Identification and Estimation of "Probabilities of Causation" in the Presence of Confounding and Selection Bias

Ryusei Shingaki **Haruka Yoshida** **Manabu Kuroki**
Graduate School of Engineering Science, Yokohama National University
79-5 Tokiwadai, Hodogaya-ku, Yokohama 240-8501 Japan
shingaki.ryusei@gmail.com, yoshida-haruka-kt@ynu.jp, manabu-kuroki-zm@ynu.ac.jp

Abstract

Probabilities of causation are valuable concepts for explainable artificial intelligence (XAI) and personalized decision-making. Pearl (2009) defined the probabilities of causation from the viewpoint of "necessity", "sufficiency", and "necessity and sufficiency" in the context of structural causal models. In addition, Tian and Pearl (2000) and Kuroki and Cai (2011) provided the identification conditions of the probabilities of causation under the monotonicity assumption. However, these identification conditions are described based on "the joint probabilities of observed random variables" and/or "causal risks" without selection biases. Thus, they are not applicable to studies in the presence of confounding and selection biases. To address this problem, this paper provides novel identification conditions for the probabilities of causation by using (i) two proxy covariates and (ii) an instrumental variable and a proxy covariate. When the probabilities of causation can be evaluated through the proposed identification conditions, new plug-in estimators for these probabilities are presented. Finally, we illustrate the application of our results on a real-world dataset.

1 INTRODUCTION

A selection bias is defined as

> Bias of the estimated effect of an exposure on an outcome due to conditioning on a common effect of the exposure and the outcome (Porta, 2014).

A selection bias arises when the selected population is not representative of the target population. It causes the phenomenon that the statistical association between the exposure variable and the outcome variable in the selected population differs from the statistical association in the target population, which often leads to serious misjudgments about the cause–effect relationship between the exposure variable and the outcome variable. Thus, to address this issue, there has been vigorous discussion regarding the selection bias problem in statistical causal inference.

To establish causation in the presence of confounding and selection biases, this paper focuses on the identification and estimation problems of the probabilities of causation because, in recent decades, these problems have received increasing attention in the fields of artificial intelligence, epidemiology, legal reasoning, policy analysis, risk analysis, and statistical science (e.g., Byrne, 2019; Cai and Kuroki, 2005; Cox, 1984; Dawid et al., 2017; Galhotra et al., 2021; Greenland, 1987; Kommiya Mothilal et al., 2021; Pearl, 2009; Robins and Greenland, 1989; Tian and Pearl, 2000; Watson et al., 2021). For example, recently, in the field of explainable artificial intelligence (XAI), it has been noted that "necessity", "sufficiency", and "necessity and sufficiency" are important explanatory concepts, and the probabilities of causation play an important role in establishing these concepts from probabilistic aspects (Galhotra et al., 2021; Kommiya Mothilal et al., 2021; Watson et al., 2021).

Pearl (2009) developed formal semantics for probabilities of causation based on structural causal models and presented formal definitions for the probability of necessity (PN), the probability of sufficiency (PS), and the probability of necessity and sufficiency (PNS). These probabilities of causation are formulated based on the joint probabilities of two potential outcome

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

variables. Since one cannot simultaneously observe the results of the same subjects exposed and unexposed in reality, even successful randomized experiments have failed to identify these quantities. To solve this problem, Pearl (2009), Tian and Pearl (2000), and Kuroki and Cai (2011) show how to bound the probabilities of causation from statistical data. Pearl (2009) and Tian and Pearl (2000) also stated that the probabilities of causation are identifiable if monotonicity can be assumed, and if the causal risks of interest are identifiable.

However, the studies mentioned above are restricted to situations where the available statistical data are provided in the form of "joint probabilities of observed random variables" and/or "causal risks" without a selection bias. This means that the current identification conditions do not hold in the presence of confounding and selection biases. Although there have been significant requirements to establish causation in the presence of confounding and selection biases, much less has been discussed about how to identify the probabilities of causation.

To address this problem, this paper provides novel identification conditions for the probabilities of causation by using (i) two proxy covariates and (ii) the instrumental variable and a proxy covariate. Here, covariates are random variables not affected by an exposure variable, and proxy covariates are a type of covariates used instead of a set of covariates of interest when such a set cannot be measured directly. In addition, the instrumental variable is a variable that (i) is not associated with all the observed and unobserved covariates, and (ii) is associated with the exposure variable but has no effect on the outcome variable except through its association with the exposure variable (Greenland, 2000).

Unlike Tian and Pearl (2000) and Kuroki and Cai (2011), who discussed the evaluation problem of the probabilities of causation in studies without selection bias, the proposed identification conditions allow us to estimate the probabilities of causation from observed data in the presence of both confounding and selection bias, without assuming monotonicity. Thus, our results extend Shingaki and Kuroki (2021, 2023) and Kawakami (2021), who addressed the evaluation problem of the probabilities of potential outcome types in settings without selection bias. Moreover, it is noteworthy that, in certain situations, the probabilities of the potential outcome types are also identifiable by using an instrumental variable together with a proxy covariate, even in studies with confounding and selection bias.

When the probabilities of causation can be evaluated through the proposed identification conditions, the estimation problem of the probabilities of causation reduces to that of singular models, and thus, they can not be evaluated by standard statistical estimation methods, such as the maximum likelihood estimation method. To solve the problem, new plug-in estimators of these probabilities are also presented. Finally, we illustrate the application of our results on a real-world dataset. The results of this paper extend the applicability of statistical causal inference to practical science. Given space constraints, the proofs and numerical experiments are provided in the Supplementary Material.

2 PRELIMINARIES

This section introduces the potential outcome variables used to address our problems. Here, we assume that the readers are familiar with the basic theory of statistical causal inference (Imbens and Rubin, 2015; Pearl, 2009). For details on the basic theory of statistical causal inference and graph-theoretic terminology used in this paper, we refer readers to Imbens and Rubin (2015) and Pearl (2009).

Letting N be the sample size, we assume that X and Y represent the observed dichotomous exposure variable and the observed dichotomous outcome variable, respectively. For the exposure status $x \in \{x_0, x_1\}$ (where x_1 is the occurrence of an event, and where x_0 is the non-occurrence of an event) and the outcome status $y \in \{y_0, y_1\}$ (where y_1 is the occurrence of an outcome, and where y_0 is the non-occurrence of an outcome), let $p(X = x, Y = y) = p(x, y)$ be the joint probability of $(X, Y) = (x, y)$, $p(X = x) = p(x)$ be the marginal probability of $X = x$, and $p(Y = y \,|\, X = x) = p(y \,|\, x) = p(x, y)/p(x)$ be the conditional probability of $Y = y$ given $X = x$. A similar notation is used for other probabilities. In principle, for $x \in \{x_0, x_1\}$, the i-th of the N subjects has a potential outcome variable $Y_x(i)$ that would have resulted if X had been x for the i-th subject. Here, $Y_x(i) = y$ denotes the counterfactual statement that "Y would be y had X been x for the i-th subject." For each subject, only the observed outcome $Y(i)$ is available, reflecting the fact that counterfactual outcomes are not jointly observable. Separately, we assume the consistency property (Pearl, 2009), which is formally defined as

$$X(i) = x \quad \Longrightarrow \quad Y_x(i) = Y(i) \tag{1}$$

for the i-th subject.

In this work, we assume a stable unit treatment value assumption (Imbens and Rubin, 2015), which can be summarized as follows: (i) the exposure status of any subject does not affect the outcome status of the other subjects, and (ii) the exposure statuses of all subjects are comparable. When N subjects in the study are considered as random samples from the population of interest, $Y_x(i)$ is referred to as the value of a random variable, Y_x. The causal risk of $X = x$ on $Y = y$ is defined as $p(Y_x = y)$.

When an ideal randomized experiment with X is feasible, X is independent of (Y_{x_0}, Y_{x_1}), i.e., $p(x, Y_{x_0} = y, Y_{x_1} = y') = p(x)p(Y_{x_0} = y, Y_{x_1} = y')$ holds for $x \in \{x_0, x_1\}$ and $y, y' \in \{y_0, y_1\}$. The causal risk $p(Y_x = y)$ is then identifiable and is given by $p(Y_x = y) = p(y \mid x)$. Here, "identifiable" means that the causal quantities, such as $p(Y_x = y)$, can be estimated consistently from a probability of observed random variables. In contrast, when it is difficult to conduct ideal randomized experiments, we can still evaluate the causal risk according to the conditionally ignorable treatment assignment condition (Imbens and Rubin, 2015) or, graphically, the back-door criterion (Pearl, 2009). In other words, for the exposure variable X, if there exists a set $\boldsymbol{Z}$ of random variables such that X is conditionally independent of (Y_{x_0}, Y_{x_1}) given $\boldsymbol{Z}$, i.e.,

$$\begin{aligned} &p(x, Y_{x_0} = y, Y_{x_1} = y' \mid \boldsymbol{Z} = \boldsymbol{z}) \\ &= p(x \mid \boldsymbol{Z} = \boldsymbol{z}) p(Y_{x_0} = y, Y_{x_1} = y' \mid \boldsymbol{Z} = \boldsymbol{z}) \end{aligned}$$

holds for $x \in \{x_0, x_1\}$, $y, y' \in \{y_0, y_1\}$ and any value of $\boldsymbol{z}$ taken by $\boldsymbol{Z}$, then we say that the treatment assignment is conditionally ignorable given $\boldsymbol{Z}$, or $\boldsymbol{Z}$ is a sufficient set of confounders. When a sufficient set $\boldsymbol{Z}$ of confounders is observed, the causal risks $p(Y_x = y)$ are identifiable and are given by

$$p(Y_x = y) = \sum_{\boldsymbol{z}} p(y \mid x, \boldsymbol{z}) p(\boldsymbol{z}). \tag{2}$$

Generally, a confounding bias arises when a sufficient set of confounders is unavailable, and often leads to severe misjudgments about the cause–effect relationship between the exposure and outcome variables. Although other identification conditions can be used to evaluate causal risk in such situations (e.g., Pearl, 2009; Tian and Pearl, 2002), we do not cover them in this paper because of space constraints.

For the potential outcome variables Y_{x_1} and Y_{x_0}, according to Lash et al. (2021), we can consider the following four potential outcome types:

$u_1 = (Y_{x_1} = y_1, Y_{x_0} = y_1)$ represents a "doomed" situation where the exposure status is irrelevant in the sense that the outcome of interest occurs $(Y = y_1)$ with or without the event of interest.

$u_2 = (Y_{x_1} = y_1, Y_{x_0} = y_0)$ represents a "causative" situation where the outcome of interest occurs $(Y = y_1)$ if and only if the event of interest occurs $(X = x_1)$.

$u_3 = (Y_{x_1} = y_0, Y_{x_0} = y_1)$ represents a "preventive" situation where the outcome of interest occurs $(Y = y_1)$ if and only if the event of interest does not occur $(X = x_0)$.

$u_4 = (Y_{x_1} = y_0, Y_{x_0} = y_0)$ represents an "immune" situation where the exposure status is irrelevant in the sense that the outcome of interest does not occur $(Y = y_0)$ with or without the event of interest.

Here, the assumption of $p(u_3) = 0$ is called monotonicity by Pearl (2009). The central aim of this paper is to solve the identification and estimation problems of the probabilities of causation without such an assumption.

Following Pearl (2009), we define three probabilities of causation, namely, the probability of necessity (PN), the probability of sufficiency (PS), and the probability of necessity and sufficiency (PNS).

The PN is defined as

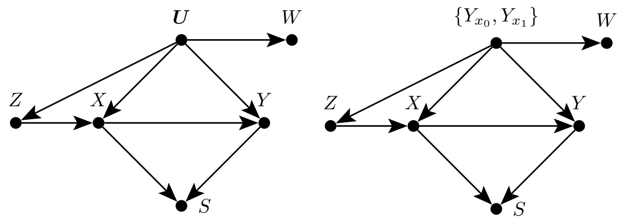
$$\mathrm{PN} = p(Y_{x_0} = y_0 \mid x_1, Y_{x_1} = y_1) = p(Y_{x_0} = y_0 \mid x_1, y_1), \tag{3}$$

which stands for the probability that the outcome of interest would not have occurred $(Y = y_0)$ in the absence of the event of interest $(X = x_0)$, given that $X = x_1$ and $Y = y_1$ did, in fact, occur. PN has applications in artificial intelligence, epidemiology, and legal reasoning. Epidemiologists have long been concerned with estimating the probability that a certain case of a disease is attributable to a particular exposure, which is normally interpreted counterfactually as "the probability that the disease would not have occurred in the absence of exposure, given that the disease and exposure did in fact occur." This probability is evaluated by the PN.

The PS is defined as

$$\mathrm{PS} = p(Y_{x_1} = y_1 \mid x_0, Y_{x_0} = y_0) = p(Y_{x_1} = y_1 \mid x_0, y_0), \tag{4}$$

which stands for the probability that the outcome of interest would have occurred $(Y = y_1)$ in the presence of the event of interest $(X = x_1)$, given that $X = x_0$ and $Y = y_0$ did, in fact, occur. The PS has applications in artificial intelligence, policy analysis, and risk analysis. Policy-makers may be interested in the dangers that a certain exposure may present to the healthy population (Khoury et al., 1989). Counterfactually, this notion is interpreted as the "probability that a healthy unexposed subject would have gotten



(a) Unobserved covariates (b) Potential outcomes

Figure 1: Graphical representation for the case of two proxy covariates.

the disease had they been exposed." This probability is evaluated by the PS.

The PNS is defined as

$$\mathrm{PNS} = p(Y_{x_0} = y_0, Y_{x_1} = y_1), \tag{5}$$

which represents the probability that $X = x_1$ is a necessary and sufficient cause for $Y = y_1$. Therefore, it measures both the sufficiency and the necessity of $X = x_1$ to produce $Y = y_1$.

Since the PN, the PS, and the PNS involve joint probabilities of two potential outcome variables, they are not estimable from statistical data, even under ideal randomized experiments, without any further causal information (Pearl, 2009).

3 IDENTIFICATION

3.1 Two proxy covariates

According to Pearl (2009), letting $\boldsymbol{U}$ be the set of all discrete and continuous covariates that could affect X and Y, both observed and unobserved, we consider the identification problem of the probabilities of causation based on the directed acyclic graph shown in Figure 1a.

Covariates Z and W, in Figure 1a, are measured as proxy covariates of $\boldsymbol{U}$. Here, note that both Z and W can be a set of discrete and/or continuous covariates. In Figure 1a, the directed edge from X to Y ($X \to Y$) indicates that X could affect Y. In addition, the absence of a directed edge from Y to X indicates that Y cannot be a cause of X, and the directed path from $\boldsymbol{U}$ to Y through X ($\boldsymbol{U} \to X \to Y$) indicates that some elements of $\boldsymbol{U}$ could affect Y mediated by X. In the context of the structural causal model, Figure 1a also graphically represents the data-generating process (cause–effect relationships)

$$X = g_x(\boldsymbol{U}, Z, \epsilon_x), \quad Y = g_y(X, \boldsymbol{U}, \epsilon_y),$$
$$Z = g_z(\boldsymbol{U}, \epsilon_z), \quad W = g_w(\boldsymbol{U}, \epsilon_w), \quad S = g_s(X, Y, \epsilon_s),$$

where ϵ_x, ϵ_y, ϵ_z, ϵ_w, and ϵ_s are independent random disturbances and are independent of $\boldsymbol{U}$.

The situation, such as Figure 1a, is also discussed in Kuroki and Pearl (2014) and Pearl (2010). However, in this paper, unlike Kuroki and Pearl (2014) and Pearl (2010), $\boldsymbol{U}$ can include an uncertain number of discrete and continuous covariates. Thus, in many situations, it is reasonable to assume the existence of proxy covariates Z and W, and that W is conditionally independent of $\{X, Y, Z\}$ given $\boldsymbol{U}$.

In the situation shown in Figure 1a, S shows a selection variable that selects subjects from the target population in such a way that the ideal randomized experiment is not successful (s_1 indicates the selected population and s_0 indicates the non-selected population). When the statistical data from the selected population labeled $S = s_1$ but not the target population are available, such statistical data would generate the statistical association between X and Y that differs in the target population. However, such a selection through S never causes X or Y because S could be a common effect of X and Y. Taking this into account, the impact of $\boldsymbol{U} \cup \{\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w\}$ on Y is classified into the following four potential outcome types, u_1 ("doomed"), u_2 ("causative"), u_3 ("preventive"), and u_4 ("immune"). When it is re-defined as a random variable U taking a value $u \in \{u_1, u_2, u_3, u_4\}$, similar to the setting of the non-compliance problem by Pearl (2009, Ch.8), throughout the paper, we assume that $p(x, y, s, z, w)$ is given as

$$\begin{aligned} &p(x, y, s, z, w) \\ &= p(s \mid x, y) \\ &\times \sum_{i=1}^{4} p(y \mid x, u_i) p(x \mid z, u_i) p(z \mid u_i) p(w \mid u_i) p(u_i) \end{aligned} \tag{6}$$

according to Figure 1b. It is also assumed that other joint probabilities of X, Y, S, Z, and W can be rewritten in a similar way. Then, this factorization implicitly implies that X is conditionally independent of W, given U, denoted as $X \perp\!\!\!\perp W \mid U$, which is used in the context of the impact evaluation problem of the program (Nguyen, 2016). In addition, from the consistency property, we have

$$\begin{aligned} p(x_1, y_0, z, w, u_i \mid s_1) &= 0 \quad \text{for } i = 1, 2, \\ p(x_1, y_1, z, w, u_i \mid s_1) &= 0 \quad \text{for } i = 3, 4, \\ p(x_0, y_0, z, w, u_i \mid s_1) &= 0 \quad \text{for } i = 1, 3, \\ p(x_0, y_1, z, w, u_i \mid s_1) &= 0 \quad \text{for } i = 2, 4. \end{aligned} \tag{7}$$

Thus, for example, the conditional probabilities $p(x_1, y_1, z, w \mid s_1)$ can be rewritten as

$$\begin{aligned} &p(x_1, y_1, z, w \mid s_1) \\ &= \sum_{i=1}^{2} p(w \mid u_i) p(z \mid x_1, y_1, u_i) p(u_i \mid x_1, y_1) p(x_1, y_1 \mid s_1). \end{aligned} \tag{8}$$

In addition, the other conditional probabilities of X, Y, Z, and W given $S = s_1$ can be rewritten similarly.

To solve the identification and estimation problems of the probabilities of causation, we derive the following theorem:

Theorem 1. *Letting S, Z, and W be random variables taking the values $s \in \{s_0, s_1\}$ for S, $k(\geq 2)$ values $z \in \{z_1, z_2, \ldots, z_k\}$ for Z, and $l(\geq 3)$ values $w \in \{w_1, w_2, \ldots, w_l\}$ for W, respectively, $p(Y_x = y \,|\, x', y')$ are identifiable for $x, x' \in \{x_0, x_1\}$ and $y, y' \in \{y_0, y_1\}$ from the positive observed conditional probabilities $p(x, y, z, w \,|\, s_1)$ of $(X, Y, Z, W) = (x, y, z, w)$ given $S = s_1$ if the following conditions are satisfied:*

Condition 1. *For $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_1, z_2, \ldots, z_k\}$, $w \in \{w_1, w_2, \ldots, w_l\}$ and $u \in \{u_1, u_2, u_3, u_4\}$,*

$$\begin{aligned} &p(x, y, s_1, w, z, u) \\ &= p(s_1 \,|\, x, y)p(y \,|\, x, u)p(x \,|\, z, u)p(w \,|\, u)p(z \,|\, u)p(u) \end{aligned} \tag{9}$$

hold.

Condition 2. *For $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_1, z_2, \ldots, z_k\}$, and $w \in \{w_1, w_2, \ldots, w_l\}$,*

$$Q_{zw \cdot xy} = \begin{pmatrix} 1 & p(z \,|\, x, y, s_1) \\ p(w \,|\, x, y, s_1) & p(z, w \,|\, x, y, s_1) \end{pmatrix} \tag{10}$$

are invertible.

Condition 3. *For $z \in \{z_1, z_2, \ldots, z_k\}$,*

$$\begin{aligned} &\frac{|Q_{zw_1 \cdot x_1 y_1}|}{|Q_{zw_2 \cdot x_1 y_1}|} \neq \frac{|Q_{zw_1 \cdot x_0 y_1}|}{|Q_{zw_2 \cdot x_0 y_1}|}, \ \frac{|Q_{zw_1 \cdot x_1 y_1}|}{|Q_{zw_2 \cdot x_1 y_1}|} \neq \frac{|Q_{zw_1 \cdot x_0 y_0}|}{|Q_{zw_2 \cdot x_0 y_0}|}, \\ &\frac{|Q_{zw_1 \cdot x_1 y_0}|}{|Q_{zw_2 \cdot x_1 y_0}|} \neq \frac{|Q_{zw_1 \cdot x_0 y_1}|}{|Q_{zw_2 \cdot x_0 y_1}|}, \ \frac{|Q_{zw_1 \cdot x_1 y_0}|}{|Q_{zw_2 \cdot x_1 y_0}|} \neq \frac{|Q_{zw_1 \cdot x_0 y_0}|}{|Q_{zw_2 \cdot x_0 y_0}|}, \end{aligned} \tag{11}$$

where $|\cdot|$ denotes the determinant.

The proof is given in the Supplementary Material. Theorem 1 shows that both the PN and the PS are identifiable under Conditions 1, 2, and 3. In addition, for $p(s_1) = 1$, Theorem 1 includes Shingaki and Kuroki (2021), which discussed the evaluation problem of probabilities of potential outcome types in observational studies with covariate information. In contrast, for $p(s_1) \in (0, 1)$, note that the PNS is not identifiable by Theorem 1 because $p(x, y)$ in the target population, not $p(x, y \,|\, S = s_1)$, is not identifiable under the selection mechanism in Figure 1a. When $p(x, y)$ are given for $x \in \{x_0, x_1\}$ and $y \in \{y_0, y_1\}$ in Theorem 1, the probabilities of potential outcome types $p(u_1)$, $p(u_2)$, $p(u_3)$, and $p(u_4)$ are identifiable, that is, the PNS is also identifiable.

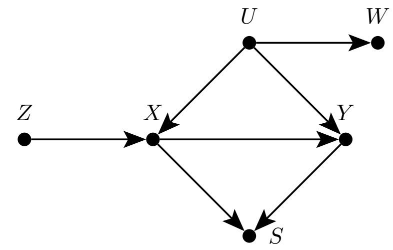


Figure 2: Graphical representation for the case of the IV together with one proxy covariate.

3.2 Proxy covariate with an instrumental variable

Kawakami (2021) noted that the combination of the instrumental variable (IV) and a proxy covariate plays a powerful role in identifying probabilities of potential outcome types from statistical data. In this section, we show that the IV, together with a proxy covariate, also enables us to identify the PN, the PS and the PNS from statistical data in the presence of confounding and selection biases. The main points are that (i) we use the IV (Z) with three or more categories as a covariate, and (ii) we use a proxy covariate (W) with three or more categories.

To see this, for the directed acyclic graph shown in Figure 2, we assume that X and Y are observed dichotomous random variables, where X represents the treatment actually received ($x \in \{x_1, x_0\}$; x_1 is the occurrence of an event, while x_0 is the non-occurrence of an event) and Y represents the observed outcome variable ($y \in \{y_1, y_0\}$; y_1 is the occurrence of an outcome, while y_0 is the non-occurrence of an outcome). Z is seen analogously to the randomized treatment assignment ($z \in \{z_1, z_2, z_3\}$). W and U represent observed proxy covariates with three or more categories ($w \in \{w_1, w_2, w_3\}$) and a random variable indicating potential outcome types ($u \in \{u_1, u_2, u_3, u_4\}$), that are not affected by X or Y, respectively. In this situation, counterfactually, Z satisfies the instrumental variable (IV) assumption: (i) Z is not independent of X, (ii) Z is independent of U, and (iii) for positive observed conditional probabilities $p(y, z \,|\, x, u)$ ($x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_1, z_2, z_3\}$, $u \in \{u_1, u_2, u_3, u_4\}$),

$$p(y, z \,|\, x, u) = p(z \,|\, x, u)p(y \,|\, x, u) \tag{12}$$

holds.

In this setting, equations (6), (7), and (8) also hold in Figure 2 under the constraints $p(y \,|\, x, z, u) = p(y \,|\, x, u)$, $p(z \,|\, u) = p(z)$ and $p(z \,|\, x, y, u) = p(z \,|\, x, u)$ for $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_1, z_2, z_3\}$ and $u \in \{u_1, u_2, u_3, u_4\}$). This observation shows that if

Conditions 2 and 3 of Theorem 1 and

$$
\begin{aligned}
&p(x, y, s_1, w, z, u) \\
&= p(s_1 \mid x, y) p(y \mid x, u) p(x \mid z, u) p(w \mid u) p(z) p(u)
\end{aligned} \tag{13}
$$

hold in Figure 2, then both the PN and the PS are identifiable from the statistical data in the presence of confounding and selection biases, which is shown as follows:

Corollary 1. *Letting S, Z, and W be random variables taking the values $s \in \{s_0, s_1\}$ for S, $k(\geq 3)$ values $z \in \{z_1, z_2, \ldots, z_k\}$ for Z, and $l(\geq 3)$ values $w \in \{w_1, w_2, \ldots, w_l\}$ for W, respectively, $p(Y_x = y \mid x', y')$ are identifiable for $x, x' \in \{x_0, x_1\}$ and $y, y' \in \{y_0, y_1\}$ from the positive observed conditional probabilities $p(x, y, w, z \mid s_1)$ of $(X, Y, Z, W) = (x, y, z, w)$ given $S = s_1$, if Conditions 2 and 3 of Theorem 1 and equation (13) hold for $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_1, z_2, \ldots, z_k\}$, $w \in \{w_1, w_2, \ldots, w_l\}$, and $u \in \{u_1, u_2, u_3, u_4\}$.*

The proof is achieved via a similar procedure to that of Theorem 1.

Remarkably, different from the previous section where the PNS is not identifiable in Figure 1b because $p(x, y)$ is not available for any x and y, Figure 2 represents that the PNS is identifiable through the combination of the instrumental variable Z with three categories and a proxy covariate W with three categories. Notably, Z and W are independent of each other in Figure 2. Then, we have

$$
\begin{aligned}
&\sum_{x,y} p(z, w \mid x, y) p(x, y) \\
&= \left(\sum_{x,y} p(w \mid x, y) p(x, y) \right) \left(\sum_{x,y} p(z \mid x, y) p(x, y) \right)
\end{aligned} \tag{14}
$$

for any z and w. Here, note that $p(z, w \mid x, y)$, $p(z \mid x, y)$ and $p(w \mid x, y)$ are estimable from $p(z, w \mid x, y, s_1)$ since S is conditionally independent of $\{Z, W\}$ given $\{X, Y\}$. Then, equation (14) is regarded as a set of equations for three unknown parameters $p(x_1, y_1)$, $p(x_1, y_0)$, and $p(x_0, y_1)$ ($p(x_0, y_0) = 1 - p(x_1, y_1) - p(x_1, y_0) - p(x_0, y_1)$). Thus, when equation (14) includes at least three non-degenerate equations, Bézout's theorem (e.g., Cox et al., 2015) implies that there exist at most eight solutions for $p(x, y)$. Equation (14) can be solved by systematic polynomial elimination methods (e.g., Gröbner bases), rather than by naive substitution. In addition, as demonstrated in Section B of the Supplementary Material, noting that the degrees of freedom for the $k \times l$ contingency table of Z and W given $S = s_1$ are $(k-1) \times (l-1)$, if it contains non-degenerate equations satisfying $(k-1)(l-1) > 8$, equation (14) admits a unique solution for $p(x, y)$. This result establishes that the identification problem of the probabilities of potential outcome types can also be addressed by using an instrumental variable together with a proxy covariate. For instance, the PNS can be expressed as

$$
\mathrm{PNS} = \mathrm{PN} \times p(x_1, y_1) + \mathrm{PS} \times p(x_0, y_0). \tag{15}
$$

In contrast, when equation (14) is degenerate (for example, when W and Z are independent of X and Y), equation (14) admits either infinitely many solutions or none. Importantly, whether equation (14) is non-degenerate is empirically testable, since $p(z, w \mid x, y)$, $p(z \mid x, y)$, and $p(w \mid x, y)$ can be consistently estimated from $p(z, w \mid x, y, s_1)$. For further details, see Section B of the Supplementary Material.

4 ESTIMATION

When the probabilities of causation are identifiable through the proposed conditions, according to the proofs of Theorem 1 (refer to the Supplemental Material A), we propose new plug-in estimators of PN and PS from statistical data in the presence of confounding and selection biases.

To do this, for $x \in \{x_1, x_0\}$, $y \in \{y_1, y_0\}$, $z \in \{z_1, z_2, \ldots, z_k\}$, and $w \in \{w_1, w_2, \ldots, w_l\}$, let $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$, and $\widehat{p}(z, w \mid x, y, s_1)$ be the maximum likelihood estimators of $p(z \mid x, y) = p(z \mid x, y, s_1)$, $p(w \mid x, y) = p(w \mid x, y, s_1)$, and $p(z, w \mid x, y) = p(z, w \mid x, y, s_1)$, respectively. Let

$$
\begin{aligned}
a_{xy} &= \frac{|Q_{zw_1 \cdot xy}|}{|Q_{zw_2 \cdot xy}|}, \\
b_{xy} &= \frac{1}{|Q_{zw_2 \cdot xy}|} \{ p(z, w_2 \mid x, y, s_1) p(w_1 \mid x, y, s_1) \\
&\qquad - p(z, w_1 \mid x, y, s_1) p(w_2 \mid x, y, s_1) \}.
\end{aligned}
$$

Note that the consistent estimators $\widehat{a}_{xy}$ and $\widehat{b}_{xy}$ of a_{xy} and b_{xy}, respectively, are derived by replacing $p(z \mid x, y, s_1)$, $p(w \mid x, y, s_1)$, and $p(z, w \mid x, y, s_1)$ of a_{xy} and b_{xy} with the maximum likelihood estimators $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$, and $\widehat{p}(z, w \mid x, y, s_1)$, respectively ($x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_1, z_2, \ldots, z_k\}$, $w \in \{w_1, w_2, \ldots, w_l\}$). Then, from equations (A.28)–(A.31) in the Supplementary Material A, the consistent estimators $\widehat{p}(w \mid u)$ of $p(w \mid u)$ are given by

$$\begin{pmatrix} 1 & -\widehat{a}_{x_1y_1} \\ 1 & -\widehat{a}_{x_0y_1} \end{pmatrix} \begin{pmatrix} \widehat{p}(w_1 \mid u_1) \\ \widehat{p}(w_2 \mid u_1) \end{pmatrix} = \begin{pmatrix} \widehat{b}_{x_1y_1} \\ \widehat{b}_{x_0y_1} \end{pmatrix},$$
$$\begin{pmatrix} 1 & -\widehat{a}_{x_1y_1} \\ 1 & -\widehat{a}_{x_0y_0} \end{pmatrix} \begin{pmatrix} \widehat{p}(w_1 \mid u_2) \\ \widehat{p}(w_2 \mid u_2) \end{pmatrix} = \begin{pmatrix} \widehat{b}_{x_1y_1} \\ \widehat{b}_{x_0y_0} \end{pmatrix},$$
$$\begin{pmatrix} 1 & -\widehat{a}_{x_1y_0} \\ 1 & -\widehat{a}_{x_0y_1} \end{pmatrix} \begin{pmatrix} \widehat{p}(w_1 \mid u_3) \\ \widehat{p}(w_2 \mid u_3) \end{pmatrix} = \begin{pmatrix} \widehat{b}_{x_1y_0} \\ \widehat{b}_{x_0y_1} \end{pmatrix}, \tag{16}$$
$$\begin{pmatrix} 1 & -\widehat{a}_{x_1y_0} \\ 1 & -\widehat{a}_{x_0y_0} \end{pmatrix} \begin{pmatrix} \widehat{p}(w_1 \mid u_4) \\ \widehat{p}(w_2 \mid u_4) \end{pmatrix} = \begin{pmatrix} \widehat{b}_{x_1y_0} \\ \widehat{b}_{x_0y_0} \end{pmatrix}.$$

Since we now derive the consistent estimators $\widehat{p}(w \mid u)$, $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$ and $\widehat{p}(z, w \mid x, y, s_1)$ of $p(w \mid u)$, $p(z \mid x, y, s_1)$, $p(w \mid x, y, s_1)$ and $p(z, w \mid x, y, s_1)$, respectively, by replacing $p(w \mid u)$, $p(z \mid x, y, s_1)$, $p(w \mid x, y, s_1)$ and $p(z, w \mid x, y, s_1)$ in equation (A.32) in the Supplementary Material A with $\widehat{p}(w \mid u)$, $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$ and $\widehat{p}(z, w \mid x, y, s_1)$, respectively ($x \in \{x_0, x_1\}, y \in \{y_0, y_1\}, w \in \{w_1, w_2, \ldots, w_l\}, u \in \{u_1, u_2, u_3, u_4\}$), the consistent estimator $\widehat{p}(u \mid x, y)$ of $p(u \mid x, y)$ is given by

$$\begin{aligned}
\widehat{p}(u_2 \mid x_1, y_1) &= \frac{\widehat{p}(w \mid x_1, y_1, s_1) - \widehat{p}(w \mid u_1)}{\widehat{p}(w \mid u_2) - \widehat{p}(w \mid u_1)}, \\
\widehat{p}(u_3 \mid x_0, y_1) &= \frac{\widehat{p}(w \mid x_0, y_1, s_1) - \widehat{p}(w \mid u_1)}{\widehat{p}(w \mid u_3) - \widehat{p}(w \mid u_1)}, \\
\widehat{p}(u_3 \mid x_1, y_0) &= \frac{\widehat{p}(w \mid u_4) - \widehat{p}(w \mid x_1, y_0, s_1)}{\widehat{p}(w \mid u_4) - \widehat{p}(w \mid u_3)}, \\
\widehat{p}(u_2 \mid x_0, y_0) &= \frac{\widehat{p}(w \mid u_4) - \widehat{p}(w \mid x_0, y_0, s_1)}{\widehat{p}(w \mid u_4) - \widehat{p}(w \mid u_2)}
\end{aligned} \tag{17}$$

and

$$\begin{aligned}
\widehat{p}(u_1 \mid x_1, y_1) &= 1 - \widehat{p}(u_2 \mid x_1, y_1), \\
\widehat{p}(u_1 \mid x_0, y_1) &= 1 - \widehat{p}(u_3 \mid x_0, y_1), \\
\widehat{p}(u_4 \mid x_1, y_0) &= 1 - \widehat{p}(u_3 \mid x_1, y_0), \\
\widehat{p}(u_4 \mid x_0, y_0) &= 1 - \widehat{p}(u_2 \mid x_0, y_0).
\end{aligned} \tag{18}$$

Thus, from equations (A.33) in Appendix A, the consistent estimators of $p(Y_x = y \mid x', y')$ ($x, x' \in \{x_0, x_1\}, y, y' \in \{y_0, y_1\}$) are given by

$$\begin{aligned}
\widehat{p}(Y_{x_0} = y_0 \mid x_1, y_1) &= \widehat{p}(u_2 \mid x_1, y_1), \\
\widehat{p}(Y_{x_1} = y_0 \mid x_0, y_1) &= \widehat{p}(u_3 \mid x_0, y_1), \\
\widehat{p}(Y_{x_0} = y_1 \mid x_1, y_0) &= \widehat{p}(u_3 \mid x_1, y_0), \\
\widehat{p}(Y_{x_1} = y_1 \mid x_0, y_0) &= \widehat{p}(u_2 \mid x_0, y_0).
\end{aligned} \tag{19}$$

In Appendix C, under the situation of Figure 2, we also provide the estimation method of the PNS through estimating $p(x, y)$ by using (14).

5 CASE STUDY

This section illustrates our results through a real-world dataset obtained from GEUVADIS (Genetic European Variation in Disease) data, reported by Lappalainen et al. (2013), and re-analyzed by Badsha and Fu (2019) and Badsha et al. (2021)[1]. The GEUVADIS project (Lappalainen et al., 2013) was conducted to characterize regulatory variation in different human populations. An average of 49M paired-end reads were generated from the total RNA extracted from 465 lymphoblastoid cell lines collected as part of the 1000 Genomes project. In this section, we analyzed the gene SBF2-AS1 in the eQTL gene set #50 with the single nucleotide polymorphism (SNP) rs7124238, and the SWAP70 gene for 358 Europeans and 87 Africans. The variables of interest are as follows:

- X**:** Gene expression for SWAP70 (x_1: upregulation; x_0: downregulation),
- Y**:** Gene expression for SBF2-AS1 (y_1: upregulation; y_0: downregulation),
- Z**:** Number of alternative alleles for SNP rs7124238 (z_1: homozygous for the reference allele; z_2: heterozygous; z_3: homozygous for the alternative allele),
- W**:** Population (w_1: British in England and Scotland, Finnish in Finland, or Toscani in Italy; w_2: Utah residents with Northern and Western European ancestry; w_3: Yoruba in Ibadan, Nigeria),
- S**:** A selection variable that selects subjects from the target population with probabilities $p(s_1 \mid x_0, y_0) = 0.25$, $p(s_1 \mid x_1, y_0) = 0.50$, $p(s_1 \mid x_0, y_1) = 0.75$, and $p(s_1 \mid x_1, y_1) = 1.00$.

Badsha et al. (2021) treat Z as an instrumental variable (IV) and argue that the data-generating process $Z \to X \to Y$ is a plausible underlying mechanism for the GEUVADIS data. In line with Badsha et al. (2021), we assume that the data-generating process is represented as in Figure 2, and we additionally introduce the selection variable S to demonstrate the applicability of the proposed estimation method. The variable W represents coarse population labels (e.g., British/Finnish/Toscani, CEU, Yoruba). These labels are not biological determinants of gene expression for SWAP70 or SBF2-AS1. Rather, they capture differences in haplotype structure, linkage disequilibrium (LD) blocks, and allele-frequency patterns—features

[1]The data are included in R package `MRPC` `https://cran.r-project.org/package=MRPC` (License: GPL-2 | GPL-3)

that are naturally interpreted as components of the unobserved variable $\boldsymbol{U}$. Because $\boldsymbol{U}$ is latent, using W as a proxy for $\boldsymbol{U}$ is consistent with standard proxy-variable approaches for unmeasured confounding (e.g., Miao et al., 2018; Alabdulmohsin et al., 2023). In such frameworks, a proxy variable may carry information about an unobserved confounder without exerting direct causal influence on either the exposure or the outcome. This corresponds exactly to the relations encoded in Figure 2 and assumed in Theorem 1.

Considering that Conditions 1, 2, and 3 of Theorem 1 hold, the bootstrap estimates of the PN, PS, and PNS via the proposed method based on the selection bias data are 0.450, 0.506, and 0.269, respectively. Thus, the estimated PN implies that, for subjects whose SWAP70 and SBF2-AS1 expression were actually upregulated, 45.0% would have a downregulated SBF2-AS1 expression if the SWAP70 expression had been downregulated. In other words, the probability that the upregulation of SWAP70 is a necessary cause of the upregulation of SBF2-AS1 is 0.450. Similarly, the estimated PS implies that, for subjects whose SWAP70 and SBF2-AS1 expression were actually downregulated, 50.6% of the subjects would have an upregulated SBF2-AS1 expression if the SWAP70 expression had been upregulated. In other words, the probability that the upregulation of SWAP70 is a sufficient cause of the upregulation of SBF2-AS1 is 0.506. Here, the estimated PNS shows that the probability that the upregulation of SWAP70 is a necessary and sufficient cause of the upregulation of SBF2-AS1 is 0.269. In contrast, the bootstrap estimates of the PN, PS, and PNS based on the data without selection bias are 0.516, 0.560, and 0.346, respectively. There seems to be no significant difference between the estimated PN, PS, and PNS with, and the estimated PN, PS, and PNS without a selection bias.

Table 1 shows the basic statistics of 100,000 bootstrapped replications of the estimates of the PN, PS, and PNS. From Table 1, it seems that the estimated PN and PS are distributed uniformly but the estimated PNSs are relatively skewed. In addition, since many outliers are generated, the estimated PN and PS have a wide sample range, and the minimum and maximum values provide no information on PN and PS.

Table 1: Basic statistics of bootstrapped estimates for "probabilities of causation" from the GEUVADIS data.

	without selection bias			with selection bias		
	PN	PS	PNS	PN	PS	PNS
Minimum	0.000	0.000	0.000	0.000	0.000	0.000
1st Quantile	0.277	0.357	0.101	0.217	0.282	0.123
Median	0.523	0.584	0.156	0.429	0.512	0.191
Mean	0.516	0.560	0.167	0.450	0.506	0.200
3rd Quantile	0.757	0.780	0.220	0.670	0.733	0.268
Maximum	1.000	1.000	0.889	1.000	1.000	0.893
s.e.	0.283	0.265	0.092	0.275	0.275	0.101

6 DISCUSSION

Probabilities of causation often appear in various fields of practical science, e.g., identification problems of prevented and preventable proportions (Yamada and Kuroki, 2017) and proportion of benefit and harm (Mueller et al., 2022), XAI problems (Galhotra et al., 2021; Kommiya Mothilal et al., 2021; Watson et al., 2021), evaluation problems of the traffic conflicts (Yamada and Kuroki, 2019), personalized decision making problems (Mueller and Pearl, 2023), and unit selection problems (Li and Pearl, 2019). These applications show that the identification and estimation problems of these probabilities have been important topics in statistical causal inference. To solve the problems in the presence of confounding and selection biases, this paper provided novel identification conditions for the probabilities of causation by using (i) two proxy covariates and (ii) an instrumental variable (IV) and a proxy covariate. To the best of our knowledge, in the presence of confounding and selection biases, although there is a significant requirement to solve these problems without the assumption of monotonicity, there has not been much discussion on how we evaluate the probabilities of causation. In addition, to solve the estimation problem of these probabilities, we propose new plug-in estimators for the PN, PS, and PNS based on the proposed identification conditions.

A natural question is whether point identification offers a practical advantage over the sharp bounds for PN, PS, and PNS developed by Tian and Pearl (2000) and later refined by Li and Pearl (2019). These bounds are informative and often tight under weaker assumptions. However, even narrow bounds may include values that reverse the optimal decision (e.g., zero for PN or PNS), which renders decision-making inherently unstable. This issue persists even with arbitrarily large sample sizes, because the residual uncertainty reflects a lack of identification rather than sampling variability. To illustrate, consider a setting in which the true probability of causation is small but positive, while the identifiable bounds include zero. In such cases, a decision maker who accounts for worst-case scenarios may rationally conclude that the intervention has no effect and refrain from acting. By contrast, under the proposed identification conditions, the relevant probabilities become point-identified, and the corresponding estimators converge to their true values as the sample size increases. This removes the ambiguity induced by bounds that include outcome-reversing values and can therefore lead to qualitatively different decisions.

Finally, to extend the identifiability of probabilities of causation in the presence of confounding and selection biases, it would be necessary to discuss how to use (a) an intermediate variable and (b) external information. Regarding (a), one of our ideas is to extend the mediation-based identification conditions of the probabilities of causation in the presence of a confounding bias (Shingaki and Kuroki, 2024). Regarding (b), it would be better to refer to Bareinboim and Tian (2015); Correa et al. (2018, 2019) to derive novel identification conditions of the probabilities of causation. Such extensions will be left for future work.

Acknowledgements

This research was partially supported by a Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science, Grant Numbers 24K14851. We thank the anonymous reviewers for their insightful comments.

References

Alabdulmohsin, I., Chiou, N., D'Amour, A., Gretton, A., Koyejo, S., Kusner, M. J., Pfohl, S. R., Salaudeen, O., Schrouff, J., and Tsai, K. (2023). Adapting to latent subgroup shifts via concepts and proxies. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 9637–9661. PMLR.

Badsha, M. B. and Fu, A. Q. (2019). Learning causal biological networks with the principle of mendelian randomization. *Frontiers in Genetics*, 10.

Badsha, M. B., Martin, E. A., and Fu, A. Q. (2021). MRPC: An R package for inference of causal graphs. *Frontiers in Genetics*, 12.

Bareinboim, E. and Tian, J. (2015). Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pages 3475–3481.

Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6276–6282. International Joint Conferences on Artificial Intelligence Organization.

Cai, Z. and Kuroki, M. (2005). Variance estimators for three "probabilities of causation". *Risk Analysis*, 25(6):1611–1620.

Correa, J., Tian, J., and Bareinboim, E. (2018). Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 6355–6342.

Correa, J. D., Tian, J., and Bareinboim, E. (2019). Identification of causal effects in the presence of selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2744–2751.

Cox, D. A., Little, J., and O'Shea, D. (2015). *Ideals, Varieties, and Algorithms*. Undergraduate Texts in Mathematics. Springer, fourth edition.

Cox, L. A. (1984). Probability of causation and the attributable proportion risk. *Risk Analysis*, 4(3):221–230.

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

Driscoll, T. A. and Braun, R. J. (2017). *Fundamentals of Numerical Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

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, page 577–590, New York, NY, USA. Association for Computing Machinery.

Greenland, S. (1987). Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. *Statist. Med.*, 6(6):701–708.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kawakami, Y. (2021). Instrumental variable-based identification for causal effects using covariate information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12131–12138.

Khoury, M. J., Flanders, W. D., Greenland, S., and Adams, M. J. (1989). On the measurement of susceptibility in epidemiologic studies. *American Journal of Epidemiology*, 129(1):183–190.

Kommiya Mothilal, R., Mahajan, D., Tan, C., and Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021*

AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, pages 652–663, New York, NY, USA. Association for Computing Machinery.

Kuroki, M. and Cai, Z. (2011). Statistical analysis of 'probabilities of causation' using co-variate information. *Scand. J. Statist.*, 38(3):564–577.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Lash, T. L., VanderWeele, T. J., Haneuse, S., and Rothman, K. J. (2021). *Modern Epidemiology*. Lippincott Williams & Wilkins, 4th edition.

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*, pages 1793–1799. International Joint Conferences on Artificial Intelligence Organization.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Mueller, S., Li, A., and Pearl, J. (2022). Causes of effects: Learning individual responses from population data. In Raedt, L. D., editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2712–2718. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Mueller, S. and Pearl, J. (2023). Personalized decision making – a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050.

Nguyen, C. (2016). An introduction to alternative methods in program impact evaluation. *Journal of Economic and Social Thought*, 3(3):349–375.

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

Pearl, J. (2010). On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432, Arlington, Virginia, USA. AUAI Press.

Porta, M. (2014). *A Dictionary of Epidemiology*. Oxford University Press.

Robins, J. and Greenland, S. (1989). The probability of causation under a stochastic model for individual risk. *Biometrics*, 45(4):1125–1138.

Shingaki, R. and Kuroki, M. (2021). Identification and estimation of joint probabilities of potential outcomes in observational studies with covariate information. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26475–26486. Curran Associates, Inc.

Shingaki, R. and Kuroki, M. (2023). Probabilities of potential outcome types in experimental studies: Identification and estimation based on proxy covariate information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12287–12294.

Shingaki, R. and Kuroki, M. (2024). Identification and estimation of "causes of effects" using covariate-mediator information. In *International Conference on Artificial Intelligence and Statistics*, pages 3574–3582. PMLR.

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

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 567–573, USA. American Association for Artificial Intelligence.

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

Yamada, K. and Kuroki, M. (2017). Counterfactual-based prevented and preventable proportions. *Journal of Causal Inference*, 5.

Yamada, K. and Kuroki, M. (2019). New traffic conflict measure based on a potential outcome model. *Journal of Causal Inference*, 7(1).

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Identification and Estimation of “Probabilities of Causation” in the Presence of Confounding and Selection Bias (Supplementary Material)

A Proof of Theorem 1

For $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_1, z_2, \dots, z_k\}$, and $w \in \{w_1, w_2, \dots, w_l\}$, since we have

$$p(z, w, u \mid x, y, s_1) = p(z, w, u \mid x, y) \tag{A.20}$$

from Condition 1 in Theorem, $p(z, w, u \mid x, y, s_1)$ is equivalent to $p(z, w, u \mid x, y)$. Then, we have

$$\begin{aligned}
p(z, w \mid x_1, y_1) &= p(z, w \mid x_1, y_1, s_1) = p(z \mid u_1, x_1, y_1)p(w \mid u_1)p(u_1 \mid x_1, y_1) + p(z \mid u_2, x_1, y_1)p(w \mid u_2)p(u_2 \mid x_1, y_1),\\
p(z, w \mid x_0, y_1) &= p(z, w \mid x_0, y_1, s_1) = p(z \mid u_1, x_0, y_1)p(w \mid u_1)p(u_1 \mid x_0, y_1) + p(z \mid u_3, x_0, y_1)p(w \mid u_3)p(u_3 \mid x_0, y_1),\\
p(z, w \mid x_1, y_0) &= p(z, w \mid x_1, y_0, s_1) = p(z \mid u_3, x_1, y_0)p(w \mid u_3)p(u_3 \mid x_1, y_0) + p(z \mid u_4, x_1, y_0)p(w \mid u_4)p(u_4 \mid x_1, y_0),\\
p(z, w \mid x_0, y_0) &= p(z, w \mid x_0, y_0, s_1) = p(z \mid u_2, x_0, y_0)p(w \mid u_2)p(u_2 \mid x_0, y_0) + p(z \mid u_4, x_0, y_0)p(w \mid u_4)p(u_4 \mid x_0, y_0)
\end{aligned} \tag{A.21}$$

for $z \in \{z_1, z_2, \dots, z_k\}$ and $w \in \{w_1, w_2, \dots, w_l\}$, since, by the consistency property,

$$\begin{aligned}
p(z, u_i \mid x_1, y_1, s_1) &= 0 \quad \text{for } i = 3, 4,\\
p(z, u_i \mid x_0, y_1, s_1) &= 0 \quad \text{for } i = 2, 4,\\
p(z, u_i \mid x_1, y_0, s_1) &= 0 \quad \text{for } i = 1, 2,\\
p(z, u_i \mid x_0, y_0, s_1) &= 0 \quad \text{for } i = 1, 3
\end{aligned} \tag{A.22}$$

hold for $z \in \{z_1, z_2, \dots, z_k\}$, $w \in \{w_1, w_2, \dots, w_l\}$, and $u \in \{u_1, u_2, u_3, u_4\}$. Then, we prove Theorem 1 through the following steps.

Step 1: Identification of $p(w \mid u)$

From equations (A.21), regarding $p(z \mid x_1, y_1, s_1)$, $p(w \mid x_1, y_1, s_1)$, and $p(z, w \mid x_1, y_1, s_1)$, we have

$$\begin{aligned}
&\begin{pmatrix} 1 & p(z \mid x_1, y_1, s_1) \\ p(w \mid x_1, y_1, s_1) & p(z, w \mid x_1, y_1, s_1) \end{pmatrix}\\
&= \begin{pmatrix} 1 & 1 \\ p(w \mid u_1) & p(w \mid u_2) \end{pmatrix} \begin{pmatrix} p(u_1 \mid x_1, y_1) & 0 \\ 0 & p(u_2 \mid x_1, y_1) \end{pmatrix} \begin{pmatrix} 1 & p(z \mid u_1, x_1, y_1) \\ 1 & p(z \mid u_2, x_1, y_1) \end{pmatrix}
\end{aligned} \tag{A.23}$$

for $w \in \{w_1, w_2\}$. From Condition 2 in Theorem 1, note that

$$\begin{pmatrix} 1 & p(z \mid x_1, y_1, s_1) \\ p(w \mid x_1, y_1, s_1) & p(z, w \mid x_1, y_1, s_1) \end{pmatrix}$$

is invertible for $z \in \{z_1, z_2\}$ and $w \in \{w_1, w_2, w_3\}$. We derive

$$\begin{aligned}
&\begin{pmatrix} 1 & p(z \mid x_1, y_1, s_1) \\ p(w_1 \mid x_1, y_1, s_1) & p(z, w_1 \mid x_1, y_1, s_1) \end{pmatrix} \begin{pmatrix} 1 & p(z \mid x_1, y_1, s_1) \\ p(w_2 \mid x_1, y_1, s_1) & p(z, w_2 \mid x_1, y_1, s_1) \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 \\ p(w_2 \mid u_1) & p(w_2 \mid u_2) \end{pmatrix}\\
&= \begin{pmatrix} 1 & 0 \\ b_{x_1y_1} & a_{x_1y_1} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ p(w_2 \mid u_1) & p(w_2 \mid u_2) \end{pmatrix}\\
&= \begin{pmatrix} 1 & 1 \\ b_{x_1y_1} + p(w_2 \mid u_1)a_{x_1y_1} & b_{x_1y_1} + p(w_2 \mid u_2)a_{x_1y_1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ p(w_1 \mid u_1) & p(w_1 \mid u_2) \end{pmatrix},
\end{aligned}$$

i.e.,

$$p(w_1 \mid u_1) = b_{x_1y_1} + p(w_2 \mid u_1)a_{x_1y_1}, \quad p(w_1 \mid u_2) = b_{x_1y_1} + p(w_2 \mid u_2)a_{x_1y_1}, \tag{A.24}$$

where

$$\begin{aligned} a_{xy} &= \frac{p(z, w_1 \mid x, y, s_1) - p(z \mid x, y, s_1)p(w_1 \mid x, y, s_1)}{p(z, w_2 \mid x, y, s_1) - p(z \mid x, y, s_1)p(w_2 \mid x, y, s_1)}, \\ b_{xy} &= \frac{p(z, w_2 \mid x, y, s_1)p(w_1 \mid x, y, s_1) - p(z, w_1 \mid x, y, s_1)p(w_2 \mid x, y, s_1)}{p(z, w_2 \mid x, y, s_1) - p(z \mid x, y, s_1)p(w_2 \mid x, y, s_1)} \end{aligned}$$

for $x \in \{x_0, x_1\}$ and $y \in \{y_0, y_1\}$. Similarly, from equations (A.21), we have

$$p(w_1 \mid u_1) = b_{x_0y_1} + p(w_2 \mid u_1)a_{x_0y_1}, \quad p(w_1 \mid u_3) = b_{x_0y_1} + p(w_2 \mid u_3)a_{x_0y_1}, \tag{A.25}$$

$$p(w_1 \mid u_3) = b_{x_1y_0} + p(w_2 \mid u_3)a_{x_1y_0}, \quad p(w_1 \mid u_4) = b_{x_1y_0} + p(w_2 \mid u_4)a_{x_1y_0}, \tag{A.26}$$

$$p(w_1 \mid u_2) = b_{x_0y_0} + p(w_2 \mid u_2)a_{x_0y_0}, \quad p(w_1 \mid u_4) = b_{x_0y_0} + p(w_2 \mid u_4)a_{x_0y_0}. \tag{A.27}$$

Thus, from Condition 3 in Theorem 1, $p(w_1 \mid u_1)$ and $p(w_2 \mid u_1)$ are identifiable from equations (A.24) and (A.25) as

$$\begin{pmatrix} p(w_1 \mid u_1) \\ p(w_2 \mid u_1) \end{pmatrix} = \begin{pmatrix} 1 & -a_{x_1y_1} \\ 1 & -a_{x_0y_1} \end{pmatrix}^{-1} \begin{pmatrix} b_{x_1y_1} \\ b_{x_0y_1} \end{pmatrix}. \tag{A.28}$$

Similarly, $p(w_1 \mid u_2)$, $p(w_2 \mid u_2)$, $p(w_1 \mid u_3)$, $p(w_2 \mid u_3)$, $p(w_1 \mid u_4)$, and $p(w_2 \mid u_4)$ are identifiable from equations (A.24)–(A.27) as

$$\begin{pmatrix} p(w_1 \mid u_2) \\ p(w_2 \mid u_2) \end{pmatrix} = \begin{pmatrix} 1 & -a_{x_1y_1} \\ 1 & -a_{x_0y_0} \end{pmatrix}^{-1} \begin{pmatrix} b_{x_1y_1} \\ b_{x_0y_0} \end{pmatrix}, \tag{A.29}$$

$$\begin{pmatrix} p(w_1 \mid u_3) \\ p(w_2 \mid u_3) \end{pmatrix} = \begin{pmatrix} 1 & -a_{x_1y_0} \\ 1 & -a_{x_0y_1} \end{pmatrix}^{-1} \begin{pmatrix} b_{x_1y_0} \\ b_{x_0y_1} \end{pmatrix}, \tag{A.30}$$

$$\begin{pmatrix} p(w_1 \mid u_4) \\ p(w_2 \mid u_4) \end{pmatrix} = \begin{pmatrix} 1 & -a_{x_1y_0} \\ 1 & -a_{x_0y_0} \end{pmatrix}^{-1} \begin{pmatrix} b_{x_1y_0} \\ b_{x_0y_0} \end{pmatrix}, \tag{A.31}$$

respectively.

Step 2: Identification of $p(u \mid x, y)$

First, equations (A.21) provide

$$\begin{aligned}
&\begin{pmatrix} 1 & p(w \mid u_1) \\ 1 & p(w \mid u_2) \end{pmatrix}^{-\top} \begin{pmatrix} 1 & p(z \mid x_1, y_1, s_1) \\ p(w \mid x_1, y_1, s_1) & p(z, w \mid x_1, y_1, s_1) \end{pmatrix} = \begin{pmatrix} p(u_1 \mid x_1, y_1) & p(z \mid u_1, x_1, y_1)p(u_1 \mid x_1, y_1) \\ p(u_2 \mid x_1, y_1) & p(z \mid u_2, x_1, y_1)p(u_2 \mid x_1, y_1) \end{pmatrix}, \\
&\begin{pmatrix} 1 & p(w \mid u_1) \\ 1 & p(w \mid u_3) \end{pmatrix}^{-\top} \begin{pmatrix} 1 & p(z \mid x_0, y_1, s_1) \\ p(w \mid x_0, y_1, s_1) & p(z, w \mid x_0, y_1, s_1) \end{pmatrix} = \begin{pmatrix} p(u_1 \mid x_0, y_1) & p(z \mid u_1, x_0, y_1)p(u_1 \mid x_0, y_1) \\ p(u_3 \mid x_0, y_1) & p(z \mid u_3, x_0, y_1)p(u_3 \mid x_0, y_1) \end{pmatrix}, \\
&\begin{pmatrix} 1 & p(w \mid u_3) \\ 1 & p(w \mid u_4) \end{pmatrix}^{-\top} \begin{pmatrix} 1 & p(z \mid x_1, y_0, s_1) \\ p(w \mid x_1, y_0, s_1) & p(z, w \mid x_1, y_0, s_1) \end{pmatrix} = \begin{pmatrix} p(u_3 \mid x_1, y_0) & p(z \mid u_3, x_1, y_0)p(u_3 \mid x_1, y_0) \\ p(u_4 \mid x_1, y_0) & p(z \mid u_4, x_1, y_0)p(u_4 \mid x_1, y_0) \end{pmatrix}, \\
&\begin{pmatrix} 1 & p(w \mid u_2) \\ 1 & p(w \mid u_4) \end{pmatrix}^{-\top} \begin{pmatrix} 1 & p(z \mid x_0, y_0, s_1) \\ p(w \mid x_0, y_0, s_1) & p(z, w \mid x_0, y_0, s_1) \end{pmatrix} = \begin{pmatrix} p(u_2 \mid x_0, y_0) & p(z \mid u_2, x_0, y_0)p(u_2 \mid x_0, y_0) \\ p(u_4 \mid x_0, y_0) & p(z \mid u_4, x_0, y_0)p(u_4 \mid x_0, y_0) \end{pmatrix},
\end{aligned} \tag{A.32}$$

where the notation $-\top$ represents a transposed inverse matrix. Comparing the first rows of both sides in the above equations. We see that $p(u_1 \mid x_1, y_1)$, $p(u_2 \mid x_1, y_1)$, $p(u_1 \mid x_0, y_1)$, $p(u_3 \mid x_0, y_1)$, $p(u_3 \mid x_1, y_0)$, $p(u_4 \mid x_1, y_0)$, $p(u_2 \mid x_0, y_0)$, and $p(u_4 \mid x_0, y_0)$ are identifiable. Note that $p(u_3 \mid x_1, y_1)$, $p(u_4 \mid x_1, y_1)$, $p(u_2 \mid x_0, y_1)$, $p(u_4 \mid x_0, y_1)$, $p(u_1 \mid x_1, y_0)$, $p(u_2 \mid x_1, y_0)$, $p(u_1 \mid x_0, y_0)$, and $p(u_3 \mid x_0, y_0)$ are evaluated by zeros from equations (A.22). From this consideration, we know that $p(u \mid x, y)$ are identifiable for $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, and $u \in \{u_1, u_2, u_3, u_4\}$.

Therefore, $p(Y_x = y \,|\, x', y')$ is identifiable for $(x, y, x', y') = (x_1, y_1, x_0, y_1)$, (x_1, y_1, x_0, y_0), (x_1, y_0, x_0, y_1), (x_1, y_0, x_0, y_0) as

$$\begin{aligned}
p(Y_{x_0} = y_0 \,|\, x_1, y_1) &= p(u_2 \,|\, x_1, y_1) + p(u_4 \,|\, x_1, y_1) = p(u_2 \,|\, x_1, y_1),\\
p(Y_{x_1} = y_0 \,|\, x_0, y_1) &= p(u_3 \,|\, x_0, y_1) + p(u_4 \,|\, x_0, y_1) = p(u_3 \,|\, x_0, y_1),\\
p(Y_{x_0} = y_1 \,|\, x_1, y_0) &= p(u_1 \,|\, x_1, y_0) + p(u_3 \,|\, x_1, y_0) = p(u_3 \,|\, x_1, y_0),\\
p(Y_{x_1} = y_1 \,|\, x_0, y_0) &= p(u_1 \,|\, x_0, y_0) + p(u_2 \,|\, x_0, y_0) = p(u_2 \,|\, x_0, y_0).
\end{aligned} \tag{A.33}$$

B Identification of $p(x, y)$ in Section 3.2

In this section, we note that the degrees of freedom for the $k \times l$ contingency table of Z and W given $S = s_1$ are $(k-1)(l-1)$. Then, we show that if equation (14) corresponds to a non-degenerate system of equations satisfying $(k-1)(l-1) > 8$, then equation (14) admits a unique solution for $p(x, y)$.

To do this, let $p_1 = p(x_0, y_0)$, $p_2 = p(x_1, y_0)$, $p_3 = p(x_0, y_1)$, and $p_4 = p(x_1, y_1)$. We then obtain the following $k \times l + 1$ equations in p_1, p_2, p_3, and p_4:

$$\begin{aligned}
&p(z, w \,|\, x_0, y_0)p_1 + p(z, w \,|\, x_1, y_0)p_2 + p(z, w \,|\, x_0, y_1)p_3 + p(z, w \,|\, x_1, y_1)p_4\\
&= (p(z \,|\, x_0, y_0)p_1 + p(z \,|\, x_1, y_0)p_2 + p(z \,|\, x_0, y_1)p_3 + p(z \,|\, x_1, y_1)p_4)\\
&\quad \times (p(w \,|\, x_0, y_0)p_1 + p(w \,|\, x_1, y_0)p_2 + p(w \,|\, x_0, y_1)p_3 + p(w \,|\, x_1, y_1)p_4)\,, && \text{(B.34)}\\
&p_1 + p_2 + p_3 + p_4 = 1, && \text{(B.35)}
\end{aligned}$$

for $z \in \{z_1, z_2, \ldots, z_k\}$ and $w \in \{w_1, w_2, \ldots, w_l\}$, where $k \geq 2$, $l \geq 2$, and $(k-1)(l-1) > 8$. Then, by substituting $p_4 = 1 - p_1 - p_2 - p_3$ into equation (B.34), we define

$$H = \begin{pmatrix} \boldsymbol{h}(z_1, w_1)^\top \\ \vdots \\ \boldsymbol{h}(z_{k-1}, w_1)^\top \\ \boldsymbol{h}(z_1, w_2)^\top \\ \vdots \\ \boldsymbol{h}(z_{k-1}, w_2)^\top \\ \vdots \\ \boldsymbol{h}(z_1, w_{l-1})^\top \\ \vdots \\ \boldsymbol{h}(z_{k-1}, w_{l-1})^\top \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} p_1^2 \\ p_2^2 \\ p_3^2 \\ p_1p_2 \\ p_1p_3 \\ p_2p_3 \\ p_1 \\ p_2 \\ p_3 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} -(a_0(z_1)b_0(w_1) - c_0(z_1, w_1)) \\ \vdots \\ -(a_0(z_1)b_0(w_{l-1}) - c_0(z_1, w_{l-1})) \\ -(a_0(z_2)b_0(w_1) - c_0(z_2, w_1)) \\ \vdots \\ -(a_0(z_2)b_0(w_{l-1}) - c_0(z_2, w_{l-1})) \\ \vdots \\ -(a_0(z_{k-1})b_0(w_1) - c_0(z_{k-1}, w_1)) \\ \vdots \\ -(a_0(z_{k-1})b_0(w_{l-1}) - c_0(z_{k-1}, w_{l-1})) \end{pmatrix},$$

where

$$\begin{aligned}
\boldsymbol{h}(z, w) = (&a_1(z)b_1(w), a_2(z)b_2(w), a_3(z)b_3(w), a_1(z)b_2(w) + a_2(z)b_1(w), a_1(z)b_3(w) + a_3(z)b_1(w),\\
&a_2(z)b_3(w) + a_3(z)b_2(w), a_1(z)b_0(w) + a_0(z)b_1(w) - c_1(z, w), a_2(z)b_0(w) + a_0(z)b_2(w) - c_2(z, w),\\
&a_3(z)b_0(w) + a_0(z)b_3(w) - c_3(z, w))^\top,
\end{aligned}$$

$$a_1(z) = p(z \mid x_0, y_0) - p(z \mid x_1, y_1), \quad a_2(z) = p(z \mid x_1, y_0) - p(z \mid x_1, y_1),$$
$$a_3(z) = p(z \mid x_0, y_1) - p(z \mid x_1, y_1), \quad a_0(z) = p(z \mid x_1, y_1),$$

$$b_1(w) = p(w \mid x_0, y_0) - p(w \mid x_1, y_1), \quad b_2(w) = p(w \mid x_1, y_0) - p(w \mid x_1, y_1),$$
$$b_3(w) = p(w \mid x_0, y_1) - p(w \mid x_1, y_1), \quad b_0(w) = p(w \mid x_1, y_1),$$

$$c_1(z, w) = p(z, w \mid x_0, y_0) - p(z, w \mid x_1, y_1), \quad c_2(z, w) = p(z, w \mid x_1, y_0) - p(z, w \mid x_1, y_1),$$
$$c_3(z, w) = p(z, w \mid x_0, y_1) - p(z, w \mid x_1, y_1), \quad c_0(z, w) = p(z, w \mid x_1, y_1).$$

Then, $k \times l + 1$ equations (B.34) and (B.35) in p_1, p_2, p_3, and p_4 can be reduced to the following $(k-1) \times (l-1)$ equations in p_1, p_2, and p_3:

$$H\boldsymbol{x} = \boldsymbol{b}. \tag{B.36}$$

Then, when $\text{rank}(H) = \text{rank}(H, \boldsymbol{b}) = 9$, which we refer to as a "non-degenerate equation", a unique solution to equation (B.36) exists and can be expressed as

$$\boldsymbol{x} = (H^\top H)^{-1} H^\top \boldsymbol{b}.$$

Letting $\boldsymbol{x} = (x_1, x_2, \ldots, x_9)$, if each element of $\boldsymbol{x}$ lies within the interval $(0,1)$ and the following constraints are satisfied:

$$x_7 = x_1^2, \quad x_8 = x_2^2, \quad x_9 = x_3^2, \quad x_4 = x_7 x_8, \quad x_5 = x_7 x_9, \quad x_6 = x_8 x_9, \quad x_7 + x_8 + x_9 < 1,$$

then p_1, p_2, and p_3 are identifiable from x_7, x_8, and x_9, respectively, and p_4 can be evaluated as $p_4 = 1 - x_7 - x_8 - x_9$.

C Estimation of the probabilities of potential outcome types

When the probabilities of causation are identifiable through the proposed conditions, according to the proofs of Theorem 1, we propose new plug-in estimators of PN and PS from statistical data in the presence of confounding and selection biases.

For $x \in \{x_1, x_0\}$, $y \in \{y_1, y_0\}$, $z \in \{z_1, z_2, \ldots, z_k\}$, and $w \in \{w_1, w_2, \ldots, w_l\}$, let $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$, and $\widehat{p}(z, w \mid x, y, s_1)$ be the maximum likelihood estimators of $p(z \mid x, y) = p(z \mid x, y, s_1)$, $p(w \mid x, y) = p(w \mid x, y, s_1)$, and $p(z, w \mid x, y) = p(z, w \mid x, y, s_1)$, respectively. Let

$$\begin{aligned}
a_{xy} &= \frac{|Q_{zw_1 \cdot xy}|}{|Q_{zw_2 \cdot xy}|}, \\
b_{xy} &= \frac{1}{|Q_{zw_2 \cdot xy}|} (p(z, w_2 \mid x, y, s_1) p(w_1 \mid x, y, s_1) - p(z, w_1 \mid x, y, s_1) p(w_2 \mid x, y, s_1)).
\end{aligned}$$

Note that the consistent estimators $\widehat{a}_{xy}$ and $\widehat{b}_{xy}$ of a_{xy} and b_{xy}, respectively, are derived by replacing $p(z \mid x, y, s_1)$, $p(w \mid x, y, s_1)$, and $p(z, w \mid x, y, s_1)$ of a_{xy} and b_{xy} with the maximum likelihood estimators $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$, and $\widehat{p}(z, w \mid x, y, s_1)$, respectively ($x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_1, z_2, \ldots, z_k\}$, $w \in \{w_1, w_2, \ldots, w_l\}$). Then, from equations (A.28)–(A.31), the consistent estimators of $p(w \mid u)$ are given by

$$\begin{aligned}
&\begin{pmatrix} \widehat{p}(w_1 \mid u_1) \\ \widehat{p}(w_2 \mid u_1) \end{pmatrix} = \begin{pmatrix} 1 & -\widehat{a}_{x_1y_1} \\ 1 & -\widehat{a}_{x_0y_1} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{b}_{x_1y_1} \\ \widehat{b}_{x_0y_1} \end{pmatrix}, \quad \begin{pmatrix} \widehat{p}(w_1 \mid u_2) \\ \widehat{p}(w_2 \mid u_2) \end{pmatrix} = \begin{pmatrix} 1 & -\widehat{a}_{x_1y_1} \\ 1 & -\widehat{a}_{x_0y_0} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{b}_{x_1y_1} \\ \widehat{b}_{x_0y_0} \end{pmatrix}, \\
&\begin{pmatrix} \widehat{p}(w_1 \mid u_3) \\ \widehat{p}(w_2 \mid u_3) \end{pmatrix} = \begin{pmatrix} 1 & -\widehat{a}_{x_1y_0} \\ 1 & -\widehat{a}_{x_0y_1} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{b}_{x_1y_0} \\ \widehat{b}_{x_0y_1} \end{pmatrix}, \quad \begin{pmatrix} \widehat{p}(w_1 \mid u_4) \\ \widehat{p}(w_2 \mid u_4) \end{pmatrix} = \begin{pmatrix} 1 & -\widehat{a}_{x_1y_0} \\ 1 & -\widehat{a}_{x_0y_0} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{b}_{x_1y_0} \\ \widehat{b}_{x_0y_0} \end{pmatrix}.
\end{aligned} \tag{C.37}$$

Since we now derive the consistent estimators $\widehat{p}(w \mid u)$, $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$ and $\widehat{p}(z, w \mid x, y, s_1)$ of $p(w \mid u)$, $p(z \mid x, y, s_1)$, $p(w \mid x, y, s_1)$ and $p(z, w \mid x, y, s_1)$, respectively, by replacing $p(w \mid u)$, $p(z \mid x, y, s_1)$, $p(w \mid x, y, s_1)$ and $p(z, w \mid x, y, s_1)$ in equation (A.32) with $\widehat{p}(w \mid u)$, $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$ and $\widehat{p}(z, w \mid x, y, s_1)$, respectively ($x \in \{x_0, x_1\}, y \in \{y_0, y_1\}, w \in \{w_1, w_2, w_3\}, u \in \{u_1, u_2, u_3, u_4\}$), the consistent estimator $\widehat{p}(u \mid x, y)$ of $p(u \mid x, y)$ is given by

$$\begin{aligned}
\widehat{p}(u_2 \mid x_1, y_1) &= \frac{\widehat{p}(w \mid x_1, y_1, s_1) - \widehat{p}(w \mid u_1)}{\widehat{p}(w \mid u_2) - \widehat{p}(w \mid u_1)}, \quad \widehat{p}(u_3 \mid x_0, y_1) = \frac{\widehat{p}(w \mid x_0, y_1, s_1) - \widehat{p}(w \mid u_1)}{\widehat{p}(w \mid u_3) - \widehat{p}(w \mid u_1)}, \\
\widehat{p}(u_3 \mid x_1, y_0) &= \frac{\widehat{p}(w \mid u_4) - \widehat{p}(w \mid x_1, y_0, s_1)}{\widehat{p}(w \mid u_4) - \widehat{p}(w \mid u_3)}, \quad \widehat{p}(u_2 \mid x_0, y_0) = \frac{\widehat{p}(w \mid u_4) - \widehat{p}(w \mid x_0, y_0, s_1)}{\widehat{p}(w \mid u_4) - \widehat{p}(w \mid u_2)}
\end{aligned} \tag{C.38}$$

and

$$\begin{aligned}
\widehat{p}(u_1 \mid x_1, y_1) &= 1 - \widehat{p}(u_2 \mid x_1, y_1), \quad \widehat{p}(u_1 \mid x_0, y_1) = 1 - \widehat{p}(u_3 \mid x_0, y_1), \\
\widehat{p}(u_4 \mid x_1, y_0) &= 1 - \widehat{p}(u_3 \mid x_1, y_0), \quad \widehat{p}(u_4 \mid x_0, y_0) = 1 - \widehat{p}(u_2 \mid x_0, y_0).
\end{aligned} \tag{C.39}$$

Thus, from equations (A.33), the consistent estimators of $p(Y_x = y \mid x', y')$ ($x, x' \in \{x_0, x_1\}, y, y' \in \{y_0, y_1\}$) are given by

$$\begin{aligned}
\widehat{p}(Y_{x_0} = y_0 \mid x_1, y_1) &= \widehat{p}(u_2 \mid x_1, y_1), \quad \widehat{p}(Y_{x_1} = y_0 \mid x_0, y_1) = \widehat{p}(u_3 \mid x_0, y_1), \\
\widehat{p}(Y_{x_0} = y_1 \mid x_1, y_0) &= \widehat{p}(u_3 \mid x_1, y_0), \quad \widehat{p}(Y_{x_1} = y_1 \mid x_0, y_0) = \widehat{p}(u_2 \mid x_0, y_0).
\end{aligned} \tag{C.40}$$

Note that the estimated PN and PS are given by $\widehat{p}(u_2 \mid x_1, y_1)$ and $\widehat{p}(u_2 \mid x_0, y_0)$, respectively.

Here, in the setting of Section 3.2, we also propose estimation methods of $p(x_0, y_0)$, $p(x_0, y_1)$, $p(x_1, y_0)$, and $p(x_0, y_0)$ via the IV, together with a proxy covariate from statistical data in the presence of confounding and selection biases.

Under the assumption that both the PN and PS can be estimated by equation (C.40), Z is independent of W. Then, letting $p_1 = p(x_0, y_0)$, $p_2 = p(x_1, y_0)$, $p_3 = p(x_0, y_1)$, and $p_4 = p(x_1, y_1)$, from equation (14) of the main paper, we have

$$\begin{aligned}
&p(z_1, w_1 \mid x_0, y_0)p_1 + p(z_1, w_1 \mid x_1, y_0)p_2 + p(z_1, w_1 \mid x_0, y_1)p_3 + p(z_1, w_1 \mid x_1, y_1)p_4 \\
&= (p(z_1 \mid x_0, y_0)p_1 + p(z_1 \mid x_1, y_0)p_2 + p(z_1 \mid x_0, y_1)p_3 + p(z_1 \mid x_1, y_1)p_4) \\
&\quad \times (p(w_1 \mid x_0, y_0)p_1 + p(w_1 \mid x_1, y_0)p_2 + p(w_1 \mid x_0, y_1)p_3 + p(w_1 \mid x_1, y_1)p_4)\,, \\
&p(z_2, w_1 \mid x_0, y_0)p_1 + p(z_2, w_1 \mid x_1, y_0)p_2 + p(z_2, w_1 \mid x_0, y_1)p_3 + p(z_2, w_1 \mid x_1, y_1)p_4 \\
&= (p(z_2 \mid x_0, y_0)p_1 + p(z_2 \mid x_1, y_0)p_2 + p(z_2 \mid x_0, y_1)p_3 + p(z_2 \mid x_1, y_1)p_4) \\
&\quad \times (p(w_1 \mid x_0, y_0)p_1 + p(w_1 \mid x_1, y_0)p_2 + p(w_1 \mid x_0, y_1)p_3 + p(w_1 \mid x_1, y_1)p_4)\,, \\
&p(z_1, w_2 \mid x_0, y_0)p_1 + p(z_1, w_2 \mid x_1, y_0)p_2 + p(z_1, w_2 \mid x_0, y_1)p_3 + p(z_1, w_2 \mid x_1, y_1)p_4 \\
&= (p(z_1 \mid x_0, y_0)p_1 + p(z_1 \mid x_1, y_0)p_2 + p(z_1 \mid x_0, y_1)p_3 + p(z_1 \mid x_1, y_1)p_4) \\
&\quad \times (p(w_2 \mid x_0, y_0)p_1 + p(w_2 \mid x_1, y_0)p_2 + p(w_2 \mid x_0, y_1)p_3 + p(w_2 \mid x_1, y_1)p_4)\,, \\
&p_1 + p_2 + p_3 + p_4 = 1.
\end{aligned}$$

Then, $p(x, y)$ can be estimated as a solution to these equations by replacing $p(z \mid x, y)$, $p(w \mid x, y)$, and $p(z, w \mid x, y)$ to $\widehat{p}(z \mid x, y, s_1)$, $\widehat{p}(w \mid x, y, s_1)$, and $\widehat{p}(z, w \mid x, y, s_1)$, respectively, i.e.,

$$\begin{aligned}
&\widehat{p}(z_1, w_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_1, w_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_1, w_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_1, w_1 \mid x_1, y_1, s_1)p_4 \\
&= (\widehat{p}(z_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_1 \mid x_1, y_1, s_1)p_4) \\
&\quad \times (\widehat{p}(w_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(w_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(w_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(w_1 \mid x_1, y_1, s_1)p_4)\,, \\
&\widehat{p}(z_2, w_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_2, w_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_2, w_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_2, w_1 \mid x_1, y_1, s_1)p_4 \\
&= (\widehat{p}(z_2 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_2 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_2 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_2 \mid x_1, y_1, s_1)p_4) \\
&\quad \times (\widehat{p}(w_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(w_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(w_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(w_1 \mid x_1, y_1, s_1)p_4)\,, \\
&\widehat{p}(z_1, w_2 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_1, w_2 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_1, w_2 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_1, w_2 \mid x_1, y_1, s_1)p_4 \\
&= (\widehat{p}(z_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_1 \mid x_1, y_1, s_1)p_4) \\
&\quad \times (\widehat{p}(w_2 \mid x_0, y_0, s_1)p_1 + \widehat{p}(w_2 \mid x_1, y_0, s_1)p_2 + \widehat{p}(w_2 \mid x_0, y_1, s_1)p_3 + \widehat{p}(w_2 \mid x_1, y_1, s_1)p_4)\,, \\
&p_1 + p_2 + p_3 + p_4 = 1.
\end{aligned}$$

Since it is difficult to obtain the closed form of these equations, we apply the Newton–Raphson method to obtain the numerical solution (e.g., Driscoll and Braun, 2017). For $\boldsymbol{p} = (p_1, p_2, p_3, p_4)^\top$, letting

$$\begin{aligned}
f_1(\boldsymbol{p}) &= \widehat{p}(z_1, w_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_1, w_1 \mid x_1, y_0, s_1)p_2 \\
&\quad + \widehat{p}(z_1, w_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_1, w_1 \mid x_1, y_1, s_1)p_4 \\
&\quad - (\widehat{p}(z_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_1 \mid x_1, y_1, s_1)p_4) \\
&\quad\quad \times (\widehat{p}(w_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(w_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(w_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(w_1 \mid x_1, y_1, s_1)p_4)\,, \\
f_2(\boldsymbol{p}) &= \widehat{p}(z_2, w_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_2, w_1 \mid x_1, y_0, s_1)p_2 \\
&\quad + \widehat{p}(z_2, w_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_2, w_1 \mid x_1, y_1, s_1)p_4 \\
&\quad - (\widehat{p}(z_2 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_2 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_2 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_2 \mid x_1, y_1, s_1)p_4) \\
&\quad\quad \times (\widehat{p}(w_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(w_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(w_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(w_1 \mid x_1, y_1, s_1)p_4)\,, \\
f_3(\boldsymbol{p}) &= \widehat{p}(z_1, w_2 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_1, w_2 \mid x_1, y_0, s_1)p_2 \\
&\quad + \widehat{p}(z_1, w_2 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_1, w_2 \mid x_1, y_1, s_1)p_4 \\
&\quad - (\widehat{p}(z_1 \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_1 \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_1 \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_1 \mid x_1, y_1, s_1)p_4) \\
&\quad\quad \times (\widehat{p}(w_2 \mid x_0, y_0, s_1)p_1 + \widehat{p}(w_2 \mid x_1, y_0, s_1)p_2 + \widehat{p}(w_2 \mid x_0, y_1, s_1)p_3 + \widehat{p}(w_2 \mid x_1, y_1, s_1)p_4)\,, \\
f_4(\boldsymbol{p}) &= p_1 + p_2 + p_3 + p_4 - 1,
\end{aligned}$$

we define $F(\boldsymbol{p}) = (f_1(\boldsymbol{p}), f_2(\boldsymbol{p}), f_3(\boldsymbol{p}), f_4(\boldsymbol{p}))^\top$, and then the Jacobian matrix of $F\colon \mathbb{R}^4 \to \mathbb{R}^4$ is given by

$$J_F(\boldsymbol{p}) = \begin{pmatrix} \frac{\partial f_1}{\partial p_1} & \frac{\partial f_1}{\partial p_2} & \frac{\partial f_1}{\partial p_3} & \frac{\partial f_1}{\partial p_4} \\ \frac{\partial f_2}{\partial p_1} & \frac{\partial f_2}{\partial p_2} & \frac{\partial f_2}{\partial p_3} & \frac{\partial f_2}{\partial p_4} \\ \frac{\partial f_3}{\partial p_1} & \frac{\partial f_3}{\partial p_2} & \frac{\partial f_3}{\partial p_3} & \frac{\partial f_3}{\partial p_4} \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

where

$$\begin{aligned} &\frac{\partial f_{m+2n-2}}{\partial p_{i+2j+1}} \\ &= \widehat{p}(z_m, w_n \mid x_i, y_j, s_1) \\ &\quad - \widehat{p}(z_m \mid x_i, y_j, s_1)\left(\widehat{p}(w_n \mid x_0, y_0, s_1)p_1 + \widehat{p}(w_n \mid x_1, y_0, s_1)p_2 + \widehat{p}(w_n \mid x_0, y_1, s_1)p_3 + \widehat{p}(w_n \mid x_1, y_1, s_1)p_4\right) \\ &\quad - \left(\widehat{p}(z_m \mid x_0, y_0, s_1)p_1 + \widehat{p}(z_m \mid x_1, y_0, s_1)p_2 + \widehat{p}(z_m \mid x_0, y_1, s_1)p_3 + \widehat{p}(z_m \mid x_1, y_1, s_1)p_4\right)\widehat{p}(w_n \mid x_i, y_j, s_1) \end{aligned}$$

for $i, j \in \{0, 1\}$ and $(m, n) \in \{(1,1), (2,1), (1,2)\}$. To estimate $\boldsymbol{p}$, we start with the initial value $\boldsymbol{p}^{(0)}$ and update the k-th estimate via the following equation:

$$\boldsymbol{p}^{(k+1)} = \boldsymbol{p}^{(k)} - J_F(\boldsymbol{p}^{(k)})^{-1} F(\boldsymbol{p}^{(k)}).$$

Here, we adopt $\boldsymbol{p}^{(0)} = (\widehat{p}(x_0, y_0 \mid s_1), \widehat{p}(x_1, y_0 \mid s_1), \widehat{p}(x_0, y_1 \mid s_1), \widehat{p}(x_1, y_1 \mid s_1))^\top$ as the initial value. If any component of $\boldsymbol{p}^k$ is not included in $(0,1)$, then we terminate the above iteration step.

Combining the estimate $\widehat{\boldsymbol{p}}$ with (C.40), we can estimate $p(u_1)$, $p(u_2)$, $p(u_3)$, and $p(u_4)$ as follows:

$$\begin{aligned} \widehat{p}(u_1) &= \widehat{p}(Y_{x_1} = y_1, Y_{x_0} = y_1) = \widehat{p}(Y_{x_1} = y_1 \mid x_0, y_1)\widehat{p}_3 + \widehat{p}(Y_{x_0} = y_1 \mid x_1, y_1)\widehat{p}_4 = \widehat{p}(u_1 \mid x_0, y_1)\widehat{p}_3 + \widehat{p}(u_1 \mid x_1, y_1)\widehat{p}_4, \\ \widehat{p}(u_2) &= \widehat{p}(Y_{x_1} = y_1, Y_{x_0} = y_0) = \widehat{p}(Y_{x_1} = y_1 \mid x_0, y_0)\widehat{p}_1 + \widehat{p}(Y_{x_0} = y_0 \mid x_1, y_1)\widehat{p}_4 = \widehat{p}(u_2 \mid x_0, y_0)\widehat{p}_1 + \widehat{p}(u_2 \mid x_1, y_1)\widehat{p}_4, \\ \widehat{p}(u_3) &= \widehat{p}(Y_{x_1} = y_0, Y_{x_0} = y_1) = \widehat{p}(Y_{x_1} = y_0 \mid x_0, y_1)\widehat{p}_3 + \widehat{p}(Y_{x_0} = y_1 \mid x_1, y_0)\widehat{p}_2 = \widehat{p}(u_3 \mid x_0, y_1)\widehat{p}_3 + \widehat{p}(u_3 \mid x_1, y_0)\widehat{p}_2, \\ \widehat{p}(u_4) &= \widehat{p}(Y_{x_1} = y_0, Y_{x_0} = y_0) = \widehat{p}(Y_{x_1} = y_0 \mid x_0, y_0)\widehat{p}_1 + \widehat{p}(Y_{x_0} = y_0 \mid x_1, y_0)\widehat{p}_2 = \widehat{p}(u_4 \mid x_0, y_0)\widehat{p}_1 + \widehat{p}(u_4 \mid x_1, y_0)\widehat{p}_2, \end{aligned} \tag{C.41}$$

which shows that the PNS is also estimable on the basis of the IV framework.

D Numerical experiments

In this section, we investigate more properties of our proposed estimators of the probabilities of the potential outcome types, i.e., $p(u_1)$, $p(u_2)$, $p(u_3)$ and $p(u_4)$, through more numerical experiments in addition to Section 3.2 because the probabilities of the potential outcome types are fundamental components of causal inference in the sense that they enable us to evaluate the probabilistic aspects of "necessity", "sufficiency", and "necessity and sufficiency", which are important concepts of successful explanation (Watson et al., 2021).

For simplicity, letting X, Y, Z, W, and U be discrete variables, we consider the directed acyclic graph shown in Fig. 2, where the conditional probabilities regarding (X, Y, Z, W, U, S) are given according to Table 2. Under the situation where $(X, Y, Z, W, S = s_1)$ can be observed but U or (X, Y, Z, W, s_0) cannot, the properties of the proposed estimators $\widehat{p}(u_1)$, $\widehat{p}(u_2)$, $\widehat{p}(u_3)$ and $\widehat{p}(u_4)$ of $p(u_1)$, $p(u_2)$, $p(u_3)$ and $p(u_4)$, respectively, are verified as follows:

Step 1 we generate the dataset $\{(X_n, Y_n, Z_n, W_n, S_n)\}_{n=1}^N$ with the size $N = 1000$, 5000, 10000, and 50000,

Step 2 based on the dataset $\{(X_n, Y_n, Z_n, W_n, S_n = s_1)\}_{n=1}^N$, we verify the properties of the proposed estimators for the following cases:

(a) $(p(u_1), p(u_2), p(u_3), p(u_4)) = (1/4, 1/4, 1/4, 1/4)$,

(b) $(p(u_1), p(u_2), p(u_3), p(u_4)) = (5/16, 5/16, 5/16, 1/16)$,

Table 2: Parameter setting in the numerical experiments.

	(a) $p(W \mid U)$			(b) $p(Y = y_1 \mid X, U)$		(c) $p(X = x_1 \mid Z, U)$		
	w_1	w_2	w_3	x_0	x_1	z_1	z_2	z_3
u_1	7/10	1/5	1/10	1	1	7/10	1/5	1/10
u_2	3/20	1/20	4/5	0	1	3/20	1/20	4/5
u_3	5/10	2/5	1/10	1	0	5/10	2/5	1/10
u_4	1/10	1/5	7/10	0	0	1/10	1/5	7/10

	(e) $p(Z)$
z_1	1/3
z_2	1/3
z_3	1/3

	(f) $p(S = s_1 \mid X, Y)$	
	y_0	y_1
x_0	1/4	3/4
x_1	1/2	1

(c) $(p(u_1), p(u_2), p(u_3), p(u_4)) = (3/8, 3/8, 1/8, 1/8)$,

(d) $(p(u_1), p(u_2), p(u_3), p(u_4)) = (13/16, 1/16, 1/16, 1/16)$.

Tables 3-6 and Figure 3 present the basic statistics and the box plots of the estimated $p(u_1)$, $p(u_2)$, $p(u_3)$, and $p(u_4)$) for $5,000$ replications with the given sample size N. Here, note that we estimate the probabilities of potential outcome types using the bootstrapped replication not to take values outside the range $[0, 1]$, because if equation (10) or (11) is close to zero then it may be judged that Condition 2 or 3 is violated.

From Table 3 and Figure 3, the sample means of the estimated probabilities of potential outcome types are close to the true values, and the sample standard errors are smaller as the sample size is larger. In addition, the interquartile ranges of the estimated probabilities of potential outcome types are narrower and still include the true values even if the sample size is large. In contrast, it seems that the estimated probabilities of potential outcome types seem to be asymmetry distributed as the true values are away from 0.5. In such situations, the asymptotic normality may not hold for the finite sample sizes. Thus, it seems that the proposed estimation method provides consistent estimators of the probabilities of potential outcome types, but a larger sample size may be required to show the asymptotic normality for some situations.

Table 3: Basic statistics of estimates based on the proposed method in the case where $(p(u_1), p(u_2), p(u_3), p(u_4)) = (1/4, 1/4, 1/4, 1/4)$.

	$\widehat{p}(u_1)$				$\widehat{p}(u_2)$			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.012	0.017	0.021	0.049	0.019	0.026	0.015	0.070
1st Quantile	0.245	0.230	0.221	0.202	0.185	0.187	0.190	0.211
Median	0.339	0.309	0.297	0.266	0.226	0.221	0.221	0.239
Mean	0.334	0.303	0.288	0.262	0.231	0.226	0.224	0.240
3rd Quantile	0.425	0.373	0.359	0.325	0.271	0.256	0.253	0.270
Maximum	0.807	0.682	0.589	0.454	0.727	0.630	0.557	0.437
s.e.	0.123	0.104	0.095	0.077	0.077	0.073	0.063	0.045
	$\widehat{p}(u_3)$				$\widehat{p}(u_4)$			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.011	0.010	0.043	0.094	0.004	0.004	0.020	0.095
1st Quantile	0.180	0.221	0.228	0.218	0.118	0.131	0.139	0.184
Median	0.257	0.283	0.278	0.251	0.154	0.158	0.168	0.241
Mean	0.268	0.286	0.282	0.255	0.166	0.181	0.203	0.248
3rd Quantile	0.352	0.347	0.334	0.288	0.192	0.201	0.248	0.301
Maximum	0.748	0.715	0.707	0.480	0.639	0.700	0.655	0.540
s.e.	0.118	0.094	0.083	0.054	0.083	0.086	0.093	0.077

Table 4: Basic statistics of estimates based on the proposed method in the case where $(p(u_1), p(u_2), p(u_3), p(u_4)) = (5/16, 5/16, 5/16, 1/16)$.

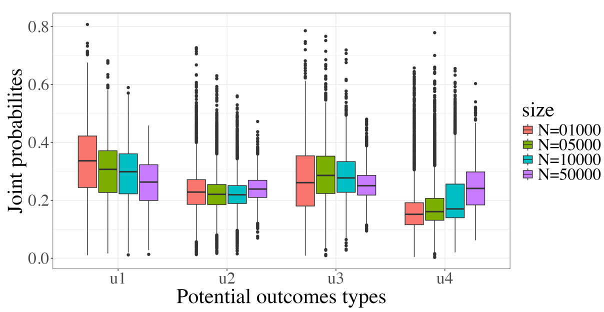
	$\widehat{p}(u_1)$				$\widehat{p}(u_2)$			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.010	0.024	0.027	0.106	0.003	0.013	0.010	0.081
1st Quantile	0.255	0.250	0.240	0.277	0.191	0.204	0.210	0.251
Median	0.365	0.335	0.314	0.312	0.232	0.239	0.249	0.293
Mean	0.364	0.332	0.314	0.309	0.236	0.252	0.269	0.289
3rd Quantile	0.468	0.417	0.385	0.344	0.272	0.282	0.320	0.329
Maximum	0.861	0.727	0.626	0.479	0.711	0.770	0.767	0.531
s.e.	0.141	0.118	0.102	0.056	0.084	0.094	0.099	0.063
	$\widehat{p}(u_3)$				$\widehat{p}(u_4)$			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.010	0.019	0.039	0.097	0.001	0.002	0.003	0.012
1st Quantile	0.212	0.248	0.251	0.268	0.040	0.038	0.042	0.057
Median	0.313	0.328	0.323	0.305	0.063	0.061	0.067	0.077
Mean	0.321	0.332	0.330	0.311	0.079	0.081	0.087	0.094
3rd Quantile	0.425	0.414	0.402	0.350	0.092	0.097	0.111	0.113
Maximum	0.913	0.849	0.805	0.583	0.639	0.565	0.615	0.399
s.e.	0.144	0.121	0.115	0.065	0.064	0.066	0.068	0.055

Table 5: Basic statistics of estimates based on the proposed method in the case where $(p(u_1), p(u_2), p(u_3), p(u_4)) = (3/8, 3/8, 1/8, 1/8)$.

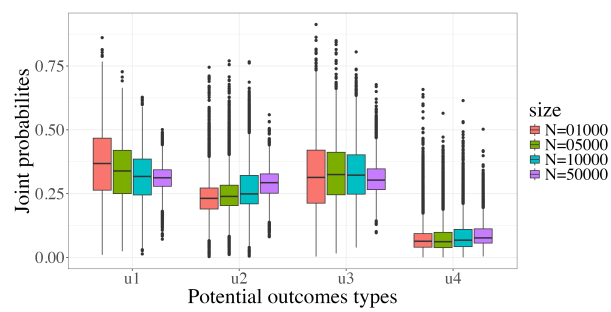
	$\widehat{p}(u_1)$				$\widehat{p}(u_2)$			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.004	0.027	0.005	0.110	0.011	0.044	0.054	0.108
1st Quantile	0.318	0.327	0.341	0.334	0.239	0.264	0.274	0.293
Median	0.430	0.429	0.442	0.422	0.289	0.297	0.300	0.312
Mean	0.410	0.408	0.416	0.407	0.289	0.301	0.305	0.321
3rd Quantile	0.511	0.505	0.505	0.496	0.336	0.332	0.329	0.353
Maximum	0.797	0.752	0.682	0.608	0.750	0.708	0.639	0.506
s.e.	0.133	0.125	0.115	0.100	0.090	0.077	0.066	0.050
	$\widehat{p}(u_3)$				$\widehat{p}(u_4)$			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.002	0.003	0.009	0.036	0.002	0.005	0.009	0.021
1st Quantile	0.096	0.105	0.105	0.109	0.060	0.065	0.067	0.073
Median	0.159	0.153	0.144	0.132	0.091	0.091	0.090	0.106
Mean	0.183	0.170	0.159	0.142	0.112	0.121	0.121	0.133
3rd Quantile	0.247	0.220	0.199	0.165	0.135	0.146	0.148	0.184
Maximum	0.697	0.622	0.537	0.432	0.698	0.725	0.610	0.452
s.e.	0.113	0.091	0.076	0.049	0.084	0.087	0.082	0.073

Table 6: Basic statistics of estimates based on the proposed method in the case where $(p(u_1), p(u_2), p(u_3), p(u_4)) = (13/16, 1/16, 1/16, 1/16)$.

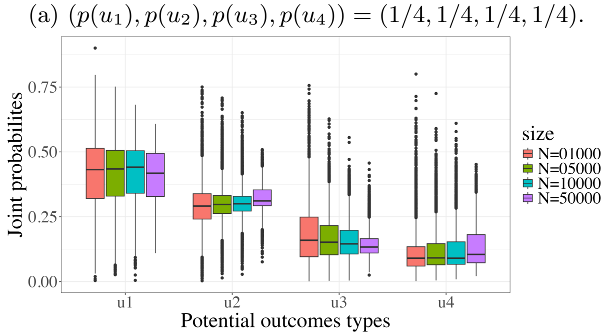
	$\widehat{p}(u_1)$				$\widehat{p}(u_2)$			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.029	0.041	0.161	0.206	0.000	0.002	0.003	0.009
1st Quantile	0.631	0.722	0.752	0.782	0.045	0.038	0.039	0.040
Median	0.754	0.813	0.830	0.848	0.071	0.051	0.050	0.048
Mean	0.717	0.773	0.791	0.817	0.087	0.062	0.060	0.058
3rd Quantile	0.835	0.864	0.873	0.879	0.106	0.069	0.066	0.065
Maximum	0.969	0.964	0.946	0.941	0.603	0.377	0.355	0.312
s.e.	0.152	0.131	0.121	0.091	0.066	0.044	0.039	0.031
	$\widehat{p}(u_3)$				$\widehat{p}(u_4)$			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.001	0.002	0.002	0.002	0.002	0.002	0.001	0.001
1st Quantile	0.061	0.052	0.049	0.047	0.021	0.021	0.022	0.025
Median	0.121	0.092	0.080	0.065	0.031	0.029	0.029	0.032
Mean	0.156	0.121	0.105	0.074	0.041	0.041	0.043	0.049
3rd Quantile	0.214	0.155	0.128	0.092	0.043	0.037	0.038	0.063
Maximum	0.735	0.886	0.765	0.523	0.528	0.497	0.422	0.467
s.e.	0.126	0.102	0.090	0.041	0.045	0.046	0.048	0.042



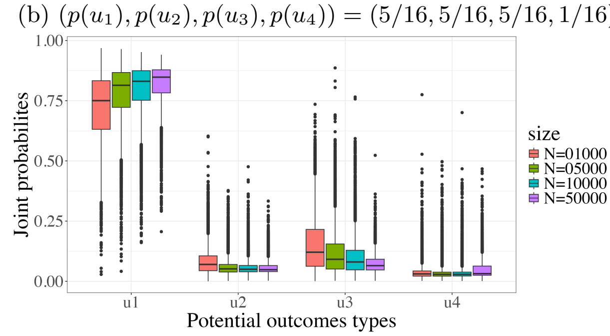
(a) $(p(u_1), p(u_2), p(u_3), p(u_4)) = (1/4, 1/4, 1/4, 1/4)$.



(b) $(p(u_1), p(u_2), p(u_3), p(u_4)) = (5/16, 5/16, 5/16, 1/16)$.



(c) $(p(u_1), p(u_2), p(u_3), p(u_4)) = (3/8, 3/8, 1/8, 1/8)$.



(d) $(p(u_1), p(u_2), p(u_3), p(u_4)) = (13/16, 1/16, 1/16, 1/16)$.

Figure 3: Boxplots of estimates based on the proposed method.

Table 7: Basic statistics of estimates based on the proposed method in the case where $(p(u_1), p(u_2), p(u_3), p(u_4)) = (1/4, 1/4, 1/4, 1/4)$.

	PN				PS			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.001	0.013	0.008	0.234	0.000	0.002	0.000	0.000
1st Quantile	0.445	0.457	0.460	0.465	0.303	0.329	0.350	0.421
Median	0.534	0.507	0.506	0.503	0.543	0.513	0.502	0.499
Mean	0.540	0.516	0.517	0.517	0.525	0.494	0.480	0.483
3rd Quantile	0.627	0.566	0.558	0.550	0.747	0.661	0.622	0.564
Maximum	1.000	0.986	0.998	0.996	1.000	0.999	0.998	0.832
s.e.	0.150	0.104	0.096	0.083	0.272	0.228	0.197	0.116

Table 8: Basic statistics of estimates based on the proposed method in the case where $(p(u_1), p(u_2), p(u_3), p(u_4)) = (5/16, 5/16, 5/16, 1/16)$.

	PN				PS			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.004	0.007	0.045	0.237	0.000	0.000	0.002	0.004
1st Quantile	0.419	0.446	0.454	0.476	0.558	0.650	0.681	0.685
Median	0.499	0.490	0.492	0.499	0.793	0.821	0.834	0.819
Mean	0.504	0.492	0.491	0.503	0.706	0.744	0.762	0.749
3rd Quantile	0.588	0.535	0.527	0.523	0.909	0.910	0.912	0.880
Maximum	1.000	1.000	0.993	0.989	1.000	1.000	1.000	1.000
s.e.	0.141	0.089	0.071	0.053	0.258	0.230	0.213	0.200

Table 9: Basic statistics of estimates based on the proposed method in the case where $(p(u_1), p(u_2), p(u_3), p(u_4)) = (3/8, 3/8, 1/8, 1/8)$.

	PN				PS			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.001	0.075	0.058	0.061	0.000	0.000	0.001	0.001
1st Quantile	0.452	0.467	0.477	0.483	0.422	0.532	0.569	0.658
Median	0.522	0.508	0.507	0.504	0.672	0.719	0.724	0.740
Mean	0.532	0.513	0.513	0.511	0.618	0.665	0.674	0.707
3rd Quantile	0.599	0.550	0.540	0.528	0.846	0.841	0.825	0.793
Maximum	1.000	0.998	0.997	0.996	1.000	1.000	1.000	0.995
s.e.	0.128	0.077	0.066	0.050	0.267	0.230	0.210	0.135

Table 10: Basic statistics of estimates based on the proposed method in the case where $(p(u_1), p(u_2), p(u_3), p(u_4)) = (13/16, 1/16, 1/16, 1/16)$.

	PN				PS			
	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$	$N = 1000$	$N = 5000$	$N = 10000$	$N = 50000$
Minimum	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000
1st Quantile	0.078	0.065	0.066	0.069	0.234	0.259	0.296	0.327
Median	0.143	0.091	0.085	0.078	0.460	0.528	0.546	0.522
Mean	0.185	0.110	0.098	0.087	0.472	0.513	0.526	0.506
3rd Quantile	0.229	0.127	0.111	0.092	0.710	0.766	0.760	0.683
Maximum	1.000	0.974	0.866	0.918	1.000	1.000	0.999	1.000
s.e.	0.166	0.090	0.064	0.044	0.283	0.288	0.275	0.237