner's guide. *Annual review of public health*, 37(1): 17–32.

VanderWeele, T. J.; and Knol, M. J. 2014. A Tutorial on Interaction. *Epidemiologic Methods*, 3(1): 33–72.

Vinokur, A. D.; and Schul, Y. 1997. Mastery and inoculation against setbacks as active ingredients in the JOBS intervention for the unemployed. *Journal of consulting and clinical psychology*, 65(5): 867.

Watson, D. S.; Gultchin, L.; Taly, A.; and Floridi, L. 2021. Local explanations via necessity and sufficiency: unifying theory and practice. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, 1382–1392. PMLR.

Wright, S. 1921. Correlation and causation. *Journal of agricultural research*, 20(7): 557–585.

Wright, S. 1934. The method of path coefficients. *The annals of mathematical statistics*, 5(3): 161–215.

Xia, F.; and Chan, K. C. G. 2022. Decomposition, identification and multiply robust estimation of natural mediation effects with multiple mediators. *Biometrika*, 109(4): 1085–1100.

Zhou, X.; and Yamamoto, T. 2023. Tracing Causal Paths from Experimental and Observational Data. *The Journal of Politics*, 85(1): 250–265.