Review Article

Zhi Geng, Chao Zhang, Xueli Wang, Chunchen Liu, and Shaojie Wei*

Prospective and retrospective causal inferences based on the potential outcome framework

https://doi.org/10.1515/jci-2023-0063
received September 30, 2023; accepted June 28, 2024

Abstract: In this article, we discuss both prospective and retrospective causal inferences, building on Neyman's potential outcome framework. For prospective causal inference, we review criteria for confounders and surrogates to avoid the Yule–Simpson paradox and the surrogate paradox, respectively. For retrospective causal inference, we introduce the concepts of posterior causal effects given observed evidence to quantify the causes of effects. The posterior causal effects provide a unified framework for deducing both effects of causes in prospective causal inference and causes of effects in retrospective causal inference. We compare the medical diagnostic approaches based on Bayesian posterior probabilities and posterior causal effects for classification and attribution.

Keywords: causal inference, cause of effect, effect of cause, potential outcome, surrogate paradox, Yule–Simpson paradox

MSC 2020: 62D20

1 Introduction

Causal inference has a solid theoretical foundation based on the potential outcome framework, which was first proposed by Neyman (1923) for experimental studies [1] and later extended by Rubin (1974) to observational studies [2]. This framework allows causal concepts and questions to be formally defined and represented mathematically. Without this formal framework, causal relationships are often conflated with correlational relationships, leading to mistaken inferences. By grounding inference in the potential outcome framework, we move beyond simply observing correlations between variables. This allows us to define causal effects more precisely, make essential assumptions to identify these effects from observational data, and develop estimators with desirable statistical properties. In this way, the framework enables rigorous causal inference that aims to uncover genuine causal relationships from both experimental and observational data.

Causal inference involves not only evaluating the effects of causes in a prospective causal inference, but also deducing the causes of effects in a retrospective causal inference. In epidemiology, both prospective and

*** Corresponding author: Shaojie Wei,** School of Mathematics and Statistics, Beijing Technology and Business University, Beijing 100048, P. R. China, e-mail: 20221207@btbu.edu.cn
Zhi Geng: School of Mathematics and Statistics, Beijing Technology and Business University, Beijing 100048, P. R. China, e-mail: zhigeng@pku.edu.cn
Chao Zhang: School of Mathematics and Statistics, Beijing Technology and Business University, Beijing 100048, P. R. China, e-mail: chzhang@st.btbu.edu.cn
Xueli Wang: School of Mathematics and Statistics, Beijing Technology and Business University, Beijing 100048, P. R. China, e-mail: xlwang@btbu.edu.cn
Chunchen Liu: LingYang, Alibaba Group, Hangzhou, P. R. China, e-mail: chencang.lcc@alibaba-inc.com

retrospective studies concern the design stage. Prospective and retrospective causal inferences concern the analysis stage. Prospective causal inference is to evaluate effects of causes, which is typically forward-looking, while retrospective causal inference is to deduce causes of effects, which is typically backward-looking [3,4]. For example, it is a prospective causal problem to determine whether a drug will have the effect of lowering blood pressure, while it is a retrospective causal problem when we know that a person died, and we retrospectively ask whether the death was caused by a particular drug. Dawid et al. [5] highlighted an important distinction between effects of causes and causes of effects. Statistical causality emphasizes evaluating the effects of causes rather than the causes of effects [6,7]. Randomized experiments are the gold standard for evaluating causal effects in prospective causal inference. However, for retrospective causal inference, even under randomized experiments, identifying the causes of effects is difficult.

In observational studies, confounding poses a major threat to valid causal inference about effects. The Yule–Simpson paradox provides a striking example of how ignoring a confounder between treatment and outcome can completely reverse an association. Similarly, the surrogate paradox can arise if there is a confounder between surrogate and true endpoints. Unless certain criteria are met, using the surrogate as a substitute for the true endpoint in assessing treatment effects can be misleading. To avoid inferential paradoxes and biases, careful consideration must be given to potential confounders and surrogates when making causal claims from observational data. In this article, we discuss the precise criteria that must be met for a variable to be a confounder or a valid surrogate for reviewing prospective causal inference and highlighting the contributions of the potential outcome framework. Understanding these criteria will allow more rigorous prospective causal inferences to be made from observational studies.

While statistical causality has focused more on prospective causal inference, deducing causes from observed effects is also an important causal reasoning task. Retrospective causal inference aims to determine the causes behind a specific effect or event that has already occurred, based on the observed data and causal assumptions. Dawid and Musio [7] highlighted that counterfactual reasoning is unnecessary for analyzing effects of causes prospectively, but essential for retrospective inference about individual-level causes. For example, whether a particular individual's lung cancer is caused by smoking requires imagining the counterfactual scenario where the person did not smoke and assessing the probability they would still have developed cancer. Retrospective causal inference is more challenging than prospective inference for several reasons. Confounding can be more complex because conditioning on the occurred effect or outcome may induce additional biases not present in a prospective design [8]. Randomization and "no unobserved confounders" assumptions are often insufficient to eliminate this bias. Moreover, causal effects may be heterogeneous across individuals, so group-level estimates may not apply to a specific individual case with occurred effects. Despite these difficulties, retrospective causal inference has many vital applications including attributing causes in epidemiology and legal cases [5,9–12].

We discuss posterior causal effects to formally unify prospective and retrospective causal inference problems. Posterior causal effects are causal effects conditioned on observed evidence, which may include observed effect variables [13,14]. They thus measure the effects of causes in a subpopulation restricted by the evidence. Depending on the evidence, posterior effects can be used for both prospective and retrospective inferences. When the evidence excludes effect variables, posterior effects evaluate causes prospectively. For instance, the posterior effect of smoking on lung cancer given age and gender evidence defines the causal effect in that age/gender population. This evaluates the prospective effect in a specific subpopulation. In contrast, when the evidence includes effect variables, posterior effects deduce causes retrospectively. The posterior effect of smoking on lung cancer given occurred lung cancer evidence defines the effect in lung cancer patients. This can judge the possibility that smoking caused lung cancer retrospectively. Posterior causal effects also explain the causal meaning of population attributable risks commonly used in public health and epidemiology. These measure the proportion of cases attributable to an exposure. Posterior effects formally connect attributable risks to causal effects conditioned on observed effects.

In this article, we offer a review of some topics in prospective and retrospective causal inferences, based on the potential outcome framework. The remainder of this article is organized as follows. In Section 2, we review contribution of the potential outcome framework to prospective causal inference, focusing on confounders and surrogate endpoints. We discuss criteria for confounders and surrogate endpoints that help avoid the Yule–Simpson paradox and the surrogate paradox, respectively. Section 3 then covers retrospective causal inference based on the framework. We introduce probabilities of causation and posterior causal effects based on counterfactual reasoning. We also interpret population attributable risks in epidemiology through the lens of posterior causal effects. Finally, we compare medical diagnostic approaches based on Bayesian posterior probabilities versus posterior causal effects.

2 Prospective causal inference

In prospective causal inference, a goal is to evaluate the effect of a cause event that occurred earlier on an outcome event that occurred later. Suppose that all variables presented below are binary, 1 denotes presence and 0 absence. Let X denote an observed cause variable (e.g., smoking) that happened earlier at time t_1, and Y denote an observed effect variable (e.g., lung cancer) that occurred at time t_2 ($t_1 < t_2$). The association between smoking and lung cancer can be measured using the observed data of X and Y, such as Pearson's correlation or relative risks. However, causation between smoking and lung cancer cannot be well defined only by the notation of two observed variables X and Y. To describe the causation, Neyman [1] and Rubin [2] proposed the following notation of potential outcomes. Let Y_x denote the potential outcome that would occur at time t_2 if an individual were exposed to the cause $X = x$ at time t_1. The individual causal effect of cause X on response Y is defined as $Y_1 - Y_0$, and the average causal effect is $E(Y_1 - Y_0)$. Focusing on the treated population where $X = 1$, the average treatment effect on the treated is expressed as $E(Y_1 - Y_0|X = 1)$. Generally, the probabilistic causal effect of X on Y for the treated population can be evaluated by comparing $\text{pr}(Y_1 = 1|X = 1)$ to the unobserved counterfactual probability $\text{pr}(Y_0 = 1|X = 1)$, which is not identifiable without any assumption. Randomized experiments are the gold standard approach for identifying $\text{pr}(Y_0 = 1|X = 1)$ by the probability $\text{pr}(Y = 1|X = 0)$ of observed variables. For observational studies, identification requires some untestable assumptions.

2.1 Confounders

The problem about confounders has been explored for a long time, especially in epidemiology. However, the criteria for assessing confounders and confounding in the epidemiological literature have been inconsistent [15–21]. Using Neyman's potential outcome framework, confounding bias B is defined as the difference between the counterfactual probability of potential outcome without exposure in the exposed population and the probability of observed outcome in the unexposed population [18,22], i.e.,

$$B = \text{pr}(Y_0 = 1|X = 1) - \text{pr}(Y_0 = 1|X = 0). \tag{1}$$

By adjusting the distribution $\text{pr}(C = k|X = 0)$ of covariate C in the unexposed population to $\text{pr}(C = k|X = 1)$, a standardized probability $\text{pr}_\Delta(Y_0 = 1|X = 0)$ is defined as

$$\text{pr}_\Delta(Y_0 = 1|X = 0) = \sum_{k=1}^{K} \text{pr}(Y_0 = 1|X = 0, C = k)\text{pr}(C = k|X = 1).$$

A confounder is defined as a risk factor whose control can reduce the confounding bias [15,23–26]. Replacing the counterfactual probability $\text{pr}(Y_0 = 1|X = 0)$ in (1) by the adjusted $\text{pr}_\Delta(Y_0 = 1|X = 0)$, Geng et al. [25] defined a confounder as a covariate C for which

$$|\text{pr}(Y_0 = 1|X = 1) - \text{pr}_\Delta(Y_0 = 1|X = 0)| < |B|.$$

This definition states that the standardized probability $\text{pr}_\Delta(Y_0 = 1|X = 0)$ adjusted for a confounder C is closer to the counterfactual probability $\text{pr}(Y_0 = 1|X = 1)$ than the observed probability $\text{pr}(Y = 1|X = 0)$. For a case of C with multiple covariates, C may be recategorized by a single categorical variate C' with the same number of categories as C. VanderWeele and Shpitser [26] considered a similar definition of confounders for the overall effect of the exposure on the whole population rather than the effect of the exposure on the exposed population. Note that these definitions of confounders do not need any assumption such as subpopulation-comparability or a known causal diagram. With this definition, we can determine that a covariate is not a confounder when $\mathrm{pr}_\Delta(Y_0 = 1|X = 0) = \mathrm{pr}(Y_0 = 1|X = 0)$, but we cannot confirm that it is a confounder since $\mathrm{pr}(Y_0 = 1|X = 1)$ is not identifiable without further assumptions.

2.2 Surrogate endpoints

In many scientific studies, true endpoint variables cannot be measured or observed due to being expensive, inconvenient, or impractical within a short time span. For example, in clinical trials, CD4 count is used as a surrogate endpoint for survival time in acequired immune deficiency syndrome (AIDS) studies, and bone mass is used as a surrogate endpoint for fracture in osteoporosis studies. However, Fleming and Demets [27] pointed out that in many real clinical trials, surrogates failed to evaluate the treatment effects on true endpoints.

Chen et al. [28] introduced and formulated the surrogate paradox, where a treatment has a positive effect on a surrogate endpoint, which in turn has a positive effect on a true endpoint, but the treatment has a negative effect on the true endpoint. Even by conducting two randomized experiments, we can separately prove probabilistically both that a variate X has a positive causal effect on a variable Y and that the variate Y has a positive causal effect on a variable Z, but we cannot judge that X has a positive causal effect on Z, even if X does not have any direct causal effect on Z. Even if the intermediate variable Y breaks all causal paths from X to Z, the probabilistic causal relationships may not be transitive, although the individual causal relationships may be. The surrogate paradox implies that the sign of treatment effect on the endpoint cannot be predicted by the sign of treatment effect on the surrogate and the sign of causal effect of surrogate on the endpoint. Therefore, the logical reasoning may not be applied to the probabilistic results of causal inference. Jiang et al. [29] also discussed the transitivity of different associations under the conditional independence of X and Z given Y and showed that the finer an association measure is, the stronger its transitivity is.

Prentice [30] proposed a criterion for a statistical surrogate Y, which requires both a strong association between treatment X and the surrogate Y and the conditional independence of the true endpoint Z and treatment X given Y, denoted as $X \perp\!\!\!\perp Z|Y$ in the notation by Dawid [31]. The conditional independence means that the surrogate Y can break the association between the treatment X and the endpoint Z, and thus, $X \perp\!\!\!\perp Y$ implies $X \perp\!\!\!\perp Z$. Frangakis and Rubin [32] presented the criterion for a principal surrogate Y, which should possess the causal necessity: a treatment X has a causal effect on an endpoint Z only if the treatment X has a causal effect on the surrogate Y. Lauritzen [33] proposed the criterion for a strong surrogate Y, which breaks all causal paths from X to Z in a causal diagram (Figure 1). Chen et al. [28] showed that the surrogate paradox cannot be avoided, even by such strong criteria of the statistical surrogates, the principal surrogates, and the strong surrogates. A surrogate Y is an intermediate variable in a causal path from X to Z, and variable X is an instrumental variable when Y is a strong surrogate. A more proper name for the paradox may be "the intermediate variable paradox" to reflect its wider generality. The same paradox applies to other situations. For example, the paradox can be called "the instrumental paradox" in the use of the instrumental variable. The surrogate paradox also points out an issue of the transitivity of causal effects on a causal path. Jiang et al. [34] proposed approaches to identifying the principal stratification causal effects by multiple trials and provided the criteria for surrogates that avoid the surrogate paradox.

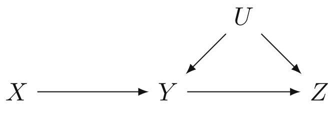

Figure 1: Criterion for a strong surrogate.

Moore [35] provided a real-world example of the surrogate paradox. Doctors knew that irregular heartbeat was a risk factor for sudden death and presumed that correcting irregular heartbeat would prevent sudden death. Therefore, they used "correction of heartbeat" as a surrogate, and several drugs (Enkaid, Tambocor and Ethmozine) were approved by FDA (Food and Drug Administration). However, the Cardiac Arrhythmia Suppression Trial [36] showed that these drugs did not improve survival times but increased mortality.

Chen et al. [28], Ju and Geng [37], Wu et al. [38], and VanderWeele and Shpitser [26] proposed consistent surrogates and criteria to avoid the surrogate paradox. These criteria apply to single surrogates only. However, in many applications, a treatment may affect the endpoint through multiple pathways, and thus, a single surrogate cannot break all of these pathways. For example, a drug may reduce the risk of death from AIDS through two pathways: by decreasing human immunodeficiency virus type 1 ribonucleic acid (HIV-1 RNA) concentrations and by increasing CD4 count. In this case, a single surrogate may not satisfy any criterion of the statistical, principal, strong, or consistent surrogates. Both HIV-1 RNA concentrations and CD4 count should be used as multiple surrogates for the risk of death from AIDS. Joffe [39] suggested that it is meaningful to generalize the criteria for a single surrogate to multiple surrogates. Luo et al. [40] proposed a criterion for multiple surrogates $Y = (Y_1, \ldots, Y_p)$ based on stochastic orders of random vectors. All of these criteria for surrogates require some knowledge of causality or associations among the observed variables X and Y and the unobserved variable Z. Therefore, these criteria are not falsifiable without untestable assumptions or observed data for Z.

3 Retrospective causal inference

When evaluating the effects of causes, we prospectively predict the results of an intervention in a population. However, when deducing the causes of effects, we explore the causes of happened effects for a specific individual retrospectively. In doing so, we may have to imagine the potential outcomes if causes would happen in counterfactual scenarios. As will be seen below, counterfactual reasoning can be well described by the potential outcome framework. Retrospective causal inference can be used for causal attribution, medical diagnosis, and blame assignment. For example, scientific studies have evaluated the causal effects of benzene and ionizing radiation exposure on leukemia using data from experimental and observational studies. When we observe that a leukemia patient has been exposed to both benzene and ionizing radiation, we would like to know how much of the patient's leukemia is attributable to benzene exposure and how much is attributable to ionizing radiation exposure, which is a problem about causes of effects.

Evaluating effects of causes is the main focus of most existing causal inference approaches, while deducing causes of effects is the focus of a few approaches. As Dawid [12] pointed out, assessing causes of effects is more challenging than assessing effects of causes, because the former is mainly a counterfactual inference problem for a single individual. For a counterfactual situation, measures for the probabilities of causation are generally not identifiable even when we use the gold standard approach of randomized experiments and there are no unobserved confounders.

3.1 Probabilities of causation

Pearl [11] provided counterfactual definitions of causation to capture how necessary and/or sufficient a cause is capable of producing a given effect or outcome. Dawid et al. [5] highlighted the distinction between the effects of causes and the causes of effects, and proposed the probability of causation to make inference about the causes of effects. Inferring the causes of effects requires more subtle logic and stronger assumptions than inferring the effects of causes. In the following, we focus on binary variables. First, we introduce probabilities of causation for the case of a single effect variable Y and a single cause X. To measure how possible X is a cause of an effect Y, Dawid et al. [5] defined the probability of causation as

$$\mathrm{PC}(X \Rightarrow Y) = \mathrm{pr}(Y_{X=0} = 0 | Y_{X=1} = 1).$$

To measure how necessary X is a cause of an occurred effect $Y = 1$, Pearl [41] defined the probability of necessary causation as

$$\mathrm{PN}(X \Rightarrow Y) = \mathrm{pr}(Y_{X=0} = 0 | X = 1, Y = 1).$$

Lu et al. [13] proposed the posterior causal effects given observed evidence to measure the probabilities of causes and treated evaluating effects of causes and discovering causes of effects from the same perspective. Let C denote a pretreatment variable prior to treatment X (i.e., a covariate). Let $O = o$ denote the observed evidence for the target individual, where o is an observed value of O. For an individual case, we can sometimes observe only a subset O of variables $\{X, Y, C\}$. Li et al. [14] defined the posterior total causal effect given $O = o$ as

$$\mathrm{PostTCE}(X \Rightarrow Y | O = o) = E(Y_{X=1} - Y_{X=0} | O = o).$$

For the evidence $O = (C = c)$, the posterior total causal effect $E(Y_{X=1} - Y_{X=0} | C = c)$ is an average causal effect conditional on $C = c$; for the evidence $O = (X = 1)$, the posterior total causal effect $E(Y_{X=1} - Y_{X=0} | X = 1)$ is an average causal effect in a treated subpopulation; for the evidence $O = (X = 1, Y = 1)$, the posterior total causal effect is equal to PN:

$$\mathrm{PostTCE}(X \Rightarrow Y | X = 1, Y = 1) = 1 - \mathrm{pr}(Y_{X=0} = 1 | X = 1, Y = 1) = \mathrm{PN}.$$

Thus, the posterior total causal effect can be used not only to evaluate effects of causes for a prospective causal inference, but also to assess causes of effects for a retrospective causal inference. For example, let X denote smoking and Y lung cancer. By the posterior total causal effect $E(Y_{X=1} - Y_{X=0} | X = 1, Y = 1)$, we evaluate the causal effect of smoking on lung cancer in the subpopulation of smokers with lung cancer, which measures the probability that individuals in the subpopulation would not have developed lung cancer if they had not smoked.

The posterior intervention causal effect of X on Y proposed by Zhao et al. [14] given the observed evidence $O = o$ is defined as

$$\mathrm{PostICE}(X \Rightarrow Y | O = o) = E(Y - Y_{X=0} | O = o).$$

Different from PostTCE, PostICE measures the change of Y's expectation if X is removed. When the observed evidence $O = o$ contains $X = 1$, we have $\mathrm{PostTCE}(X \Rightarrow Y | X = 1, \ldots) = \mathrm{PostICE}(X \Rightarrow Y | X = 1, \ldots)$. When the observed evidence $O = o$ only contains $Y = 1$, we have $\mathrm{PostICE}(X \Rightarrow Y | Y = 1) = \mathrm{PN}(\mathrm{X} \Rightarrow \mathrm{Y})\mathrm{pr}(X = 1 | Y = 1)$. Therefore, PostICE is different from PN. In disease diagnosis, PostICE considers not only the probability that the disease X is the cause of symptoms Y but also the posterior probability of the disease given the occurrence of symptoms.

Next, we extend the aforementioned case to the case of multiple effect variables $Y = (Y_1, \ldots, Y_q)$ and multiple cause variables $X = (X_1, \ldots, X_p)$, and we define the posterior causal effect of simultaneously intervening on a subset of X on Y. In real applications, available evidence may include multiple observed effect variables, and thus, they can be used simultaneously to more accurately deduce the causes. For example, in medical diagnosis, the more symptoms of a patient are available, the more accurately a doctor can diagnose the patient's disease. Without loss of generality, we assume that the causes are arranged in a topological order such that X_l is not a cause of X_k for $k < l$, and that $Y_1, \ldots, Y_q$ subsequent to X are arranged in a topological order such that Y_j is not a cause of Y_m for $m < j$. Let $g(y)$ be a known function weighting the importance of multiple effects in Y. For example, an additive weighting function is $g(Y) = \sum_{i=1}^{q} a_i \times Y_i$, where a_i is a weight for Y_i. Let X_S denote a subvector of X, where S is the subset of indexes $\{1, \ldots, p\}$, and let $x_S^1 \succcurlyeq x_S^0$ denote $x_i^1 \geq x_i^0$ for each $i \in S$. For a treated group of $X_S = x_S \neq \mathbf{0}$ versus a control group of $X_S = \mathbf{0}$, where $\mathbf{0}$ denotes $(0, \ldots, 0)$, we define the posterior total causal effect of multiple causes $X_S = x_S$ on multiple effects $Y = (Y_1, \ldots, Y_q)$ as

$$\mathrm{PostTCE}[X_S(x_S) \Rightarrow Y|O = o] = E[g(Y_{x_S}) - g(Y_{X_S=\mathbf{0}})|O = o].$$

Differing from the conditional counterfactual causal effect, defined by Zhao et al. [42], which restricts $x_S = (1, \ldots, 1)$, the aforementioned definition does not require this restriction. Comparing PostTCE of X_S on Y with different values x_S and x'_S, we can obtain various posterior controlled direct causal effects and interaction effects. For example, given the observed evidence $o = (X_1 = 1, X_2 = 1, X_3 = 1, Y = 1)$, a controlled direct causal effect of a set (X_1, X_2) on Y by controlling for $X_3 = x_3$ can be measured by

$$\mathrm{PostTCE}[X_{\{1,2,3\}}(1, 1, x_3) \Rightarrow Y|O = o] - \mathrm{PostTCE}[X_{\{1,2,3\}}(0, 0, x_3) \Rightarrow Y|O = o].$$

Comparing PostTCE across different subsets X_S and $X_{S'}$, we contrast whether an event should be attributed more to X_S or to $X_{S'}$. For example, given the observed evidence $o = (X_1 = 1, X_2 = 1, X_3 = 1, Y = 1)$, comparing $\mathrm{PostTCE}[X_1(1) \Rightarrow Y|O = o]$, $\mathrm{PostTCE}[X_2(1) \Rightarrow Y|O = o]$, and $\mathrm{PostTCE}[X_3(1) \Rightarrow Y|O = o]$, we can argue that the event $Y = 1$ should be attributed most to X_1, X_2, or X_3.

It can be shown that for an additive weighting function $g(Y) = \sum_{j=1}^{q} a_j \times Y_j$,

$$\mathrm{PostTCE}[X_S(x_S) \Rightarrow Y|O = o] = \sum_{j=1}^{q} a_j \times \mathrm{PostTCE}[X_S(x_S) \Rightarrow Y_j|O = o].$$

In terms of PostTCE, we can do the attributions of multiple effects to multiple causes with interaction effects.

The posterior intervention causal effect of X_S on Y is defined as

$$\mathrm{PostICE}(X_S \Rightarrow Y|O = o) = E[g(Y) - g(Y_{X_S=\mathbf{0}})|O = o].$$

When the evidence O includes some $(Y_j = 1)$'s, comparing $\mathrm{PostICE}(X_S \Rightarrow Y|O = o)$ with $\mathrm{PostICE}(X_{S'} \Rightarrow Y|O = o)$, we can retrospectively judge which of X_S and $X_{S'}$ might make the happened effects more likely. For $g(y_1, \ldots, y_q) = \sum_{j=1}^{q} y_j$, the posterior intervention causal effect $\mathrm{PostICE}(X_S \Rightarrow Y|O = o)$ measures the expected number of outcomes eliminated by removing risk factors in X_S.

3.2 Identification assumptions of posterior causal effects

Let $(X, Y) = (V_1, \ldots, V_{p+q})$ be arranged in a causal order and $V_{r:s} = (V_r, V_{r+1}, \ldots, V_s)$ be a subvector of V for $r \leq s$. Let $(V_s)_{v_{1:s-1}}$ denote the potential outcome of V_s if $V_{1:s-1}$ were intervened to $v_{1:s-1}$ To identify these posterior causal effects, we need to follow the monotonicity and no-confounding assumptions [13].

Assumption 1. (Monotonicity) For $s = 2, \ldots, p + q$, the potential outcomes of V_s satisfy the monotonicity relation: $(V_s)_{v^*_{1:s-1}} \leq (V_s)_{v_{1:s-1}}$ whenever $v^*_{1:s-1} \preccurlyeq v_{1:s-1}$.

This assumption is often expressed as "no prevention" in epidemiology and states that no individual can be helped by exposure to a risk factor. For example, let V_1, V_2, and V_3 denote poor diet, high blood pressure and stroke, respectively. The monotonicity assumption means that poor diet and high blood pressure are two potential risk factors for stroke. Exposures to them are not preventive for stroke, and a poor diet is also not preventive for high blood pressure. The validity of monotonicity cannot be directly tested, but this assumption imposes testable restrictions on the probability distribution of observed data in certain cases. Similar assumptions are often made in studies of imperfect compliance of treatment.

Assumption 2. (No confounding)

(i) There is no confounding between V_s and $V_{1:s-1}$, i.e., $(V_s)_{v_{1:s-1}} \perp\!\!\!\perp V_{1:s-1}$ for all $v_{1:s-1}$ and $s = 2, \ldots, p + q$;

(ii) The elements in $\{(V_s)_{v_{1:s-1}}\}_{s=2}^{p+q}$ are mutually independent for any given $v_{1:p+q-1}$.

Assumption 2 (i) means that the potential outcomes of each variable are independent of its precedent variables arranged in the causal order. If V_s has a causal structural model $V_s = f_s(V_{1:s-1}, \varepsilon_s)$ and an error variable $\varepsilon_s \perp\!\!\!\perp \varepsilon_{1:s-1}$, then Assumption 2 is equivalent to the absence of latent confounders. Assumption 2 excludes the presence of unobserved confounders between variables X and Y. However, each variable X_k may still confound the relationships between Y and X_l or between X_l and X_s for where $k < l, s$. When there exists a set C containing observed background variables that are not influenced by X, the independence in Assumption 2 can be relaxed to those conditional on C.

When the evidence does not contain any effect variable Y_i, the identification of posterior causal effects only requires Assumption 2 of no confounding. But when the evidence contains some effect variables, the identification of posterior causal effects requires both Assumptions 1 and 2. First, consider the case with a single X and a single Y. For the case of a single X and a single Y, Assumption 1 of monotonicity means $Y_{X=0} \leq Y_{X=1}$, and PN has the following equation:

$$\mathrm{PN} = \frac{\mathrm{pr}(Y = 1|X = 1) - \mathrm{pr}(Y_0 = 1|X = 1)}{\mathrm{pr}(Y = 1|X = 1)}.$$

The numerator is the treatment effect on treated. Under Assumption 2 of no confounding, we have $\mathrm{pr}(Y_0 = 1|X = 1) = \mathrm{pr}(Y = 1|X = 0)$, and thus, PN is identifiable. Similarly, under Assumptions 1 and 2 of monotonicity and no-confounding, it can be shown that the aforementioned posterior causal effects defined by intervening on a single cause X_k are identifiable [13,14]. When simultaneously intervening on a set $X_S = x_S$ of multiple causes, for the identification of $\mathrm{PostTCE}[X_S(x_S) \Rightarrow Y|O = o]$, we further need the restriction on the relationship between x_S and o. Let $X_{S^*} = X_S \cap O$, and x_{S^*} and x'_{S^*} are the value of X_{S^*} in x_S and o, respectively. Let $X_{S'} = X_S \backslash X_{S^*}$, and $x_{S'}$ is the value of $X_{S'}$. When $q = 1$, i.e., $Y = Y_1$, and the evidence contains the effect variable $Y = y$, we have the following theorem.

Theorem 1. *Suppose that Assumptions* 1 *and* 2 *hold.* $\mathrm{PostTCE}[X_S(x_S) \Rightarrow Y|O = o]$ *is identifiable if one of the following conditions holds*:
(1) $x_{S^*} \succcurlyeq x'_{S^*}$ *and* $x_{S'} = (1, \ldots, 1)$;
(2) $x'_{S^*} \succcurlyeq x_{S^*}$ *and* $x_{S'} = (0, \ldots, 0)$.

For the case of $q > 1$ and that the evidence O includes Y_k, let $X_O = O \cap X$ and $Y_O = O \cap \{Y_1, \ldots, Y_{k-1}\}$. The following equality holds from Zhao et al. [42]:

$$\mathrm{PostTCE}[X_S(x_S) \Rightarrow Y_k|O = o] = \mathrm{PostTCE}[X_S(x_S) \Rightarrow Y_k|x_O, y_O, Y_k = y].$$

Therefore, for an additive function $g(Y)$, we have

$$\mathrm{PostTCE}[X_S(x_S) \Rightarrow Y|O = o] = \sum_{j=1}^{q} a_j \times \mathrm{PostTCE}[X_S(x_S) \Rightarrow Y_j|x_O, y_O, Y_k = y].$$

When each item of the aforementioned equation is identifiable, $\mathrm{PostTCE}[X_S(x_S) \Rightarrow Y|O = o]$ is identifiable.

3.3 Relationship between posterior causal effect and population attributable risk

Greenland [10] pointed out that there are many incorrect equations regarding the probabilities of causation and the population attributable risks. The population attributable risks are used to measure the proportional amounts by which a disease risk would be reduced if risk factors were eliminated from a population [43]. For example, how much of the disease burden due to leukemia in a population could be eliminated if the exposures of benzene and ionizing radiation were eliminated from the population. In the following, we explain the relation of the posterior causal effects to the population attributable risks. For a case of a single Y and multiple causes $X = (X_1, \ldots, X_p)$, the population attributable risk is defined by Bruzzi et al. [44] as follows:

$$\mathrm{AR} = \frac{\mathrm{pr}(Y = 1) - \mathrm{pr}(Y = 1|X_1 = 0, \ldots, X_p = 0)}{\mathrm{pr}(Y = 1)}.$$

It measures the proportional amount by which a disease risk would be reduced if all risk factors were eliminated from a population. Under Assumption 2 of no confounding and a weak monotonicity assumption that $Y_{X=\mathbf{0}} \leq Y_{X=x} \leq Y_{X=\mathbf{1}}$ for any x, the population attributable risk is equal to the posterior causal effect of multiple causes X on Y given the evidence of $Y = 1$, i.e.,

$$\mathrm{AR} = \mathrm{PostTCE}(X \Rightarrow Y|Y = 1) = E(Y_{X=\mathbf{1}} - Y_{X=\mathbf{0}}|Y = 1). \tag{2}$$

AR does not measure how much the disease Y is attributed to a specified risk factor X_k. The adjusted attributable risk for X_k is defined by adjusting for the remaining risk factors $X_{-k} = X \backslash \{X_k\}$ [44]:

$$\mathrm{AR}(X_k|X_{-k}) = \frac{\mathrm{pr}(Y = 1) - \sum_{x_{-k}} \mathrm{pr}(Y = 1|X_k = 0, x_{-k})\mathrm{pr}(x_{-k})}{\mathrm{pr}(Y = 1)}.$$

Note that the set X_{-k} should not contain any intermediate factor between X_k and Y since eliminating X_k can affect the intermediate factors in the condition X_{-k}. Let $A_k = (X_1, \ldots, X_{k-1})$ denote the variable set, which is prior to X_k in a topological causal order, and thus, $\mathrm{AR}(X_k|A_k)$ is a proper adjusted attributable risk. Lu et al. [13] showed that under Assumption 2 of no confounding and a weak monotonicity assumption $Y_{X_k=0} \leq Y_{X_k=1}$, the attributable risk of X_k on Y adjusted for the set A_k is equal to the posterior total effect of X_k on Y given the evidence of $Y = 1$:

$$\mathrm{AR}(X_k|A_k) = \mathrm{PostTCE}(X_k \Rightarrow Y|Y = 1). \tag{3}$$

Equations (2) and (3) show the relationships between the posterior causal effects and the population and adjusted attributable risks, and they explain the causal meaning of the attributable risks in terms of the potential outcome framework. The equations also give other identification equations of posterior causal effects $\mathrm{PostTCE}(X \Rightarrow Y|Y = 1)$ and $\mathrm{PostTCE}(X_k \Rightarrow Y|Y = 1)$ under the weaker monotonicity assumptions than Assumption 1 of monotonicity.

3.4 Diagnostic approaches based on Bayesian posterior probabilities and posterior causal effects

In the following, we discuss the problem about whether the medical diagnosis should be based on Bayesian posterior probabilities or the posterior causal effects. Bayesian posterior probabilities measure the uncertainty of past events given the later observed evidence, but they do not capture the causal relationships between these events. Posterior causal effects, on the other hand, measure the uncertainty of past events that have causal effects on the later happened evidence.

In the field of medical diagnosis, a probabilistic expert system based on Bayesian posterior probabilities was developed by Lauritzen and Spiegelhalter [45] and Spiegelhalter et al. [46]. This system computes the posterior probabilities of diseases given the observed symptoms, which depend on the prior probabilities of the diseases. The diagnosis based on the maximum posterior probability minimizes the misdiagnosis error [47]. However, this approach does not account for the causal relationships between diseases and symptoms.

As Encyclopaedia Britannica [48] defines, the diagnostic process is the method by which health professionals select one disease over another, identifying one as the most likely cause of a person's symptoms. Richens et al. [49] pointed out that most existing diagnostic algorithms, including Bayesian model-based and deep learning methods, rely on associative inference, and they identified diseases based on how correlated they are with a patient's symptoms and medical history. This contrasts with how doctors perform medical diagnosis, selecting the diseases that offer the best causal explanations for the patient's symptoms. They argued that disease diagnostic reasoning should satisfy three principles concerning not only the posterior probability, but also causality and simplicity. They proposed an approach based on a noisy-operation model, but their model is restricted to the case where neither diseases nor symptoms can affect each other. Li et al. [14] proposed a medical diagnostic approach based on posterior intervention causal effects PostICE, which satisfies the aforementioned principles for medical diagnosis. For disease diagnosis, the evidence O contains some symptoms and backgrounds of a patient, but does not contain the status of diseases X_k. Since $\mathrm{PostICE}(X_k \Rightarrow Y|O = o, X_k = 0) = 0$, we can obtain

$$\mathrm{PostICE}(X_k \Rightarrow Y|o) = \mathrm{PostICE}(X_k \Rightarrow Y|o, X_k = 1) \times \mathrm{pr}(X_k = 1|o). \tag{4}$$

This equation means that the diagnostic approach based on PostICE considers not only Bayesian posterior probability $\mathrm{pr}(X_k = 1|o)$, but also the posterior causal effect of the disease X_k on the symptoms Y in the subpopulation with the disease $X_k = 1$, which is ignored by the approach based on Bayesian posterior probability. For a patient given the evidence $O = o$, we diagnose the patient with a disease X_k, which has the largest value in $\{\mathrm{PostICE}(X_j \Rightarrow Y|o), \forall j\}$. It means that the number of symptoms could be eliminated at most if the patient had not gotten the disease X_k.

When a patient may have multiple diseases simultaneously, Bayesian approach diagnoses the patient with multiple diseases based on the maximum posterior probabilities $\mathrm{pr}(X = x|O = o)$. The approach based on posterior causal effects uses $\mathrm{PostICE}(X_S \Rightarrow Y|O = o)$ for diagnosis. For a patient given the evidence $O = o$, we diagnose the patient with multiple diseases X_S, which has the largest value in $\{\mathrm{PostICE}(X_{S'} \Rightarrow Y|o), \forall S' \subseteq \{1, \ldots, p\}\}$. Bayesian posterior probabilities and posterior intervention causal effects have the following equation:

$$\mathrm{PostICE}(X_S \Rightarrow Y|O = o) = \sum_{x_s} \mathrm{PostICE}(X_S \Rightarrow Y|X_S = x_s, O = o) \times \mathrm{pr}(X_S = x_s|O = o).$$

If the diagnostic result is used further for eliminating the symptoms in the future, the Bayesian diagnostic approach may not be optimal, and the diagnostic approach based on posterior causal effects requires an assumption of invariance maybe require the following assumption of invariant relationships between causes and effects in the past and the future.

Assumption 3. (Invariance) Let W and Z denote the future diseases and symptoms, respectively. The potential outcomes of the past and future symptoms are the same (i.e., $Y_x = Z_w$) if the statuses of the past and future diseases are the same (i.e., $X = W$).

This assumption can be weakened to that the average causal effects of the diseases on symptoms are invariant across time in the subpopulations of $O = o$, i.e., $E(Y_{X=1} - Y_{X=0}|O = o) = E(Z_{W=1} - Z_{W=0}|O = o)$. This invariance assumption may hold for many real scenarios, e.g., in a specified room, X denotes a switch on or off, and Y a light on or not. But the assumption may not hold in some real scenarios, e.g., X denotes that Jack drank poison and $Y = 1$ denotes that he died.

In the following, we use a numerical example of medical diagnosis to compare the approach based on Bayesian posterior probabilities with that based on the posterior causal effects.

Example 1. Let X_1 and X_2 denote two diseases and Y a symptom caused by X_1. Consider the causal mechanism described by the diagram in Figure 2, where X_1 is the cause of disease X_2 and symptom Y, but X_2 is not the cause of Y. Suppose that the causal diagram has the following probabilities:

$$\mathrm{pr}(X_1 = 1) = 0.400, \quad \mathrm{pr}[(X_2)_{x_1=0} = 1] = 0.550, \quad \mathrm{pr}[(X_2)_{x_1=1} = 1] = 0.622,$$
$$\mathrm{pr}(Y_{x_1=0} = 1) = 0.401, \quad \mathrm{pr}(Y_{x_1=1} = 1) = 0.500.$$

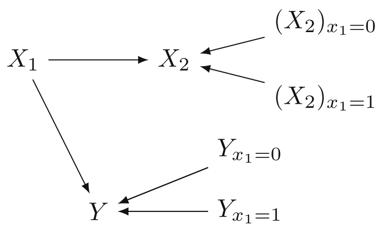

Figure 2: Causal diagram of two diseases X_1 and X_2 and a symptom Y.

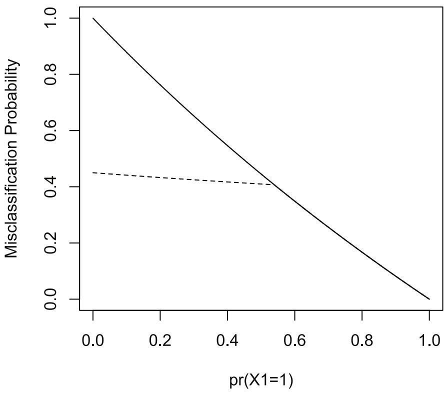

Figure 3: Misclassification probabilities of two approaches.

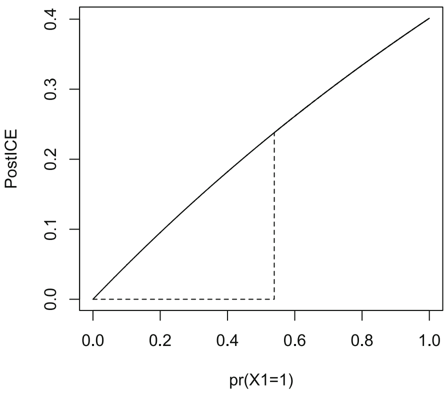

Figure 4: PostICEs of two approaches.

Thus, the observed variables X_2 and Y are generated by

$$X_2 = X_1 \times (X_2)_{X_1=1} + (1 - X_1) \times (X_2)_{X_1=0},$$
$$Y = X_1 \times Y_{X_1=1} + (1 - X_1) \times Y_{X_1=0}.$$

From the probabilities, we can obtain the posterior probabilities and causal effects given the symptom $Y = 1$:

$$\begin{aligned} \mathrm{pr}(X_1 = 0, X_2 = 0|Y = 1) &= 0.246, & \mathrm{pr}(X_1 = 0, X_2 = 1|Y = 1) &= 0.300, \\ \mathrm{pr}(X_1 = 1, X_2 = 0|Y = 1) &= 0.171, & \mathrm{pr}(X_1 = 1, X_2 = 1|Y = 1) &= 0.283, \\ \mathrm{PostICE}[(X_1, X_2) \Rightarrow Y|Y = 1] &= 0.272, & \mathrm{PostICE}(X_{S=\emptyset} \Rightarrow Y|Y = 1) &= 0.000, \\ \mathrm{PostICE}(X_1 \Rightarrow Y|Y = 1) &= 0.272, & \mathrm{PostICE}(X_2 \Rightarrow Y|Y = 1) &= 0.000. \end{aligned}$$

For Bayesian approach based on posterior probabilities, the diagnostic results based on the maximum joint and marginal posterior probabilities of $\mathrm{pr}(x_1, x_2|Y = 1)$ and $\mathrm{pr}(x_k|Y = 1)$ are $(X_1, X_2) = (0, 1)$ and $(X_2 = 1)$, respectively. Neither of these results identifies the true cause X_1 of symptom Y. In contrast, by the maximum intervention posterior causal effects $\mathrm{PostICE}(X_S \Rightarrow Y|Y = 1)$, the diagnostic results are $X_S = (X_1, X_2) = (1, 1)$ and $X_S = X_1 = 1$, respectively, since they have the maximum value of 0.182. Either diagnostic result finds the true cause X_1 of symptom Y, and by simplicity, the diagnosis prefers X_1.

In the following, we first compare the misclassification probabilities of the two diagnostic approaches. The diagnostic results of Bayesian approach are $(X_1, X_2) = (0, 1)$ and $(X_2 = 1)$. Overall, we consider the diagnostic result to be taking on the disease X_2, so the individuals without the disease X_2 ($X_2 = 0$), which include individuals with $(X_1 = 1, X_2 = 0)$ and $(X_1 = 0, X_2 = 0)$, are misclassified. Thus, the misclassification probability of Bayesian diagnostic approach for the population of $Y = 1$ is

$$\mathrm{pr}[(X_1, X_2) = (0, 0)|Y = 1] + \mathrm{pr}[(X_1, X_2) = (1, 0)|Y = 1] = \mathrm{pr}(X_2 = 0|Y = 1) = 0.417.$$

The diagnostic results of posterior causal effect approach are $(X_1, X_2) = (1, 1)$ and $(X_1 = 1)$. We consider the diagnostic result to be taking on the disease X_1, and the individuals without the disease X_1 $(X_1 = 0)$, which include individuals with $(X_1 = 0, X_2 = 0)$ and $(X_1 = 0, X_2 = 1)$, are misclassified. Thus, the misclassification probability of diagnostic approach based on posterior causal effects for the population of $Y = 1$ is

$$\mathrm{pr}[(X_1, X_2) = (0, 0)|Y = 1] + \mathrm{pr}[(X_1, X_2) = (0, 1)|Y = 1] = \mathrm{pr}(X_1 = 0|Y = 1) = 0.546.$$

Bayesian diagnostic approach has a lower misclassification probability than the posterior causal effect approach.

Furthermore, we vary the prior probability $\mathrm{pr}(X_1 = 1)$ of disease X_1, and we show the misclassification probabilities of the two diagnostic approaches in Figure 3, where the solid line is the misclassification probability for the posterior causal effect approach and the dotted line is that for Bayesian approach. It can be seen that as $\mathrm{pr}(X_1 = 1)$ increases, the misclassification probability of the posterior causal effect approach decreases since it always diagnoses patients with disease $X_1 = 1$. For a lower prior probability $\mathrm{pr}(X_1 = 1)$, Bayesian approach diagnoses patients with the disease $X_2 = 1$, and its misclassification probability is lower than that of the posterior causal effect approach. When $\mathrm{pr}(X_1 = 1)$ increases to a certain extent, Bayesian approach changes the diagnosis $X_2 = 1$ to $X_1 = 1$, and thus, it has the same misclassification probability as the posterior causal effect approach.

Next, we compare the causal effect of treating the diagnosed disease on the elimination of symptoms. To evaluate the causal effects of the treatment after diagnosis, we make Assumption 3 of invariance. Let X_k denote the diagnosed disease, and then, the posterior intervention causal effect $\mathrm{PostICE}(X_k \Rightarrow Y|Y = 1)$ measures the elimination of symptoms attributed to the intervention on X_k. $\mathrm{PostICE}(X_k \Rightarrow Y|Y = 1)$s for the two diagnostic approaches are shown in Figure 4. The posterior causal effect approach always diagnoses patients with disease $X_1 = 1$ and its $\mathrm{PostICE}(X_1 \Rightarrow Y|Y = 1)$ increases as $\mathrm{pr}(X_1 = 1)$ increases. Bayesian approach diagnoses the patients with disease $X_2 = 1$ for a lower $\mathrm{pr}(X_1 = 1)$, and $\mathrm{PostICE}(X_2 \Rightarrow Y|Y = 1) = 0$, which is much less than $\mathrm{PostICE}(X_1 \Rightarrow Y|Y = 1)$ obtained by the posterior causal effect approach. When $\mathrm{pr}(X_1 = 1)$ increases to a certain extent, Bayesian approach changes the diagnostic result to disease $X_1 = 1$, and then, it has the same value of $\mathrm{PostICE}(X_1 \Rightarrow Y|Y = 1)$ as that of the posterior causal effect approach.

In this example, the posterior causal effect approach represents a white-box method with complete knowledge of the causal mechanisms. It would never diagnose patients with symptom $Y = 1$ as suffering from disease X_2, which is non-causative of the symptom. The Bayesian posterior probability approach can be viewed as a black-box method without any knowledge of the underlying causal mechanisms. Although it may diagnose some patients with symptom $Y = 1$ as having disease X_2 despite its non-causal relationship, this approach has the minimum probability of misdiagnosis overall.

To diagnose possible diseases, regardless of whether or not they are the causes of the occurred symptoms, Bayesian diagnosis based on posterior probabilities always has the minimum misclassification probability [47]. One drawback of the Bayesian diagnostic approach is that it may not identify the causes of occurred symptoms. To identify the causes, we argue that the approach based on posterior causal effects is a better choice. In the aforementioned numerical example, we assume that the probabilities of the causal mechanism are known. A limitation of the approach based on posterior causal effects is that the identifiability of the posterior causal effects requires Assumptions 1 and 2 of monotonicity and no confounding. Under Assumption 1 of monotonicity, patients with symptom $Y = 1$ are always diagnosed with certain diseases and are not diagnosed with no disease since $\mathrm{PostICE}(X_{S=\varnothing} \Rightarrow Y|Y = 1) = 0$ has the least value.

4 Discussion

Prospective and retrospective causal inferences investigate causality from different perspectives. Prospective inference reasons forward from causes to effects. Randomized experiments represent the gold standard for prospective causal inference. In contrast, retrospective inference works backward from observed outcomes to infer their potential causes. However, there is currently no established gold standard methodology for retrospective causal analysis. Posterior causal effects can be utilized for prospective causal inference when the available evidence lacks known outcome details. For example, let X denote smoking and Y denote lung cancer. The average treatment effect on the treated, $E(Y_{X=1} - Y_{X=0}|X = 1)$, evaluates the causal influence of smoking on lung cancer risk in an exposed population. However, it cannot definitively conclude whether smoking causes lung cancer. Conversely, posterior causal effects are employed in retrospective causal analysis when the evidence includes known outcome information. The posterior causal effect $E(Y_{X=1} - Y_{X=0}|X = 1, Y = 1)$ assesses the causal impact of smoking on lung cancer in the subpopulation of smokers diagnosed with lung cancer. It estimates the probability that lung cancer patients in this group would not have developed lung cancer had they not smoked. Similarly, the posterior causal effect $E(Y_{X=1} - Y_{X=0}|Y = 1)$ evaluates the causal effect of smoking on lung cancer in the overall lung cancer patient population. It gauges the probability that these patients would not have had lung cancer without smoking. Thus, posterior causal effects can quantify the attributable risks of smoking within specific patient groups.

Confounding poses a challenge in causal inference, as identifying confounders is difficult using only observational data. Without untestable assumptions, we cannot definitively determine if a covariate is a confounder. The surrogate paradox further demonstrates that probabilistic causal effects are generally non-transitive. Specifically, it shows that the signs or directions of causal impacts cannot be logically deduced from the probabilistic outputs of causal analyses. In other words, logistic reasoning does not necessarily apply to the probabilistic results of causal inference.

For a diagnostic problem, Bayesian posterior probability approach may minimize misclassification probability, while the posterior causal effect approach may identify the causes of occurred symptoms. The suitable diagnostic method depends on whether the goal is to predict diseases or uncover the causes of presented symptoms, or even eliminate those symptoms.

Similar to posterior probabilities, posterior causal effects also derive from Bayesian thinking and use potential outcome framework. In retrospective causal analysis, potential outcomes are essential for expressing counterfactual scenarios. However, the posterior causal effects differ from the Bayesian posterior distributions of causal effects. The posterior causal effects are the expectations of the causal effects conditional on the observed evidence, rather than posterior distributions.

Causal graphs may be learned from observed data, which depict causal relationships among variables in a population. But a causal graph cannot be used to deduce the causes of effects for a specific individual because different individuals in the population may have different causes that depend on the evidence.

Many open questions remain regarding retrospective causal inference. It shares numerous topics with prospective causal inference, but may also involve unique considerations specific to reasoning backward from effects to causes. For a given set of evidence about occurred outcomes, there can be different conceptual causes depending on the research objective. In some real-world applications, identifying the root causes of effects may be of greater interest.

Acknowledgement: We would like to thank the editors and the three reviewers for their very helpful and valuable comments, which led to a significant improvement of this manuscript.

Funding information: This research was partially supported by the National Natural Science Foundation of China (No. 12071015), the Disciplinary Funding of Beijing Technology and Business University (No. 50500101002), the Joint Key Research Project funded by the Beijing Municipal Education Commission and the Beijing Municipal Natural Science Foundation (No. 23JA0006), the Research Foundation for Advanced Talents of Beijing Technology and Business University (No. 19008024084), and a joint research project of Alibaba group.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

References

[1] Neyman JS. On the application of probability theory to agricultural experiments. Stat Sci. 1923;5:465–80 (1990). http://www.jstor.org/stable/2245382.

[2] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701. doi: 10.1037/h0037350.

[3] Halpern JY. Actual causality. London: MIT Press; 2016.

[4] Pearl J, Mackenzie D. The book of why: the new science of cause and effect. New York: Basic Books; 2018.

[5] Dawid AP, Faigman DL, Fienberg SE. Fitting science into legal contexts: Assessing effects of causes or causes of effects? Soc Meth Res. 2014;43(3):359–90. doi: 10.1177/0049124113515188.

[6] Holland PW. Statistics and causal inference. J Amer Statist Assoc. 1986;81(396):945–60. doi: 10.1080/01621459.1986.10478354.

[7] Dawid AP, Musio M. Effects of causes and causes of effects. Annu Rev Stat Appl. 2022;9:261–87. doi: 10.1146/annurev-statistics-070121-061120.

[8] Dawid AP, Faigman DL, Fienberg SE. On the causes of effects: Response to Pearl. Soc Meth Res. 2015;44(1):165–74. doi: 10.1177/0049124114562613.

[9] Robins J, Greenland S. The probability of causation under a stochastic model for individual risk. Biometrics. 1989;45:1125–38. doi: 10.2307/2531765.

[10] Greenland S. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. Am J Public Health. 1999;89(8):1166–9. doi: 10.2105/AJPH.89.8.1166.

[11] Pearl J. Probabilities of causation: three counterfactual interpretations and their identification. Synthese. 1999;121:93–149. doi: 10.1023/A:1005233831499.

[12] Dawid AP. Causal inference without counterfactuals. J Amer Statist Assoc. 2000;95(450):407–24. doi: 10.2307/2669377.

[13] Lu Z, Geng Z, Li W, Zhu S, Jia J. Evaluating causes of effects by posterior effects of causes. Biometrika. 2023;110(2):449–65. doi: 10.1093/biomet/asac038.

[14] Li W, Lu Z, Jia J, Xie M, Geng Z. Retrospective causal inference with multiple effect variables. Biometrika. 2024;111(2):573–89. doi: 10.1093/biomet/asad056.

[15] Miettinen OS, Cook EF. Confounding: essence and detection. Am J Epidemiol. Oct 1981;114(4):593–603. doi: 10.1093/oxfordjournals.aje.a113225.

[16] Boivin JF, Wacholder S. Conditions for confounding of the risk ratio and of the odds ratio. Am J Epidemiol. 1985;121(1):152–8. doi: 10.1093/oxfordjournals.aje.a113977.

[17] Grayson D. Confounding confounding. Am J Epidemiol. 1987;126(3):546–53. doi: 10.1093/oxfordjournals.aje.a114687.

[18] Greenland S, Holland PW, Mantel N, Wickramaratne PJ, Holford TR. Confounding in epidemiologic studies. Biometrics. 1989;45(4):1309–22. doi: 10.2307/2531783.

[19] Weinberg CR. Toward a clearer definition of confounding. Am J Epidemiol. 1993;137(1):1–8. doi: 10.1093/oxfordjournals.aje.a116591.

[20] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10:37–48. https://www.jstor.org/stable/3702180.

[21] Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. Statist Sci. 1999;14:29–46. doi: 10.1214/ss/1009211805.

[22] Wickramaratne PJ, Holford TR. Confounding in epidemiologic studies: the adequacy of the control group as a measure of confounding. Biometrics. 1987;43(4):751–65. http://www.jstor.org/stable/2531530.

[23] Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. New York: Van Nostrand Reinhold; 1982.

[24] Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15:413–9. doi: 10.1093/ije/15.3.413.

[25] Geng Z, Guo J, Fung WK. Criteria for confounders in epidemiological studies. J R Stat Soc Ser B (Stat Methodol). 2002;64(1):3–15. doi: 10.1111/1467-9868.00321.

[26] VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Statist. 2013;41(1):196–220. doi: 10.1214/12-aos1058.

[27] Fleming TR, Demets DL. Surrogate end points in clinical trials: Are we being misled? Ann Intern Med. 1996;125(7):605–13. doi: 10.7326/0003-4819-125-7-199610010-00011.

[28] Chen H, Geng Z, Jia J. Criteria for surrogate end points. J R Stat Soc Ser B (Stat Methodol). 2007;69(5):919–32. doi: 10.1111/j.1467-9868.2007.00617.x.

[29] Jiang Z, Ding P, Geng Z. Qualitative evaluation of associations by the transitivity of the association signs. Statist Sinica. 2015;25(3):1065–79. http://www.jstor.org/stable/24721221.

[30] Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. 1989;8(4):431–40. doi: 10.1002/sim.4780080407.

[31] Dawid AP. Conditional independence in statistical theory. J R Stat Soc Ser B (Stat Methodol). 1979;41(1):1–15. doi: 10.1111/j.2517-6161.1979.tb01052.x.

[32] Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics. 2002;58:21–9. doi: 10.1111/j.0006-341X.2002.00021.x.

[33] Lauritzen S. Discussion on causality. Scand J Stat. 2004;31(2):189–93. doi: 10.1111/j.1467-9469.2004.03-200A.x.

[34] Jiang Z, Ding P, Geng Z. Principal causal effect identification and surrogate end point evaluation by multiple trials. J R Stat Soc Ser B Stat Meth. Nov 2016;78(4):829–48. doi: 10.1111/rssb.12135.

[35] Moore T. Deadly medicine: why tens of thousands of patients died in America's worst drug disaster. New York: Simon & Schuster; 1995.

[36] Investigators CASTC. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. New Engl J Med. 1989;321:406–12. doi: 10.1056/NEJM198908103210629.

[37] Ju C, Geng Z. Criteria for surrogate end points based on causal distributions. J R Stat Soc Ser B (Stat Methodol). 2010;72(1):129–42. doi: 10.1111/j.1467-9868.2009.00729.x.

[38] Wu Z, He P, Geng Z. Sufficient conditions for concluding surrogacy based on observed data. Stat Med. 2011;30(19):2422–34. doi: 10.1002/sim.4273.

[39] Joffe M. Discussion on "Surrogate measures and consistent surrogates". Biometrics. 2013;69(3):572–5. https://www.jstor.org/stable/24538121.

[40] Luo P, Cai Z, Geng Z. Criteria for multiple surrogates. Statist Sinica. 2019;29(3):1343–66. https://www.jstor.org/stable/26706005.

[41] Pearl J. Probabilities of Causation: Three Counterfactual Interpretations and Their Identification. 1st ed. New York: Association for Computing Machinery; 2022. p. 317–72. doi: 10.1145/3501714.3501735.

[42] Zhao R, Zhang L, Zhu S, Lu Z, Dong Z, Zhang C, et al. Conditional counterfactual causal effect for individual attribution. In: Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence. vol. 216 of Proceedings of Machine Learning Research. PMLR; 2023. p. 2519–28. https://proceedings.mlr.press/v216/zhao23a.html.

[43] Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health. 1998;88(1):15–9. doi: 10.2105/ajph.88.1.15.

[44] Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol. 1985;122(5):904–14. doi: 10.1093/oxfordjournals.aje.a114174.

[45] Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. J R Stat Soc Ser B (Stat Methodol). 1988;50(2):157–94. doi: 10.1111/j.2517-6161.1988.tb01721.x.

[46] Spiegelhalter DJ, Dawid AP, Lauritzen SL, Cowell RG. Bayesian analysis in expert systems. Statist Sci. 1993;8(3):219–47. https://www.jstor.org/stable/2245959.

[47] Berger JO. Statistical decision theory and Bayesian analysis. New York: Springer Science & Business Media. 2013.

[48] Rakel RE. Diagnosis. 2023. https://www.britannica.com/science/diagnosis.

[49] Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun. 2020;11(1):1–9. doi: 10.1038/s41467-020-17419-7.