Econometric Causality: How to Express it and Why it Matters.

James Heckman and Rodrigo Pinto [*]

Revised April 5, 2021

[*] James Heckman is the Henry Schultz Distinguished Service Professor of Economics and Public Policy at the University of Chicago; Professor of Science and Society, University College Dublin; and Research Fellow at the American Bar Foundation. Rodrigo Pinto is an Assistant Professor in the Department of Economics at the University of California - Los Angeles. We thank XXX for useful comments. All errors are our own.

Abstract

Frisch's 1930 seminal lectures lay the fundamentals of econometric causality. A century later, Frisch's ideas remain the basis of causal thinking. Although little has changed on the meaning of causality, the literature had major advances in expressing and manipulating causal inquiries. Several causal frameworks translate causal concepts into a mathematical language. However, the literature offers little guidance on which framework is best suited to investigate a given causal model. We revisit the core concepts of econometric causality and discuss the merits and limitations of several causal frameworks. We employ popular models of policy evaluation to illustrate the advantages and drawbacks of the frameworks. We show that the choice of causal framework matters. The interpretation and scrutiny of a causal model are substantially affected depending on the adopted causal framework.

Key words: Causality, Identification, Causal Calculus, Directed Acyclic Graphs, Simultaneous Treatment Effects

JEL codes: C10, C18

James Heckman
The University of Chicago
Department of Economics
1126 E. 59th St.
Chicago, IL 60637
(773) 702-0634
jjh@uchicago.edu

Rodrigo Pinto
University of California at Los Angeles
Department of Economics
315 Portola Plaza, Room 8385
Los Angeles, CA 90095
(310) 825-0849
rodrig@econ.ucla.edu

1 Introduction

A primary goal of econometric causality is to study the causal effect of an observed variable, often called the treatment, on an outcome. Economists have long realized that the co-movement between the treatment and the outcome is not sufficient to claim causation. This understanding dates back to Yule (1895), who coined the maxim "correlation is not causation". A century later, Yule's motto has become a mantra among economists and social scientists.

Despite its popularity, econometric causality is seldom a central subject in the formation of young economists. Causality is often conflated with the study of econometric models. Causal effects are usually addressed as statistical properties of some models such as randomized controlled trials or the instrumental variable model. This generates some confusion as statistical theory is void of causal concepts, and the fundamentals of causal inference apply to any causal model Heckman (2005). Indeed, Econometric causality is based on counterfactual concepts that are not well-defined in either statistics or probability theory.

The foundation of econometric causality stems from the seminal ideas of Frisch (1930), who explains that the concept of causality lies outside the realm of standard statistic theory. Its formalization requires a causal framework, that consists of additional mathematical machinery that enable the researcher to define and manipulate causal concepts. The literature on causality offers several options of causal frameworks. For instance, a researcher may opt to investigate a the nonparametric identification of a causal inquiry through a model defined in terms of structural equations. Instead, he/she could employ the potential outcome framework of Holland (1986), also known as Rubin-Holland causal model. The researcher could use the hypothetical model approach of Heckman and Pinto (2012) or apply the do-calculus of Pearl (2009a).

Causal frameworks can be broadly understood as alternative methods to formally express Frisch's original ideas on causality. Although they share the same causal concepts, the frameworks differ substantially in terms of the complexity of their setup and the mechanics of their analysis. The language of potential outcomes is the simplest and the most popular framework among those cited. It describes a causal model by statistical independence relations between potential (or counterfactual) counterpart of observed variables. Each nonparametric model described by the language of potential outcomes can be equivalently expressed in terms of structural equations. The hypothetical model (HM) framework and the do-calculus (DoC) are built upon these structural equations. HM invokes an alternative (hypothetical) model in which the treatment variable causing the outcome is exogenous. Identification is secured by expressing the probability the outcome conditioned on the treatment of the hypothetical model in terms of the observed data generated by the original model. DoC explores the fact that the structural equations can be expressed as a Directed Acyclic Graph (DAG). The method consists of several DAG-related rules that enables the researcher to investigate the statistical relation among counterfactual variables. The method is complete, that is to say that if a causal effect is identified, then it can be obtained by the iterative use of the DoC rules (Shpitser and Pearl, 2006, Tian and Pearl, 2002b). The Settable Systems of White and Chalak (2009) extends the DoC to include features of central interest to economists such as optimization, equilibrium, and game theory.

The literature on causality offers little guidance on deciding which framework is most suitable to investigate a causal model. This paper fills this gap by comparing the properties, benefits and drawbacks among causal frameworks. We revisit key principles of causal inference and discuss how each causal framework embodies these principles. We clarify the advantages that a more complex causal framework has when compared to a simpler approach. We illustrate cases where the additional complexity is justified and when it is not. We compare the machinery of each causal framework when examining well-known econometric models, such as matching, instrumental variable model, and the mediation model. We also examine more complex settings that combine the features across these econometric models. We investigate the merits of causal frameworks in terms of notational simplicity, tractability, intuition. We also discuss the spectrum of causal analysis empowered by each framework.

Most importantly, we show that the choice of causal framework matters. Distinct causal frameworks that are suitable to investigate the same causal model must arrive at the same identification results. Nevertheless, this fact does not imply that the frameworks are equally useful. Causal frameworks may substantially affect the interpretability of the causal model and severely limit the sophistication of causal analysis. The choice of causal frameworks is particularly critical when the researcher seeks to investigate model properties.

Section 2 recalls seminal ideas of causality made by early researches. Section 3 presents the causal model. Section 4 describes several causal frameworks. Sections (5)–(7) illustrates how the choice of causal languages influences the analysis of causal models that are commonly used by economists. Section (5) investigates the simple case of matching. Section (6) investigates the instrumental variable model. Section (7) investigates more complex settings such as the mediation model. Section (9) concludes.

2 Some Historical Background

A typical research activity in natural sciences consists of testing a theory using the controlled environment of a laboratory. An empiricist accustomed to using controlled environments is naturally inclined to define causality based on predictability. It sounds reasonable to state that a causal law is established when it is possible to determine the later state of system of variables based on its early onset.

It may come as a surprise that the notion of causality based on predictability is flawed. The problem with this causal notion is not its deterministic nature, but its lack of directionality. This issue is particularly relevant in physics. For instance, Newtonian equations of gravitational interaction determines the relation between two bodies but does not establish its causal direction. Earth mass attracts a falling apple as well as the falling fruit attracts the planet. Ampre's law determines that a changing magnetic field induces an electric field and vice versa. The law establishes an accurate equality but not a causal relation. This causal dilemma was investigate by the influential paper "On the Notion of Cause" of Bertrand Russell (1912), who denies the usefulness of any notion causality in physics.

The notion that predictability renders causality is clearly misguided. Yet, it is intuitive to postulate that a controlled environment is an ideal setting to examine causality. The assertion is correct. The key feature that renders causality in a controlled environment not the predictability of an experiment, but rather the possibility to vary a single variable of the system at the onset while holding others variables constant. Marshall (1890) proposes a notion of causality that captures the selective variation of inputs variables. He uses the term "ceteris paribus" to denote the causal effect of an *input* variable on an *output* variable when holding all remaining inputs constant.

Frisch (1930) made a substantial contribution to the notion of causality by refuting the notion of physical or actual manipulation of input variables altogether. He explains that the notion of causality must be *necessarily* defined in terms of a hypothetical instead of actual manipulation of variables. He stated that "causality is in the mind". By this, he meant:

> "*. . . we think of a cause as something imperative which exists in the* **exterior world.** *In my opinion this is fundamentally* ***wrong****. If we strip the word cause of its animistic mystery, and leave only the part that science can accept, nothing is left except* a certain way of thinking, [T]he scientific . . . problem of **causality** is essentially a problem regarding our **way of thinking,** not a problem regarding the nature of the exterior world." (Frisch 1930, p. 36, published 2011)

Frisch (1938) makes the case that causality has to be defined within a system of variables governed autonomous functions. That is to say that functions that maintain their shape as their inputs vary. Haavelmo (1944) formalized Frisch's insight. He defined causal operations that capture the fundamentals of econometric causality. Haavelmo introduced the causal operation of fixing a variable, which entails causal concepts such as counterfactuals and ceteris paribus.

A causal model consists of a system of structural (autonomous) equations that determine the causal relations among observed and unobserved variables. Counterfactual outcomes are defined by a hypothetical experiment of *fixing* the value of one or some input variables that are arguments of these equations. Causal effects are determined by a weighted difference between counterfactuals.

Consider a simple example where an outcome Y is caused by observed inputs X_1, X_2 and an unobserved variable U. According to of Haalvelmo's rationale, these causal relations are characterised by the structural equation $Y = f(X_1, X_2, U)$. In the case of a linear model, we have that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U.$$

The counterfactual outcomes are defined by fixing the inputs of function f. The counterfactual outcome Y when variables (X_1, X_2, U) are fixed at values (x_1, x_2, u) is given by $Y(x_1, x_2, u) = \beta_0 + \beta_1 x_1 + \beta_2 x2 + u$. The causal effect of an unit increase in input X_1 on outcome Y is the expected difference between the counterfactual outcomes Y when X_1 is fixed to values $x_1 + 1$ and x_1. Ceteris Paribus means that the remaining inputs of Y, that is X_2 and U, are fixed at constant values. This setup generates the following effect:

$$\begin{aligned}
&Y(x_1 + 1, x_2, u) - Y(x_1, x_2, u) \\
&\quad = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \\
&\quad = \beta_1(x_1 + 1 - x_1) = \beta_1
\end{aligned}$$

Suppose U has mean zero and is (mean) independent of (X_1, X_2). Then the conditional expectation $E(Y|X_1 = x_1, X_2 = x_2)$ is given by:

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

The conditional expectation $E(Y|X_1 = x_1, X_2 = x_2)$ and the counterfactual outcome $Y(x_1, x_2, 0)$ are mathematically equivalente but conceptually different. The counterfactual $Y(x_1, x_2, 0)$ is a thought experiment that hypothetically fixes the values to the inputs of outcome Y. It is a causal statement that takes into account the causal direction of a system of equations. It describes the outcome when input variables are fixed. On the other hand, the conditional expectation $E(Y|X_1 = x_1, X_2 = x_2)$ is a random variable that describes the observed data. It is a statistical operation that depends on the joint probability the variables. The expectation does not require nor impose any causal relation among observed variables. The expectation $E(X_1|Y = y, X_2 = x_2)$ is well-defined, while the counterfactual $X(y, x_2, 0)$ is not.

Heckman (2006) disentangles the econometric approach to causality into three tasks displayed in Table 1. The first task uses scientific theory to determine a causal model. The second task uses causal analysis to study the identification of causal effects. This task requires a causal framework that enables to define and manipulate causal concepts. A causal parameter is identified if it can expressed as a function of the observed variables of the causal model. The last task is the estimation, which employs statistical theory to evaluate model parameters using observed data. The evaluation of a causal parameters is not necessarily contingent on a particular statistical method. This logical sequence of tasks is often blurred when the parameter is defined by a statistical method such as the difference-in-difference model or when proving the statistical consistency of an estimator.

Table 1: Tasks of the Econometric Approach to Causality

Task	**Description**	**Requirements**
1	Defining Causal Models	A Scientific Theory A Mathematical Framework
2	Identifying Causal Parameters from Known Population Distribution Functions of Data	Mathematical Analysis Connect Hypothetical Variation with Data Generating Process (Identification in the Population)
3	Estimating Parameters from Real Data	Statistical Analysis Estimation and Testing Theory

Our primary goal is to examine the second task of econometric analysis. Namely, examine frameworks that enable to define and manipulate causal concepts. To this end, it is useful to formally define a causal model.

3 The Causal Model

The Structural Causam Model (SCM) dates back to Haavelmo (1944), who describes a causal model characterised by a set $\mathcal{T}$ of observed and unobserved variables such that for each variable $Y \in \mathcal{T}$ we have an associated autonomous (or structural) function f_Y and an exogenous error term ϵ_Y which are not observed.[1] Arguments of the function f_Y are the variables in $\mathcal{T}$ that cause Y. Without loss of generality, we can assume that terms are statistically independent.[2] All variables are defined in common probability space $(\mathcal{I}, \mathcal{F}, P)$.

A common goal in policy evaluations to evaluate the causal effect of the treatment T on and outcome Y. The identification of causal effects can be broadly understood as methods to control for unobserved confounding variables V that cause both T and Y. Consider a simple model where an endogenous treatment T that causes an unobserved abilities A, and both T, A cause the outcome Y. Agent's unobserved variable V plays the role of a confounder that generates selection bias. V causes both the treatment choice T and the unobserved skill A. The model is displayed in the first column of Table 2 and it is not identified without additional assumptions.

The second column of Table 2 displays the model as a Directed Acyclic Graph (DAG).[3] Causal links are denoted by directed arrows, observed variables are displayed by squares and unobserved variables by circles. The model is acyclic, i.e. non-recursive, because no causal path leads a variable to cause itself. In this case, we can generate a series of conditional independent relations using the Local Markov Condition (LMC) of Kiiveri et al. (1984). The condition states that *a variable is independent of its non-descendants condition its parents.* Parents of a variable Y are the argument of its function f_Y and non-descendants of Y consists of variables that are not directly or indirectly caused by Y. The third column of

[1] Autonomy means deterministic functions that are "invariant" to changes in their arguments (Frisch. 1938). Hurwicz (1962) prefers the term "structural" to denote autonomous equations.

[2] This assumption comes without a loss of generality because any correlation structure among error terms can be modeled by adding unobserved variables to the model. See (Heckman and Pinto, 2015a) for a discussion.

[3] See Lauritzen (1996) for the theory of Bayesian Networks.

Table 2: Causal Model, DAG Representation and Conditional Independence Relationships

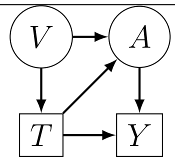
Causal Model	**DAG**	**Local Markov Conditions**
$V = f_V(\epsilon_V)$		$V \perp\!\!\!\perp \varnothing \vert \varnothing$
$A = f_A(T, V, \epsilon_A)$		$A \perp\!\!\!\perp \varnothing \vert (T, V)$
$T = f_T(V, \epsilon_T)$		$T \perp\!\!\!\perp \varnothing \vert V$
$Y = f_Y(T, A, \epsilon_Y)$		$Y \perp\!\!\!\perp V \vert (T, A)$

Table 2 applies the LMC to each variable.

A non-parametric causal model is equivalently described by its structural equations, its DAG or the relationships generated by the LMC. Note that LMC generates no independence relationship for V, A or T. So the model is characterized by a single independence relationship, that is, $Y \perp\!\!\!\perp V|(T, A)$.[4]

Fixing is a primary concept used to define counterfactuals. It is a causal exercise that hypothetically assigns values to inputs of autonomous equation. The counterfactual abilities A when the treatment is fixed at a value $t \in \mathit{supp}(T)$ is given by $A(t) \equiv f_A(t, V, \epsilon_A)$ while the counterfactual outcome is given by $Y(t) \equiv f_Y(t, A(t), V, \epsilon_Y)$. is obtained by *fixing* the input T of the outcome function to a value $t \in \mathit{supp}(T)$. The average Causal Effects of T on Y when T is *fixed* at t, t' is $ATE = \mathrm{E}(Y(t) - Y(t'))$. The independence of error terms $\epsilon_V, \epsilon_A, \epsilon_T, \epsilon_Y$ implies that $Y(t) \perp\!\!\!\perp T|V$. [5]

[4] The Graphoid Axioms of Dawid (1976) may generate additional conditional independence relations. These consist of six rules that apply for any disjoint sets of variables $X, W, Z, Y \subseteq \mathcal{T}$:

$$\begin{aligned}
\text{Weak Union: } & X \perp\!\!\!\perp (W, Y)|Z \Rightarrow X \perp\!\!\!\perp Y|(W, Z). \\
\text{Contraction: } & X \perp\!\!\!\perp W|(Y, Z) \text{ and } X \perp\!\!\!\perp Y|Z \Rightarrow X \perp\!\!\!\perp (W, Y)|Z. \\
\text{Intersection: } & X \perp\!\!\!\perp W|(Y, Z) \text{ and } X \perp\!\!\!\perp Y|(W, Z) \Rightarrow X \perp\!\!\!\perp (W, Y)|Z \\
\text{Symmetry: } & X \perp\!\!\!\perp Y|Z \Rightarrow Y \perp\!\!\!\perp X|Z. \\
\text{Decomposition: } & X \perp\!\!\!\perp (W, Y)|Z \Rightarrow X \perp\!\!\!\perp Y|Z. \\
\text{Redundancy: } & X \perp\!\!\!\perp Y|X.
\end{aligned}$$

[5] Note that if the error terms are mutually independent than $(\epsilon_Y, \epsilon_A) \perp\!\!\!\perp \epsilon_T|\epsilon_V$ holds, and in particular, $(\epsilon_Y, \epsilon_A) \perp\!\!\!\perp \epsilon_T|V$ holds. This implies that the relationship $f_Y(t, f_A(t, V, \epsilon_A), \epsilon_Y) \perp\!\!\!\perp f_T(V, \epsilon_T)|V$ also holds, which means that $Y(t) \perp\!\!\!\perp T|V$.

Fixing T does not affect the distribution of V, while statistical conditioning does.[6] Indeed, conditioning is a statistical operator that lacks directionality and affects the distribution of all correlated variables. Fixing, on the other hand, is a causal operator that embodies causal direction of a model and only affects the distribution of the variables caused by the input being fixed.

The lack of directionality in standard statistical and probability theory generates some confusion.[7] Causal concepts such as fixing are not well-defined. This fact foments several causal frameworks that seek to make statistics and probability to converse with causality. We discuss some of these frameworks next.

4 Causal Frameworks

The Language of Potential Outcomes

The most commonly used causal framework is the language of potential outcomes (PO), which is also called the Rubin-Holland causal model (Holland, 1986). The method does not employ structural equations. It describes a causal model by stating statistically independence relationships among counterfactuals of observed variables.

The PO framework typically described in terms of the unit of analysis $\imath \in \mathcal{I}$ that usually represents an economic agent. The common setup relies on three observed variables: baseline variables X, a treatment T, and an outcome Y. X_i, T_i, Y_i stand for realizations of agent $\imath$. $Y_i(t)$ is the unobserved potential outcome Y for agent i when the treatment T takes the value t and the causal effect of a change in the treatment status from t' to t for unit i is

[6] For instance, the factorization of the joint distribution of the variables of the model under when conditioning on X is given by $P(Y,V,A|T=t) = P(Y|A,V,T=t)P(A|V,T=t)P(V|T=t)$, while the joint distribution under fixing is given by $P(Y,V,A|T \text{ fixed at } t) = P(Y|V,T=t)P(A|V,T=t)P(V)$.

[7] See Pearl (2009b) and Spirtes et al. (2000) for discussions.

$Y_1(t) - Y_1(t')$. The observed outcome is given by:

$$Y_i = \sum_{t \in supp(T)} Y_i(t) \cdot \mathbf{1}[T_i = t] \;\equiv\; Y_i(T_i),$$

where $supp(T)$ denotes the support of the treatment and $\mathbf{1}[\cdot]$ is the standard indicator function. The model characterized by independence relationships among counterfactuals of observed variables. For example, if the unobserved confounding variable V of the model in Table 2 where replaced by the observed baseline variable X, we would have that $Y(t) \perp\!\!\!\perp T|X$. This is often called the *matching* or *exogeneity* assumption.

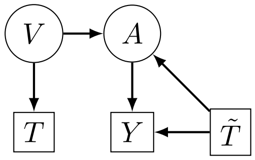
Imbens (2019) lists several reasons why the PO language is a popular policy evaluation framework. The PO assumptions can include monotonicity and shape restrictions are easily assessed and implemented in economic contexts. The method allows for heterogeneity in treatment effects and it lends itself to traditional economic models, such as supply and demand settings, where non-causal and causal variables are distinctive. The method is particularly suitable for identification strategies generally analyze causal model that stem from a small pool of variables, which have been exhaustively analyzed.

The PO framework is lauded for its simplicity and has been widely implemented in psychology, sociology, and economics. PO simplicity is beneficial when studying simple models such as the Randomized Control Trials or causal models that invoke the matching assumption. We show that its simplicity however becomes a hindrance when examining more complex models. A primary drawback of the PO framework is that the model is defined only in terms of observed variables. The lack of unobserved variables poses substantial limitations when examining identification assumptions of the instrumental variable model. Most important, the framework does not explicitly states structural equations. This drastically reduces the interpretation of the causal relations among variables. We illustrate this fact using the mediation model.

The Hypothetical Model Framework

Heckman and Pinto (2015b) provides a framework that permits to examine causal con-

Table 3: Associated Hypothetical Model

Hypothetical Model	**DAG**	**Local Markov Conditions**
$V = f_V(\epsilon_V)$		$V \perp\!\!\!\perp \tilde{T}$
$\tilde{T} = f_{\tilde{T}}(\epsilon_{\tilde{T}})$		$\tilde{T} \perp\!\!\!\perp (T, A, V)$
$A = f_A(\tilde{T}, V, \epsilon_A)$		$A \perp\!\!\!\perp T \vert (V, \tilde{T})$
$T = f_T(V, \epsilon_T)$		$T \perp\!\!\!\perp \tilde{T} \vert V$
$Y = f_Y(T, A, \epsilon_Y)$		$Y \perp\!\!\!\perp (V, T) \vert (\tilde{T}, A)$

cepts using standard probability theory. It is inspired by the seminal ideas of Frisch (2010) and Haavelmo (1944). The framework invokes the same structural equations of the causal model. It copes with the causal operator of fixing by generating a *hypothetical model* that replaces the variable we ought to fix by an exogenous *hypothetical variable.* The hypothetical model defined in this fashion has a desired property of independence of the input aim to fix.

Consider the causal model in Table 2. In the language of Heckman and Pinto (2015b), this represents the *empirical model* that generates the observed distribution of the data. The *hypothetical model* used to examine the causal effect of treatment T is characterised by four properties:

1. It uses the same autonomous functions f_V, f_A, f_T, f_Y of the empirical model.
2. It uses the same error terms $\epsilon_V, \epsilon_A, \epsilon_T, \epsilon_Y$ of the empirical model.
3. It appends a variable $\tilde{T}$ which is exogenous, that is to say that it is not caused by any variable of the system.
4. It replaces the T-input of the skill equation and the outcome equation by the hypothetical variable $\tilde{T}$.

Standard statistical tools apply to both the empirical and hypothetical models. Table 3 display the hypothetical model using structural equations, a DAG and via the LMCs.

The primary motivation for introducing the hypothetical model is to make statistics converse with causality. Although the causal operation of fixing is outside statistical realm, fixing translates to standard statistical conditioning in the hypothetical model. Formally, let P_e, E_e be the probability and expectation for the empirical model and P_h, E_h for the hypothetical model. The probability (or expectation) of the counterfactual outcome in the empirical model is equal to the probability (or expectation) of the outcome conditioned on the hypothetical variable in the hypothetical model. For any set of variables $W \subset \mathcal{T}$ and any set $\mathcal{Y} \subset \mathit{supp}(Y)$, we have that:

$$P_e(Y(t) \in \mathcal{Y}|W) = E_h(Y \in \mathcal{Y}|\tilde{T} = t, W) \text{ and } E_e(Y(t)|W) = E_h(Y|\tilde{T} = t|W), \tag{1}$$

In the example of Table 3, we have that $E_e(Y(t)) = E_h(Y|\tilde{T} = t)$ as well as $E_e(Y(t)|A) = E_h(Y|\tilde{T} = t, A)$ and $E_e(Y(t)|V) = E_h(Y|\tilde{T} = t, V)$.

The hypothetical model is an abstract model derived from the original model that generates data. Causal parameters are defined as conditional probabilities in the hypothetical model P_h are said to be identified if they can be expressed in terms of the probabilities of observed data generated by the empirical model P_e. Thus identification analysis requires us to connect the hypothetical and the empirical models. Two rules suffice for the study of identification. For any disjoint set of variables Y, W in $\mathcal{T}$ we have that:

$$Y \perp\!\!\!\perp \widetilde{T}|(T, W) \Rightarrow \mathbf{P}_h(Y|\widetilde{T}, T = t', W) = \mathbf{P}_e(Y|T = t', W) \tag{2}$$

$$Y \perp\!\!\!\perp T|(\widetilde{T}, W) \Rightarrow \mathbf{P}_h(Y|\widetilde{T} = t, T, W) = \mathbf{P}_e(Y|T = t, W) \tag{3}$$

Rules (2)–(3) state that we can migrate from the hypothetical model to the empirical model whenever we have independence conditions that employ the treatment T and the hypothetical treatment $\tilde{T}$, that is, $Y \perp\!\!\!\perp \widetilde{T}|(T, W)$ or $Y \perp\!\!\!\perp T|(\widetilde{T}, W)$. Consider the example of Table 3. The independence relations $A \perp\!\!\!\perp T|(V, \tilde{T})$ (from the LMC of A) and $Y \perp\!\!\!\perp T|(A, V, \tilde{T})$ (from the LMC of A) imply $Y \perp\!\!\!\perp T|(\tilde{T}, V)$.[8] According to (3), we have that $E_h(Y|\tilde{T} = t, V) = E_e(Y|T = t, V)$, which is equivalent to stating $E_e(Y(t)|V) = E_e(Y|T = t, V)$.

The hypothetical model appends a substantial machinery to the original causal model. A natural question is whether this additional structure is justified to examine the identification of a causal model. The simple answer is that the hypothetical model is useful when examining for causal models that are more complex than our leading example. Section 7 investigates the

[8] Due to the Graphoid Axiom called contraction.

mediation model and illustrate how the hypothetical framework can substantially simplify the identification analysis.

A notable benefit of the hypothetical model is to clarify the concept of causality. The framework is intuitive. It clearly disentangle each of the tasks of causal analysis in Table 1. Scientific knowledge is necessary o define the causal model. Causal effects are assessed using the hypothetical model, which is an abstraction of the causal model that generates data. The hypothetical model formalizes Frisch motto "*Causality is in the Mind*". Finally, the identification of causal effects require us to connect the hypothetical model that characterise causal effects with the empirical model that generates the observed data.

The do-Calculus

The *do*-calculus of Pearl (2009b) is a powerful framework that uses a graphical approach to examine causality. Its name stems from using the term $do(X) = x$ for fixing a variable X at a value $x \in supp(X)$. The method consists of rules that uses Directed Acyclic Graphs manipulations to investigate if causal effects are identified.[9] The starting point of the *do*-calculus is a causal model represented by a DAG. The method consists of three rules. Each rule combines a graphical condition and a conditional independence relation that, when satisfied, imply a probability equality. Some notation is necessary to explain these rules.

Let G denotes a DAG that represents the causal model and X, W, Z denote sets of variables in $\mathcal{T}$. Let $Z(W)$ be the variables in Z that do not directly or indirectly cause W. Let $G_{\bar{X}}$ denotes the DAG that deletes all causal arrows arriving at X in the original DAG G. $G_{\underline{Z}}$ is the DAG that deletes all causal arrows emerging from Z. $G_{\overline{X},\underline{Z}}$ deletes all arrows arriving at X and emerging from Z. $G_{\overline{X,Z(W)}}$ deletes all arrows arriving at X in addition to arrows arriving at $Z(W)$, namely, arriving at variables in Z that are not ancestors of W. Now let X, Y, Z, W be any disjoint sets of variables in $\mathcal{T}$. In this notation, the three rules of

[9] For a recent book on the graphical approach to causality, see Peters et al. (2017), and for related works on causal discovery, see Glamour et al. (2014), Heckman and Pinto (2015a), Hoyer et al. (2009), and Lopez-Paz et al. (2017).

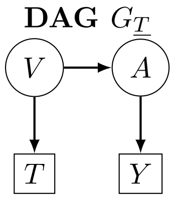
the do-calculus are then can be stated as:

1. if $Y \perp\!\!\!\perp Z|(X,W)$ holds in $G_{\overline{X}}$, then $P(Y|do(X),Z,W) = P(Y|do(X),W)$,
2. if $Y \perp\!\!\!\perp Z|(X,W)$ holds in $G_{\overline{X},\underline{Z}}$, then $P(Y|do(X),do(Z),W) = P(Y|do(X),Z,W)$,
3. if $Y \perp\!\!\!\perp Z|(X,W)$ holds in $G_{\overline{X},\overline{Z(W)}}$, then $P(Y|do(X),do(Z),W) = P(Y|do(X),W)$,

where X,W denote disjoint sets of variables in $\mathcal{T}$. The process of checking if a causal effect is identified requires the reiterative use of these rules.

Consider applying the do-calculus in our leading example of Table 2 to show that $Y(t) \perp\!\!\!\perp T|V$ holds. In the notation of the *do*-calculus, the statement is expressed as $P(Y|do(T),V) = P(Y|T,V)$. Rule 2 is the best candidate to prove the equality. It is useful to rewrite Rule 2 by setting variable X to nil, Z to T, and W to V. Under this transformations, Rule 2 is restated as: if $Y \perp\!\!\!\perp T|V$ holds in $G_{\underline{T}}$, then $P(Y|do(T),V) = P(Y|T,V)$. Table 4 display the DAG $G_{\underline{T}}$ and its associated LMC for each of the variables. The LMC for T in the DAG $G_{\underline{T}}$ generates $T \perp\!\!\!\perp (Y,A)|V$. Therefore the modified Rule 2 above implies that $P(Y|do(T),V) = P(Y|T,V)$ holds as intended.

Table 4: DAG for the *do*-calculus



LMC
$V \perp\!\!\!\perp \varnothing|\varnothing$
$A \perp\!\!\!\perp T|V$
$T \perp\!\!\!\perp (Y,A)|V$
$Y \perp\!\!\!\perp V|(T,A)$

A great feet of the do-calculus is that it is complete. This means that the iterative use of the rules of the do-calculus will always deliver the an identification equation in the case of an identified counterfactual outcome or it will indicate if the causal parameter is not identified.

The do-calculus differs from the hypothetical framework in several aspects. The first difference is conceptual, do-calculus employ DAG manipulations that lie outside standard probability theory while the hypothetical framework refrains from using machinery outside probability theory.

Another difference is that the hypothetical framework targets causal links while the do-calculus focus on variables. Specifically, the hypothetical framework introduces a new variable $\tilde{T}$ that can replace one, several or all the T-inputs of the model. The do-calculus fix the variable T itself which comprises all the T-inputs.

A limitation of the do calculus is that it is hermetic in the sense that it is based on rules that only apply to nonparametric models that can be expressed by a DAG. The identification stems from the causal relationship among the variables in a model. The do-calculus is not intended to examine identification assumptions based on functional form restrictions such as monotonicity conditions, which attains to the form of structural functions instead of the causal relationship among variables.

5 The Matching Assumption

The most common identification approach is to invoke the *matching* assumption. It states that the treatment choice T is independent of counterfactual outcomes $Y(t)$ when conditioning on observed *matching variables* X, that is, $Y(t) \perp\!\!\!\perp T|X$.[10] The matching assumption solves the problem of selection bias comparing (matching) individuals with different treatment statuses that share the same values of baseline characteristics X. The average causal effect of a binary treatment $T \in \{t_0, t_1\}$ is identified by a weighted average of the conditional

[10] In the language of Bayesian Networks of Pearl (2009a), it is said that X *d-separates* $Y(t)$ and T.

difference in means between treated and control participants across the values of X :[11]

$$E(Y(t_1) - Y(t_0)) = \int \Big(E(Y(t_1)|T = t_1, X = x) - E(Y(t_0)|T = t_0, X = x) \Big) dF_X(x) \quad (4)$$

$$= \int \Big(E(Y|T = t_1, X = x) - E(Y|T = t_0, X = x) \Big) dF_X(x) \quad (5)$$

Our leading example in Table 2 helps to interpret the matching assumption. Unobserved variable V plays the role of a matching variable. Thus, a natural interpretation of the matching assumption is that the analyst observes a sufficiently rich set of baseline variables X that cause Y and T that justify ignoring any remaining unobserved confounders V. The matching assumption is justified in the case of randomized controlled trials. Variables X are those used in the randomization protocol and the assumption is assured by the design of the experiment. On the other hand, the matching assumption is easily criticized in observational studies (Heckman, 2008; Heckman and Navarro, 2004).

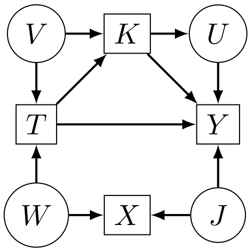
The matching assumption is usually invoked within the PO framework, which suppresses structural equations. This limitation incites misleading conclusions regarding matching variables Rubin (2008); Shrier (2008). The simplest causal model that generates the matching assumption is the one in which X causes T, Y and T causes Y. This induces the researcher to conclude that the greater the number of matching variables, the greater credibility the matching assumption. This assessment is wrong.[12] Table 5 illustrate a causal model where conditioning on pre-program variables X induces bias while conditioning on post-treatment variable K renders $Y(t) \perp\!\!\!\perp T|K$.

The propensity score matching is a celebrated result for the binary choice models. Let $P(X) \equiv P(T = t_1|X); T \in \{t_0, t_1\}$ be the propensity score, that is, the probability of being treated given X. Rosenbaum and Rubin (1983) show that if the matching assumption holds

[11] Heckman et al. (1998) investigate several estimation methods that invoke the matching assumption. They incorporated additive separability between observable and unobservable variables as well as exogeneity conditions that isolate outcomes and treatment participation into the matching framework. Additionally, they compare various types of estimation methods to show that kernel-based matching and propensity score matching have similar treatment of the variance of the resulting estimator.

[12] See Pearl (2009c) and Greenland et al. (1999) for examples of causal models where augmenting the set of matching variables generates bias.

Table 5: Associated Hypothetical Model

Causal Model	DAG	Independence Relationships
$V = f_V(\epsilon_V)$		
$J = f_J(\epsilon_J)$		
$W = f_W(\epsilon_W)$		
$V = f_V(\epsilon_V)$		$Y(t) \perp\!\!\!\perp T \vert K$
$T = f_T(V, W, \epsilon_T)$		$Y(t) \not\perp\!\!\!\perp T \vert X$
$K = f_K(T, V, \epsilon_K)$		$Y(t) \not\perp\!\!\!\perp T \vert (K, X)$
$U = f_U(K, \epsilon_U)$		
$X = f_K(W, J, \epsilon_X)$		
$Y = f_Y(T, K, U, J, \epsilon_Y)$		

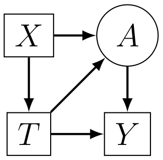
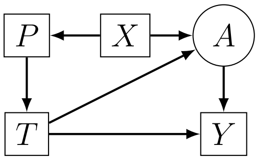
for matching variables X, it must also holds for the propensity score $P(X)$:

$$Y(t) \perp\!\!\!\perp T|X \Rightarrow Y(t) \perp\!\!\!\perp T|P(X). \tag{6}$$

The major benefit of (6) is to reduce the dimensionality of matching variables X.

Rosenbaum and Rubin (1983) describe the matching model using PO and derive (6) using a statistical rationale. It is useful to examine the model using the language of structural equations. The left-side of the Table 6 presents the matching model where X causes T. The propensity score is a sufficient statistic that fully characterises the distribution of a binary treatment T. In terms of causal relations, it means that $P(X)$ causes T and is the only parent of the treatment T. The right-side of the Table 6 displays the propensity score model. The independence relationship $Y(t) \perp\!\!\!\perp T|P$ now arises from the causal relation among the variables. Using the same DAG, it is easy to see that Rosenbaum and Rubin (1983) result extends to any sufficient statistics that fully characterise the distribution of T, such as the λ-parameter in the case of a poison distribution. It is easy to prove that $Y(t) \perp\!\!\!\perp T|P$ would also hold if an exogenous variable (observed or not) were to cause P or T.

Table 6: The Matching Model and The Propensity Score Matching Model

The Matching Causal Model	**The Propensity Score Model**
	
$Y(t) \perp\!\!\!\perp T \mid X$	$Y(t) \perp\!\!\!\perp T \mid P(X)$

6 Instrumental Variable Model

The instrumental variable (IV) model is a more credible alternative to the matching assumption. The IV model is typically described in terms of the PO framework. The simplest model consists of three observed variables: an instrumental variable Z that causes a treatment T, which in turn causes an outcome Y.

The IV model is characterised by three core properties. The exclusion restriction states that Z only causes outcome Y through T. Notationally, we have that $Y_i(t,z) = Y_i(t,z')$ for all agents $i \in \mathcal{I}$. The IV relevance states that Z is not statistically independent of T, that is $Z \not\perp\!\!\!\perp T$. The exogeneity condition states that Z is statistically independent of the choice and outcome counterfactuals: $Z \perp\!\!\!\perp (Y(t), T(z))$.

The core properties of the IV model are not sufficient to identify causal effects. According to the do-calculus, the model is not identified. However, a vast literature exists on the additional assumptions to secure identification of the IV model. These identifications assumptions stem from restrictions on the support, distribution or functional form of the model variables. These type of restrictions are outside the realm investigated the do-calculus, which focus only on the causal direction among model variables.

A popular identification assumption in the PO framework is the monotonicity condition of Imbens and Angrist (1994). The assumption applies to the case of a binary treatment

$T \in \{0,1\}$ and states that a change in the instrument may induce agents to change their treatment choice towards the same direction. Notationally, for any $z, z' \in supp(Z)$, we have that:

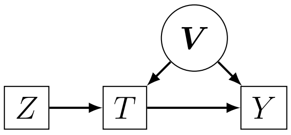
$$T_i(z) \geq T_i(z') \forall i \in \mathcal{I} \text{ or } T_i(z) \leq T_i(z') \forall i \in \mathcal{I} \tag{7}$$

In the case of a binary instrumental variable $Z \in \{z_0, z_1\}$, the monotonicity (7) secures the identification of the Local Average Treatment Effect (LATE), which is the causa effect for those who switch their choice as the instrument changes: $LATE = E(Y(1) - Y(0)|T(z_1) \neq T(z_0))$.

The PO framework is intuitive, simple and parsimonious. Unfortunately, it offers limited tools to further explore the properties of the IV model. Its main drawback is that it does not explicitly defines an unobserved confounding variable that renders the treatment endogenous. The lack of a confounder prevents the further investigation of the model. It does not precludes the identification of causal effect, but substantially confines its interpretation, comprehension and manipulation.

We now describe the same IV model as a SCM. The treatment is endogenous, that is to say that there exists a confounding variable represented by an unobserved random vector $\boldsymbol{V}$ that causes T and Y. The instrument is exogenous, which implies that $Z \perp\!\!\!\perp \boldsymbol{V}$. The simplest causal model using IV is described in Table 8.

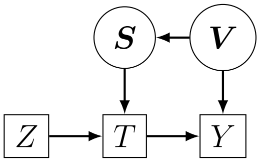
Table 7: Instrumental variable Model

Causal Model	**DAG**	**LMC**	**Counterfactuals**
$\boldsymbol{V} = f_V(\epsilon_V)$		$\boldsymbol{V} \perp\!\!\!\perp Z$	
$Z = f_Z(\epsilon_Z)$		$Z \perp\!\!\!\perp \boldsymbol{V}$	
$T = f_T(Z, V, \epsilon_T)$		$T \perp\!\!\!\perp -\|(Z, \boldsymbol{V})$	$T(z) = f_T(z, V, \epsilon_T)$
$Y = f_Y(T, \boldsymbol{V}, \epsilon_Y)$		$Y \perp\!\!\!\perp Z\|(T, \boldsymbol{V})$	$Y(t) = f_Y(t, V, \epsilon_Y)$

The core IV properties arise as consequence of the causal model itself. The exclusion restriction $Y_i(t,z) = Y_i(t,z')$ holds because Z does not directly cause Y. The IV relevance $Z \not\perp\!\!\!\perp T$ holds as Z causes T and the exogeneity condition $Z \perp\!\!\!\perp (Y(t), T(z))$ is a consequence of the independence of error terms.

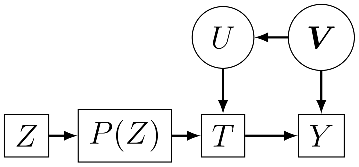
Heckman and Pinto (2018) discuss the identification of the IV model with multiple treatments $T \in \{t_1, ..., t_{N_T}\}$ and and categorical instrument $Z \in \{z_1, ..., z_{N_Z}$. A central variable in their analysis is the response variable $\boldsymbol{S} = [T(z_1), ..., T(z_{N_Z})]$, which stands for unobserved vector of counterfactual choices across all values in the support of the instrument. The response variable is a function only of the confounding variable $\boldsymbol{V}$ as $T(z) = f_T(z, \boldsymbol{V}, \epsilon_T)$. Thus it does not add any additional information to the model. Nevertheless, $\boldsymbol{S}$ helps to understand the identification problem. The vectors $\boldsymbol{s} \in \mathit{supp}(\boldsymbol{S})$ that $\boldsymbol{S}$ can take are called response-types.

Table 8: Instrumental variable Model with a Response Variable

Causal Model	**DAG**	**LMC**	**Counterfactuals**
$\boldsymbol{V} = f_V(\epsilon_V)$		$\boldsymbol{V} \perp\!\!\!\perp Z$	
$\boldsymbol{S} = f_S(\boldsymbol{V})$		$\boldsymbol{S} \perp\!\!\!\perp Z \vert \boldsymbol{V}$	
$Z = f_Z(\epsilon_Z)$		$Z \perp\!\!\!\perp \boldsymbol{V}$	
$T = f_T(Z, V, \epsilon_T)$		$T \perp\!\!\!\perp - \vert (Z, \boldsymbol{V})$	$T(z) = f_T(z, V, \epsilon_T)$
$Y = f_Y(T, \boldsymbol{V}, \epsilon_Y)$		$Y \perp\!\!\!\perp Z \vert (T, \boldsymbol{V})$	$Y(t) = f_Y(t, V, \epsilon_Y)$

Given $\boldsymbol{S}$, the treatment choice T depends on Z, which is independent of $Y(t)$. Therefore, we have that $Y(t) \perp\!\!\!\perp T|\boldsymbol{S}$ holds. The response variable $\boldsymbol{S}$ can be understood as a balancing score for $\boldsymbol{V}$, namely, a coarse transformation of $\boldsymbol{V}$ that maintains the matching property $Y(t) \perp\!\!\!\perp T|\boldsymbol{V}$. In essence, the identification inquiry consists in evaluating counterfactual outcome means $E(Y(t)|\boldsymbol{S} = \boldsymbol{s}); \boldsymbol{s} \in \mathit{supp}(S)$ from observed outcome means $E(Y|T = t, Z = z); z \in \mathit{supp}(Z)$. An identification problem arises because, without additional assumptions, as the total number of vectors $\boldsymbol{s} \in \mathit{supp}(\boldsymbol{S})$ grows exponentially in N_Z, that is $|\,\mathit{supp}(\boldsymbol{S})| = N_T^{N_Z}$, while the number of observed outcome mean grows linearly $N_T \cdot N_Z$.

Table 9: Instrumental variable Model under Monotonicity

Causal Model	**DAG**	**LMC**
$\boldsymbol{V} = f_V(\epsilon V)$		$\boldsymbol{V} \perp\!\!\!\perp Z$
$U = f_U(\boldsymbol{V}) \sim Uni[0,1]$		$U \perp\!\!\!\perp Z \vert \boldsymbol{V}$
$Z = f_Z(\epsilon_Z)$		$P(Z) \perp\!\!\!\perp \boldsymbol{V} \vert Z$
$P(Z) = P(T = t_1 \vert Z)$		$Z \perp\!\!\!\perp \boldsymbol{V}$
$T = \mathbf{1}[P(Z) \geq U]$		$T \perp\!\!\!\perp \boldsymbol{V} \vert (P(Z), U)$
$Y = f_Y(T, \boldsymbol{V}, \epsilon_Y)$		$Y \perp\!\!\!\perp Z \vert (T, \boldsymbol{V})$

The identification of counterfactual outcomes require assumptions that limit the number of possible response-types. Consider the case of a binary choice $T \in \{0,1\}$ where $|\,supp(Z)| = N_Z$. There are N_Z outcome means $E(Y|T = 0, Z = z); z \in supp(Z)$ for the choice $T = 0$. However, the total number of possible response-types that contain the choice 0 is $N_Z^2 - 1$. The monotonicity assumption 7 enables to eliminate response-types such that the final number of response-types that contain the choice 0 is precisely N_Z. The model is just identified.

Vytlacil (2002) has shown that the monotonicity (7) is equivalent to stating that the treatment choice $T \in \{0,1\}$ is governed by a separable equation on Z and $\boldsymbol{V}$, that is $T = \mathbf{1}[\phi(Z) \geq \xi(\boldsymbol{V})]$. This separable equation can be conveniently restated as $T = \mathbf{1}[P(Z) \geq U]$ where $P(Z)$ is the propensity score and $U = F_{\xi(V)}(\xi(V))$. Moreover, for absolutely continuous variable $\boldsymbol{V}$, U has a uniform density in $[0,1]$, that is $U \sim Uni[0,1]$. The final causal model is presented in Table 9. In summary, the binary choice IV model under monotonicity can be equivalently expressed by the causal model of Table 9.

Heckman and Vytlacil (2005) term the causal model in Table 9 as the Generalized Roy Model. The model delivers far more information than the IV assumption of the PO framework without cost of generality. This is possible because the structural model explicitly defines an unobserved variable $\boldsymbol{V}$, which, in turn, enables restating the monotonicity criteria in terms of the confounding variable. This renders the unobserved variable U, and the separability condition. Variable U is particularly useful as it entails a range of novel parameters

Table 10: Some Causal Parameters as Weighted Average the MTE

Causal Parameters	MTE Representation	Weights
$ATE = E(Y(t_1) - Y(t_0)$	$= \int_0^1 MTE(p) W^{ATE}(p) dp$	$W^{ATE}(p) = 1$
$TT = E(Y(t_1) - Y(t_0) \| T = t_1)$	$= \int_0^1 MTE(p) W^{TT}(p) dp$	$W^{TT}(p) = \frac{1 - F_P(p)}{\int_0^1 \left(1 - F_P(t)\right) dt}$
$TUT = E(Y(t_1) - Y(t_0) \| T = t_0)$	$= \int_0^1 \Delta^{MTE}(p) W^{TUT}(p) dp$	$W^{TUT}(p) = \frac{F_P(p)}{\int_0^1 \left(1 - F_P(t)\right) dt}$
$TSLS = \frac{Cov(Y,Z)}{Cov(T,Z)}$	$= \int_0^1 MTE(p) W^{TSLS}(p) dp$	$W^{TSLS}(p) = \frac{\int_p^1 \left(t - E(P)\right) dF_P(t)}{\int_0^1 \left(t - E(P)\right)^2 dF_P(t)}$
$LATE = \frac{E(Y \| Z = z_1) - E(Y \| Z = z_0)}{P(z_1) - P(z_0)}$	$= \int_{P(z_0)}^{P(z_1)} MTE(p) W^{LATE}(p) dp$	$W^{LATE}(p) = \frac{1}{P(z_1) - P(z_0)}$

that cannot be studied within the PO framework.

Heckman and Vytlacil (2005) define the marginal treatment effect $MTE(p) = E(Y(1) - Y(0)|U = p)$, which stands for the causal effect of T on Y for the share of the population that is indifferent among treatment statuses when U is equal to a value $p \in [0, 1]$. A primary contribution of Heckman and Vytlacil (2005) is to show that a range of causal parameters can be expressed as a weighted average of the MTE. A few of those are presented in Table 10.

It is worth remembering that the binary IV model described by the PO and SCM frameworks is equivalent. Although IV model share the same assumptions in both frameworks, the power of analysis generated by switching from the PO framework to SCM equations cannot be overstated. The MTE enables a far richer analysis of the IV properties. It also enables various extension of the original IV model. For instance, Brinch et al. (2017); Mogstad and Torgovitsky (2018) overcome the limitation of categorical IVs by extrapolating the MTE while Mogstad et al. (2018) use the MTE machinery to study sharp bounds of causal parameters under similar settings.

7 Examining the Mediation Model

The mediation model seeks to decompose the effect of the treatment on the outcome into sub-components associated to the treatment effects on intermediate variables. It enables the researcher to study the sources of the treatment effects. The mediation model goes beyond the study of "effect of a cause", as it examines the "causes of the effect" (Gelman and Imbens, 2013).

The mediation model stems from three main observed variables: a treatment T that causes a mediator M and an outcome Y that is caused by both T and M. The goal is to identify and quantify each of these causal relations. We suppress baseline variables X for sake of notational simplicity. All analysis can be understood as conditioned on X.

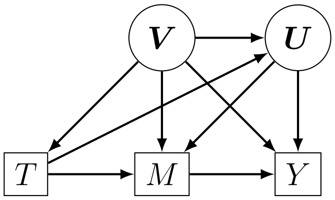
There are three counterfactuals of interest: $Y(t,m)$ is the counterfactual outcome when T, M are fixed at $(t,m) \in \mathit{supp}(T) \times \mathit{supp}(M)$, $M(t)$ is the counterfactual mediator when T is fixed at $t \in \mathit{supp}(T)$ and $Y(t, M(t'))$ which stands for the the counterfactual outcome when T is fixed at t and the mediator is fixed to the counterfactual variable $M(t')$. $Y(t, M(t'))$ enables us to disentangle the total effect of the treatment on the outcome into the direct and indirect effects. The average total effect for the case of binary treatment $T \in \{t_0, t_1\}$ is given by $E(Y(t_1) - Y(t_0))$. The direct effect, $DE(t) = E(Y(t_1, M(t)) - Y(t_0, M(t)))$, accounts for the effect of the treatment while keeping the mediator fixed. The indirect effect, $IE(t) = E(Y(t, M(t_1)) - Y(t, M(t_0)))$, is the effect of the treatment that operates through the mediator.

Table 11 presents a general mediation model. $\boldsymbol{V}$ denotes an unobserved confounding variable that causes T, M, Y. $\boldsymbol{U}$ is an unobserved post-treatment variable that causes M, Y. It can be understood as an unobserved mediator. Table 12 illustrates the DAGs corresponding to the direct, indirect and total effects using the hypothetical framework.

Table 11: General Mediation Model

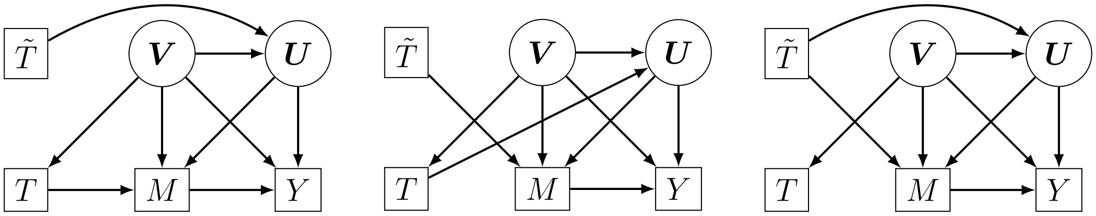
Causal Model

$\boldsymbol{V} = f_V(\epsilon_V)$
$T = f_T(\boldsymbol{V}, \epsilon_T)$
$\boldsymbol{U} = f_U(T, \boldsymbol{V}, \epsilon_U)$
$M = f_M(T, \boldsymbol{U}, \boldsymbol{V}, \epsilon_M)$
$Y = f_Y(M, \boldsymbol{U}, \boldsymbol{V}, \epsilon_Y)$



DAG

Table 12: Hypothetical Models for Direct, Indirect and Total Effects



Direct Effect **Indirect Effect** **Total Effect**

A large literature in PO invoke the Sequential Ignorability Assumption of Imai et al. (2010) to identify the direct and indirect effects f the mediation model. The assumption is displayed in (8).[13]

$$\big(Y(t',m), M(t),\big) \perp\!\!\!\perp T, \tag{8}$$

$$Y(t',m) \perp\!\!\!\perp M(t)|T, \tag{9}$$

It is easy to show that, under Sequential Ignorability (8)–(9), the distributions of counterfactual variables are identified by $P(Y(t,m)) = P(Y|T=t, M=m)$ and $P(M(t)) = P(M|T=t)$ and thereby the mediating causal effects can be expressed as:

$$DE(t) = \int \big(E(Y|T=t_1, M=m) - E(Y|T=t_0, M=m)\big) dF_{M|T=t}(m) \tag{10}$$

$$IE(t) = \int E(Y|T=t, M=m) dF_{M|T=t_1}(m) - \int E(Y|T=t, M=m) dF_{M|T=t_0}(m). \tag{11}$$

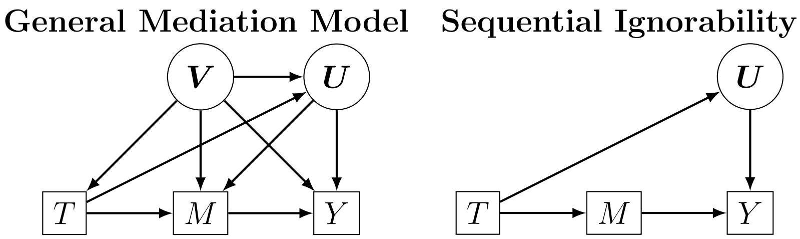
Assumptions (8)-(9) are rather strong. The first independence relationship is a matching-type assumption stating that the treatment is as good as random. The second relationship

[13] See Imai et al. (2011) for a comprehensive discussion on the Assumption (8).

assumes no confounders between Y and M. Assumption (8) would arise if the treatment T were randomly assigned. Assumption (9), on the other hand, does not arise even when T and M are a fully randomized. None of those assumptions are testable.

Table 13 compares the DAGs of the general causal model and the causal model induced by the sequential ignorability (8). The assumptions eliminates the unobserved variable $\boldsymbol{V}$ and implies that $\boldsymbol{U}$ does not cause M. Otherwise stated, sequential ignorability (8) assumes that there are no confounding variables and that given a treatment T, all unobserved mediators are independent of the observed mediator M.

Table 13: General Mediation Model



General Mediation Model **Sequential Ignorability**

Finding palatable assumptions that identify the mediation model is not a simple task. The advent of an instrumental variable helps. The mediator model can arise from a standard IV model. A typical empirical setting of an IV model consists of a treatment, an instrumental variable and multiple outcomes. An intermediate outcome can play the role of a mediator variable that causes a final outcome.

Table 16 presents the general mediation model with a categorical instrumental variable Z. The DAG displays the unobserved response-variable $\boldsymbol{S}$ that plays the role of a balancing score for the unobserved confounding variable $\boldsymbol{V}$. As mentioned, the inclusion of response-variable $\boldsymbol{S}$ does not incur in loss of generality. The sub-model generated by variables $Z, T, \boldsymbol{S}, \boldsymbol{V}, M$ consists on the same IV model displayed in Table 8. Thus the IV properties $Z \perp\!\!\!\perp (M(t), T(z))$ and $T \perp\!\!\!\perp M(t)|\boldsymbol{S}$ hold. The sub-model generated by variables $Z, T, \boldsymbol{S}, \boldsymbol{V}, Y$ is also an IV model where properties $Z \perp\!\!\!\perp (Y(t), T(z))$ and $T \perp\!\!\!\perp Y(t)|\boldsymbol{S}$ hold.

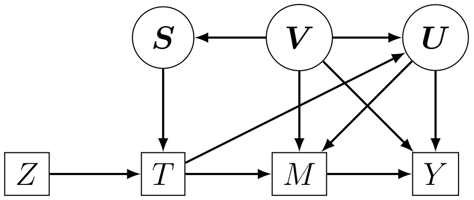
Indeed, M and Y can be interpreted as outcomes of a standard IV model and Z can be used to identify the causal effect of T on M and the total effect of T on Y. The main challenge to identify mediation effects if to identify the causal effect of M on Y.

Table 14: General Mediation Model with IV

Structural Equations

$$
\begin{aligned}
&\boldsymbol{V} = f_V(\epsilon_V)\\
&Z = f_Z(\epsilon_Z)\\
&\boldsymbol{S} = f_S(\boldsymbol{V}) = [T(z_1), ..., T(z_{N_Z})]'\\
&T = f_T(\boldsymbol{S}, Z)\\
&\boldsymbol{U} = f_U(T, \boldsymbol{V}, \epsilon_U)\\
&M = f_M(T, \boldsymbol{U}, \boldsymbol{V}, \epsilon_M)\\
&Y = f_Y(M, \boldsymbol{U}, \boldsymbol{V}, \epsilon_Y)
\end{aligned}
$$

General DAG with IV



Yamamoto (2013) proposes an interesting solution to identify mediation effects using instrumental variables. He investigates the mediation model with a binary treatment and a binary instrument. He uses the PO framework to offer an identification assumption that combines the sequential ignorability (8) with LATE analysis of Imbens and Angrist (1994). Let $T \in \{0, 1\}$ and $\boldsymbol{S} = [T(z_0), T(z_1)]'$ be the response variable. The response-type $\boldsymbol{s}_c = [0, 1]'$ denotes the compliers, who chose treatment assigned to z_1 and control otherwise. Yamamoto (2013) invokes an identification assumption termed local average causal mediation effects (LACME). In our notation, his assumption can be stated as:

$$(Y(t, m), M(t')) \perp\!\!\!\perp T | (\boldsymbol{S} = \boldsymbol{s}_c), \tag{12}$$

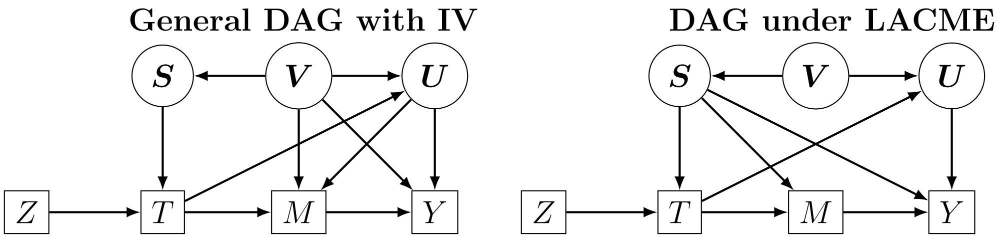
$$Y(t, m) \perp\!\!\!\perp M(t') | (T, \boldsymbol{S} = \boldsymbol{s}_c). \tag{13}$$

Assumptions (12)–(13) differ from (8)-(9) only by conditioning on $\boldsymbol{S}$. Assumption (12) is an extension of the IV-model property $Y(t) \perp\!\!\!\perp T | \boldsymbol{S}$. Yamamoto (2013) shows that the monotonicity assumption (7) identifies the direct and indirect mediation effects for compliers.

Yamamoto (2013) is best understood as a clever combination of the sequential ignorability assumption with the PO properties of the IV model. It is relatively easy to comprehend the rationale of merging two PO assumptions. It is much harder to figure out the causal relations implied by merging these assumptions. The task of assessing the causal model generated by assumptions (12)–(13) is not trivial. The difficulty in translating PO assumptions into causal relation between model variables impairs our ability to judge the plausibility of the assumptions.

Table 15 compares the general mediation model with IV with the mediation model induced by assumptions (12)–(13). The key property of the mediation model generated by (12)–(13) is that the response variable $\boldsymbol{S}$ plays the role of the unobserved confounding variable for the causal relation between Y and M. Otherwise stated, $\boldsymbol{S}$ subsumes the role of $\boldsymbol{V}$ and $\boldsymbol{U}$ and becomes a matching variable for the causal relation $M \to Y$. The assumptions prevent $\boldsymbol{V}$ to cause M, Y and imply that $\boldsymbol{S}$ directly causes M, Y. Although we can assume that $\boldsymbol{S}$ causes T without loss of generality, there is no justification for $\boldsymbol{S}$ to cause M and Y.

Table 15: General Mediation Model under LACME Assumption



The second DAG in Table 15 illustrates the danger of seeking identification strategies using the PO framework. Namely, the PO framework relies on independence assumptions instead of causal relations. The framework can induce the researcher to generate assumptions that are statistically sound but often hard to be interpreted. These assumptions can easily induce a causal model that is seldom justifiable.

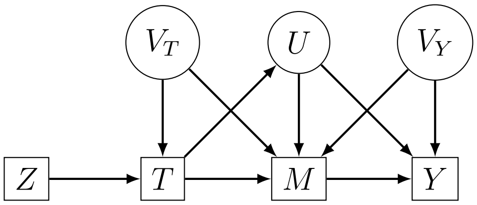
Dippel et al. (2020) investigate whether it is possible to identify mediation effects amongst outcomes in a standard IV model without revoking the endogeneity of the treatment with respect to intermediate and final outcomes. They show that mediation effects can be identified when the IV model is partially confounded, i.e. the unobserved confounding variables that cause the treatment and the intermediate outcome are independent of the confounders that cause the intermediate and final outcomes. Their model is displayed in Table **??**.

Table 16: Partially Confounded Mediation Model

Structural Equations

$Z = f_Z(\epsilon_Z)$
$V_T = f_{V_T}(\epsilon_{V_T})$
$V_Y = f_{V_Y}(\epsilon_{V_Y})$
$Z = f_Z(\epsilon_Z)$
$T = f_T(Z, V_T)$
$U = f_U(T, \epsilon_U)$
$M = f_M(T, V_T, V_Y, \epsilon_M)$
$Y = f_Y(M, U, V_Y, \epsilon_Y)$

DAG



The partially confounded model preserves the endogenous properties of the IV model. That is to say that T remains endogenous with respect to Y, M and M is endogenous with respect to Y. Standard properties of the instrumental variable imply that $Z \perp\!\!\!\perp (Y(t), M(t))$ holds. The novel condition generated by the model is $Z \perp\!\!\!\perp Y(m)|T$. In other words, the instrumental variable Z can be used to identify the causal effect of M on Y when conditioning on T. While $Z \perp\!\!\!\perp Y(m)|T$ holds, $Z \perp\!\!\!\perp Y(m)$ does not. The relationship arises from the fact that T is caused by both Z and V_T. The unobserved confounder V_T and the observed instrument Z are unconditionally statistically independent. However, conditioning on T induces a correlation between Z and V_T, though V_T causes M and does not (directly) cause Y. Therefore, conditioned on T, Z affects M (via V_T) and does not affect Y by any channel other than M.

The partially confounded splits the unobserved counfounding variable V into two statistically independent variables: V_T that causes T, M; and V_Y that causes M, Y. The model does not broadly applies to IV settings with multiple outcomes. Dippel et al. (2020) clarify the type of empirical settings that justify the use of the model and those which do not.

8 Beyond Mediation

[Not Finished Yet]

9 Conclusion

This paper presents the key concepts regarding econometric causality and discusses the benefits and limitations of several causal frameworks. This paper clarifies that causal inference is based on three distinguish tasks: the adoption of a causal model, the identification analysis and the estimation of causal parameters. It revisits the seminal ideas of Frisch (1930); Haavelmo (1944) who describe a causal model by a system of autonomous equations. Counterfactuals are generated by *fixing*, which is a causal operator that assigns values to the inputs of structural equations associated to the variable. Fixing is a primary concept in causal analysis as it embodies the concept of causal direction.

Fixing is ill-defined in standard probability theory, which lacks directionality. This fact motivates the existence of several causal frameworks that combine causal concepts with probability and statistics. This paper uses popular causal models in the literature of policy evaluation to compare the benefits and drawbacks of causal frameworks.

A popular framework is the language of potential outcomes (PO), also known as the Rubin-Holland model. The framework suppresses the structural equations of the underlying causal model that generates the distribution of observed variables. Instead, the framework simply describes the properties of the causal model in terms of statistical independence relationships among the counterfactuals of observed variables. The trade-off between simplicity and accuracy poses a few limitations on causal analysis. The lack of structural equations harms the interpretation of the underlying causal models that justifies the identification assumptions commonly invoked in the PO framework. Lack of structural equations can generated misguided conclusions regarding the underlying causal model. We illustrate this fact using the matching assumption.

Another drawback of the framework is that it does not assess unobserved variables. Counterfactual outcomes are only based on observed variables. This fact poses a great limitation in advancing the understanding of a causal model. We clarify this fact using the instrumental variable model. We also show that recovering the underlying causal model based on a set of independence relationship is a complex task. The PO framework invokes independence relationships that often obscures strong causal assumptions that would be clearly understood if the causal model were described by structural equations.

The hypothetical framework copes with the concept of fixing by generating a hypothetical model where the treatment variable has de desired property of being exogenous. The framework clearly formalises the key concepts of causal inference. It makes probability converse with causality as the causal operator of fixing becomes the statistical conditioning in the hypothetical model. It also makes a clear distinction between the definition of causal parameters and its identification from observed data.

The gain in clarity of the hypothetical framework comes at a cost of additional structure. The framework not only invokes the structural equations that describe a causal model, but generates an additional (hypothetical) model that stems from the empirical causal model that generates the observed data. The framework requires rules that connect hypothetical an empirical models. We illustrate that this additional structure is justified when examining causal models with a complex causal relation among observed and unobserved variables.

The do-calculus requires definition of new graphical/statistical rules outside of standard probability theory. These are not needed when the hypothetical model is used, which leads to a simpler and less cumbersome approach. We illustrate this fact e limitations of the do-calculus in analyzing the instrumental variable model. It is identified under standard conditions. It is not identified using the do-calculus.

Pearl's framework cannot accommodate the fundamentally non-recursive simultaneous equations model. The hypothetical model readily accommodates an analysis of causality in the simultaneous equations model. The framework of simultaneous equations is fundamentally non-recursive and falls outside of the framework of Bayesian causal nets and DAGs. The rigorous definition of causality in a variety of models including the simultaneous equations framework and the identification of causal parameters, are central and enduring contributions of Haavelmo (1944).

References

Brinch, C. N., M. Mogstad, and M. Wiswall (2017). Beyond late with a discrete instrument. *Journal of Political Economy 125*(4), 985–1039.

Dawid, A. P. (1976). Properties of diagnostic data distributions. *Biometrics 32*(3), 647–658.

Dippel, C., R. Gold, S. Heblich, and R. Pinto (2020). Mediation analysis in iv settings with a single instrument. *Unpublished Manuscript*.

Frisch, R. (1930). A dynamic approach to economic theory: Lectures by Ragnar Frisch at Yale University. Lectures at Yale University beginning September, 1930. Mimeographed, 246 pp.

Frisch, R. (1930, published 2010). In O. Bjerkholt and D. Qin (Eds.), *A Dynamic Approach to Economic Theory: The Yale Lectures of Ragnar Frisch, 1930*. New York, New York: Routledge.

Frisch, R. (1938). Autonomy of economic relations: Statistical versus theoretical relations in economic macrodynamics. Paper given at League of Nations. Reprinted in D.F. Hendry and M.S. Morgan (1995), *The Foundations of Econometric Analysis*, Cambridge University Press.

Gelman, A. and G. Imbens (2013, November). Why ask why? forward causal inference and reverse causal questions.

Glamour, C., R. Scheines, and P. Spirtes (2014). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press.

Greenland, S., J. Pearl, and J. Robins (1999). Causal diagrams for epidemiologic research. *Epidemiology 10 1*, 37–48.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica 12*(Supplement), iii–vi and 1–115.

Heckman, J. (2006, June). The scientific model of causality. *Sociological Methodology 10*.

Heckman, J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *65*, 261–294.

Heckman, J. and R. Pinto (2018). Unordered monotonicity. *Econometrica 86*, 1–35.

Heckman, J. J. (2005, August). The scientific model of causality. *Sociological Methodology 35*(1), 1–97.

Heckman, J. J. (2008). The principles underlying evaluation estimators with an application to matching. *Annales d'Economie et de Statistiques 91–92*, 9–73.

Heckman, J. J. and S. Navarro (2004, February). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics 86*(1), 30–57.

Heckman, J. J. and R. Pinto (2012). Causal analysis after Haavelmo: Definitions and a unified analysis of identification. Unpublished manuscript, University of Chicago.

Heckman, J. J. and R. Pinto (2015a). Causal analysis after Haavelmo. *Econometric Theory 31*(1), 115–151.

Heckman, J. J. and R. Pinto (2015b). Causal analysis after Haavelmo. *Econometric Theory 31*(1), 115–151.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Holland, P. W. (1986, December). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Hoyer, P. O., D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf (2009). Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21*, pp. 689–696. Curran Associates, Inc.

Hurwicz, L. (1962). On the structural form of interdependent systems. In E. Nagel, P. Suppes, and A. Tarski (Eds.), *Logic, Methodology and Philosophy of Science*, pp. 232–239. Stanford University Press.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Imai, K., L. Keele, and T. Yamamoto (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review 105*, 765–789.

Imbens, G. (2019, July). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Kiiveri, H., T. P. Speed, and J. B. Carlin (1984). Recursive causal models. *Journal of the Australian Mathematical Society (Series A)*, 30–52.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford, UK: Clarendon Press.

Lopez-Paz, D., R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou (2017). Discovering causal signals in images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6979–6989.

Marshall, A. (1890). *Principles of Economics.* New York: Macmillan and Company.

Mogstad, M., A. Santos, and A. Torgovitsky (2018). Using instrumental variables for inference about policy relevant treatment effects. *Econometrica 86*(5), 1589–1619.

Mogstad, M. and A. Torgovitsky (2018). Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics 2*, 577–613.

Pearl, J. (2009a). *Causality: Models, Reasoning and Inference* (Second ed.). Cambridge University Press.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference* (2nd ed.). New York: Cambridge University Press.

Pearl, J. (2009c). Myth, confusion, and science in causal analysis. *Technical Report, UCLA, Department of Statistics*.

Peters, J., D. Jazzing, and B. Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms.* Cambridge, MA: MIT Press.

Rosenbaum, P. R. and D. B. Rubin (1983, April). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. (2008). Authors reply (to ian shriers letter to the editor). *Statistics in Medicine 27*, 27412742.

Shrier, I. (2008). Letter to the editor. *Statistics in Medicine 27*, 27402741.

Spirtes, P., C. N. Glymour, and R. Scheines (2000). *Causation, Prediction and Search* (2 ed.). Cambridge, MA: MIT Press.

Vytlacil, E. J. (2002, January). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Yamamoto, T. (2013). Identification and estimation of causal mediation effects with treatment noncompliance. *Unpublished Manuscript, MIT Department of Political Science*.

Yule, G. U. (1895). On the correlation of total pauperism with proportion of out-relief. *The Economic Journal 5*(20), 603–611.

Online Appendix

A Definitions Details