



**HAL**  
open science

# On the (in)compatibility between potential outcomes and structural causal models and its signification in counterfactual inference

Lucas de Lara

► **To cite this version:**

Lucas de Lara. On the (in)compatibility between potential outcomes and structural causal models and its signification in counterfactual inference. 2024. hal-04203003v3

**HAL Id: hal-04203003**

**<https://hal.science/hal-04203003v3>**

Preprint submitted on 17 Mar 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# On the (in)compatibility between potential outcomes and structural causal models and its signification in counterfactual inference

Lucas De Lara\*

Institut de Mathématiques de Toulouse, Université Paul Sabatier

## Abstract

Most of the scientific literature on causal modeling considers the structural framework of Pearl and the potential-outcome framework of Rubin to be formally equivalent, and therefore interchangeably uses the do-notation and the potential-outcome subscript notation to write counterfactual outcomes. In this paper, we agnostically superimpose the two causal models to specify under which mathematical conditions structural counterfactual outcomes and potential outcomes need to, do not need to, can, or cannot be equal (almost surely or law). Our comparison reminds that a structural causal model and a Rubin causal model compatible with the same observations do not have to coincide, and highlights real-world problems where they even cannot correspond. Then, we examine common claims and practices from the causal-inference literature in the light of these results. In doing so, we aim at clarifying the relationship between the two causal frameworks, and the interpretation of their respective counterfactuals.

**Keywords:** causality, structural causal models, potential outcomes

## 1 Introduction

Understanding causation between phenomena rather than mere association is a fundamental scientific challenge. Over the last decades, two mathematical frameworks using a terminology based on random variables have become the gold standards to address this problem.

On the one hand, the notorious *structural account* of Pearl [2009] rests on the knowledge of a *structural causal model* (SCM) which specifies all cause-effect equations between observed random variables (often depicted by a graph). The interest of such equations comes from the possibility of carrying out *do-interventions*: forcing a variable to take a given value while keeping the rest of the mechanism untouched. More concretely, let  $T$  and  $Y$  be observed variables of the model such that we would like to understand the downstream effect of  $T$  onto  $Y$ . Replacing the formula generating  $T$  by  $T = t$  for a given possible value  $t \in \mathcal{T}$  and propagating this change through the other equations defines the altered variable  $Y_{T=t}$ , representing  $Y$  had  $T$  been equal to  $t$ .

On the other hand, the widely-used *potential-outcome account* of Rubin [1974] mathematically formalizes causal inference in clinical trials. Letting  $T$  denote a *treatment status* (e.g., taking a drug or not) and  $Y$  an *outcome* of interest (e.g., recovering or not), a so-called *Rubin causal model* (RCM) postulates the existence of *potential outcomes*  $(Y_t)_{t \in \mathcal{T}}$  representing what the outcome would be were  $T$  equal to  $t$  for any  $t \in \mathcal{T}$ . The *fundamental problem of causal inference* [Holland, 1986] refers to the fact that in practice we cannot observe simultaneously all the potential outcomes, rendering unidentifiable the causal effect of  $T$  onto  $Y$ . Nevertheless, causal inference can still be achieved thanks to a mix of untestable assumptions and statistical tools: adjusting on a set of available covariates  $X$  containing all possible *confounders* between the treatment and the potential outcomes notoriously permits to identify the law of counterfactual outcomes.

---

\*E-mail: lucas.de\_lara@math.univ-toulouse.fr

Each of these causal theories enables one to carry out *counterfactual reasoning*, that is answering contrary-to-fact questions such as “Had they taken the drug, would have they recovered?”: by applying do-interventions on an SCM, one can compute the outcomes  $Y_{T=t}$  for all possible treatment statuses  $t \in \mathcal{T}$ ; using an RCM with appropriate hypotheses, one can infer the law of the every potential outcome  $Y_t$ . Both approaches involve variables describing counterfactual outcomes, more precisely outcomes *had the variable  $T$  taken a certain value*. This naturally raises the question: are these outcome variables equal (almost surely or in law) across frameworks? We believe the literature on causal inference to be strongly misleading on this matter. A plethora of scientific books and survey papers interchangeably use Pearl’s do-notation and the potential-outcome subscript notation to write outcomes after interventions, suggesting that the corresponding definitions of counterfactuals are identical and differ only from their perspectives [Imbens, 2020, Makhlof et al., 2020, Neal, 2020, Barocas et al., 2023, Colnet et al., 2024]. To justify this, they often refer to Pearl, who argued that “the two frameworks can be used interchangeably and symbiotically”.<sup>1</sup> However, influential works on equivalences between the two causal frameworks have mostly focused on translating conditional-independence restrictions into graphical assumptions instead of actually proving whether counterfactual outcomes were strictly equal across models, or implicitly addressed specific cases. Notably, both [Pearl, 2009, Chapter 7] and [Richardson and Robins, 2013]—acclaimed references unifying both causal frameworks—consider *ex nihilo* the exchangeability of the two notations.

In this paper, we essentially aim at clarifying in which sense using interchangeably two distinct causal models is appropriate. To this end, we compare a potential-outcome model and a structural causal model compatible with a same distribution of observations from an agnostic perspective. We introduce three levels of comparisons, corresponding to different degrees of counterfactual reasoning, and neutrally ask under which conditions the models are (un)distinguishable at these levels. This analysis crucially reminds that the models are *not* mathematically bound to correspond, meaning that using them symbiotically generally rests on a *choice*. Moreover, it classifies real-world scenarios where the models can(not) be considered equivalent, depending on their respective assumptions. Then, we interpret the counterfactual statements and causal effects respectively induced by  $(Y_t)_{t \in \mathcal{T}}$  and  $(Y_{T=t})_{t \in \mathcal{T}}$  when the models do not coincide, and explain how such results relate to the formal equivalence between causal frameworks accepted by the causal-inference community. Proofs of intermediary results are deferred to Appendix B.

In a similar vein, Ibeling and Icard [2024] recently provided a in-depth theoretical comparison of the two frameworks by adopting a neutral viewpoint, notably proving that a well-behaved RCM can always be represented by structural causal model. Altogether, our contributions supplement their work by specifying graphical assumptions that must satisfy an equivalent structural causal model, and by giving a real-world interpretation to these assumptions. In particular, analyzing theoretical equivalence results through the prism of practically relevant problems enables us to point out overlooked paradoxes in the causal inference literature. On the basis of our results and discussions, we call the community to rigorously justify their exchanges of notations across frameworks, as it could lead to misleading conclusions. In doing so, we hope to further clarify the role of each causal modeling in the past, current, and future causal-inference research.

## 2 Preliminaries

This section provides the necessary background on structural causal models and potential outcomes. It is meant to keep the paper self-contained. Section 2.1 introduces generic mathematical notations; Section 2.2 presents Pearl’s causal framework; Section 2.3 explains Rubin’s causal framework.

### 2.1 Basic mathematical notations

Throughout, we consider a probability space  $(\Omega, \Sigma, \mathbb{P})$  with  $\Omega$  a sample space,  $\Sigma$  a  $\sigma$ -algebra, and  $\mathbb{P} : \Sigma \rightarrow [0, 1]$  a probability measure. This space does not necessarily have a physical interpretation; it abstractly represents the possible underlying states of the world. Crucially, it serves as the common mathematical basis to define and compare random variables.

---

<sup>1</sup><http://causality.cs.ucla.edu/blog/index.php/2012/12/03/judea-pearl-on-potential-outcomes/>

A *random variable*  $W$  (including *random vectors*) is a measurable function from  $\Omega$  to an Euclidean space equipped with the Borel  $\sigma$ -algebra. It produces a probability distribution on its output space: we write  $\mathcal{L}(W) := \mathbb{P} \circ W^{-1}$  and  $\mathbb{E}[W] := \int W(\omega) d\mathbb{P}(\omega)$  for respectively the *law and expectation under*  $\mathbb{P}$  of a random variable  $W$ . We emphasize that the laws of  $\mathbb{R}$ -valued random variables can be completely general in this paper; we do not suppose them to be either Lebesgue-absolutely continuous or discrete. For any Borel set  $F$ , we use the common probability-textbook notation  $\{W \in F\}$  for the set  $\{\omega \in \Omega \mid W(\omega) \in F\} \in \Sigma$ . Two variables  $W_1$  and  $W_2$  are  *$\mathbb{P}$ -almost surely equal*, denoted by  $W_1 \stackrel{a.s.}{=} W_2$ , if  $\mathbb{P}(W_1 = W_2) = 1$ ; they are *equal in law under*  $\mathbb{P}$ , denoted by  $\mathcal{L}(W_1) = \mathcal{L}(W_2)$ , if  $\mathbb{P}(W_1 \in F) = \mathbb{P}(W_2 \in F)$  for every Borel set  $F$ . The notation  $W_1 \perp\!\!\!\perp W_2$  means that  $W_1$  and  $W_2$  are *independent under*  $\mathbb{P}$ , that is  $\mathbb{P}(W_1 \in F_1, W_2 \in F_2) = \mathbb{P}(W_1 \in F_1) \cdot \mathbb{P}(W_2 \in F_2)$  for all Borel sets  $F_1, F_2$ .

We denote by  $\mathbb{P}(\cdot \mid W = w)$  the *regular conditional probability measure with respect to*  $\{W = w\}$ , which exists and is unique for  $\mathcal{L}(W)$ -almost every  $w$ . Then, whenever they are well-defined, we write  $\mathcal{L}(W_2 \mid W_1 = w_1) := \mathbb{P}(\cdot \mid W_1 = w_1) \circ W_2^{-1}$  and  $\mathbb{E}[W_2 \mid W_1 = w_1] := \int W_2(\omega) d\mathbb{P}(\omega \mid W_1 = w_1)$  for respectively the *law and expectation of*  $W_2$  *conditional to*  $W_1 = w_1$ . The expression  $W_1 \perp\!\!\!\perp W_2 \mid W_3$  means that  $W_1$  and  $W_2$  are *independent conditional to*  $W_3$  *under*  $\mathbb{P}$ , namely that  $W_1$  and  $W_2$  are independent under  $\mathbb{P}(\cdot \mid W_3 = w_3)$  for  $\mathcal{L}(W_3)$ -almost every  $w_3$ .

Moreover, for any tuple  $w := (w_i)_{i \in \mathcal{I}}$  indexed by a finite index set  $\mathcal{I}$  and any subset  $I \subseteq \mathcal{I}$  we write  $w_I := (w_i)_{i \in I}$ . Similarly, we define the Cartesian product  $\mathcal{W}_I := \prod_{i \in I} \mathcal{W}_i$  for any collection of spaces  $(\mathcal{W}_i)_{i \in \mathcal{I}}$ .

## 2.2 Pearl's causal framework

Pearl's causal modeling mathematically formalizes associations that standard probability calculus cannot describe through the notions of structural causal models and do-interventions [Pearl, 2009]. This section recalls the basics on this topic, borrowing the introduction proposed in [Blom et al., 2020, Bongers et al., 2021].

### 2.2.1 Structural causal models

A *structural causal model* (SCM) represents the causal relationships between the studied variables. It is the cornerstone of Pearl's causal framework.

**Definition 1** (Structural causal model). Let  $\mathcal{I}$  and  $\mathcal{J}$  be two disjoint finite index sets, and write  $\mathcal{V} := \prod_{i \in \mathcal{I}} \mathcal{V}_i \subseteq \mathbb{R}^{|\mathcal{I}|}$ ,  $\mathcal{U} := \prod_{i \in \mathcal{J}} \mathcal{U}_i \subseteq \mathbb{R}^{|\mathcal{J}|}$  for two measurable product spaces. A *structural causal model*  $\mathcal{M}$  is a couple  $\langle U, G \rangle$  where:

1.  $U : \Omega \rightarrow \mathcal{U}$  is a vector of mutually independent random variables, sometimes called the *random noise*;
2.  $G = \{G_i\}_{i \in \mathcal{I}}$  is a collection of measurable  $\mathbb{R}$ -valued functions, where for every  $i \in \mathcal{I}$  there exist two subsets of indices  $\text{Endo}(i) \subseteq \mathcal{I}$  and  $\text{Exo}(i) \subseteq \mathcal{J}$ , respectively called the *endogenous* and *exogenous parents* of  $i$ , such that  $G_i$  is from  $\mathcal{V}_{\text{Endo}(i)} \times \mathcal{U}_{\text{Exo}(i)}$  to  $\mathcal{V}_i$ .<sup>2</sup>

A random vector  $V : \Omega \rightarrow \mathcal{V}$  is a solution of  $\mathcal{M}$  if for every  $i \in \mathcal{I}$ ,

$$V_i \stackrel{a.s.}{=} G_i(V_{\text{Endo}(i)}, U_{\text{Exo}(i)}). \quad (1)$$

The collection of equations defined by (1) and characterized by  $G$  and  $U$  are called the *structural equations*.

Such a model explains how some *endogenous* variables  $V$ , representing observed data, are generated from *exogenous* variables  $U$ , describing background factors. The structural equations quantify the causal dependencies between all these variables and are frequently illustrated by the directed graph with nodes  $\mathcal{I} \cup \mathcal{J}$ , and such that a directed edge points from node  $k$  to node  $l$  if and only if  $k \in \text{Endo}(l) \cup \text{Exo}(l)$  (we say in this case that  $k$  is a parent of  $l$ ). For convenience, we make the common assumption that the studied models are *acyclic*, which means that their associated graphs do not contain any cycles.

**Assumption 1** (Acyclicity).  $\mathcal{M}$  induces a directed *acyclic* graph.

<sup>2</sup>This definition tolerates that distinct endogenous variables share the same exogenous parents, that is  $\text{Exo}(i) \cap \text{Exo}(i') \neq \emptyset$  for some  $i \neq i'$ . Therefore, the  $(U_{\text{Exo}(i)})_{i \in \mathcal{I}}$  are not necessarily mutually independent.

Not only acyclicity simplifies the interpretation of causal dependencies, but it entails *unique solvability* of the SCM: according to [Bongers et al., 2021, Proposition 3.4], Equation (1) admits a unique solution up to  $\mathbb{P}$ -negligible sets. We will abusively refer to such a solution as *the* solution of the SCM.

The purpose of causal structures is to capture the assumption that features are not independently manipulable. As we detail next, they enable to understand the downstream effect of fixing some variables to certain values onto nonintervened variables.

### 2.2.2 Do-intervention

A *perfect (or deterministic) do-intervention* is an operation forcing a set of endogenous variables to take predefined values while keeping all the rest of the causal mechanism equal.

**Definition 2** (Perfect do-intervention). Let  $\mathcal{M} = \langle U, G \rangle$  be an SCM,  $I \subseteq \mathcal{I}$  a subset of endogenous variables, and  $v_I \in \mathcal{V}_I$  a value. The action  $\text{do}(I, v_I)$  defines the modified model  $\mathcal{M}_{\text{do}(I, v_I)} = \langle U, \tilde{G} \rangle$  where  $\tilde{G}$  is given by

$$\tilde{G}_i := \begin{cases} ((v', u') \mapsto v_i) & \text{if } i \in I, \\ G_i & \text{if } i \in \mathcal{I} \setminus I. \end{cases}$$

Do-interventions preserve acyclicity, and therefore unique solvability. As a consequence, if  $V$  is the solution of an acyclic  $\mathcal{M}$ , one can define (up to  $\mathbb{P}$ -negligible sets) its post-intervention counterpart  $V_{\text{do}(I, v_I)}$  solution to  $\mathcal{M}_{\text{do}(I, v_I)}$ . It describes an alternative world where every  $V_i$  for  $i \in I$  is set to value  $v_i$ . In the rest of the paper, we simply write  $\text{do}(V_I = v_I)$  for the operation  $\text{do}(I, v_I)$ , and use the subscript  $V_I = v_I$  to indicate results of this operation. For instance, we write  $\mathcal{M}_{V_I = v_I}$  for  $\mathcal{M}_{\text{do}(I, v_I)}$  and  $V_{V_I = v_I}$  for  $V_{\text{do}(I, v_I)}$ . Crucially, intervening does not amount to conditioning in general, that is  $\mathcal{L}(V \mid V_I = v_I) \neq \mathcal{L}(V_{V_I = v_I})$ .

The next proposition provides a general expression of the solution before and after intervention, and will play a key role throughout this paper.

**Lemma 1** (Do-calculus on variables). *Let  $\mathcal{M} = \langle U, G \rangle$  be an SCM satisfying acyclicity (Assumption 1) with solution  $V$ , and consider a partition  $\{I, J\}$  of  $\mathcal{I}$ . There exists a deterministic measurable function  $F_J$  such that*

$$V_J \stackrel{a.s.}{=} F_J(V_{\text{Endo}(J) \setminus J}, U_{\text{Exo}(J)}).$$

Moreover, for any intervention  $\text{do}(V_I = v_I)$  the solution  $\tilde{V}$  of  $\mathcal{M}_{V_I = v_I}$  verifies

$$\begin{aligned} \tilde{V}_J &\stackrel{a.s.}{=} F_J(v_{\text{Endo}(J) \setminus J}, U_{\text{Exo}(J)}), \\ \tilde{V}_I &\stackrel{a.s.}{=} v_I. \end{aligned}$$

Importantly, this is the same deterministic function  $F_J$  that generates  $V_J$  and its post-intervention counterpart  $\tilde{V}_J$ , the only change being the assignment  $V_I = v_I$ . Slightly abusing notations, we will sometimes artificially extend the input variables of  $F_J$  to write  $V_J \stackrel{a.s.}{=} F_J(V_I, U_{\text{Exo}(J)})$  and  $\tilde{V}_J \stackrel{a.s.}{=} F_J(v_I, U_{\text{Exo}(J)})$ .

### 2.2.3 Counterfactual inference with structural causal models

Counterfactual inference aims at predicting outcomes had a certain event occurred given some factual observations. Typically, it addresses what-if questions such as “Would I have been cured had I taken medicine?”. Do-calculus combined with conditioning provide a natural probabilistic framework to address counterfactual queries. Let for instance  $V := (T, X, Y)$  be the solution to an acyclical SCM  $\mathcal{M} := \langle U, G \rangle$ . Pearl [2009] answers the question “had  $T$  been equal to  $t$ , what would have been the value of  $Y$  given the factual context  $X = x$ ” by using the so-called *three-step procedure*:

1. **(Abduction)** compute  $\mathcal{L}(U \mid X = x)$ , the posterior distribution of  $U$ -values compatible with the context  $\{X = x\}$ ;
2. **(Action)** carry out do-calculus on  $\mathcal{M}$  to obtain the intervened causal mechanism  $G_{T=t}$  of  $\mathcal{M}_{T=t}$ ;
3. **(Prediction)** pass the posterior distribution  $\mathcal{L}(U \mid X = x)$  through  $G_{T=t}$  to generate the distribution  $\mathcal{L}(Y_{T=t} \mid X = x)$  of counterfactual outcomes.

More generally, an SCM enables one to sample from probability distributions of counterfactual outcomes for any choices of context, variables to alter by do-intervention, and outcomes of interest.

## 2.3 Rubin’s causal framework

The potential-outcome framework, also known as *Neyman-Rubin causal modeling* [Neyman, 1923, Rubin, 1974], was designed to understand the causal effect of a treatment onto an outcome of interest, for instance when one aims at assessing the contribution of a drug to recovering from some disease in clinical trials. In this section, we introduce this widely-used framework in the specific case of a binary treatment.

### 2.3.1 Potential outcomes

Let  $T : \Omega \rightarrow \{0, 1\}$  represent a binary *treatment status*, typically such that  $T(\omega) = 0$  indicates the absence of treatment and  $T(\omega) = 1$  indicates a treatment. More generally, it can encode any distinction between some groups (e.g., men and women). Assuming *no interference between units*,<sup>3</sup> this framework postulates two *potential outcomes*  $Y_0 : \Omega \rightarrow \mathbb{R}$  and  $Y_1 : \Omega \rightarrow \mathbb{R}$ , one for each treatment status. These potential outcomes as well as the treatment may depend on some covariates  $X : \Omega \rightarrow \mathbb{R}^d$  (such as the patient’s weight, height, or historical data in clinical trials). Critically, we cannot observe simultaneously  $Y_0(\omega)$  and  $Y_1(\omega)$  for a same  $\omega$ : a problem referred as the *fundamental problem of causal inference* [Holland, 1986]. We only have access to the realized *outcome variable*  $Y : \Omega \rightarrow \mathbb{R}$  which is supposed to be *consistent* with  $(Y_0, Y_1)$ , that is satisfying  $Y = (1 - T) \cdot Y_0 + T \cdot Y_1$ . Concretely, if  $T(\omega) = 1$  for some  $\omega \in \Omega$ , then  $Y(\omega) = Y_1(\omega)$ , and  $Y_0(\omega)$  becomes unidentifiable by mere observations. In this case,  $Y_1(\omega)$  is called the *factual* outcome while  $Y_0(\omega)$  is called the *counterfactual* outcome. We refer to the random vector  $(T, X, Y_0, Y_1)$  as the Rubin causal model (RCM), which is an augmented version of  $(T, X, Y)$  due to consistency.

Understanding the causal relationship between the treatment and the outcome in this framework consists in answering counterfactual questions such as “What would have been the value of  $Y(\omega)$  had  $T(\omega)$  been equal to 1 instead of 0 for a specific  $\omega$  (such that  $X(\omega) = x$ )?”. This cannot be answered since the value of either  $Y_0(\omega)$  or  $Y_1(\omega)$  will always be missing. Instead, in practice, one estimates and compares under some assumptions features of  $\mathcal{L}(Y_1)$  and  $\mathcal{L}(Y_0)$ , or  $\mathcal{L}(Y_1 | X = x)$  and  $\mathcal{L}(Y_0 | X = x)$ . People commonly focus on computing the *average treatment effect*  $\mathbb{E}[Y_1 - Y_0]$  or the *conditional average treatment effect*  $\mathbb{E}[Y_1 - Y_0 | X = x]$ . The main challenge lies in the fact that *association is not causation* in general. In particular, the quantity  $\mathbb{E}[Y | T = t]$  does not necessarily coincide with the quantity  $\mathbb{E}[Y_t]$  for  $t \in \{0, 1\}$ . Typically, if some medical treatment is more likely to be taken by weaker patients, we may observe a lower rate of recovery among the treated group compared to the nontreated group due to the health condition even though the medicine does increase recovery all other things being kept equal: we would observe  $\mathbb{E}[Y | T = 1] < \mathbb{E}[Y | T = 0]$  while  $\mathbb{E}[Y_1] > \mathbb{E}[Y_0]$  (a phenomenon that can be seen as a consequence of *Simpson’s paradox* [Blyth, 1972]). In this case, the health condition is called a *confounder*: a variable associated with both the distribution of the treatment and the outcome. However, causal inference from observational data is still possible, as explained next.

### 2.3.2 Counterfactual inference with fundamental assumptions

Estimating a feature of  $(T, X, Y_0, Y_1)$  under  $\mathbb{P}$ , like a treatment effect, can be achieved by expressing it in terms of  $(T, X, Y)$  whose law generates the empirical observations. This requires two fundamental assumptions. The first one goes by many names through the literature: *conditional ignorability*, *conditional exchangeability*, *conditional exogeneity*, and *conditional unconfoundedness* (among others). Originally formulated by Rosenbaum and Rubin [1983], it states that the potential outcomes are independent of the treatment conditional to the covariates, that is  $Y_t \perp\!\!\!\perp T | X$  for  $t \in \{0, 1\}$ .<sup>4</sup> Said differently, it ensures that all confounders between the treatment and the potential outcomes are included in the covariates. Note that this assumption is untestable, as it would require to observe simultaneously the two potential outcomes. The second key hypothesis is *positivity*, which ensures that all units can be exposed to both treatment statuses, that is  $0 < \mathbb{P}(T = 1 | X = x) < 1$  for  $\mathcal{L}(X)$ -almost every  $x \in \mathbb{R}^d$ . It readily follows from positivity that the probability distribution  $\mathcal{L}(Y | X = x, T = t)$  is well defined for  $\mathcal{L}(X)$ -almost every  $x \in \mathbb{R}^d$  and every  $t \in \{0, 1\}$ , and from conditional ignorability that it coincides with  $\mathcal{L}(Y_t | X = x)$ ,

<sup>3</sup>Paraphrasing Rubin [2010], a *unit* refers to a study object (like a person). This assumption excludes cases where the treatment of one unit may affect the outcome of another.

<sup>4</sup>There exist several versions of the conditional-ignorability assumption with slightly different implications. The original formulation actually demands  $(Y_0, Y_1) \perp\!\!\!\perp T | X$ , which is stronger than  $Y_t \perp\!\!\!\perp T | X$  for all  $t \in \{0, 1\}$ . Nevertheless, most causal-inference methods only require the weaker form.

meaning that association-based outcomes have a causal interpretation. Several statistical methods coexist to estimate the (conditional) average causal effect, all building upon this implication (see for instance [Imbens, 2004, Yao et al., 2021]). We do not detail them for concision and clarity since it is not the topic of this paper. We only point out that, similarly to SCMs, the potential-outcome framework enables one to carry out counterfactual inference.

### 3 Problem setup

This section precises the problem we address: analyzing the mathematical similarities and differences between two models respectively derived from the two causal frameworks. Section 3.1 introduces a potential-outcome model and a structural causal model compatible with a same dataset, and formalizes the assumptions we may place upon them. Section 3.2 defines notions of equivalence between the two models, corresponding to different levels of comparison.

#### 3.1 Causal models and assumptions

Let  $N, d, p \geq 1$  be integers, and define three random variables  $T : \Omega \rightarrow \mathcal{T} := \{0, 1, \dots, N\}$ ,  $X : \Omega \rightarrow \mathbb{R}^d$ , and  $Y : \Omega \rightarrow \mathbb{R}^p$ . We refer to the random vector  $(T, X, Y)$  as the *observational vector*. In order to compare the frameworks, we consider a superimposed construction where the observational vector concurrently governed by an RCM and an SCM. We emphasize that we adopt an agnostic approach where there is no presumed equivalence between the two causal models. This is meant to highlight when equality (possibly in law) between potential and structural counterfactual outcomes is a mathematical necessity or the result of specific assumptions.

##### 3.1.1 Potential-outcome model

On the one hand, we assume that  $Y$  is the outcome of interest,  $T$  the treatment status, and  $X$  some covariates in a potential-outcome framework. This amounts to postulating  $N + 1$  random vectors  $(Y_t)_{t \in \mathcal{T}}$  satisfying the *consistency rule*:

$$Y \stackrel{a.s.}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t.$$

For any  $t \in \mathcal{T}$ ,  $Y_t$  is referred as the *potential outcome* had the treatment been equal to  $t$ . The random vector  $(T, X, (Y_t)_{t \in \mathcal{T}})$  characterizes the Rubin (or potential-outcome) model. Note that we address a more general framework than in Section 2, considering a nonbinary treatment and a multivariate outcome. In this setting, the first fundamental assumptions for causal inference can be written as follows.

**Assumption 2** (Positivity). For all  $t \in \mathcal{T}$ , and  $\mathcal{L}(X)$ -almost every  $x$ ,  $0 < \mathbb{P}(T = t \mid X = x) < 1$ .

Note that, strictly speaking, positivity in an hypothesis on the observational vector rather than the RCM. But people typically suppose it in the context of potential outcomes. We distinguish two formulations for the second fundamental assumption, namely conditional ignorability.

**Assumption 3** (Cross-world conditional ignorability).  $(Y_t)_{t \in \mathcal{T}} \perp\!\!\!\perp T \mid X$ .

**Assumption 4** (Single-world conditional ignorability). For all  $t \in \mathcal{T}$ ,  $Y_t \perp\!\!\!\perp T \mid X$ .

The single-world/cross-world vocabulary is inspired from [Richardson and Robins, 2013]. Note that Assumption 3 implies Assumption 4. In our main results, we will clearly specify which form of conditional ignorability is required.

##### 3.1.2 Structural causal model

On the other hand, we assume that these variables are generated by a latent, unknown SCM: the random vector  $V := (T, X, Y)$  is the solution to an acyclical SCM  $\mathcal{M} = \langle U, G \rangle$  where  $U_T, U_X$  and  $U_Y$  denote the exogenous parents of respectively  $T, X$ , and  $Y$ . Moreover, we suppose that  $\mathcal{M}$  satisfies:

**Assumption 5** (Outcome).  $G$  and  $U$  are such that:

- (i) For all  $i \in \{1, \dots, d\}$ ,  $Y_{\text{Endo}(T)} = Y_{\text{Endo}(X_i)} = \emptyset$ ;

(ii)  $U_Y \perp (U_T, U_X)$ .

The first item of Assumption 5 is a graphical condition that formally defines the variable  $Y$  as the *outcome*; it changes in response to  $X$  and  $T$  but not the contrary. Through Lemma 1, it permits to write

$$\begin{aligned} T &\stackrel{a.s.}{=} F_T(X, U_T), \\ X &\stackrel{a.s.}{=} F_X(T, U_X), \\ Y &\stackrel{a.s.}{=} F_Y(T, X, U_Y), \end{aligned}$$

where  $F_X, F_T$  and  $F_Y$  are deterministic measurable functions derived from  $G$ . The artificial cycle in these formulas (i.e.,  $X$  and  $T$  are both functions of each other) merely serves to consider all configurations of causal links between  $T$  and  $X$  (see Figure 1); strictly,  $\mathcal{M}$  satisfies Assumption 1. The following lemma clarifies the role of second item in Assumption 5.

**Lemma 2** (Random noise). *Let  $V := (T, X, Y)$  be the solution of an SCM  $\mathcal{M}$  satisfying acyclicity (Assumptions 1) and 5 where  $U_T, U_X$  and  $U_Y$  denote the exogenous parents of respectively  $T, X$  and  $Y$ . Then  $U_Y \perp (T, X)$ .*

It guarantees that all potential confounders between  $T$  and  $Y$ —except  $T$  itself—are included in  $X$ . All in all, Assumption 5 simply means through Lemma 2 that the randomness of the outcome  $Y \stackrel{a.s.}{=} F_Y(T, X, U_Y)$  can be divided into three sources: the direct effect of the treatment status  $T$ , the direct effect of the covariates  $X$ , and any other possible effects  $U_Y$  independent to  $T$  and  $X$ .

To conclude this setup, recall that Lemma 1 also enables one to define for every  $t \in \mathcal{T}$  the post-intervention outcome under  $\text{do}(T = t)$  as

$$Y_{T=t} \stackrel{a.s.}{=} F_Y(t, X_{T=t}, U_Y),$$

where the altered covariates are  $X_{T=t} \stackrel{a.s.}{=} F_X(t, U_X)$ . We refer to  $Y_{T=t}$  as the *structural counterfactual outcome* had the treatment been equal to  $t$ . Similarly to potential outcomes, structural counterfactuals satisfy the consistency rule.<sup>5</sup>

**Lemma 3** (Consistency rule for structural counterfactuals). *Let  $V := (T, X, Y)$  be the solution of an SCM  $\mathcal{M}$  satisfying acyclicity (Assumptions 1).<sup>6</sup> Then,  $(Y_{T=t})_{t \in \mathcal{T}}$  verifies the consistency rule,*

$$Y \stackrel{a.s.}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_{T=t}.$$

As a direct consequence of Lemma 3, structural counterfactuals form an RCM  $(T, X, (Y_{T=t})_{t \in \mathcal{T}})$ , since the only requirement to be admissible potential outcomes is to follow the consistency rule. However, this does not signify that  $(T, X, (Y_{T=t})_{t \in \mathcal{T}})$  necessarily coincide (almost surely or in law) with  $(T, X, (Y_t)_{t \in \mathcal{T}})$ . Addressing such comparisons between causal models is precisely the goal of this paper, as formalized next.

### 3.2 Notions of equivalence between causal models

We aim at studying the mathematical similarities and differences between potential outcomes and structural counterfactuals from a theoretically neutral perspective. We consider three levels of comparison that will guide our analysis throughout the paper.

**Definition 3** (Equivalences between causal models). Let  $\mathcal{M} := \langle U, G \rangle$  be an SCM satisfying acyclicity (Assumption 1) with solution  $V := (T, X, Y)$ . It induces via do-calculus the structural counterfactuals  $(Y_{T=t})_{t \in \mathcal{T}}$ . Additionally, let  $(Y_t)_{t \in \mathcal{T}}$  be random variables verifying  $Y \stackrel{a.s.}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t$ . They define the potential-outcome model  $(T, X, (Y_t)_{t \in \mathcal{T}})$ . We say that:

- (i) the models are *almost-surely equivalent* if  $Y_{T=t} \stackrel{a.s.}{=} Y_t$  for any  $t \in \mathcal{T}$ ;

<sup>5</sup>As explained by Pearl [2010], consistency is *theorem* for structural counterfactuals and a constitutive *assumption* for potential outcomes.

<sup>6</sup>We point out that Assumption 5 is not required.



(ii) the models are *cross-world equivalent* if  $\mathcal{L}((T, X, (Y_{T=t})_{t \in \mathcal{T}})) = \mathcal{L}((T, X, (Y_t)_{t \in \mathcal{T}}))$ ;

(iii) the models are *single-world equivalent* if  $\mathcal{L}((T, X, Y_{T=t})) = \mathcal{L}((T, X, Y_t))$  for every  $t \in \mathcal{T}$ .

Remark that (i)  $\implies$  (ii)  $\implies$  (iii). The motivation for Definition 3 comes from different levels at which people reason counterfactually.

The almost-sure level focuses on counterfactual questions at the scale of  $\omega \in \Omega$ . It asks (for example) “What would have been the value of  $Y(\omega)$  had  $T(\omega)$  been equal to 1 instead of 0?”. The answer is deterministic in both causal approaches, given by  $Y_1(\omega)$  in the potential-outcome framework and by  $Y_{T=1}(\omega)$  in the SCM. Should the models be almost-surely equivalent, the answer would be identical for almost-every  $\omega \in \Omega$ . We emphasize that this level is methodologically inessential: in practice, one does not have access to the random variables of the models themselves but to some realizations of their laws. Nevertheless, comparing causal models on this almost-sure baseline is theoretically important to completely understand their mathematical differences.

We now turn to the cross-world level. Because counterfactual questions at the scale of  $\omega$  cannot be answered, people rather address surrogate queries like “What would have been the value of  $Y$  had  $T$  been equal to 1 instead of 0 given that  $X = x$ ?”. In both causal approaches, the answer is generally not deterministic (that is uniquely determined). Possible answers along with their probabilities are respectively described by  $\mathcal{L}(Y_1 | X = x)$  (which can be estimated using the techniques mentioned in Section 2.3.2) in the potential-outcome framework and by  $\mathcal{L}(Y_{T=1} | X = x)$  (which can be inferred via the three-step procedure) in the SCM. More generally, all answers to imaginable counterfactual questions in the structural framework and in the potential-outcome framework are respectively characterized by the joint probability distributions  $\mathcal{L}((T, X, (Y_{T=t})_{t \in \mathcal{T}}))$  and  $\mathcal{L}((T, X, (Y_t)_{t \in \mathcal{T}}))$ . Therefore, should the models be cross-world equivalent, they would yield the same conclusions when reasoning counterfactually at this level.

The single-world level resembles the cross-world level in the sense that it also focuses on distributions of outcomes; it differs by being mathematically weaker. It is motivated by the fact that researchers and practitioners predominantly ask counterfactual questions involving single counterfactual outcomes instead of joint counterfactual outcomes, as in the above paragraph. As such, knowing only the marginal laws  $\{\mathcal{L}((T, X, Y_t))\}_{t \in \mathcal{T}}$  and  $\{\mathcal{L}((T, X, Y_{T=t}))\}_{t \in \mathcal{T}}$  suffices to answer most of the practically relevant counterfactual queries. This is why it is arguably the most critical level in practice. For illustration, consider average treatment effects like  $\mathbb{E}[Y_1 - Y_0]$  or  $\mathbb{E}[Y_1 - Y_0 | X = x]$ , distributional treatment effects like  $D(\mathcal{L}(Y_1), \mathcal{L}(Y_0))$  or  $D(\mathcal{L}(Y_1 | X = x), \mathcal{L}(Y_0 | X = x))$  where  $D$  is a discrepancy between probability measures [Muandet et al., 2021, Park et al., 2021], or the counterfactual-fairness condition:  $\mathcal{L}(Y_{T=t'} | X = x, T = t) = \mathcal{L}(Y_{T=t} | X = x, T = t)$  for every  $t \in \mathcal{T}$  [Kusner et al., 2017]. They all concern the single-world level. Therefore, should the models be single-world equivalent, they would yield the same results in these state-of-the-art methodologies.

To summarize, when two causal models compatible with a same observational vector verify an equivalence condition from Definition 3, they can be used interchangeably at a certain level of counterfactual reasoning. In the next section, we study from a purely mathematical perspective whether structural counterfactuals and potential outcomes need to, do not need to, can, or cannot be equivalent under different degrees of assumptions.

## 4 Main results

In this section, we compare the generic RCM and SCM introduced in Section 3.1 according to the three levels presented in Section 3.2. More precisely, Section 4.1 firstly reminds that equivalence does not necessarily hold whatever the level, then Section 4.2 identifies and studies cases where equivalence does (not) hold at the single-world level.

### 4.1 Equivalence does not generally hold

We start by a crucial reminder justifying why it is relevant to compare the models at the aforementioned levels: an RCM and an SCM compatible with a same observational vector  $(T, X, Y)$  are not necessarily equivalent.

### 4.1.1 General case

The proposition below formalizes this claim in the most general scenario.

**Proposition 1** (Equivalence is not necessary). *Consider any SCM  $\mathcal{M} := \langle U, G \rangle$  satisfying acyclicity (Assumption 1) with solution  $V := (T, X, Y)$  such that  $0 < \mathbb{P}(T = t)$  for all  $t \in \mathcal{T}$ . It defines via do-calculus the structural counterfactual outcomes  $(Y_{T=t})_{t \in \mathcal{T}}$ . Then, there exists random variables  $(Y_t)_{t \in \mathcal{T}}$  verifying  $Y \stackrel{a.s.}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t \stackrel{a.s.}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_{T=t}$  such that for any  $t \in \mathcal{T}$ :*

$$\mathbb{P}(Y_{T=t} \neq Y_t) > 0 \text{ and } \mathcal{L}(Y_{T=t}) \neq \mathcal{L}(Y_t).$$

Therefore,  $\mathcal{M}$  and  $(T, X, (Y_t)_{t \in \mathcal{T}})$  are equivalent in none of the senses from Definition 3.

To introduce the proof of this result, we trace back to how potential outcomes and structural counterfactuals are respectively defined. As noted by Pearl [2010], the two considered causal models differ fundamentally in their constructions of counterfactual outcomes. The potential outcomes  $(Y_t)_{t \in \mathcal{T}}$  are “undefined *primitives*” of the RCM, not related to any formal of measurable quantities, while the intervened outcomes  $(Y_{T=t})_{t \in \mathcal{T}}$  are “*derivatives*” of the SCM by application of do-calculus. Said differently, the firsts are inputs *defining* the causal model, whereas the seconds are post-intervention outputs *defined by* the causal model.

Critically, this input/output difference is more than just conceptual. Because an input can be arbitrarily chosen whereas an output is a necessary consequence, it feels that we could easily find settings where they are not equal. Of course, potential outcomes are not completely arbitrary: they must follow the consistency rule, that is  $Y \stackrel{a.s.}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t$ . But this property does not fully characterize the outcomes as there is no unique choice of  $(Y_t)_{t \in \mathcal{T}}$  satisfying the consistency rule. More precisely, while necessarily  $Y_t(\omega) = Y(\omega) = Y_{T=t}(\omega)$  on  $\{T = t\}$  for  $t \in \mathcal{T}$  (up to  $\mathbb{P}$ -negligible subsets) due to Lemma 3, there is no constraint on  $Y_t(\omega)$  over  $\Omega \setminus \{T = t\}$ ; it could take any value over it without violating the consistency rule. In contrast,  $Y_{T=t}$  is defined (almost) everywhere through the altered SCM  $\mathcal{M}_{T=t}$ . The proof of Proposition 1 exploits this identification issue of potential outcomes: for any  $t \in \mathcal{T}$ ,  $Y_t$  can be any function on  $\{T \neq t\}$ , thereby can be chosen distinct to  $Y_{T=t}$ .

**Proof of Proposition 1** Define the potential outcomes  $(Y_t)_{t \in \mathcal{T}}$  as follows. For any  $t \in \mathcal{T}$ ,

$$Y_t := \mathbf{1}_{\{T=t\}} Y_{T=t} + \mathbf{1}_{\{T \neq t\}} (Y_{T=t} + y),$$

where  $y \in \mathbb{R}^p$  is not the null vector. The tuple  $(Y_t)_{t \in \mathcal{T}}$  satisfies the consistency rule according to Lemma 3. Moreover, for any  $t \in \mathcal{T}$ ,  $Y_t$  is not almost-surely equal nor equal in law to  $Y_{T=t}$  due to  $\mathbb{P}(T \neq t) > 0$  and  $Y_{T=t} + y \neq Y_{T=t}$ . ■

It may seem counter-intuitive to design oneself the potential outcomes (as done in the proof and in upcoming examples), since in practice they are externally imposed. We emphasize that Proposition 1 is theoretically neutral, and only serves to remind that equality between counterfactual outcomes across causal frameworks does not *necessarily* hold.

### 4.1.2 Causal-inference setting

Proposition 1 focuses on the most general setting where the potential outcomes verify only consistency. Said differently, it simply shows that consistency alone is not enough to guarantee equality (almost-surely or in law) between counterfactual outcomes. Nevertheless, people suppose most often that the potential outcomes also satisfy conditional ignorability in order to apply causal-inference techniques. This raises the question whether such an additional hypothesis could render the causal models equivalent. The next theorem ensures that equivalence does not always hold even under the fundamental assumptions of causal inference.

**Proposition 2** (Equivalence is not necessary under the fundamental assumptions of causal inference). *There exists an SCM verifying acyclicity (Assumption 1) with solution  $V := (T, X, Y)$  satisfying positivity (Assumption 2), defining via do-calculus the structural counterfactual outcomes  $(Y_{T=t})_{t \in \mathcal{T}}$ , and there exist random variables  $(Y_t)_{t \in \mathcal{T}}$  satisfying  $Y \stackrel{a.s.}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t$  and conditional ignorability (Assumption 3 or Assumption 4) such that for any  $t \in \mathcal{T}$ :*

$$\mathbb{P}(Y_{T=t} \neq Y_t) > 0 \text{ and } \mathcal{L}(Y_{T=t}) \neq \mathcal{L}(Y_t).$$

Therefore,  $\mathcal{M}$  and  $(T, X, (Y_t)_{t \in \mathcal{T}})$  are equivalent in none of the senses from Definition 3.

One can readily understand why positivity and conditional ignorability cannot completely remedy the identification issue of potential outcomes: they do not constrain the potential outcomes almost surely but only in law. Therefore, an SCM is not necessarily almost-surely equivalent to an RCM in the causal-inference setting. This advances us to comparing causal models at the single-world and cross-world levels. Note that the fundamental assumptions of causal inference permit to fully determine the law of each potential outcome: they entail that  $\mathcal{L}(Y_t | X = x) = \mathcal{L}(Y | X = x, T = t)$  for any  $t \in \mathcal{T}$ . But closer examination of this distributional constraint shows that it does not suffice to make the two causal models cross-world equivalent nor single-world equivalent.

**Proof of Proposition 2** We address the specific case where  $\mathcal{T} := \{0, 1\}$  and both  $X$  and  $Y$  are  $\mathbb{R}$ -valued; one can readily generalize the proof. Consider the following SCM:

$$\begin{aligned} T &\stackrel{a.s.}{=} U_T, \\ X &\stackrel{a.s.}{=} T + U_X \\ Y &\stackrel{a.s.}{=} T + X + U_Y, \end{aligned}$$

where  $U_T$  follows a Bernoulli distribution with parameter  $1/2$ , while  $U_X$  and  $U_Y$  both follow the centered Gaussian distribution with unit variance, such that  $U_T, U_X$  and  $U_Y$  are mutually independent.

According to the rules of do-calculus,  $X_{T=t} \stackrel{a.s.}{=} t$  and  $Y_{T=t} \stackrel{a.s.}{=} t + X_{T=t} + U_Y$  for any  $t \in \mathcal{T}$ . Therefore,  $Y_{T=0} \stackrel{a.s.}{=} U_X + U_Y$  and  $Y_{T=1} \stackrel{a.s.}{=} 2 + U_X + U_Y$ . Then, we write  $U'_Y = -U_Y$  and define potential outcomes as:

$$\begin{aligned} Y_0 &:= (1 - T) \cdot (X + U_Y) + T \cdot (X + U'_Y), \\ Y_1 &:= (1 - T) \cdot (1 + X + U'_Y) + T \cdot (1 + X + U_Y). \end{aligned}$$

After simplification,

$$\begin{aligned} Y_0 &= (1 - T) \cdot (T + U_X + U_Y) + T \cdot (T + U_X + U'_Y) = T + (1 - 2T) \cdot U_Y + U_X, \\ Y_1 &= (1 - T) \cdot (1 + T + U_X + U'_Y) + T \cdot (1 + T + U_X + U_Y) = (1 + T) - (1 - 2T) \cdot U_Y + U_X. \end{aligned}$$

Let us check the required assumptions. Firstly,  $(Y_0, Y_1)$  clearly satisfy  $Y \stackrel{a.s.}{=} (1 - T) \cdot Y_0 + T \cdot Y_1$ . Secondly, for  $\mathcal{L}(X)$ -almost every  $x \in \mathbb{R}$  and any  $t \in \mathcal{T}$ ,  $\mathbb{P}(T = t | X = x) = \frac{1}{2} \frac{\varphi(x-t)}{\varphi(x)} > 0$  where  $\varphi$  is the density function of the centered Gaussian distribution with unit variance. Therefore, positivity holds. Thirdly,  $\mathcal{L}((Y_0, Y_1) | X = x, T = 0) = \mathcal{L}((x + U_Y, 1 + x - U_Y) | X = x, T = 0)$  by definition. Next, Lemma 2 ensures that  $U_Y \perp (T, X)$ , thereby  $\mathcal{L}((Y_0, Y_1) | X = x, T = 0) = \mathcal{L}((x + U_Y, 1 + x - U_Y)) = \mathcal{L}((x + U'_Y, 1 + x - U'_Y))$  since  $\mathcal{L}(U_Y) = \mathcal{L}(U'_Y)$ . Then, notice that

$$\mathcal{L}((Y_0, Y_1) | X = x, T = 1) = \mathcal{L}((x + U'_Y, 1 + x - U_Y)) = \mathcal{L}((Y_0, Y_1) | X = x, T = 0)$$

using once again  $U_Y \perp (T, X)$ . Therefore, cross-world conditional ignorability holds.

To conclude the proof, note that:

$$\begin{aligned} \mathcal{L}(Y_0 | T = 0) &= \mathcal{L}(Y_{T=0} | T = 0), \\ \mathcal{L}(Y_0 | T = 1) &= \mathcal{L}(1 - U_Y + U_X) \neq \mathcal{L}(U_Y + U_X) = \mathcal{L}(Y_{T=0} | T = 1), \\ \mathcal{L}(Y_1 | T = 1) &= \mathcal{L}(Y_{T=1} | T = 1), \\ \mathcal{L}(Y_1 | T = 0) &= \mathcal{L}(1 - U_Y + U_X) \neq \mathcal{L}(2 + U_Y + U_X) = \mathcal{L}(Y_{T=1} | T = 0). \end{aligned}$$

Therefore,  $\mathcal{L}(Y_t) \neq \mathcal{L}(Y_{T=t})$  for any  $t \in \mathcal{T}$ . ■

To summarize Section 4.1: whether it be at the variable or distributional level, with or without the fundamental assumptions of causal inference, potential outcomes and structural counterfactuals are not necessarily equal. Therefore, whatever the level of counterfactual reasoning, using  $(T, X, (Y_t)_{t \in \mathcal{T}})$  and  $(T, X, (Y_{T=t})_{t \in \mathcal{T}})$  interchangeably (as commonly done in the scientific literature) must be either the manifestation of an arbitrary choice—a selection among all the pairs of equivalent causal models—or the mathematical consequence of further assumptions. Notably, Proposition 2 demands strong hypotheses on the potential outcomes but let the latent SCM almost unrestrained. This motivates a sharper analysis of the conditions (in particular on the SCM) that could imply equivalence or on the contrary make it impossible. This is precisely what the following subsection proposes.

## 4.2 Comparison of causal models in the causal-inference setting

In what follows, we aim at providing through mathematical results and concrete illustrations a better understanding of what renders (or not) an SCM and an RCM equivalent. In addition, we exemplify the meaning of reasoning counterfactually with nonequivalent models.

### 4.2.1 Single-world identification

Let us start with some theoretical results focusing on the single-world level. To contrast two causal models at this level, we need a common basis to compare laws across them. The theorem below characterizes the law of every potential outcome *through the latent SCM*  $\mathcal{M}$  under the fundamental assumptions of causal inference, thereby enabling us to compare  $\mathcal{L}(Y_t)$  with  $\mathcal{L}(Y_{T=t})$  for any  $t \in \mathcal{T}$ .

**Theorem 1** (Single-world structural identification of potential outcomes). *Let  $\mathcal{M} = \langle U, G \rangle$  be an SCM satisfying acyclicity (Assumption 1) and Assumption 5 with solution  $V := (T, X, Y)$  such that positivity (Assumption 2) holds. Denoting by  $U_Y$  the exogenous parents of  $Y$ , there notably exists a deterministic function  $F_Y$  such that  $Y \stackrel{a.s.}{=} F_Y(T, X, U_Y)$  where  $U_Y \perp\!\!\!\perp (T, X)$ . Additionally, let  $(Y_t)_{t \in \mathcal{T}}$  be random variables such that  $Y \stackrel{a.s.}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t$ , and suppose that they verify single-world conditional ignorability (Assumption 4). Then, for any  $t \in \mathcal{T}$ ,*

$$\mathcal{L}((T, X, Y_t)) = \mathcal{L}((T, X, F_Y(t, X, U_Y))).$$

The key idea of the proof is leveraging the consistency rule, as it connects the potential outcomes and the SCM through the variable  $Y$  which belongs to both causal models.

**Proof of Theorem 1** We aim at expressing the conditional distribution  $\mathcal{L}(Y_t | X = x)$  in terms of the latent SCM. Before all, recall that positivity ensures that  $\mathcal{L}(Y_t | X = x, T = t)$  is well-defined for  $\mathcal{L}(X)$ -almost every  $x \in \mathbb{R}^d$ , and thereby conditional ignorability entails that  $\mathcal{L}(Y_t | X = x, T = t) = \mathcal{L}(Y_t | X = x, T = t)$ . Then, consistency implies that

$$\mathcal{L}(Y_t | X = x) = \mathcal{L}(Y | X = x, T = t).$$

Moreover, according to  $\mathcal{M}$  the observed outcome can be written as  $Y \stackrel{a.s.}{=} F_Y(T, X, U_Y)$ , leading to

$$\mathcal{L}(Y_t | X = x) = \mathcal{L}(F_Y(T, X, U_Y) | X = x, T = t) = \mathcal{L}(F_Y(t, x, U_Y) | X = x, T = t). \quad (2)$$

Also, it follows from Assumption 5 and Lemma 2 that  $U_Y \perp\!\!\!\perp (T, X)$ . Therefore,

$$\mathcal{L}(Y_t | X = x) = \mathcal{L}(F_Y(t, x, U_Y)) d\mathbb{P}(X = x, T = t).$$

Next, let  $F \subseteq \mathcal{T} \times \mathbb{R}^d \times \mathbb{R}^p$  be a Borel set. We have,

$$\mathbb{P}((T, X, Y_t) \in F) = \int \mathbb{P}((t, x, Y_t) \in F | X = x, T = t).$$

By defining the Borel set  $F(x, t) := \{y \in \mathbb{R}^p | (t, x, y) \in F\}$  we can write

$$\begin{aligned} \mathbb{P}((T, X, Y_t) \in F) &= \int \mathbb{P}(Y_t \in F(x, t) | X = x, T = t) d\mathbb{P}(X = x, T = t) \\ &= \int \mathbb{P}(F_Y(t, x, U_Y) \in F(x, t) | X = x, T = t) d\mathbb{P}(X = x, T = t), \end{aligned}$$

where the last equality follows from Equation (2). Finally,

$$\begin{aligned} \mathbb{P}((T, X, Y_t) \in F) &= \int \mathbb{P}((T, X, F_Y(T, X, U_Y)) \in F | X = x, T = t) d\mathbb{P}(X = x, T = t) \\ &= \mathbb{P}((T, X, F_Y(T, X, U_Y)) \in F). \end{aligned}$$

This means that  $\mathcal{L}((T, X, Y_t)) = \mathcal{L}((T, X, F_Y(t, X, U_Y)))$ , which concludes the proof.  $\blacksquare$

**Remark 1** (Single-world identification v.s. cross-world identification). Theorem 1 expresses for any  $t \in \mathcal{T}$  the marginal distribution  $\mathcal{L}((T, X, Y_t))$  with SCM-based quantities under single-world conditional ignorability (Assumption 4), but does not provide a similar formula for the whole joint distribution  $\mathcal{L}((T, X, (Y_t)_{t \in \mathcal{T}}))$ . There is an explanation:  $\mathcal{L}((T, X, (Y_t)_{t \in \mathcal{T}})) \neq \mathcal{L}((T, X, (F_Y(t, X, U_Y))_{t \in \mathcal{T}}))$  in general—even under cross-world conditional ignorability (Assumption 3). To prove this point, we consider the same setting as the proof of Proposition 2, in which cross-world conditional ignorability hold.

Using the SCM notations,  $F_Y(0, X, U_Y) = X + U_Y$  and  $F_Y(1, X, U_Y) = 1 + X + U_Y$ . It then follows from  $\mathcal{L}(U'_Y) = \mathcal{L}(U_Y)$  that  $\mathcal{L}(Y_0) = \mathcal{L}(F_Y(0, X, U_Y))$  and  $\mathcal{L}(Y_1) = \mathcal{L}(F_Y(1, X, U_Y))$ , as anticipated by Theorem 1. However,  $\mathcal{L}((Y_0, Y_1)) \neq \mathcal{L}((F_Y(0, X, U_Y), F_Y(1, X, U_Y)))$ . It suffices to remark that  $\mathcal{L}(Y_1 - Y_0) = \mathcal{L}(1 + 2(2T - 1) \cdot U_Y)$  whereas  $\mathcal{L}(F_Y(1, X, U_Y) - F_Y(0, X, U_Y)) = \delta_1$  (the Dirac measure at 1).

As such, Theorem 1 enables one to compare potential outcomes and structural counterfactuals at the single-world level. It critically implies under the fundamental assumptions of causal inference that for any  $t \in \mathcal{T}$ :

$$\mathcal{L}(Y_t) = \mathcal{L}(F_Y(t, X, U_Y)) \text{ and } \mathcal{L}(Y_{T=t}) = \mathcal{L}(F_Y(t, X_{T=t}, U_Y)). \quad (3)$$

This result illuminates Proposition 2:  $Y_t$  and  $Y_{T=t}$  are not necessarily equal in law since  $\mathcal{L}(X) \neq \mathcal{L}(X_{T=t})$  in general. Moreover, it furnishes sufficient conditions on the latent SCM for single-world equivalence to hold. Observe that the probability distributions from Equation (3) are equal if: (1)  $X$  is not altered by do-interventions on  $T$ , or (2)  $Y$  is not impacted by  $X$ . The following assumption captures scenario (1), which is the most relevant in causal inference.

**Assumption 6** (Fully controllable treatment).  $G$  and  $U$  are such that for all  $i \in \{1, \dots, d\}$ ,  $T_{\text{Endo}(Y)} = T_{\text{Endo}(X_i)} = \emptyset$ .

Assumption 6 means that  $T$  has no endogenous parents. It implies that  $X_{T=t} \stackrel{\text{a.s.}}{=} X$  for any  $t \in \mathcal{T}$ , which guarantees single-world equivalence under the assumptions of Theorem 1.

**Corollary 1** (Single-world equivalence under fully controllable treatment). *Let  $\mathcal{M} = \langle U, G \rangle$  be an SCM satisfying acyclicity (Assumption 1) and Assumption 5 with solution  $V := (T, X, Y)$  such that positivity (Assumption 2) holds. Additionally, let  $(Y_t)_{t \in \mathcal{T}}$  be random variables such that  $Y \stackrel{\text{a.s.}}{=} \sum_{t \in \mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t$ , and suppose that they verify single-world conditional ignorability (Assumption 4). If Assumption 6 holds then  $\mathcal{M}$  and  $(T, X, (Y_t)_{t \in \mathcal{T}})$  are single-world equivalent but not necessarily cross-world equivalent, even under cross-world conditional ignorability (Assumption 3).*

**Proof of Corollary 1** Let  $t \in \mathcal{T}$ . Under the considered assumptions,  $\mathcal{L}((T, X, Y_t)) = \mathcal{L}((T, X, F_Y(t, X, U_Y)))$  according to Theorem 1. Moreover, recall that  $Y_{T=t} \stackrel{\text{a.s.}}{=} F_Y(t, X_{T=t}, U_Y)$ . If  $T$  is not a parent of  $X$  in  $\mathcal{M}$ , then  $X_{T=t} \stackrel{\text{a.s.}}{=} X$ , leading to  $\mathcal{L}((T, X, Y_t)) = \mathcal{L}((T, X, Y_{T=t}))$ .

We now turn to proving that cross-world equivalence does not necessarily hold, even under cross-world conditional ignorability. To this end, we consider the following SCM:

$$\begin{aligned} T &\stackrel{\text{a.s.}}{=} U_T, \\ X &\stackrel{\text{a.s.}}{=} U_X, \\ Y &\stackrel{\text{a.s.}}{=} T + X + U_Y, \end{aligned}$$

where  $U_T$  follows a Bernoulli distribution with parameter 1/2,  $U_X$  is any  $\mathbb{R}$ -valued random variable, and  $U_Y$  follows a centered Gaussian distribution with unit variance, such that  $U_T, U_X$  and  $U_Y$  are mutually independent. According to the rules of do-calculus,  $Y_{T=0} \stackrel{\text{a.s.}}{=} U_X + U_Y$  and  $Y_{T=1} = 1 + U_X + U_Y$ . Next, set  $U'_Y := -U_Y$  and define the potential outcomes as follows:

$$\begin{aligned} Y_0 &:= (1 - T) \cdot (X + U_Y) + T \cdot (X + U'_Y) = U_X + (1 - 2T) \cdot U_Y, \\ Y_1 &:= (1 - T) \cdot (1 + X + U'_Y) + T \cdot (1 + X + U_Y) = 1 + U_X - (1 - 2T) \cdot U_Y. \end{aligned}$$

Let us verify the required assumptions. Firstly, the potential outcomes clearly satisfy consistency, that is  $Y = (1 - T) \cdot Y_0 + T \cdot Y_1$ . Secondly, since  $T \perp\!\!\!\perp X$ , we have  $\mathbb{P}(T = 1 \mid X) = \mathbb{P}(T = 1) = 1/2$  which entails positivity. Thirdly,  $\mathcal{L}((Y_0, Y_1) \mid X = x, T = 1) = \mathcal{L}((Y_0, Y_1) \mid T = 1)$  because  $U_Y \perp\!\!\!\perp (T, X) \stackrel{\text{a.s.}}{=} (U_T, U_X)$ .

Additionally,  $\mathcal{L}((Y_0, Y_1) \mid T = 1) = \mathcal{L}((U_X - U_Y, 1 + U_X + U_Y)) = \mathcal{L}((U_X + U_Y, 1 + U_X - U_Y))$  since  $\mathcal{L}(U_Y) = \mathcal{L}(-U_Y)$ , and  $\mathcal{L}((U_X + U_Y, 1 + U_X - U_Y)) = \mathcal{L}((Y_0, Y_1) \mid T = 0) = \mathcal{L}((Y_0, Y_1) \mid X = x, T = 0)$  using again  $U_Y \perp\!\!\!\perp (T, X)$ . Wrapping up:  $\mathcal{L}((Y_0, Y_1) \mid X = x, T = 1) = \mathcal{L}((Y_0, Y_1) \mid X = x, T = 0)$ , meaning that cross-world conditional ignorability holds.

Therefore, the causal models are single-world equivalent. To conclude that they are not cross-world equivalent, simply note that  $\mathcal{L}(Y_1 - Y_0) = \mathcal{L}(1 + 2(2T - 1) \cdot U_Y)$  whereas  $\mathcal{L}(Y_{T=1} - Y_{T=0}) = \delta_1$ . ■

This result identifies a large class of SCMs that are single-world equivalent to an RCM verifying positivity and conditional ignorability: those in which  $T$  does not impact  $X$ . This raises the question of whether this graphical condition is necessary. Said differently: are there SCMs not satisfying Assumption 6 but also single-world equivalent to such a potential-outcome model? The answer is yes, but in actually useless configurations. Note that if  $T$  impacts  $X$ , then  $\mathbb{P}(X_{T=t} \neq X) > 0$  for some  $t \in \mathcal{T}$ , and it follows from Equation (3) that the invariance  $\mathcal{L}(Y_t) = \mathcal{L}(Y_{T=t})$  basically occurs if  $F_Y$  does not really change in response to its  $X$ -input. Therefore, excluding these pathological cases, Assumption 6 through Theorem 1 fully classifies single-world equivalence under the standard causal-inference regime of the potential-outcome framework. For simplicity, we will abusively say in the rest of the paper that single-world equivalence holds with such an RCM *only if* Assumption 6 holds.<sup>7</sup> We now turn to injecting this abstract comparison into real-world problems.

#### 4.2.2 A practical perspective on cases of (non)equivalence

Up until now, we adopted a theoretical viewpoint where the models were abstract mathematical objects. However, an SCM is not a mere impartial tool; it is meant to genuinely capture the world’s functioning. As such, practitioners do not decide the graphical relationships between endogenous variables themselves, they are imposed by nature. In contrast, an RCM introduces hypothetical counterfactual variables whose interpretation may depend on the assumptions placed upon them. This critically signifies that single-world equivalence between the *true* latent SCM and a potential-outcome model satisfying conditional ignorability is impossible in many real-world problems, due to the dichotomy ruled by Assumption 6.

To clarify this point, consider a problem where one relies on an RCM equipped with positivity and conditional ignorability to apply standard causal-inference techniques. They wonder whether this model is equivalent to the underlying SCM (always supposed to verify Assumptions 1 and 5). Answering this question amounts to translating the pivotal causal-dependency assumption between  $T$  and  $X$  (Assumption 6) and its negation into common language.

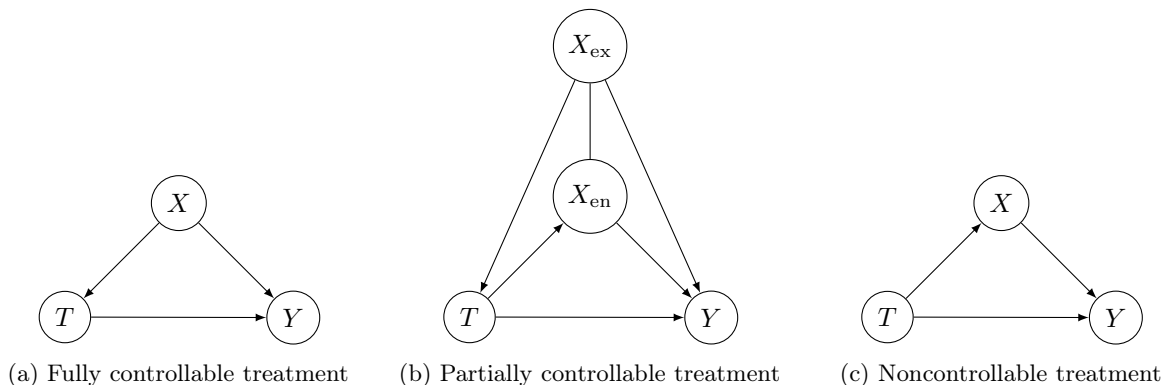


Figure 1: Three possible configurations of the treatment under Assumption 5.  $X_{\text{ex}} := X_{\text{Endo}(T)}$  denotes the parents of  $T$  in  $X$  while  $X_{\text{en}}$  are the remaining covariates. Exogenous variables are not represented. A single node can represent several variables. In (a),  $T$  does not impact  $X$ ; in (b),  $T$  may impact some  $X$ -variables and some  $X$ -variables may impact  $T$ ; in (c),  $X$  does not impact  $T$ .

Assumption 6 basically signifies that the covariates are not altered by the treatment, as illustrated in

<sup>7</sup>The “almost-true” equivalence between a graphical condition like Assumption 6 and its structural implications was already discussed in [Pearl, 2009, Section 7.4], where the so-called graphical and counterfactual criteria of exogeneity coincide after excluding incidental cases.

Figure 1a. Notably, this configuration encompasses various typical causal-inference scenarios: in clinical trials, the covariates  $X$  may influence the treatment allocation  $T$  but never the contrary. In these common situations, both the RCM and the SCM produce the same single-world counterfactuals due to Corollary 1. Nevertheless, recall that cross-world equivalence does not necessarily hold.

Importantly, the negation of Assumption 6—which states that  $T$  is a parent of several (or even all) covariates in  $\mathcal{M}$ —is also mathematically possible. Figures 1b and 1c illustrate the possible causal graphs. In these situations, Theorem 1 does not guarantee single-world equivalence. Consequently, confusion between the two causal approaches can lead to misleading results: according to Equation (3), the RCM considers counterfactual outcomes at fixed  $X$ , whereas the SCM alters the covariates into  $X_{T=t}$ . These cases are empirically relevant, since people also rely on causal inference outside the scope of clinical trials, in settings where the treatment cannot be completely manipulated and thereby impacts the covariates. For example,  $T$  drives  $X$  but not the contrary (as in Figure 1c) in emblematic causal problems such as the Berkeley’s admission paradox where  $T$  represents the sex and  $X$  the course choice [Bickel et al., 1975]. This is more generally true in the whole causal-fairness literature, where the variable to alter typically encodes the sex, the race, or the age of individuals (see for instance [Kusner et al., 2017, Barocas et al., 2023, Nilforoshan et al., 2022] for machine-learning-related research).

All in all, under the fundamental assumptions of causal inference, equivalence of single-world counterfactuals across causal frameworks depends on the relationships between the treatment and the covariates, as described in Figure 1. What distinguishes the different configurations is the nature of the so-called treatment. In particular, if the treatment can be assigned *a posteriori* (as in Figure 1a), then the two notions of counterfactuals coincide. If the treatment is an intrinsic feature (as in Figure 1c), such as individuals’ race or sex, then structural counterfactuals and potential-outcomes counterfactuals are generally not equal. Section 4.2.3 below concretely exemplifies this point by studying an SCM corresponding to Figure 1c.

#### 4.2.3 Illustration: an immutable treatment and two different kinds of counterfactuals

Counterfactual reasoning can be defined as thinking about outcomes in hypothetical worlds where some circumstances changes from what factually happened while others are kept equal. Crucially, there is not a single way of reasoning counterfactually. Theorem 1 clearly shows that the potential-outcome framework compares worlds sharing the same observed covariates  $X$  but differing in  $T$ , while the structural account compares worlds sharing the same background factors  $U$  but differing in  $T$ . Said differently, potential-outcome counterfactuals are *ceteris paribus* counterfactuals (i.e., all other things being kept equal) with respect to the covariates, whereas structural counterfactuals are *mutatis mutandis* counterfactuals (i.e., after changing what must be changed) with respect to the covariates—but *ceteris paribus* with respect to the background factors. We emphasize that both definitions are perfectly legitimate, but convey distinct meanings and thereby correspond to different causal effects. Therefore, *they should not be employed for the same purpose*. Let us illustrate their implications on a concrete case.

The following fairness-inspired example generalizes and circumstantiates the discussion from [Kusner et al., 2017, Appendix S1]. The treatment status  $T$  indicates the gender,  $T(\omega) = 0$  standing for women and  $T(\omega) = 1$  standing for men; the covariate  $X$  quantifies the level of work experience, a higher score encoding a richer experience; the outcome  $Y$  evaluates a candidate’s application for some position, a better score giving a higher probability of acceptance. Suppose that these three variables are ruled by the following SCM fitting Figure 1c:

$$\begin{aligned} T &\stackrel{a.s.}{=} U_T, \\ X &\stackrel{a.s.}{=} \alpha T + U_X, \\ Y &\stackrel{a.s.}{=} X + \beta T + U_Y, \end{aligned}$$

where  $\alpha$  and  $\beta$  are deterministic parameters quantifying the causal influence of  $T$  onto respectively  $X$  and  $Y$ , and  $U_X$  represents the hidden merit or effort of an individual. Typically, a positive parameter  $\alpha$  describes the societal inequalities leading women to have a lower level of work experience than men with equal merit  $U_X$ . Moreover, we suppose that positivity (Assumption 2) is true (for instance by choosing  $U_X$  inducing a Gaussian distribution), and that  $U_Y \perp\!\!\!\perp (U_T, U_X)$  so that Assumption 5 holds. Finally, we set two potential outcomes  $(Y_0, Y_1)$  verifying the consistency rule, that is  $Y \stackrel{a.s.}{=} (1 - T) \cdot Y_0 + T \cdot Y_1$ , and single-world conditional ignorability (Assumption 4). Note the analogy with the proof of Proposition 2.

We consider the problem of assessing the causal effect of  $T$  onto  $Y$  conditional to  $X = x$ . In the potential-outcome approach this amounts to computing the conditional average treatment effect:

$$\begin{aligned} \text{CATE}(x) &:= \mathbb{E}[Y_1 - Y_0 \mid X = x] \\ &= \mathbb{E}[(x + \beta + U_Y) - (x + U_Y)] \\ &= \beta, \end{aligned}$$

where one can either directly use Theorem 1 or go through  $\mathcal{L}(Y_t \mid X = x) = \mathcal{L}(Y \mid X = x, T = t)$  and  $U_Y \perp\!\!\!\perp (T, X)$  again. Observe that this first quantity measures only the *direct effect* of the treatment: it completely ignores the dependence of  $Y$  on  $T$  through  $X$ , as it involves only  $\beta$ . This is due to the fact that the CATE keeps the covariate  $X$  fixed, comparing two *distinct* levels of background factors with identical profiles but different genders. In contrast, Pearl’s approach assesses the following structural counterfactual effect:

$$\begin{aligned} \text{SCE}(x) &:= \mathbb{E}[Y_{T=1} - Y_{T=0} \mid X = x] \\ &= \mathbb{E}[(X_{T=1} + \beta + U_Y) - (X_{T=0} + U_Y) \mid X = x] \\ &= \mathbb{E}[X_{T=1} - X_{T=0} \mid X = x] + \beta \\ &= \mathbb{E}[(\alpha + U_X) - U_X \mid X = x] + \beta \\ &= \alpha + \beta. \end{aligned}$$

Remark that this second quantity measures the *total effect* of the treatment: it takes into account the whole path of influence of  $T$  onto  $Y$ , involving both  $\alpha$  and  $\beta$ . This comes from the fact that the SCE fixes the random seed  $U$  and not the covariates, comparing a *same* level of background factors in two alternative realities where the gender is switched. Most importantly,  $\text{CATE} \neq \text{SCE}$  if  $\alpha \neq 0$ : as expected the causal models are not single-world equivalent.

From a fairness perspective, the CATE says that if  $\beta = 0$ , that is if  $T$  is not a *direct* cause of  $Y$ , then the application process is fair; whether it is unfair towards men or women when  $\beta \neq 0$  depends on the sign of  $\beta$ . In contrast, the SCE says that if  $\beta = -\alpha$ , that is if the decision rule  $Y$  compensates the discrepancy of work experiences  $X$  across genders  $T$ , then the application process is fair. Each analysis points out a different notion of fairness: considering the SCE as a fairness criterion suggests that recruiters should correct societal inequalities by preferring women with potentially lower work experience but higher merit whereas relying on the CATE suggests it is only explicitly including the gender in the decision-rule pipeline that is unfair. Critically, if  $\alpha \neq 0$ , *practitioners mixing potential outcomes with structural counterfactuals could reach contradictory conclusions on fairness*.

**Remark 2** (Computing direct effects from an SCM). One can still compute the CATE, that is the *direct* effect, using do-interventions on the SCM. Under the same assumptions as above,

$$\begin{aligned} \mathbb{E}[Y_{T=1, X=x} - Y_{T=0, X=x}] &= \mathbb{E}[(x + \beta + U_Y) - (x + U_Y)], \\ &= \beta, \\ &= \text{CATE}(x). \end{aligned}$$

This is due to the fact that, more generally,  $\mathcal{L}(Y_{T=t, X=x}) = \mathcal{L}(Y_t \mid X = x)$  under the fundamental assumptions of causal inference. In contrast, one cannot always compute the SCE, that is the *total* effect from potential outcomes. Besides, we point out that even though the CATE can be theoretically derived from both a potential-outcome model and an SCM, the practical methods to estimate them from data differ between approaches. In Rubin’s causal framework, one only needs to estimate  $\mathbb{E}[Y \mid X = x, T = t]$  (as it equals  $\mathbb{E}[Y_t \mid X = x]$ ); in Pearl’s causal framework, one typically needs to learn the full SCM beforehand (which is a notoriously difficult task), and then to apply the three-step procedure with  $\text{do}(T = t, X = x)$ . Further practical distinctions specific to positivity (Assumption 2) are discussed in Appendix A.

To sum-up, each approach has a different signification in this immutable-treatment configuration, and therefore corresponds to a specific way of reasoning counterfactually. This signifies that the difference between frameworks does not amount to practical considerations only. Analysts and researchers should also justify the chosen model depending on the kind of causal effects they want to compute. In the final section, we include the conceptual and methodological divergences between models pointed out by this illustration into a more comprehensive reflection.



## 5 Discussion

This section examines common practices in the causal-inference literature in light of the mathematical differences between models explained in Section 4. Firstly, Section 5.1 clarifies the relation between our contribution and the notorious formal equivalence between frameworks, unveiling a fundamental dichotomy in the applications of potential outcomes. Secondly, Section 5.2 leverages this discussion to provide recommendations on the exchange of potential-outcome and do notations.

### 5.1 On the formal equivalence between frameworks

As mentioned in the introduction, many articles interchangeably use the potential-outcome notation and the do notation, invoking an “equivalence” between causal frameworks. This may seem paradoxical after reading Section 4. In the following, we explain what an example like Section 4.2.3 signifies to the relationship between SCMs and RCMs, and to how people generally manipulate these models.

#### 5.1.1 Structural representation and graphical translation

Recall that an acyclical SCM with solution  $(T, X, Y)$  always defines an RCM  $(T, X, (Y_{T=t})_{t \in \mathcal{T}})$  according to Lemma 3. Conversely, any potential-outcome model  $(T, X, (Y_t)_{t \in \mathcal{T}})$  as defined in Section 3.1 can be *represented* by an SCM in the following sense: there exists an SCM with solution  $V = (T, X, Y)$  such that  $\mathcal{L}((T, X, (Y_{T=t})_{t \in \mathcal{T}})) = \mathcal{L}((T, X, (Y_t)_{t \in \mathcal{T}}))$  [Ibeling and Icard, 2024, Proposition 1]. In other words, there always exists a cross-world equivalent SCM, referred to as a *structural representation*.

Interestingly, if one chooses to define potential outcomes as structural counterfactuals, then Rubin’s causal framework and Pearl’s causal framework become two different languages to talk about the same objects. In the RCM, assumptions for causal inference are generally framed as conditional-independence restrictions on counterfactual variables (e.g., conditional ignorability); in Pearl’s causal framework, assumptions on causal relationships are generally framed in terms of graphical conditions on observational variables (e.g., backdoor criterion, see Remark 3). Both [Pearl, 2009, Chapter 7] and [Richardson and Robins, 2013] focus on unifying these two mathematical languages by providing rules for translating assumptions and theorems from one viewpoint to the other. This ensures what people often refer as the *logical* or *formal equivalence* between frameworks (not between models, in contrast to the notions in Definition 3). It notably allows analysts to work symbiotically with an RCM and an SCM, as long as equivalent assumptions are made across the models.

**Remark 3** (Graphical interpretation of conditional ignorability). One of the most notorious translation rule is [Pearl et al., 2016, Theorem 4.3.1], which states that if  $X$  verifies the so-called *backdoor criterion* relative to  $(T, Y)$  in  $\mathcal{M}$  [Pearl, 2009, Definition 3.3.1], then  $Y_{T=t} \perp\!\!\!\perp T \mid X$  for all  $t \in \mathcal{T}$ : said differently, single-world conditional ignorability holds for the structural counterfactuals. This counterfactual interpretation of graphical conditions echoes Corollary 1. Our result states that if an SCM satisfies Assumptions 5 and 6, then it is single-world equivalent to an RCM such that  $Y_t \perp\!\!\!\perp T \mid X$  for all  $t \in \mathcal{T}$  (Assumption 4). The connection follows the fact that the graphical conditions in Assumptions 5 plus 6 (illustrated by Figure 1a) entail that  $X$  verifies the so-called *backdoor criterion* relative to  $(T, Y)$  in  $\mathcal{M}$ .

Interestingly, the perspective offered by Corollary 1 reframes Pearl’s translation rules in terms of equivalence between presumably separated SCMs and RCMs. This change of angle helps understanding some grey areas of the formal equivalence, as discussed below.

#### 5.1.2 Using the formal equivalence is a choice

There is no logical contradiction between the results from Section 4 and this formal equivalence. Nevertheless, our contribution highlights two overlooked features.

Firstly, [Ibeling and Icard, 2024, Proposition 1] simply *allows* to represent an RCM by an SCM. Ultimately, defining potential outcomes as structural counterfactuals is a *choice*—it does not rest on any proof. When Pearl writes  $Y_t := Y_{T=t}$  in [Pearl, 2009, Equation 3.51], claiming that the operation  $\text{do}(T = t)$  on the SCM gives a physical meaning to the vague “had  $T$  been  $t$ ” of the potential outcome, this is nothing more than an arbitrary choice. As demonstrated by Propositions 1 and 2, naively looking at the definitions of the Rubin’s causal framework and of Pearl’s causal framework, there is nothing that mathematically constrains potential outcomes to coincide at any level with the structural counterfactuals of a chosen

SCM.<sup>8</sup> We emphasize that [Pearl, 2009, Chapter 7] and [Richardson and Robins, 2013] show that *if* potential outcomes are chosen to be structural counterfactuals, then one can translate the assumptions made on potential outcomes into assumptions on the underlying causal graph and *vice versa*—not that they *must* be chosen as such.

Secondly, even though there always exist *theoretical* structural representations of a given RCM, nothing guarantees that they tangibly represent the real world. In particular, they could fail to describe the true causal dependencies within  $(T, X, Y)$ . Notably, Section 4.2.2 reveals through the lens of Theorem 1 cases of incompatibility between an RCM tailored to causal inference and the true SCM. Basically—after excluding the aforementioned useless configurations—the structural representation of a potential-outcome model under the fundamental assumptions of causal inference must encode a fully controllable treatment (Assumption 6).

### 5.1.3 Two fundamentally distinct paradigms for potential outcomes

We submit that the graph of an SCM is meant to correctly capture the cause-effect link between endogenous variables. This notably requires ignoring the structural representation of an RCM whenever it does not correspond to the true SCM. Under this principle, two paradigms for defining and applying potential outcomes coexist (in the sense that there is no mathematical obstacle): in synergy with the true latent SCM; as hypothetical outcomes whose structural representations may differ from the true latent SCM. Before explaining them more precisely, we crucially emphasize that we do not discuss what people should or should not do in this regard. Instead, we neutrally classify what people *actually* do, and explain the implicit meaning of these practices.

On the one hand, Pearl has firmly advocated for long to *always* use the potential-outcome framework in symbiosis with an SCM, as the latter enables to formulate its conditional-independence conditions deemed nebulous in the intelligible language of causal graphs. We refer to [Pearl, 2009, Section 7.4] for a detailed argumentation. In this paradigm, a condition on the potential outcomes like conditional ignorability does not come from a fundamental assumption but possibly follows from a graphical condition on the SCM (see Remark 3). In particular, this forbids the approach from Section 4.2.3 where the potential outcomes satisfy conditional ignorability while Assumption 6 does not hold for the SCM. Semantically, modeling potential outcomes through an SCM formally defines their “had  $T$  been  $t$ ” as the operation  $\text{do}(T = t)$  which changes the remaining endogenous variables accordingly.

On the other hand, disconnecting the two causal frameworks is mathematically doable, as detailed in Section 4.2.3. In this paradigm, fundamental assumptions placed upon potential outcomes define the semantics of their associated counterfactual statements. Recall that Theorem 1 notably showed through Equation (3) that dressing potential outcomes with conditional ignorability defines their “had  $T$  been  $t$ ” as switching  $T$  into  $t$  while keeping the remaining endogenous variables unchanged. In problems where the treatment is fully controllable, this interpretation happens to coincide at the single-world level with the do-operator (due to Corollary 1). The methodological interest of this approach comes from the possibility to compute the direct causal effects of noncontrollable treatments through statistical methods while completely disregarding the latent SCM (like in Section 4.2.3). This practice is perhaps controversial and philosophically debatable. Nevertheless, we point out that (even though it is generally implicitly done) *not* defining potential outcomes as structural counterfactuals is actually something common in the scientific literature. In particular, [Li et al., 2017, Glymour and Spiegelman, 2017, Khademi et al., 2019, Khademi and Honavar, 2020, Makhoul et al., 2020, Qureshi et al., 2020] rely on (or refer to) the potential-outcome framework equipped the fundamental assumptions of causal inference to understand the influence of a noncontrollable treatment like sex, race, and biological factors.<sup>9</sup> Therefore, if we consider this corpus of the causal-inference literature to be admissible and nonparadoxical, then we must accept that unifying potential outcomes and structural counterfactuals is not an obligation. Finally, we emphasize that even in this paradigm, the potential outcomes are not completely unaligned with the true SCM. Theorem 1 literally interprets at the single-world level potential outcomes under conditional ignorability through the structural equations. Such outcomes just do not necessarily result from a do-intervention.

<sup>8</sup>From a more philosophical angle, [Markus, 2021, Section 2.1] made a similar remark to argue that the two frameworks were *weakly* equivalent rather than *strongly* equivalent.

<sup>9</sup>Other articles, for instance [Ridgeway, 2006] and [Gaebler et al., 2022], explicitly focus on sex and race as *perceived* by a decider. Such a perception could depend on the covariates, hence not be immutable.

**Remark 4** (No manipulation without causation). The “no causation without manipulation” principle of Holland [1986] states that the causal influence of immutable variables like a person’s race is ill-defined since we cannot conceptualize an assignment mechanism for such a nonmanipulable status (in contrast, for example, to allocating a medical treatment). This notably led some researchers to consider full controllability of the treatment to be a fundamental assumption of RCMs [Imbens and Rubin, 2010]. Adopting this rule, Corollary 1 ensures that whenever the potential-outcome framework can be applied, the employed RCM is at least single-world equivalent to the latent SCM. This would mechanically rule out several of the contentious points between frameworks that we discuss in this paper. However, we recall that our work does not focus on philosophical arguments: it neutrally analyzes the literature. In particular, we do not argue in favor or against including Holland’s principle in Rubin’s framework. Rather, we acknowledge that many researchers actually study noncontrollable treatments using RCMs equipped with the fundamental assumptions of causal inference, and discuss the implications of such a choice.

It feels that Pearl’s rule to always leverage an SCM as the axiomatic characterization of potential outcomes through the do-operator possibly made unclear the existence of these two paradigms. Each approach can be legitimate; *what crucially matters is having a clear understanding of the produced counterfactuals and being transparent about the choices made*. In this sense, we conclude this article by defending a more cautious use of notations.

## 5.2 On the exchange of notations

In contrast to Section 5.1, this subsection discusses what researchers and practitioners should do. More specifically, our suggestions do not concern the methods people use, but how they *present* their approaches and their results. Concretely, we argue that exchanging the do-notation and the potential-outcome notation should be done with greater care than what is commonly done in the literature. A notation is the identification of a mathematical object. Therefore, a same notation can be used for two differently-defined objects just in case they are mathematically equal. In the first paradigm, where the potential outcomes are chosen as the true structural counterfactuals, the notational exchange with the do is valid; not in the second paradigm. In what follows, we examine books and articles referring to the formal equivalence between causal frameworks in a confusing way, to underline the importance of specifying and justifying the adopted paradigm.<sup>10</sup>

For example, [Colnet et al., 2024, Section 5] commences with potential-outcome notations and then invoke the formal equivalence to substitute them by do notations. Implicitly, this exchange engages a structural representation of their RCM which satisfies conditional ignorability. But recall that nothing guarantees that this structural representation corresponds at the single-world level to the true SCM. In their case, everything works fine precisely because only settings where the treatment is fully controllable are considered. However, they never explicitly state Assumption 6 and even less explain its crucial role. While this may seem harmless in practice, this could mislead people to wrongly believe that this equivalence holds in general: it would not hold with a noncontrollable treatment. Even from a purely logical viewpoint, working in this favorable scenario should not obviate the need for proper justification to why  $\mathcal{L}((T, X, Y_t)) = \mathcal{L}((T, X, Y_{T=t}))$  for every  $t \in \mathcal{T}$ .

Another critical point concerns fairness settings where the treatment typically encodes sex or race. As repeated, in these cases one cannot interchangeably manipulate an RCM under the fundamental assumptions for causal inference and the *true* SCM. Therefore, introducing the two causal frameworks specifically in the context of fairness with identical notations and suggesting that the appropriate choice of framework is mostly a matter of practical considerations, as done in Makhlouf et al. [2020] and [Barocas et al., 2023, Chapter 5], feels pedagogically dangerous. Section 4.2.3 notably exemplifies how mixing the notations could lead to contradictory results. If one wants to exploit the practical convenience of the RCM equipped with conditional ignorability in such fairness problems, then they must specify that they follow the second paradigm, and that they are interested in computing *direct* causal effects. In contrast, if one defines potential outcomes as the true structural counterfactuals in the same context, then  $\mathbb{E}[Y_1 - Y_0]$  and  $\mathbb{E}[Y_1 - Y_0 \mid X = x]$  do not represent direct effects. This illustrates how critical it is to understand the two paradigms.

To summarize Section 5, we think that the scope and implications of the formal equivalence between causal frameworks can be misleading. In particular, it does not mean that  $\mathcal{L}((T, X, (Y_t)_{t \in \mathcal{T}})) =$

---

<sup>10</sup>We do not suggest that these references contain erroneous claims, only that the presentation of specific aspects can be misleading. Beyond that, we strongly recommend reading them.

$\mathcal{L}((T, X, (Y_{T=t})_{t \in \mathcal{T}}))$  in general; it means that such an equality in law can hold *if equivalent assumptions are made on the counterfactual outcomes across models*. Supposing distinct axioms across models means giving distinct interpretations to their respective counterfactual outcomes, which thereby relate to distinct causal effects. Critically, using an RCM to estimate *direct* causal effects (as commonly done) requires assuming conditional ignorability, which cannot always be assumed in the *true* latent SCM due to the possibly immutable nature of the treatment. This is why we recommend to present Rubin’s causal framework and Pearl’s causal framework as distinct mechanisms for reasoning counterfactually that coincide under specific assumptions and choices, rather than merely different perspectives.

## 6 Conclusion

In this paper, we superimposed Pearl’s causal framework and Rubin’s causal framework without presuppositions to show that structural counterfactual outcomes and potential outcomes do not necessarily coincide at any levels of counterfactual reasoning. To furnish a thorough comparison at the most relevant level, we expressed the laws of potential outcomes in terms of the latent SCM under classical causal-inference assumptions. On the basis of this result, we gave a detailed interpretation of counterfactuals in each causal framework, specifying when they entailed different conclusions. More specifically, counterfactual inference with potential outcomes under conditional ignorability yields *ceteris paribus* counterfactuals with respect to the covariates, whereas counterfactual inference with a do-intervention on an SCM yields *mutatis mutandis* counterfactuals with respect to the covariates. If the cause of interest is immutable, these constructions are generally not equal in law. For these reasons, we call the community to not interchangeably use the do-notation and the potential-outcome notation, unless the justification is explicitly made.

We emphasize that our contribution is not an argument in favor of using one causal model rather than the other, or against the formal equivalence between frameworks. Instead of taking position or addressing philosophical arguments, it highlights some facts: theoretically, one can perfectly define potential outcomes as distinct to the true structural counterfactuals; empirically, researchers have actually (implicitly) worked with potential outcomes defined as distinct to the true structural counterfactuals. Thereby, our work is meant to shed light on the different mathematical choices that analysts can make when working with counterfactual outcomes, and to precise their implications in order to prevent incorrect or ambiguous conclusions in causal studies. In doing this paper, we hope to clarify the similarities and differences between the two major causal approaches.

## A Nonequivalence in applicability

In this appendix, we further compare the two causal approach from a practical viewpoint. Summing-up: to apply Pearl’s causal framework, one must postulate a plausible SCM or infer it from data to then carry out do-calculus; to apply the Rubin’s causal framework, one must find a set of covariates believed to satisfy conditional ignorability and then use statistical methods (e.g., matching, stratification, re-weighting, regression) to estimate observable quantities with causal meanings. As such, the two approaches are different in *how* they are applied; but there also exists a critical difference in *when* they can be applied due to the positivity assumption. Let us detail this point, as it also concerns a distinction between counterfactuals across frameworks.

Causal inference with an RCM requires two fundamental assumptions: conditional ignorability (Assumptions 4 or 3) and positivity (Assumption 2). While the second is testable in contrast to the first, it raises practical issues. It basically states that the distributions  $\mathcal{L}(X \mid T = t)$  for  $t \in \mathcal{T}$  share the same support, which is violated as soon as the groups represented by  $T$  bear unique properties. Consider for example that we study individuals where  $T$  encodes their genders, and that the covariates  $X$  specify their jobs (among other attributes). Positivity would not hold if gender-locked positions existed, and consequently we could not identify the counterfactual outcome *had she been a man* of every woman occupying a women-only job. In contrast, an SCM always allows such a computation. Therefore, there exist problems where the two causal models cannot be simultaneously applied to carry out counterfactual inference in practice. Not only the two kinds of counterfactuals should not be used for the same purpose (as they provably have different significations), but they cannot always be used for the same tasks.

## B Remaining proofs

**Proof of Lemma 1** Since  $\mathcal{M}$  is acyclical, there exists a topological ordering on the indices in  $\mathcal{I}$ , and therefore on the subset  $J$ . This means in particular that there exist some  $j \in J$  such that  $G_j$  takes only variables in  $V_I$  as endogenous inputs. Starting from these indices, and recursively substituting along the topological ordering produces a measurable  $F_J$  such that

$$V_J \stackrel{a.s.}{=} F_J(V_{\text{Endo}(J) \setminus J}, U_{\text{Exo}(J)}).$$

Note that  $\text{Endo}(J) \setminus J \subseteq I$ . Carrying out the same substitution on the intervened model  $\mathcal{M}_{V_I=v_I}$  with solution  $\tilde{V}$  gives

$$\tilde{V}_J \stackrel{a.s.}{=} F_J(v_{\text{Endo}(J) \setminus J}, U_{\text{Exo}(J)}),$$

while by definition  $\tilde{V}_I \stackrel{a.s.}{=} v_I$ . ■

**Proof of Lemma 2** By assumption, the random vector  $V := (T, X, Y)$  is the solution to an acyclical SCM where  $U_T, U_X$  and  $U_Y$  denote the exogenous parents of respectively  $T, X$  and  $Y$ . Since additionally  $Y_{\text{Endo}(T)} = Y_{\text{Endo}(X)} = \emptyset$ , Lemma 1 ensures the existence of a measurable function  $F_{T,X}$  such that  $(T, X) \stackrel{a.s.}{=} F_{T,X}(U_T, U_X)$ . Therefore, if  $U_Y \perp\!\!\!\perp (U_T, U_X)$ , then  $U_Y \perp\!\!\!\perp (T, X)$ . ■

**Proof of Lemma 3** Let  $t \in \mathcal{T}$ . By assumption, the random vector  $V := (T, X, Y)$  is the solution to an acyclical SCM. We write  $U_X$  and  $U_Y$  the exogenous parents of respectively  $X$  and  $Y$ . Therefore, by partitioning  $V$  into  $T$  and  $(X, Y)$ , Lemma 1 guarantees the existence of a measurable function  $F_{X,Y}$  such that

$$\begin{aligned} (X, Y) &\stackrel{a.s.}{=} F_{X,Y}(T, U_X, U_Y) \\ (X_{T=t}, Y_{T=t}) &\stackrel{a.s.}{=} F_{X,Y}(t, U_X, U_Y). \end{aligned}$$

Therefore, selecting the coordinates corresponding to  $Y$  furnishes a measurable function  $\tilde{F}_Y$  such that

$$\begin{aligned} Y &\stackrel{a.s.}{=} \tilde{F}_Y(T, U_X, U_Y) \\ Y_{T=t} &\stackrel{a.s.}{=} \tilde{F}_Y(t, U_X, U_Y). \end{aligned}$$

These identities hold on a measurable set  $\Omega^* \subseteq \Omega$  such that  $\mathbb{P}(\Omega^*) = 1$ . To conclude, simply observe that for any  $\omega \in \Omega^*$  such that  $T(\omega) = t$ , we have

$$Y(\omega) = \tilde{F}_Y(t, U_X(\omega), U_Y(\omega)) = Y_{T=t}(\omega). ■$$

## References

- S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- P. J. Bickel, E. A. Hammel, and J. W. O’Connell. Sex bias in graduate admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.
- T. Blom, S. Bongers, and J. M. Mooij. Beyond structural causal models: Causal constraints models. In R. P. Adams and V. Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 585–594. PMLR, 22–25 Jul 2020. URL <https://proceedings.mlr.press/v115/blom20a.html>.
- C. R. Blyth. On simpson’s paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

- S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.
- B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.
- J. Gaebler, W. Cai, G. Basse, R. Shroff, S. Goel, and J. Hill. A causal framework for observational studies of discrimination. *Statistics and public policy*, 9(1):26–48, 2022.
- M. M. Glymour and D. Spiegelman. Evaluating public health interventions: 5. causal inference in public health research—do sex, race, and biological factors cause health outcomes? *American journal of public health*, 107(1):81–85, 2017.
- P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.
- D. Ibeling and T. Icard. Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions. *Advances in Neural Information Processing Systems*, 36, 2024.
- G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.
- G. W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, 2020.
- G. W. Imbens and D. B. Rubin. Rubin causal model. In *Microeconometrics*, pages 229–241. Springer, 2010.
- A. Khademi and V. Honavar. Algorithmic bias in recidivism prediction: A causal perspective (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13839–13840, 2020.
- A. Khademi, S. Lee, D. Foley, and V. Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.
- M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc., 2017.
- J. Li, J. Liu, L. Liu, T. D. Le, S. Ma, and Y. Han. Discrimination detection by causal effect estimation. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1087–1094. IEEE, 2017.
- K. Makhoulouf, S. Zhioua, and C. Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.
- K. A. Markus. Causal effects and counterfactual conditionals: contrasting Rubin, Lewis and Pearl. *Economics & Philosophy*, 37(3):441–461, 2021.
- K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. *The Journal of Machine Learning Research*, 22(1):7322–7392, 2021.
- B. Neal. *Introduction to causal inference from a machine learning perspective*. bradyneal.com, 2020. [https://www.bradyneal.com/Introduction\\_to\\_Causal\\_Inference-Dec17\\_2020-Neal.pdf](https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf).
- J. Neyman. Sur les applications de la thar des probabilités aux expériences agricoles: Essay des principe. excerpts reprinted (1990) in english. *Statistical Science*, 5(463-472):4, 1923.
- H. Nilforoshan, J. D. Gaebler, R. Shroff, and S. Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pages 16848–16887. PMLR, 2022.
- J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *International Conference on Machine Learning*, pages 8401–8412. PMLR, 2021.

- J. Pearl. *Causality*. Cambridge university press, 2009.
- J. Pearl. Brief report: On the consistency rule in causal inference: " axiom, definition, assumption, or theorem?". *Epidemiology*, pages 872–875, 2010.
- J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.
- B. Qureshi, F. Kamiran, A. Karim, S. Ruggieri, and D. Pedreschi. Causal inference for social discrimination reasoning. *Journal of Intelligent Information Systems*, 54:425–437, 2020.
- T. S. Richardson and J. M. Robins. Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences*, 128(30):2013, 2013. Working Paper.
- G. Ridgeway. Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of quantitative criminology*, 22:1–29, 2006.
- P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- D. Rubin. Causal inference. In P. Peterson, E. Baker, and B. McGaw, editors, *International Encyclopedia of Education*, pages 66–71. Elsevier, Oxford, third edition edition, 2010. ISBN 978-0-08-044894-7. doi: <https://doi.org/10.1016/B978-0-08-044894-7.01313-0>. URL <https://www.sciencedirect.com/science/article/pii/B9780080448947013130>.
- D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.