TECHNICAL REPORT
R-51
December, 2019

Causal Inference and Data-Fusion in Econometrics*

Paul Hünermund† Elias Bareinboim‡

This version: 18 December 2019
First version: 17 December 2019

Learning about cause and effect is arguably the main goal in applied econometrics. In practice, the validity of these causal inferences is contingent on a number of critical assumptions regarding the type of data that has been collected and the substantive knowledge that is available about the phenomenon under investigation. For instance, unobserved confounding factors threaten the internal validity of estimates, data availability is often limited to non-random, selection-biased samples, causal effects need to be learned from surrogate experiments with imperfect compliance, and causal knowledge has to be extrapolated across structurally heterogeneous populations. A powerful causal inference framework is required in order to tackle all of these challenges, which plague essentially any data analysis to varying degrees. Building on the structural approach to causality introduced by Haavelmo (1943) and the graph-theoretic framework proposed by Pearl (1995), the artificial intelligence (AI) literature has developed a wide array of techniques for causal learning that allow to leverage information from various imperfect, heterogeneous, and biased data sources (Bareinboim and Pearl, 2016). In this paper, we discuss recent advances made in this literature that have the potential to contribute to econometric methodology along three broad dimensions. First, they provide a unified and comprehensive framework for causal inference, in which the above-mentioned problems can be addressed in full generality. Second, due to their origin in AI, they come together with sound, efficient, and complete (to be formally defined) algorithmic criteria for automatization of the corresponding identification task. And third, because of the nonparametric description of structural models that graph-theoretic approaches build on, they combine the strengths of both structural econometrics as well as the potential outcomes framework, and thus offer a perfect middle ground between these two competing literature streams.

Key words: Causal Inference; Econometrics; Directed Acylic Graphs; Data Science; Machine Learning; Artificial Intelligence
JEL classification: C01, C30, C50

*We are grateful to Carlos Cinelli, Juan Correa, Beyers Louw, Guido Imbens, Judea Pearl, participants at EEA-ESEM 2019 and seminar participants at Maastricht University for helpful comments and suggestions.

†Maastricht University, School of Business and Economics. Tongersestraat 53, 6211LM Maastricht, The Netherlands. Email: p.hunermund@maastrichtuniversity.nl

‡Columbia University, Department of Computer Science, 500 W 120th Street, New York, NY, 10027. Email: eb@cs.columbia.edu

1. INTRODUCTION

Causal inference is arguably one of the most important goals in applied econometric work. Policy-makers, legislators, and managers need to be able to forecast the likely impact of their actions in order to make informed decisions. Constructing causal knowledge by uncovering quantitative relationships in statistical data is the goal of econometrics since the beginning of the discipline (Frisch, 1933). After a steep decline of interest in the topic during the postwar period (Hoover, 2004), causal inference has recently been receiving growing attention again and was brought back to the forefront of the methodological debate by the emergence of the potential outcomes framework (Rubin, 1974; Imbens and Rubin, 2015; Imbens, 2019) and advances in structural econometrics (Heckman and Vytlacil, 2007; Matzkin, 2013; Lewbel, 2019).

Woodward (2003) defines causal knowledge as "knowledge that is useful for a very specific kind of prediction problem: the problem an actor faces when she must predict what would happen if she or some other agent were to act in a certain way [...]".[1] This association of causation with control in a stimulus-response-type relationship is likewise foundational for econometric methodology. According to Strotz and Wold (1960), "z is a cause of y if [...] it is or 'would be' possible by *controlling* z indirectly to control y, at least stochastically" (p. 418; emphasis in original).

Although implicit in earlier treatments in the field (e.g., Haavelmo, 1943), Strotz and Wold (1960) were the first to express actions and control of variables as *"wiping out"* of structural equations in an economic system (Pearl, 2009, p. 32). To illustrate this idea, consider the two-equation model

$$z = f_z(w, u_z), \tag{1.1}$$
$$y = f_y(z, w, u_y), \tag{1.2}$$

in which Y might represent earnings obtained in the labor market, Z the years of education an individual received, W other relevant socio-economic variables,

[1] Woodward continues: "[...] on the basis of observations of situations in which she or the other agent have not (yet) acted" (p. 32).

and U unobserved background factors.[2] Since W enters in both equations of the system, it creates variation between Z and Y that is not due to a causal influence of schooling on earnings. Therefore, in order to predict how Y reacts to induced changes in Z, the causal mechanism that naturally determines schooling needs to be replaced to avoid non-causal (spurious) sources of variation. In this particular example, the values that Z attains must be uncoupled from W, so that Z can freely influence Y. Symbolically, this is achieved by deleting $f_z(\cdot)$ from the model and fixing Z at a constant value z_0. The modified system thus becomes:

$$z = z_0 \tag{1.1'}$$
$$y = f_y(z_0, w, u_y). \tag{1.2'}$$

Subsequently, the quantitative impact on Y of the intervention can be traced via equation (1.2') in order to pin down Z's causal effect.

The notion of "wiping out" equations, as proposed by Strotz and Wold, eventually received central status and a formal treatment in a specific language with the definition of the *do*-operator (Pearl, 1995). Consider the task of predicting the post-intervention distribution of a random variable Y that is the result of a manipulation of X. In mathematical notation, this can be written as $Q = P(Y = y|do(X = x))$, where $do(X = x)$ denotes the replacements of whatever mechanisms were there for X, f_x, with a constant x.

In practical applications, however, simulating interventions to such a degree of granularity would either require knowledge about the precise form of the system's underlying causal mechanisms or the possibility to physically manipulate X in a controlled experiment. Both are luxuries that policy forecasters very rarely have available. In many economic settings, experiments can be difficult to implement. Likewise, exactly knowing the structural mechanisms that truly govern the data generating process is hard in the social sciences, where often only qualitative knowledge about causal relationships is available.[3] This means that the counterfactual distribution $P(y|do(x))$ will be, in general, not immediately estimable. In practice,

[2] We follow the usual notation of denoting random variables by uppercase and their realized values by lowercase letters.

[3] Quoting prominent physicist Murray Gell-Mann: *"Imagine how hard physics would be if electrons could think."* (cited in Page, 1999).

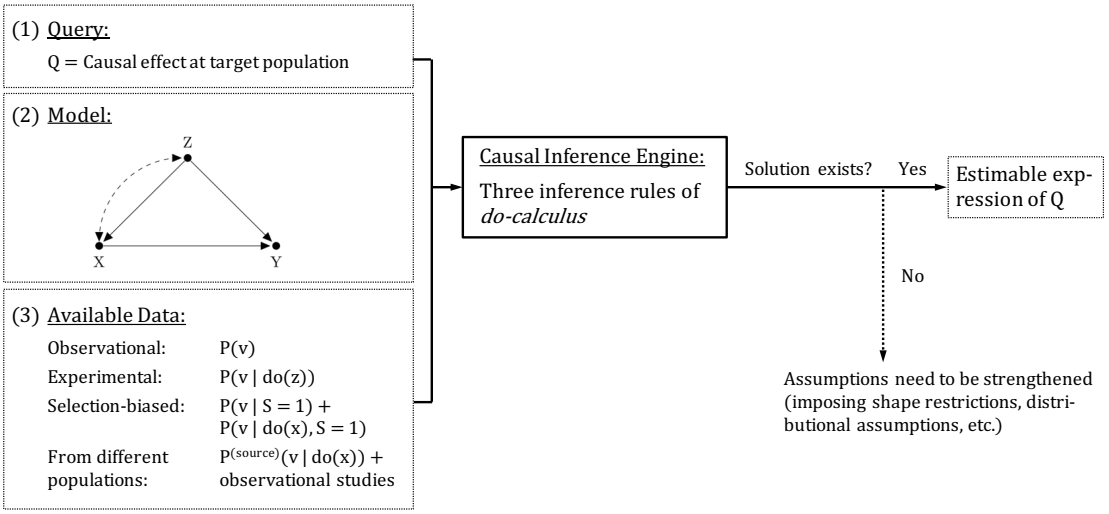


Figure 1: *Schematic illustration of the data fusion process. The causal inference engine provided by do-calculus takes three inputs: (1) a causal effect query Q, (2) a model G, and (3) the type of data, $P(v|\cdot)$, that is available. It is guaranteed to return a transformation of Q, based on G, that is estimable with the available data, whenever such a solution exists.*

instead, Q will first need to be transformed into a standard probability object that only comprises ex-post observable quantities before estimation can proceed. The symbolic language that warrants such kinds of syntactic transformations is called *do-calculus* (Pearl, 1995).

Do-calculus is a causal inference engine that takes three inputs:

1. A causal quantity Q, which is the query the researchers want to answer;

2. A model G that encodes the qualitative understanding about the structural dependencies between the economic variables under study;

3. A collection of datasets $P(v|\cdot)$ that are available to the analyst, including observational, experimental, from selection-biased samples, from different populations, and so on.

Based on these inputs, do-calculus constitutes three inference rules for transforming probabilistic sentences involving do-expressions into equivalent expressions. The inferential goal is then to re-express the causal quantity (1 above) through the repeated application of the rules of the calculus, licensed by the assumptions

in G (2 above), into expressions that are estimable by the observable probability distributions $P(v|\cdot)$ (3 above). Figure 1 provides a schematic illustration of this process.

Do-calculus complements standard tools in econometrics in two important ways. First, it builds on a mathematical formalism borrowed from graph theory, which describes causal models as a set of nodes in a network, connected by directed edges (so-called *Directed Acyclic Graphs*; Pearl, 1995). An advantage of such a description is that it does not rely on any functional-form restrictions imposed on the relationships between economic variables. Therefore, the approach provides a formal treatment of nonparametric causal inference in full generality. Second, as a subfield of artificial intelligence, the literature on graph-theoretic treatments of causality has developed algorithmic solutions for a wide variety of causal inference problems arising in applied work. These algorithms are able to carry out the syntactic transformation described above – mapping a query to the available data through the model's assumptions – fully automatically. From do-calculus, the algorithms furthermore inherit the property of *soundness* and *completeness* (Tian and Pearl, 2002a; Shpitser and Pearl, 2006b; Huang and Valtorta, 2006; Bareinboim and Pearl, 2012c; Lee et al., 2019). This means that the approach is guaranteed to return a correct solution whenever one exists. Conversely, and remarkably, if the algorithm fails to provide an answer to a causal query, it is assured that no such answer will be obtainable unless the assumptions imposed on the model are strengthened. In other words, for the class of models in which these algorithmic conditions are applicable, the identification problem is fully solved (Pearl, 2013; Bareinboim and Pearl, 2016).

The development of do-calculus gave the literature on causal inference within the field of artificial intelligence a tremendous boost, and many significant advances have been made since Pearl (2000) published his seminal contribution. The aim of this paper is to discuss these more recent developments and show how do-calculus can be utilized to solve many recurrent problems in applied econometric work. The three main topics we cover are: dealing with confounding bias (Section 3), recovering from sample selection bias (Section 4), and extrapolation of causal claims across heterogeneous settings (Section 5), which we describe in turn next.

Confounding bias (Section 3). In most applied settings, the post-interventional

distribution of Y following a manipulation of X, $P(y|do(x))$, does not coincide with the conditional distribution $P(y|x)$ – a distinction that has been popularized through the mantra *"correlation does not imply causation"* (List, 2011). This is due to confounding influence factors, which can render two variables stochastically dependent irrespective of any causal relationship between them. The inference rules of do-calculus were developed precisely to neutralize confounding bias. Syntactically, this task amounts to transforming $P(y|do(x))$ into an equivalent expression, generally different from $P(y|x)$, that is nonetheless estimable from the available data. If a reduction containing standard probability objects can be reached, the confounding problem is solvable with the help of observational data alone. Additionally, sometimes the analyst is able to experimentally manipulate a third variable Z, which is itself causally related to the treatment of interest. In such settings (one example is the classic *encouragement design*; Duflo et al., 2008), the identification problem can be relaxed, since estimable syntactic transformations of $P(y|do(x))$ reached by do-calculus can now also involve $do(z)$-distributions.

Sample selection bias (Section 4). A common threat to to the validity of inferences in practice is sample selection bias, which occurs if the analyst is only able to observe information for members of the population that possess specific characteristics or fulfill certain requirements (e.g., market wages are only observable if individuals are employed; Heckman, 1979). Selection-biased data aggravate the identification problem, as $P(y|do(x))$ needs to be transformed into an expression solely comprised of probabilities from a non-random sample (inclusion in the selected sample is usually denoted by an indicator S, which implies that only probabilities conditional on $S = 1$ are observable). The inference rules of do-calculus provide a principled and complete solution for carrying out this task.

Extrapolation of causal claims across settings (Section 5). While confounding and selection biases threaten the internal validity of estimates, another important topic in econometric practice is external validity, or generalizability of causal inferences across settings and populations. Causal knowledge is usually acquired in a specific population (e.g., for probands in a laboratory setting), but needs to be brought to productive use in other domains in order to be most valuable. What permits such a transportation of causal knowledge across settings, however, if the underlying populations differ structurally in important ways? Do-calculus provides

an answer to this question. Its inference rules can be applied in order to transform a causal query in a target population into an expression that is estimable with the help of information stemming from a source population. In its more general form, transportability theory encompasses the problem of combining causal knowledge from several, possibly heterogeneous source domains (a strategy generically known under the rubric of *"meta-analysis"*). Thereby, do-calculus opens up entirely new possibilities for leveraging results from a whole body of empirical literature in order to address policy questions arising in yet under-researched contexts.

These three thematic areas are indeed quite diverse and encompass several seemingly unrelated empirical challenges, yet they share a common structure. Data, which are created in various different ways – e.g., from observational or experimental studies, from non-random sampling, or from heterogeneous underlying populations – are combined in order to answer a causal query of interest. For this strategy of *"data fusion"* (see Figure 1) to be viable, the analyst needs to be equipped with a model of the underlying economic context under study and a powerful inference framework that license this kind of information transfer (Bareinboim and Pearl, 2016). In the remainder of the paper, we will describe such a causal modeling and inference framework in detail.

2. PRELIMINARIES: STRUCTURAL CAUSAL MODELS, CAUSAL GRAPHS, AND INTERVENTIONS

This section introduces structural causal models (SCM) and directed acyclic graphs, which form the basis for all the data fusion techniques discussed in this paper.[4] We follow the standard notation in the literature, as summarized in Pearl (2009), and define an SCM as:

Definition 2.1. *(Structural causal model; Pearl, 2009) A structural causal model is a 4-tuple* $M = \langle U, V, F, P(u) \rangle$ *where*

[4]Structural causal models are nonparametric versions of structural equation models (SEM). We purposefully will use the term SCM to avoid confusion with the vast literature on SEM that traditionally assumes parametric or even linear functional forms, and many times has confounded the inherent causal nature of structural models.

1. *U is a set of background variables (also called exogenous) that are determined by factors outside the model.*

2. *$V = \{V_1, \ldots, V_n\}$ is a set of endogenous variables that are determined by variables in the model, viz. variables in $U \cup V$.*

3. *F is a set of functions $\{f_1, \ldots, f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup PA_i$ to V_i, where $U_i \subseteq U$ and $PA_i \subseteq V \setminus V_i$ and the entire set of F forms a mapping from U to V. In other words, f_i assigns a value to the corresponding $V_i \in V$, $v_i \leftarrow f_i(pa_i, u_i)$, for $i = 1, \ldots, n$.*

4. *$P(u)$ is a probability function defined over the domain of U.*

An SCM constitutes a set of (exogenous) background factors, U, which are determined outside of the model and taken as given. Their associated (joint) probability distribution, $P(u)$, creates variation in the endogenous variables, V, whose source remains not further specified. Inside the model, the value of an endogenous variable V_i is determined by a causal process, $v_i \leftarrow f_i(pa_i, u_i)$, that maps the background factors U_i and a set of endogenous variables PA_i (so-called *parents*) into V_i. These causal processes – or mechanisms – are assumed to be invariant unless explicitly intervened on (see Section 2.1). Together with the background factors, they represent the data generating process (DGP) according to which *nature* assigns values to the (endogenous) variables under study.[5]

To emphasize the interpretation of f_i's as stimulus-response relationships, and in contrast to the standard notation in econometrics, the computer science literature uses assignment operators "$\leftarrow$" instead of equality signs (similar to the syntax of programming languages). Assignments change meaning under solution-preserving algebraic operations; i.e., $y \leftarrow ax \neq x \leftarrow y/a$ (Pearl, 2009, p. 27). This highlights the asymmetric nature of elementary causal mechanisms (Woodward, 2003; Cartwright, 2007), in the sense that if x is a cause of y, it cannot be the case that y is also a cause of x in the same instance of time.

In a fully specified SCM, $\langle U, V, F, P(u) \rangle$, any counterfactual quantity is well-defined and immediately computable from the model. In many social science

[5]Background factors correspond to what is often referred to as "error terms" in classical econometrics. However, we deliberately avoid this terminology to emphasize that the U_i's in an SCM have a causal interpretation, in contrast to the purely statistical notion of a prediction error or deviation from the conditional mean.

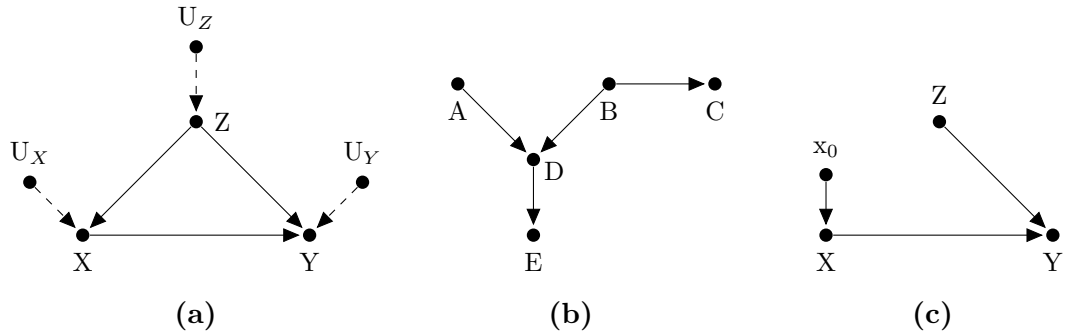


Figure 2: *(a) Directed acyclic graph corresponding to SCM in equation (2.1) with background variables U_i explicitly depicted. (b) Graphical illustration of d-separation. D acts as a collider that opens up the path between A and C, whereas B blocks it. (c) Post-intervention graph of (a) for $do(X = x_0)$.*

contexts, however, precise knowledge of the functional relationships, f_i, and the distribution of the exogenous variables, $P(u)$, governing the DGP, is not available. In the following, we will thus advocate for an approach that fully embraces and acknowledges the existence of the underlying causal mechanisms and exogenous variations in the system (i.e., that nature follows a structural causal model), but which will be much less committal regarding what the analyst needs to know about this reality in order to be able to make causal inferences. In particular, the inferences entailed by our analysis will rely on the graphical representation of the underlying structural system, which is a parsimonious way of encoding a minimalistic set of assumptions of the system necessary for identifiability.

Every SCM M defines a directed graph $G(M)$ (or G, for simplicity). Nodes in G correspond to endogenous variables in V, and directed edges point from the set of parent nodes PA_i towards V_i.[6] An example is given in Figure 2a, which refers to the following SCM:

$$\begin{aligned} z &\leftarrow f_z(u_z), \\ x &\leftarrow f_x(z, u_x), \\ y &\leftarrow f_y(x, z, u_y). \end{aligned} \tag{2.1}$$

Note that Z appears as an argument in the structural function of X, f_x. Accord-

[6] As it is standard in the field, we will use the notation of kinship relations (parents, children, ancestors, descendants, etc.) to describe the relative position of nodes in directed graphs. For instance, for the graph in Figure 2b we can read that B is a parent of D, since $B \to D$, and A is an ancestor of E, since $A \to D \to E$.

ingly, Z is a parent of X and an arrow should be added pointing from node Z to X. Similarly, X and Z appear in f_y, which means that the causal graph contains arrows from these variables to Y. For the sake of readability, we will usually not depict the U_i's explicitly, as in Figure 2a, but will omit them from the graph, whenever they affect only one endogenous variable. Background factors are by default assumed to be independent, unless otherwise specified. The presence of common unobserved parent nodes, which render two variables stochastically dependent, is represented by dashed bidirected arcs in the graph (see, e.g., Figure 3a).[7]

The graph in Figure 2a contains no sequences of edges that point from a variable back to itself (i.e., there are no feedback loops). This property is called *acyclicity*. Throughout the paper, we restrict attention to structural causal models that can be represented by directed acyclic graphs (DAG). This class of models, which economists often refer to as *recursive*, is of central importance in causal inference, because it describes economic systems in which individual causal mechanisms have a direct and autonomous stimulus-response interpretation, in accordance with the notion of causality put forward by Strotz and Wold (1960; see also Woodward, 2003; Cartwright, 2007; Pearl, 2009).[8]

Working with the graphical representation of M entails a deliberate choice by the analyst to refrain from parametric and functional form assumptions, since the shape of the f_i's and the distribution of background factors U_i remain unspecified throughout the analysis. Another way of thinking about the causal graph is that it represents the equivalence class of all structural functions sharing the same scope. Consequently, graphical models are fully nonparametric in nature. This constitutes an important distinction relative to the structural econometrics literature, which often assumes specific parametric error distributions (such as the normal or logistic distribution) or imposes shape restrictions on functions (such as separability, monotonicity, or differentiability) in order to establish identification (Heckman and

[7] A dashed bidirected arc $X \dashleftarrow\dashrightarrow Y$ serves as a shortcut notation for $X \dashleftarrow U \dashrightarrow Y$, if the set of common causes U is unobservable to the analyst.

[8] It is important to note, however, that the axioms of structural counterfactuals in SCMs (Pearl, 2009, ch. 7) also hold in nonrecursive models, see Halpern (2000). For an introduction into the literature on *cyclic directed graphs*, the interested reader is referred to Spirtes et al. (2001, ch. 12) and Pearl (2009, ch. 3.6). Appendix A.1 provides a brief discussion of the differences arising with respect to the conceptual interpretations of causality in recursive versus nonrecursive economic systems.

Vytlacil, 2007; Matzkin, 2007, 2013). In certain applications, these distributional and functional-form assumptions might be licensed by economic theory (Matzkin, 2013). If they are not, however, we concur with Manski (2003) that it is a more robust research approach to start with the most flexible model possible and only resort to parametric and functional form assumptions once the explanatory power of nonparametric approaches has been exhausted. In line with this philosophy, the techniques we present in the following explore ways to identify causal effects from data when only knowledge about the graph G is available.[9]

One key feature of DAGs is that they are falsifiable through testable implications over the observed distributions, including conditional independence relationships between variables in the model.[10] We define below such notion.

Definition 2.2. *(D-separation; Pearl, 1988) A set Z of nodes is said to block a path p if either*

1. *p contains at least one arrow-emitting node that is in Z,*
2. *p contains at least one collision node that is outside Z and has no descendant in Z.*

If Z blocks all *paths from set X to set Y, it is said to "d-separate X and Y", and then it can be shown that variables X and Y are independent given Z, written $X \perp\!\!\!\perp Y|Z$.*[11]

Conditional independence licensed by d-separation (d stands for "directional") holds for any distribution $P(v)$ over the variables in the model, which is compatible with the causal assumptions encoded in the graph. Remarkably, this is true

[9] This is indeed the case unless otherwise specified, and should constitute the starting point of any analysis. Whenever nonparametric identification is not entailed by the available knowledge, the causal graph can still be used as a computation device to analyze identifiability of entire classes of structural models. For instance, the most general identification results of structural coefficients if the system is linear are within the graphical perspective. For a survey and the latest results, please refer to Pearl (2009, Ch. 5) and Chen et al. (2017).

[10] Historically, DAGs were first introduced in the context of the AI literature in the early 1980's as efficient encoders of conditional independence constraints, and as a basis that avoided the explicit enumeration of exponentially many of such constraints. This encoding lead to a huge literature on efficient algorithms for computing and updating probabilistic relationships in data-intensive applications (Pearl, 1988).

[11] See Verma and Pearl (1988). A path refers to any consecutive sequence of edges in a graph. The orientation of edges plays no role. If the direction of edges is taken into account, one speaks of a *directed* or *causal path*: $A \to B \to C$.

regardless of the parametrization of the arrows. An example is given in Figure 2b, where the path $A \to D \leftarrow B \to C$ is blocked by $Z = \{B\}$, since B emits arrows on that path. Consequently, we can infer the conditional independencies $A \perp\!\!\!\perp C|B$ and $D \perp\!\!\!\perp C|B$. In fact, A and C are independent conditional on the empty set $\{\emptyset\}$ too. D acts as a so-called *collider* node, because of two arrows pointing into it. Therefore, according to the second condition of Definition 2.2, it blocks the path between A and C without any conditioning. Conversely, when conditioned on, a collider would open up a path that has been previously blocked; thus, $A \not\perp\!\!\!\perp C|D$. The same holds for descendants of colliders such as E in Figure 2b, yielding $A \not\perp\!\!\!\perp C|E$.

D-separation allows to systematically read off the conditional independencies implied by the structural model from the graph. As mentioned earlier, this method provides the analyst with a set of testable implications that can be benchmarked with the available data. The full list of conditional independence relations (with separator sets up to cardinality one) implied by the graph in Figure 2b is given by:

$$\begin{array}{lllll} A \perp\!\!\!\perp B; & A \perp\!\!\!\perp C; & A \perp\!\!\!\perp E|D; & B \perp\!\!\!\perp E|D; & \\ & C \perp\!\!\!\perp D|B; & C \perp\!\!\!\perp E|D; & C \perp\!\!\!\perp E|B. & \end{array} \tag{2.2}$$

These independence relations can easily be tested through statistical hypothesis testing, and if rejected, the hypothesized model can be discarded too. An advantage of such local tests, compared to global goodness-of-fit measures, for example, is that they indicate exactly where the model is incompatible with the observed data. Thus, the analyst can rely on concrete clues about where to improve the model, which facilitates an iterative process of model building.

Conditional independence assumptions constitute a main building block of causal inference – a theme that we will further pursue in Section 3. With the help of the d-separation criterion, their validity can be determined simply based on the topology of the graph. For this reason, DAGs constitute a valuable complement to the treatment effects literature, in which independence assumptions for counterfactuals, such as *ignorability*, are usually invoked without a reference to an explicit model (Imbens and Rubin, 2015). A shortcoming of such an approach is that the analyst has little to no guidance for scrutinizing the plausibility of crucial

identifying assumptions on which the whole analysis hinges on. DAGs facilitate this task significantly; in particular, because finding d-separation relations, even in complex graphs, can easily be automatized (Textor and Liśkiewicz, 2011; Textor et al., 2011). Moreover, using causal graphs increases the transparency of research designs compared to purely verbal justifications of identification strategies and thereby improves the communication between researchers and facilitates cumulative research efforts, as exemplified in future sections.

2.1. Interventions in structural causal models

The aim of causal inference is to predict the effects of interventions, such as those resulting from policy actions, social programs, and management initiatives (Woodward, 2003). Based on early ideas from the econometrics literature (Haavelmo, 1943; Strotz and Wold, 1960; Pearl, 2015b), interventions in structural causal models are carried out by deleting individual functions, f_i, from the model and fixing their left-hand side variables at a constant value.[12] As alluded earlier, this action is denoted by a mathematical operator called $do(\cdot)$. For example, in model M of equation 2.1 (with the respective graph shown in Figure 2a), the action $do(X = x_0)$ results in the post-intervention model M_{x_0}:

$$\begin{aligned} z &\leftarrow f_Z(u_z), \\ x &\leftarrow x_0, \\ y &\leftarrow f_Y(x, z, u_Y). \end{aligned} \tag{2.3}$$

The diagram associated with M_{x_0} is depicted in Figure 2c, in which all incoming arrows into X are deleted and replaced by $x \leftarrow x_0$. This captures the notion that an intervention interrupts the original data generating process and eliminates all naturally occurring causes of the manipulated variable. Because other causal paths are effectively shut off in that way, any difference between the two probability distributions associated with M_{x_0} and M_{x_1} captures the variations in outcome Y that is the result of a causal impact of $\Delta x = x_1 - x_0$. A randomized control

[12] The early literature on graphical models, including Bayesian networks and Markov random fields, relied entirely on probabilistic models, which were unable to answer causal and counterfactual queries (Pearl and Mackenzie, 2018, p. 284f.). A major intellectual breakthrough was achieved in the early 1990s by switching focus to the quasi-deterministic functional relationships of the sort that are ubiquitous in econometrics (Pearl, 2009, p. 104f.).

trial closely follows this idea. Experimentation ties the value of a variable to the outcome of a coin flip, which thus induces variation in X that is uncorrelated to any other factors or causal mechanisms.

The post-intervention distribution of Y can also be denoted in counterfactual notation as

$$P(y|do(x)) \triangleq P(Y_x = y), \tag{2.4}$$

where $Y_x = y$ should be read as "Y would be equal to y, if X had been x" (Pearl, 2009, def. 7.1.5). This definition illustrates the connection to the potential outcomes framework (Neyman, 1923; Rubin, 1974; Imbens, 2004), where counterfactuals such as Y_{x_0} and Y_{x_1} are taken as primitives. By contrast, in an SCM, counterfactuals are constructs; i.e., derivable quantities from the underlying, more fundamental causal mechanisms. Naturally, we can write explicitly,

$$Y_{x_0} \leftarrow f(x_0, z, u_Y), \tag{2.5}$$
$$Y_{x_1} \leftarrow f(x_1, z, u_Y), \tag{2.6}$$

which follow immediately from M_{x_0} and M_{x_1}, respectively. In other words, counterfactuals are *derived* from first principles in SCMs, instead of taken as axiomatic primitives.

Equipped with clear semantics for causal models in terms of the underlying mechanisms, and causal effects in terms of interventions on the naturally occurring structural processes in the system, we can now finally state the problem of nonparametric identification.[13]

Definition 2.3. *(Identifiability; Pearl, 2000) A causal query Q is identifiable ($I\!D$, for short) from distribution $P(v)$ compatible with a causal graph G, if for any two (fully specified) models M_1 and M_2 that satisfy the assumptions in G, we have*

$$P_1(v) = P_2(v) \Rightarrow Q(M_1) = Q(M_2). \tag{2.7}$$

[13]This definition of identification is not the same, but related to the one used in Matzkin (2007),. More notedly, the shared feature assumed to be available across structural systems in Matzkin are constraints in the form of (weak) functional assumptions such as monotonicity in somewhat more coarse models, with treatment, outcome, and covariates. Here, on the other hand, we do not assume constraints over the form of the structural functions, but the corresponding shared features are topological, that is, exclusion and independence restrictions are encoded in the causal graph.

This definition requires that for any two (unobserved) SCMs M_1 and M_2, if their induced distributions $P_1(v)$ and $P_2(v)$ coincide, both models need to provide the same answers to query Q. Identifiability entails that Q depends solely on $P(v)$ and the assumptions in G, and can therefore be uniquely expressed in terms of the observed distribution. This holds true *regardless* of the underlying mechanisms f_i and randomness $P(u)$, which, therefore, do not need to be known to the analyst. This is a quite remarkable result, if achieved, since while embracing and acknowledging the true, unobserved structural mechanisms, one can still make the causal statement *as if* these mechanisms were fully known, such as they would be, e.g., in many settings in physics, chemistry, or biology.

Naturally, once the post-intervention distribution $P(y|do(x))$ for any value of x is identified, the average causal effect (as well as any other quantity, such as risk ratios, odds ratios, quantile effects, etc.) can be computed as[14]

$$\mathbb{E}\left[Y|do(X=x_1)\right] - \mathbb{E}\left[Y|do(X=x_0)\right] = \sum_y y\left[P(y|do(x_1)) - P(y|do(x_0))\right]. \quad (2.8)$$

3. CONFOUNDING BIAS

One of the biggest threats to causal inference, and the one which usually receives the greatest attention from methodologists, is confounding bias. The suspicion that a correlation might not reflect a genuine causal link between two variables, but is instead driven by a set of common causes, gives rise to the maxim *"correlation does not imply causation"* (List, 2011). In the presence of confounding, the analyst needs to find a (non-trivial) mapping from a causal query Q to observables $P(v)$, in order to achieve identification. In this section, we will introduce the inference rules of *do-calculus* that allow a logical and systematic treatment of the identification problem solely based on information encoded in a directed acyclic graph G.

Before we do so, however, we will discuss two special cases for dealing with confounding bias – the backdoor and frontdoor adjustments – that are instances of the general treatment provided by do-calculus. Eventually, we will also discuss identification strategies for cases when confounding bias cannot be eliminated in purely

[14]For ease of exposition, we assume random variables to be discrete throughout the text. Summations should be replaced by integrals if variables with continuous support are considered.

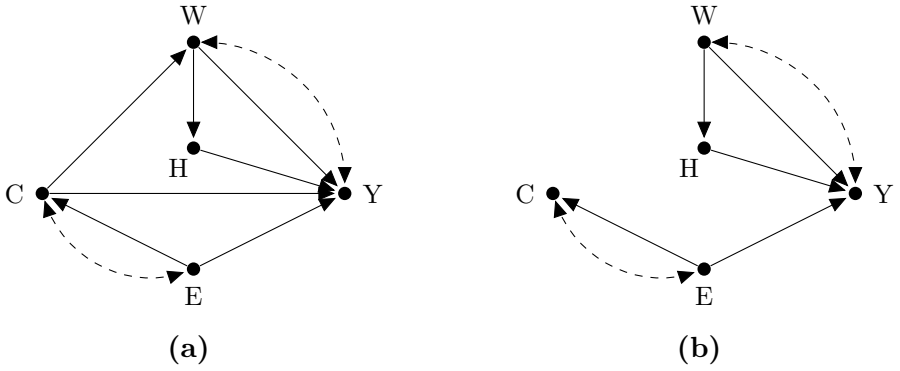


Figure 3: *(a) College wage premium example of Section 3.1. Variables: college degree (C), earnings (Y), occupation (W), work-related health (H), socio-economic factors (E). (b) Graph $G_{\underline{C}}$ obtained when all arrows emitted by C in the graph of panel (a) are deleted.*

observational data, but in which a surrogate experiment, akin to an instrumental variable that creates exogenous variation in a treatment, is available.

3.1. Covariate selection and the backdoor criterion

Consider the well-known example from labor economics of estimating the college wage premium (Angrist and Pischke, 2009, ch. 3.2.3). Let the causal relationships in the problem be represented by the causal graph G in Figure 3a. C is a dummy variable that is equal to one for individuals who obtained a college degree, and the outcome of interest, Y, refers to annual earnings. W is a dummy indicating whether an individual works in a "white-collar" or "blue-collar" job. W is causally affected by C, since many white-collar jobs require a college degree. At the same time, the effect of W is partially mediated by an individual's work-related health H. This assumption captures the idea that blue-collar jobs might be associated with higher adverse health effects, which ultimately reduce life-time earnings. Finally, E represents a set of socio-economic variables that influence both the probability to graduate from college as well as individuals' future earning potentials. Dashed bidirected arrows depict unmeasured common causes that lead to a dependence between the background characteristics U of the connected variables.

In order to estimate the causal effect of a college degree on earnings, the following graphical criterion can be used to find admissible adjustment sets that eliminate

any confounding influences between C and Y.

Definition 3.1. *(Admissible sets – the backdoor criterion; Pearl, 1995) Given an ordered pair of treatment and outcome variables (X, Y) in a causal DAG G, a set Z is backdoor admissible if it blocks every path between X and Y in the graph $G_{\underline{X}}$.*

$G_{\underline{X}}$ in definition 3.1 refers to the graph that is obtained when all edges emitted by node X are deleted in G. Figure 3b depicts $G_{\underline{C}}$ for the college wage premium example, where $C \to H$ and $C \to W$ have been removed. The intuition behind the backdoor criterion is simple. Unblocked paths between X and Y pointing into X (i.e., "entering through the backdoor") create an association between X and Y that is not due to any causal influence exerted by X.[15] By adjusting for (or conditioning on) variables along these spurious paths, this association can be canceled such that only the causal influence from X to Y remains.

In the particular example of Figure 3a, the set $Z = \{E\}$ satisfies the backdoor criterion and is thus an admissible adjustment set.[16] W can be left unaccounted for because it does not lie on a backdoor path between X and Y. In fact, the graph illustrates why conditioning on occupation would produce, rather than reduce, estimation bias. According to the d-separation criterion in Definition 2.2, W is a collider node on $C \to W \dashleftarrow\dashrightarrow Y$, and thus would open, or unblock, this path when conditioned on. As a consequence, adjusting for W would inject bias, creating a non-causal (spurious) correlation between C and Y, and would thus be a serious mistake in this example.

Whenever a backdoor set exists, the causal effect of X on Y can be estimated by adjustment, as shown next.

Theorem 3.2. *(Backdoor Adjustment Criterion) If a set of variables satisfies the backdoor criterion relative to (X, Y), the causal effect of X on Y can be identified from observational data by the adjustment formula*

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z). \tag{3.1}$$

[15] Genuine causal effects can only be transmitted "downstream" of X, via directed paths pointing from X to its descendants and eventually to Y.

[16] Note that $Z = \{E\}$ remains an admissible adjustment set even if edges pointing from E to W and H are added to the graph in Figure 3a.

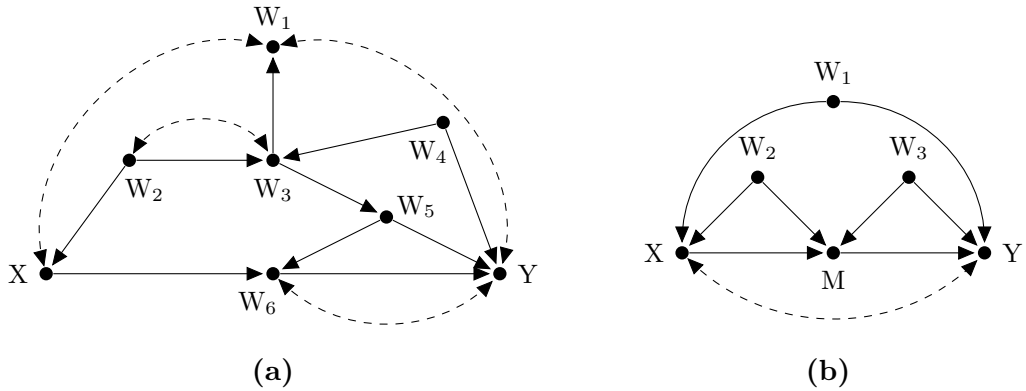


Figure 4: *(a) Application of the backdoor criterion in larger graphs. (b) The presence of M on the directed path from X to Y allows for identification via the front-door criterion.*

Practically speaking, estimation can be carried out by propensity score matching (Rosenbaum and Rubin, 1983; Heckman et al., 1998), inverse probability weighting (Robins, 1999), or deep neural networks (Shi et al., 2019), among other efficient estimation methods.

At this point, the similarity with the treatment effects literature is no coincidence, as the backdoor criterion formally implies ignorability (Rosenbaum and Rubin, 1983).

Theorem 3.3. *(Counterfactual interpretation of backdoor; Pearl, 2009) If a set Z of variables satisfies the backdoor condition relative to (X, Y), then for all x, the counterfactual Y_x is conditionally independent of X given Z*

$$Y_x \perp\!\!\!\perp X|Z. \tag{3.2}$$

In contrast to the potential outcomes framework, however, which provides the analyst with little guidance to identify biasing paths, the search for appropriate adjustment sets via the backdoor criterion can easily be automated (Textor and Liśkiewicz, 2011; Textor et al., 2011). This is particularly useful in larger graphs, such as in Figure 4a. Here, the set of all admissible adjustment sets for identifying

$P(y|do(x))$ is given by

$$\begin{aligned} Z = &\{\{W_2\}, \{W_2, W_3\}, \{W_2, W_4\}, \{W_3, W_4\}, \\ &\{W_2, W_3, W_4\}, \{W_2, W_5\}, \{W_2, W_3, W_5\}, \{W_4, W_5\}, \\ &\{W_2, W_4, W_5\}, \{W_3, W_4, W_5\}, \{W_2, W_3, W_4, W_5\}\}. \end{aligned} \tag{3.3}$$

This list of suitable covariate adjustment sets illustrates that it is neither necessary nor sufficient to adjust for all variables in a model. The analyst could, for example, decide to save costs on data collection efforts for W_4 and instead estimate the effect of X by conditioning on $\{W_2, W_3\}$. At the same time, it would be a serious mistake to condition on W_1, since that would introduce collider bias on the path $X \dashleftarrow\dashrightarrow W_1 \dashleftarrow\dashrightarrow Y$. These intricacies of finding appropriate adjustment sets – in particular in more realistic models – cast serious doubts on the possibility to judge the validity of conditional independence assumptions simply based on introspection and verbal discussions. Causal diagrams, therefore, offer an indispensable complement to any estimation approach that takes ignorability (or conditional exogeneity) as a starting point.

3.2. Frontdoor adjustment in the presence of unmeasured confounders

Identification via backdoor adjustment requires that all backdoor paths can be blocked by a set of observed nodes, which is not always feasible in many practical settings. In situations where no set of observables is backdoor admissible, another (admittedly less familiar to economists) identification strategy might be applicable. Figure 4b presents an example in which adjusting for a set of observable variables $\{W_1\}$ is not sufficient to close all backdoor paths between X and Y. The same is true for the sets $\{W_1, W_2\}$ as well as the set of all pretreatment covariates, $\{W_1, W_2, W_3\}$. For any possible adjustment set, there are unobserved confounders remaining in the graph, represented by the bidirected arc $X \dashleftarrow\dashrightarrow Y$. At the same time, the entire effect of X is assumed to be mediated by a another observed variable M. This assumption is plausible, for example, if a policy intervention in the educational sector affects the job market prospects of graduates solely by

raising test scores.[17]

Still, and perhaps surprisingly, if the data allows to adjust for the confounders at the mediator (since W_2 and W_3 in Figure 4b are assumed to be observed) the effect of X on Y is identifiable with the help of the following criterion (inspired by Pearl, 1995).

Definition 3.4. *(Conditional frontdoor criterion) A set of variables Z is said to satisfy the conditional frontdoor criterion (frontdoor, for short) relative to a triplet (X, Y, W) if*

1. *Z intercepts all directed paths from X to Y,*
2. *there is no unblocked backdoor path from X to Z given W, and*
3. *all the backdoor paths from Z to Y are blocked by $\{X, W\}$.*

Theorem 3.5. *(Conditional frontdoor adjustment) If a set of variables satisfies the conditional frontdoor criterion relative to (X, Y, W), the causal effect of X on Y can be identified from observational data by the frontdoor formula*

$$P(Y = y|do(X = x)) = \sum_{m,w} P(m|w, X = x)p(w) \sum_{x'} P(Y = y|w, m, X = x')P(X = x'|w) \tag{3.4}$$

Applying the frontdoor criterion to the graph in Figure 4b with $W = \{W_1, W_2, W_3\}$ yields the following identification expression.

$$P(y|do(x)) = \sum_{m,w} P(m|x, W = w)P(W = w) \sum_{x',w} P(y|x', m, W = w)P(x'|W = w). \tag{3.5}$$

Frontdoor adjustment amounts to a sequential application of the backdoor criterion. First, the effect of X on M can be identified by adjusting for W_2. Second, the backdoor path $M \leftarrow X \dashleftarrow\dashrightarrow Y$, which remains open after adjusting for W_3, can be blocked by conditioning on X. The frontdoor adjustment formula then chains these individual causal effect estimates together to arrive at the overall effect of X on Y. Because the frontdoor criterion is applicable even in the presence

[17] Obviously, adjusting for the mediator M will not be a viable solution either, since this would block, in the d-separation sense, part of the effect the researcher aims to estimate.

of unobserved confounders (when ignorability does not hold), it is a good example of how causal graphs can point to new identification strategies that go beyond the standard tools currently applied in econometrics.[18]

3.3. Causal calculus and the algorithmization of identification strategies

The backdoor and frontdoor criteria offer simple graphical identification rules that are easy to check. However, while definitely important, they only represent a limited subset of the overall identification results that are derivable in DAGs. In more generality, identifiability of any query of the form $P(y|do(x))$ can be decided systematically by using a symbolic causal inference engine called *do-calculus* (Pearl, 1995). Do-calculus consists of three inference rules, which allow the analyst to transform probabilistic sentences involving interventions and observations, whenever certain separation conditions hold in the causal graph G defined by model M.

Let X, Y, Z, and W be arbitrary disjoint sets of nodes in G. The mutilated graph that is obtained by removing all arrows pointing to nodes in X from G is denoted by $G_{\overline{X}}$. Similarly, $G_{\underline{X}}$ results from deleting all arrows that are emitted by X in G. Finally, the removal of both arrows incoming in X and arrows outgoing from Z is denoted by $G_{\overline{X}\underline{Z}}$. Given this notation, the following three rules – valid for every interventional distribution compatible with G – can be formulated.

Do-Calculus Rule 1. *(Insertion/deletion of observations)*

$$P(y|do(x), z, w) = P(y|do(x), w) \qquad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{3.6}$$

Do-Calculus Rule 2. *(Action/observation exchange)*

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \qquad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{3.7}$$

[18] Glynn and Kashin (2017) present an interesting application of the frontdoor criterion for evaluating the effect of the National Job Training Partnership Act program (JTPA; Heckman et al., 1997) on earnings. In their setting, captured by a graph similar to Figure 4b, X measures the (self-selected) sign-up for the program and M whether an individual actually showed up for the training. The authors are able to relax the assumptions given in Definition 3.4 by complementing the frontdoor criterion with a difference-in-differences-type identification approach that tackles potential bias stemming from unobserved confounders between M and outcome Y.

Do-Calculus Rule 3. *(Insertion/deletion of actions)*

$$P(y|do(x), do(z), w) = P(y|do(x), w) \qquad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{XZ(W)}}}, \tag{3.8}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Rule 1 is a reaffirmation of the d-separation criterion for the X-manipulated graph $G_{\overline{X}}$. Since Z is independent of Y, conditional on X and W, Z can be freely inserted or deleted in the do-expression. Rule 2 states the condition for an intervention $do(Z = z)$ to have the same effect as a passive observation $Z = z$. This condition is fulfilled if $\{X \cup W\}$ blocks all backdoor paths from Z to Y. Note that in $G_{\overline{X}\underline{Z}}$ only such backdoor paths are remaining, since edges emitted by Z are deleted from the graph. Rule 3, then indicates under which condition a manipulation of Z does not affect the probability of Y. This is the case if in the X- and Z-manipulated graph $G_{\overline{XZ}}$, Z is independent of Y conditional on X and W.[19]

Identifiability of a causal query can be decided by repeatedly applying the rules of do-calculus, until Q is transformed into a final expression that no longer contains a do-operator. This renders Q consistently estimable from nonexperimental data. In Appendix A.2.1 we demonstrate this process by showing a step-by-step do-calculus derivation (with the corresponding subgraphs shown alongside) for the college wage premium example from Figure 3a.

Do-calculus was proved sound and complete for general queries of the form $Q = P(y|do(x), z)$ (Pearl, 1995; Tian and Pearl, 2002b; Shpitser and Pearl, 2006a; Huang and Valtorta, 2006; Bareinboim and Pearl, 2012a; Lee et al., 2019). Completeness refers to the property that do-calculus is guaranteed to return a solution for the identification problem, whenever such a solution exists.[20] It implies that if no sequence of steps applying the rules of do-calculus can be found that allow to transform Q into an expression which only contains ex-post observed probabilities, the causal effect is known to be non-identifiable with observational data. If that is the case, point identification will only be achievable by imposing stronger functional-form restrictions (such as linearity, monotonicity, additivity, etc.), or by

[19] The reason for restricting the deletion to Z-nodes that are not ancestors of any W-node in rule 3 of the do-calculus is discussed in Pearl (1995).

[20] Soundness means that if do-calculus returns an answer, this answer is assured to be correct.

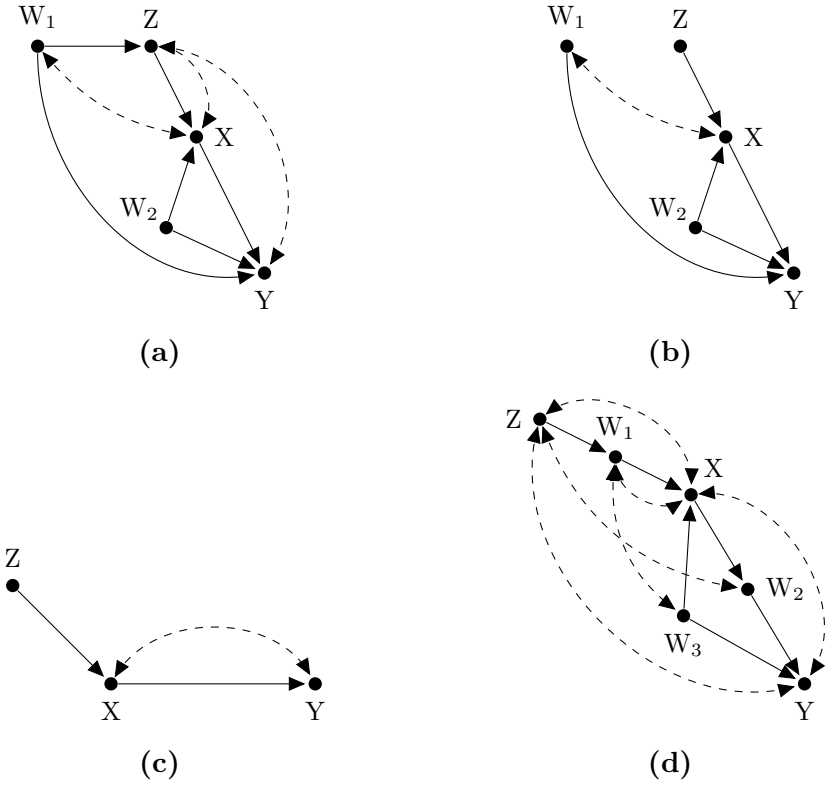


Figure 5: *(a) $P(y|do(x))$ is not identifiable with observational data alone, but z-identifiable if experimental variation in Z is available. (b) Graph $G_{\overline{Z}}$ where all arrows pointing into Z in (a) are deleted. (c) The canonical instrumental variable setting. (d) Example of zID in the presence of unobserved confounders between X and Y and Z affecting X only indirectly.*

making assumptions about the distribution of the background factors U_i. In fact, this result can also be seen algorithmically which allows one to fully automatize the often tedious task of transforming causal effect queries into do-free expressions. That way, the identification of causal effects becomes a straightforward exercise, which can be solved with the help of a computer (Tian and Pearl, 2002a).

3.4. Identification by surrogate experiments

In practice, identification of causal queries based on observational data alone often remains an unattainable goal. At the same time, conducting a randomized control trial (RCT) for the treatment of interest might likewise be infeasible due to cost,

ethical, or technical considerations. In such cases, a frequently applied strategy is to make use of experiments involving a third variable, which is only proximately linked to the treatment but more easily manipulable. In development economics and economic policy such an approach is known under the name of *"encouragement design"* (Duflo et al., 2008). An instructive example is given by Duflo and Saez (2003) who analyze the effect of financial knowledge on retirement planning decisions. They conduct an RCT that randomly allocates monetary rewards for attending an information session on tax deferred account (TDA) retirement plans to university employees. In this surrogate experiment, experimental control of a proxy variable (financial rewards) is supposed to create (or "encourage") exogenous variation in the otherwise endogenous treatment of interest (knowledge about TDA retirement plans). However, compliance remains imperfect, since not all eligible test persons will take up treatment (i.e., show up for the information session).

To make the idea of surrogate experiments even more concrete, Figure 5a presents an example in which several paths passing through Z are confounding the relationship between X and Y. Backdoor adjustment is not a viable identification strategy in this graph, since Z is a collider on $X \dashleftarrow\dashrightarrow Z \dashleftarrow\dashrightarrow Y$, and conditioning on Z would thus open up the path. Furthermore, it can be shown that any other attempt of identifying $Q = P(y|do(x))$ with purely observational data is prone to fail as well in this example. By contrast, if it is possible to manipulate Z in a randomized control trial, the causal effect of X on Y can be identified from the interventional distribution $P(v|do(z))$ instead. Generalizing this idea leads to a natural extension of the identification problem formulated earlier (see Definition 2.3).

Definition 3.6. *($\mathcal{Z}$-identifiability; Bareinboim and Pearl, 2012a) Let X, Y, Z be disjoint sets of variables, and let G be the causal diagram. The causal effect of an action $do(X = x)$ on a set of variables Y is said to be $\boldsymbol{z}$-identifiable ($\boldsymbol{z}I\mathcal{D}$, for short) from P in G, if $P(y|do(x))$ is (uniquely) computable from $P(V)$ together with the interventional distributions $P(V \setminus Z'|do(Z'))$, for all $Z' \subseteq Z$, in any model that induces G.*

Bareinboim and Pearl (2012a) show that the z-identification task can be solved in a similar fashion to the standard identification problem, by repeatedly applying

the rules of do-calculus in order to transform a causal query Q into an expression that only contains $do(z)$.

Theorem 3.7. *(Bareinboim and Pearl, 2012a) Let X, Y, Z be disjoint sets of variables, and let G be the causal diagram, and $Q = P(y|do(x))$. Q is $z\mathcal{ID}$ from P in G if the expression $P(y|do(x))$ is reducible, using the rules of do-calculus, to an expression in which only elements of Z may appear as interventional variables.*

It can further be proved that do-calculus is likewise complete for z-identification (Bareinboim and Pearl, 2012a, Corrolary 3); i.e., it reaches a solution to the $z\mathcal{ID}$ problem whenever such a solution exists.

For the sake of concreteness, however, we discuss a weaker condition, which is only sufficient but not necessary, in order to exemplify the mechanics of the z-identification problem.

Theorem 3.8. *(Sufficient condition – z-identification; Bareinboim and Pearl, 2012a) Let X, Y, Z be disjoint sets of variables and let G be the causal graph. The causal effect $Q = P(y|do(x))$ is $z\mathcal{ID}$ in G if one of the following conditions hold:*

(i) Q is identifiable in G; or.

(ii) There exists $Z' \subseteq Z$ such that the following conditions hold,

a. X intercepts all directed paths from Z' to Y and

b. Q is identifiable in $G_{\overline{Z'}}$.

Condition *(i)* is the base case for when standard identifiability is reached. Whenever this is not the case, if all directed paths from Z to Y are blocked by X, this means that Z has no effect on Y, which by the do-calculus implies $P(y|do(x)) = P(y|do(x, z))$; i.e., the effect of X on Y is the same as the effect of X, Z on Y. Condition *(ii:b)* notes that manipulation of Z leads to the post-intervention graph $G_{\overline{Z}}$, in which all incoming arrows into Z are deleted. If the effect of X can then be identified in this graph, by the removal of $do(x)$ in the expression, then z-identification is ascertained.

For example, recall that in Figure 5a the effect of X on Y is not identifiable from $P(v)$. If experimental data over Z is available, i.e., $P(v|do(z))$, then Theorem 3.8

can be applied. Note that all the directed paths from Z to Y are blocked by X, which satisfies condition *(i:a)*. It is also the case that in the graph $G_{\overline{Z}}$ (see Figure 5b), the set $\{W_1, W_2\}$ is backdoor admissible (by Theorem 3.1), which in turn satisfies condition *(ii:b)*. After all, the effect $P(Y = y|do(X = x))$ is identifiable and given by the expression:

$$\sum_{w_1,w_2} P(Y = y|do(Z = z), X = x, w_1, w_2)P(w_1, w_2|do(Z = z)). \tag{3.9}$$

As in the observational case, researchers are not required to engage in these derivations by hand, since fully automated algorithms exist for z-identification and its generalizations (see Bareinboim and Pearl, 2012a; and Lee et al., 2019, for a survey and the latest results).

Since z-identification exploits experimental variation in a surrogate variable, which causally effects the treatment of interest, it bears close resemblance to instrumental variable (IV) estimation. But the two are not exactly the same. Take the canonical IV setting (following Angrist, 1990) with an exogenous instrument and unobserved confounders between treatment and outcome, depicted in Figure 5c. In this graph, $P(y|do(x))$ is not $\boldsymbol{zID}$, because the bidirected arc between X and Y violates condition (*ii:b*) of Theorem 3.8.[21]

The fact that $P(y|do(x))$ remains unidentifiable in Figure 5c is not very surprising, however. It is a well-known result that point identification of the canonical IV estimator is not possible in the nonparametric case (Manski, 1990; Balke and Pearl, 1995). Introducing additional functional form restrictions, such as monotonicity or linearity, would likewise only permit to identify a *local average treatment effect* for the latent subgroup of compliers (Imbens and Angrist, 1994). $\mathcal{Z}$-identification, by contrast, leverages the fully nonparametric nature of the order relations expressed in causal diagrams. If a query is $\boldsymbol{zID}$, the entire post-interventional distribution, including the average treatment effect, is computable from data. Moreover, z-identification is applicable in more complicated settings than just the canonical IV. An example is given in Figure 5d, where, in addition to an unobserved con-

[21] Theorem 3.8 is only sufficient, but not necessary. Nonetheless, z-identification can also be proved impossible for the graph in Figure 5c, following the most general treatment in Lee et al. (2019).

founder between X and Y, Z exerts only an indirect effect on X.[22] For these reasons, we consider $\boldsymbol{zID}$, including Theorem 3.8, an attractive generalization of the IV strategy in fully nonparametric settings.

4. SAMPLE SELECTION BIAS

The previous section discussed strategies to control for confounding bias, which is the result of nonrandom assignment into treatment and decision-making. Apart from that, researchers often encounter another source of bias in applied empirical work that stems from preferential selection of units into the data pool. Sample selection poses a serious threat to both statistical as well as causal inference, because it jeopardizes the representativeness of the data for the underlying population. A seminal discussion of this problem in an economic context is given by Heckman (1976, 1979). He estimates a model of female labor supply in a sample of 2,253 working women interviewed in 1967. The challenge to valid inference in this setting arises due to the fact that market wages are only observable for women who actually choose to work. His model is described as follows.

$$s_i \leftarrow \quad \mathbb{1}[Z_i'\delta - \eta_i > 0] \tag{4.1}$$

$$y_i \leftarrow \begin{cases} x_i\beta + Z_i'\gamma + \varepsilon_i & \text{if } s_i = 1, \\ \text{unobserved} & \text{if } s_i = 0. \end{cases} \tag{4.2}$$

Equation (4.1) characterizes the sampling mechanism. Wages y_i for an individual i are only observed if $(Z_i'\delta - \eta_i)$ attains a value above zero, which is captured by the selection indicator variable s_i. Economically, this expresses the idea that individuals will choose to remain unemployed if the wage they are able to attain on the market (determined by the vector of socio-economic characteristics Z_i) does not exceed their reservation level η_i. Systematic bias in the coefficient of interest β for hours worked x_i can then arise if reservation wages are correlated

[22] The causal effect of X on Y is $\boldsymbol{zID}$ in Figure 5d by:

$$P(y|do(x)) = \sum_{w_2} P(w_2|x, do(z)) \sum_{x'} P(y|w_2, X = x', do(z)) P(X = x'|do(z))$$

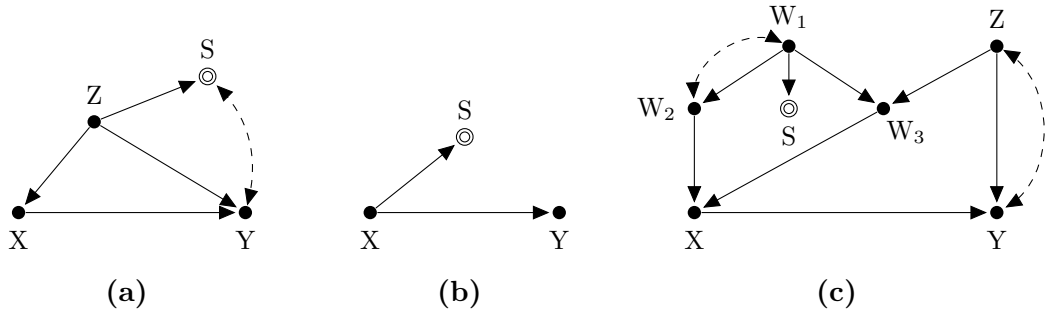


Figure 6: *(a) A model of female labor supply (Heckman, 1976, 1979). Variables: hours worked (X), earnings (Y), socio-economic factors (Z), sampling mechanism (S). (b) $P(y|do(x))$ is recoverable from selected data as $P(y|x, S = 1)$. (c) $\{W_1, W_3\}$, $\{W_2, W_3\}$ and $\{Z\}$ are all backdoor admissible, but the causal effect is only recoverable with $\{Z\}$.*

with unobservables in the market wage equation (4.2); that is, if $Corr(\eta_i, \varepsilon_i) \neq 0$.

Similar cases of sample selection are widespread in economics. Examples are discussed by Levitt and Porter (2000), who estimate the effectiveness of seatbelts and airbags in a sample of fatal crashes, and by Ihlanfeldt and Martinez-Vazquez (1986), who note the difficulty of assessing the determinants of house prices when using data on recently sold homes. Knox et al. (2019) point out another illustrative case. They critique studies which attempt to estimate the extent of racial bias in policing with the help of administrative data (Fryer, 2018). Problematic in this context is that individuals only appear in such records if police officers decided to stop and interrogate them in the first place. If this stopping decision is itself causally affected by minority status, sample selection bias might arise, since the data is not a representative sample of the overall population anymore.

In causal diagrams, cases of sample selection can be captured by explicitly modeling the sampling selection mechanism. We will realize this goal by augmenting the semantics of the causal diagram to account for the sampling mechanism, which graphically will be achieved by adding a new special variable called S. This variable S will take on two values: one, if a unit is part of the sample, and zero otherwise. If endogenous variables in the analysis affect the sampling probabilities, we will add an arrow from these variables to S, which will constitute the specification of the selection mechanism.[23] Figure 6a depicts a DAG for the fe-

[23]We will consider the case here where the sample selection nodes are only allowed to have

male labor supply example that has been augmented by such a selection node; the resulting graph is referred to as a *selection diagram* and denoted by G_S. An individual's socio-economic characteristics Z determine inclusion in the sampling pool and the bidirected dashed arc between S and Y indicates the presence of unobserved confounders that are the source of the error correlation in the model.

Simultaneously controlling for confounding and selection biases introduces a new challenge to the do-calculus. Not only is it necessary to transform interventional distributions into do-free expressions, but the probabilities that make up these expressions now also need to be conditional on $S = 1$, because that is all the analyst is able to observe. This additional restriction explains why dealing with selection bias is such a hard problem in practice. At the same time, the literature on recovering causal effects from selection-biased data (Bareinboim and Pearl, 2012b; Bareinboim et al., 2014; Bareinboim and Tian, 2015) aims at preserving the fully nonparametric nature of causal graphs in this task. Consequently, the proposed approaches refrain from making any functional form assumptions related to the selection-propensity score $P(s_i|pa_i)$ (such as monotonicity or joint normality), which are ubiquitous in the econometrics literature since early on (Angrist, 1997). Nevertheless, even with this limited set of assumptions as a starting point, several positive results for the recoverability of causal effects from selection bias can be derived.

As a first step, Bareinboim et al. (2014) provide a complete condition for recovering conditional probabilities that do not yet contain a do-operator.

Theorem 4.1. *(Bareinboim et al., 2014) The conditional distribution $P(y|t)$ is recoverable from G_S (as $P(y|t, S = 1)$) if and only if $(Y \perp\!\!\!\perp S|T)$.*

Sufficiency of this condition follows immediately. However, its necessity is less obvious and implies that if Y is not d-separated from S in G_S, its conditional distribution will not be recoverable. Combining Theorem 4.1 with do-calculus suggests a straightforward strategy for also recovering do-expressions from selection bias (Bareinboim and Tian, 2015).

incoming arrows, but will not emit arrows themselves.

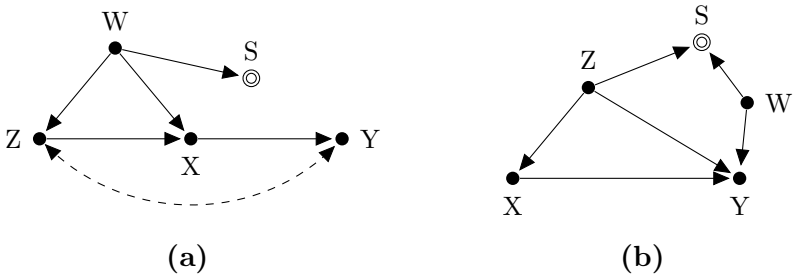
Corollary 4.2. *(Bareinboim and Tian, 2015) The causal effect $Q = P(y|do(x))$ is recoverable from selection-biased data (i.e., $P(v|S = 1)$) if using the rules of the do-calculus, Q is reducible to an expression in which no do-operator appears, and recoverability is determined by Theorem 4.1.*

Take Figure 6b as an example. Here, the relationship between X and Y is unconfounded and, therefore, $P(y|do(x)) = P(y|x)$ holds. Moreover, since S and Y are d-separated by X, we find the causal effect to be recoverable and given by $P(y|x, S = 1)$.

An immediate consequence of Theorem 4.1 is that causal effects will not be recoverable if Y is directly connected to S via an edge in the graph. Thus, without invoking stronger functional form assumptions, there is no possibility to control for selection bias in the female labor supply model of Figure 6a. In general, selection-biased data impair identification in observational studies since now the problem of both confounding and selection needs to be addressed simultaneously. An example is given by the graph in Figure 6c, which contains three backdoor admissible adjustment sets, $\{W_1, W_3\}$, $\{W_2, W_3\}$, and $\{Z\}$, that are (minimally) sufficient for controlling for confounding bias, following Theorem 3.1. However, in this case, recoverability from selection bias can only be achieved with the set $\{Z\}$. That is, because in the adjustment formula (equation 3.1), the prior distribution of the adjustment set needs to be recovered as well, and $\{Z\}$ is the only conditioning set that is marginally d-separated from S. Thus, following the strategy dictated by Corollary 4.2, the estimable backdoor adjustment expression in this case will be:

$$P(y|do(x)) = \sum_z P(y|x, z, S = 1)P(z|S = 1). \tag{4.3}$$

It is important to note, that although Theorem 4.1 provides a necessary condition for recovering conditional probabilities, the same does not hold for Corollary 4.2 with respect to do-expressions. This is exemplified by the graph in Figure 7a. Due to unobserved confounders between Z and Y, and the fact that Z is a collider in the path $X \leftarrow W \rightarrow Z \dashleftarrow\dashrightarrow Y$, identification via the backdoor criterion would require to adjust for both Z and W in order to close all backdoor paths. However, $\{Z, W\}$ is not d-separable from S (W has a direct arrow to S), and a attempt to apply Corollary 4.2 will thus fail. Nevertheless, $P(y|do(x))$ can



(a) **(b)**

Figure 7: *(a) $P(y|do(x))$ is not recoverable from selection bias following the approach laid out in Corollary 4.2. Nevertheless recovery can be achieved by applying the rules of do-calculus. (b) Adaption of the sample selection model in Figure 6a, in which the set $\{Z, W\}$ is s-backdoor admissible.*

still be recovered in Figure 7a with the help of do-calculus using a slightly more sophisticated approach.[24] To witness, note that $(S, W \perp\!\!\!\perp Y)$ in $G_{\overline{X}}$, i.e., the resulting graph when all incoming arrows in X are deleted (see Section 3.3). Then, according to the first rule of do-calculus,

$$P(y|do(x)) = P(y|do(x), w, S = 1), \tag{4.4}$$

$$= \sum_z P(y|do(x), z, w, S = 1)P(z|do(x), w, S = 1), \tag{4.5}$$

where the second line follows by conditioning on Z. Applying rule 2 of do-calculus, since $(Y \perp\!\!\!\perp X|W, Z, S)$ in $G_{\underline{X}}$, the do-operator can be removed in the first term of Equation 4.5, which can be written as:

$$= \sum_z P(y|x, z, w, S = 1)P(z|do(x), w, S = 1). \tag{4.6}$$

Finally, since $(Z \perp\!\!\!\perp X|W, S = 1)$ in $G_{\overline{X(W)}}$, rule 3 of the calculus allows us to remove the $do(x)$ from the second term, such that:

$$P(y|do(x)) = \sum_z P(y|x, z, w, S = 1)P(z|w, S = 1). \tag{4.7}$$

Note that the quantities in the final expression of $P(y|do(x))$ do not involve any do-operator, since the dataset is observational, and always contain $S = 1$, given

[24] The following do-calculus derivations are shown in more detail, with corresponding subgraphs depicted alongside, in Appendix A.2.2.

that the samples were selected preferentially. Taken together, this ensures recoverability.

Bareinboim and Tian (2015) provide algorithmic criteria for recovering interventional distributions (i.e., containing $do(x)$-operators) in arbitrary causal graphs. They permit full automatization of derivations such as the one just performed. Recently, this algorithm was also proved complete for the recovery task by Correa et al. (2019).

4.1. Combining biased and unbiased data

Another promising strategy for recovering causal quantities from sample selection is when biased and unbiased data sources are combined. For example, the distributions of socio-economic factors such as age, sex, and education can often be measured without bias from population-level statistics. To illustrate how this helps for recoverability, we revisit the female labor supply example from above, but now assume that the common parent of wages Y and the selection node S is observable as W (see Figure 7b, which is the same as Figure 7a but for the replacement of the bidirected arrow with the observed W). If that is the case, conditioning on the set $\{Z, W\}$ closes all backdoor paths between X and Y and simultaneously d-separates Y from S. From the backdoor adjustment formula discussed above (Theorem 3.2), we can thus derive

$$P(y|do(x)) = \sum_{z,w} P(y|x,z,w)P(z,w), \tag{4.8}$$

$$= \sum_{z,w} P(y|x,z,w,S=1)P(z,w), \tag{4.9}$$

where the second line follows from Theorem 4.1, since $(Y \perp\!\!\!\perp S|Z,W)$. As $P(z,w)$ cannot be recovered from selection bias, Corollary 4.2 is not applicable. However, if in addition to the selected data, unbiased measurements of $P(z,w)$ are available (e.g., from census data), equation (4.9) becomes estimable.

Bareinboim et al. (2014) leverage this idea and present the following generalization of the backdoor criterion, which can be invoked if a subset Z of the data is measured without bias.

Definition 4.3. *(Selection backdoor criterion; Bareinboim et al., 2014) Let a set Z of variables be partitioned into $Z^+ \cup Z^-$ such that Z^+ contains all non-descendants of X and Z^- the descendants of X, and let G_S stand for the graph that includes sampling mechanism S. Z is said to satisfy the selection backdoor criterion (s-backdoor, for short) if it satisfies the following conditions:*

(i) *Z^+ blocks all backdoor paths from X to Y in G_S;*

(ii) *X and Z^+ block all paths between Z^- and Y in G_S, namely, $(Z^- \perp\!\!\!\perp Y | X, Z^+)$;*

(iii) *X and Z block all paths between S and Y in G_S, namely, $(Y \perp\!\!\!\perp S | X, Z)$; and*

(iv) *Z and $Z \cup \{X, Y\}$ are measured in the unbiased and biased studies, respectively.*

The following theorem can then be proved.

Theorem 4.4. *(Bareinboim et al., 2014) If Z is s-backdoor admissible, then causal effects are identified by*

$$P(y|do(x)) = \sum_z P(y|x, z, S = 1)P(z). \tag{4.10}$$

The s-backdoor criterion is a sufficient condition for generalized adjustment, which is able to deal with confounding and selection bias simultaneously. Correa et al. (2018) substantially extend this line of work by presenting conditions that are both necessary *and* sufficient. Furthermore, Correa et al. (2019) provide a sound algorithm for recovering causal effects from a mix of biased and unbiased data in causal graphs that are arbitrary in size and shape.

5. TRANSPORTABILITY OF CAUSAL KNOWLEDGE

Extrapolating causal knowledge across settings is a fundamental problem in causal inference. Experiments are usually conducted in different populations than they are supposed to inform. Expecting experimental results to hold across populations may be fallacious, however, if domains differ structurally in important ways. Duflo et al. (2008) allude to this problem in a development economics context when

asking: *"If a program worked for poor rural women in Africa, will it work for middle-income urban men in South Asia?"*.

In this section, we discuss the conditions under which a transfer of causal knowledge across structurally heterogeneous domains is valid. This issue is known under the rubric of *"transportability"* in the computer science literature, while social scientists usually refer to it as *"external validity"* (Pearl and Bareinboim, 2014).[25] Nakamura and Steinsson (2018) discuss the challenge of external validity from a macroeconomic perspective and come to the conclusion that *"even very cleanly identified monetary and fiscal natural experiments give us, at best, only a partial assessment of how future monetary and fiscal policy actions—which may differ in important ways from those in the past—will affect the economy."* Causal diagrams, in conjunction with do-calculus, allow to formally address these kinds of concerns in a principled, general, and efficient way, eliciting the assumptions needed to analyze these settings and making precise how much can actually be learned from experiments across different domains.

In practice, it is often implicitly assumed that an experimental result obtained in population Π provides at least a good approximation for the impact of the same intervention in other settings. This assumption is made for convenience, because it allows to use results from Π for policy decisions in a different population Π^*. However, such kind of *direct transportability*, which we formally define in the following, is likely to be violated in many empirical settings.

Definition 5.1. *(Direct Transportability; Pearl and Bareinboim, 2011) A causal relation R is said to be directly transportable from Π to Π^*, if $R(\Pi^*) = R(\Pi)$.*

For an example, consider the study by Banerjee et al. (2007) that analyzes the effects of a remedial education program in two major cities in Western India: Mumbai and Vadodara. The randomized intervention provided schools with an extra teacher for tutoring children in the third and fourth grades, who had been lagging behind their peers. The program showed substantial positive effects on children's

[25] In econometrics, the term "external validity" is sometimes used in the narrower sense of extrapolating local average treatment effect estimates to the group of always- and never-takers within the same empirical domain (Kowalski, 2018). In the remainder of this section, we will focus on the more challenging task of transporting causal knowledge across domains.

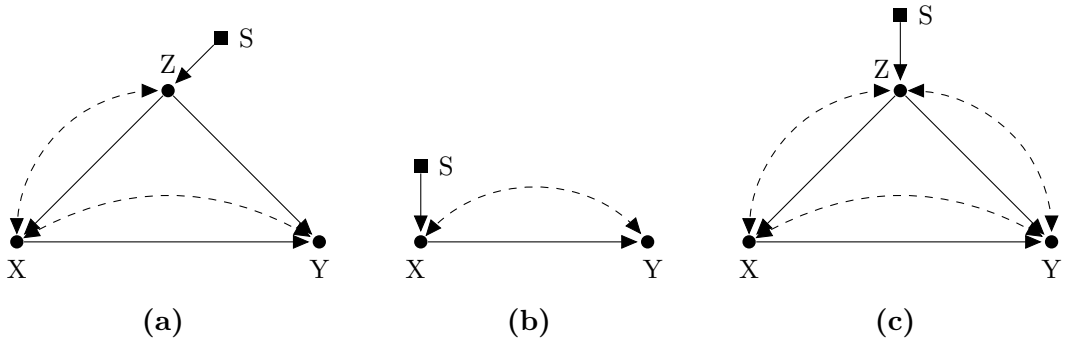


Figure 8: *(a) Z d-separates S and Y in $D_{\overline{X}}$. The causal effect of X on Y is thus transportable. (b) If S-nodes are only pointing into X, the causal effect $P^*(y|do(x))$ is directly transportable. (c) Compared to (a), a single additional unobserved confounder between Z and Y prevents transportability.*

academic achievements, at least in the short-run. Interestingly, however, while treatment effects on math scores were similar in both cities, the effect on language proficiency was weaker in Mumbai compared to Vadodara. The authors explain this finding by higher baseline reading skills in Mumbai, where families were on average wealthier and schools were better equipped. In math, by contrast, baseline skill levels did not differ significantly. As a consequence, the remedial education program, which targeted only the most basic competencies in the curriculum, was equally effective.

The graph in Figure 8a provides a graphical representation of the setting in Banerjee et al. (2007). Assume that we want to generalize experimental results from a trial conducted in Vadodara (Π) to the population in Mumbai (Π^*). We are aware, however, of the fact that income levels of families Z, which are an important determinant of children's academic achievements Y, are higher in Mumbai. In a causal diagram, we can incorporate this knowledge by adding a set of *selection nodes* S that indicate where both populations under study differ, either in the distribution of background factors $P(U)$ or due to divergent causal mechanisms f_i. These S-nodes thus locate the source of structural discrepancies that threaten transportability. Switching between two populations Π and Π^* is denoted by conditioning on different values of S.[26] Next, we define the joint graphical representa-

[26]For clarity, S nodes related to transportability are depicted by squares (■), in order to distinguish them from the selection bias case. Also note that now S is emitting arrows, whereas

tion of the corresponding structural models in the source and target populations, which is required to judge transportability.

Definition 5.2. *(Selection Diagram; Pearl and Bareinboim, 2011) Let $\langle M, M^* \rangle$ be a pair of structural causal models (see Definition 2.1) relative to domains $\langle \Pi, \Pi^* \rangle$, sharing a causal diagram G. $\langle M, M^* \rangle$ is said to induce a selection diagram D if D is constructed as follows:*

(i) Every edge in G is also an edge in D.

(ii) D contains an extra edge $S_i \rightarrow V_i$ whenever there might exist a discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between M and M^*.

The absence of an S-node in the selection diagram represents the assumption that the causal mechanism, which assigns values to the respective variable, is the same in both populations. In the extreme case, one could add S-nodes to all variables in the graph, to express the notion that the two populations are maximally structurally heterogeneous (i.e., there is no knowledge whatsoever about structural invariances). Obviously, this would undermine any hope for information exchange across domains though.

Equipped with the definition of a selection diagram, we can state the following theorem, which allows to transport experimental results obtained in a source Π to another target domain Π^*, where only passive observations are possible.[27]

Theorem 5.3. *(Pearl and Bareinboim, 2011) Let D be the selection diagram characterizing two populations, Π and Π^*, and S the set of selection variables in D. The strata-specific causal effect $P^*(y|do(x), z)$ is transportable from Π to Π^* if Z d-separates Y from S in the X-manipulated version of D, that is, Z satisfies $(Y \perp\!\!\!\perp S|Z, X)_{D_{\overline{X}}}$.*

selection nodes indicating preferential inclusion into the sample only receive incoming arrows.

[27] Note that, following Definition 5.2, both domains Π and Π^* have to share the same causal diagram G. Consequently, if a causal query Q is identifiable with observational data alone in the source domain Π (i.e., no experimental knowledge is necessary), it will also be identifiable in the target domain Π^*, and Q will thus be *trivially transportable* (Pearl and Bareinboim, 2011). Pearl and Bareinboim (2011) discuss *observational transportability* of a statistical query of the form $P(y|x)$ (e.g., a classifier) from a source domain to a target domain, where only a subset of the variables in the selection diagram are observed. Thus, statistical transportability permits the analyst to save on data collection costs. Later on, Correa and Bareinboim (2019) devised a complete algorithm for this task. We will not further pursue this topic in what follows and refer the interested reader to the respective paper.

Note that $D_{\overline{X}}$ refers to the post-intervention graph, in which all incoming arrows into X are deleted (see Section 3.3). D-separation between S-nodes and the outcome variable Y can be achieved by adjusting for a conditioning set T, as the following definition formalizes.

Definition 5.4. *(S-admissibility; Pearl and Bareinboim, 2011) A set T of variables satisfying $(Y \perp\!\!\!\perp S|T)$ in $D_{\overline{X}}$ will be called s-admissible (with respect to the causal effect of X on Y).*

The intuition behind this result is somewhat similar to the selection bias case (see Theorem 4.1), where the selection indicator was likewise required to be d-separated from Y by a set T (Pearl, 2015a). Looking at the selection diagram in Figure 8a, we note that the set Z d-separates S and Y in $D_{\overline{X}}$ (i.e., when X is experimentally manipulated). It therefore satisfies s-admissibility.

By applying the rules of do-calculus, we can now show that s-admissibility implies transportability across domains.

$$P^*(y|do(x)) = P(y|do(x), s) \tag{5.1}$$
$$= \sum_z P(y|do(x), z, s)P(z|do(x), s) \tag{5.2}$$
$$= \sum_z P(y|do(x), z, s)P(z|s) \tag{5.3}$$
$$= \sum_z P(y|do(x), z)P^*(z). \tag{5.4}$$

The first equation follows from the definition that distributions in the target domain Π^* are denoted by conditioning on S. The second line follows by conditioning. The third line is derived by using the s-admissibility of Z and recognizing the fact that X is a child of Z and, therefore, exerts no causal influence on Z (formally, rule 3 of do-calculus can be applied). The last line is then just a restatement.

As long as Figure 8a provides an accurate model for the setting in Banerjee et al. (2007), the causal effect of the remedial education program in Mumbai can thus be computed by reweighting the stratum-specific causal effect (for every income level of Z) obtained in Vadodara by the income distribution $P^*(z)$ in Mumbai. No experimental data from Mumbai is required. This result is stated in its full

generality in the following corollary.

Corollary 5.5. *(Pearl and Bareinboim, 2011) The causal effect $P^*(y|do(x))$ is transportable from Π to Π^* if there exists a set Z of observed pretreatment covariates that is s-admissible. Moreover, the transport formula is given by the weighting*

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z). \tag{5.5}$$

It is an immediate consequence of Theorem 5.3 that any S variable that points into X can be ignored. The causal effect $P(y|do(x))$ is thus directly transportable in Figure 8b. The same holds for S nodes that are d-separated by the empty set in $D_{\overline{X}}$.

As a graphical criterion, s-admissibility is easy to check. Without a reference to a causal diagram, however, the intricacies of transportability can be hard to discern. Figure 8c provides a cautionary tale in that regard. Apart from the unobserved confounder between Z and Y, it is identical to Figure 8a. Here, however, s-admissibility is violated because conditioning on Z would open up the path $S \rightarrow Z \dashleftarrow\!\dashrightarrow Y$. In can further be shown that transporting causal effects is also impossible in general in this selection diagram. Thus, the example illustrates that the absence or presence of one single edge can determine whether transportability is feasible. Recognizing such subtleties by pure introspection, without the reference to an explicit model, would be an extremely difficult undertaking.

The transport formula presented in equation (5.5) has been acknowledged in the econometrics literature (Hotz et al., 2005; Dehejia et al., 2015; Andrews and Oster, 2018). Most commonly in this literature, this formula is expressed using the potential outcomes framework, where s-admissibility is encoded through ignorability relations; i.e., domain heterogeneity S is assumed to be ignorable given pretreatment covariates X. While it is hard to judge ignorability statements, we note that this assumption is easily violated in practice; for example, by a single unobserved confounder between Z and Y in Figure 8c. Causal graphs offer valuable guidance for judging the validity of ignorability assumptions, which is missing in the potential outcomes framework. Furthermore, using the rules of do-calculus, it becomes possible to establish transportability in more general cases that are not covered by Corollary 5.5.

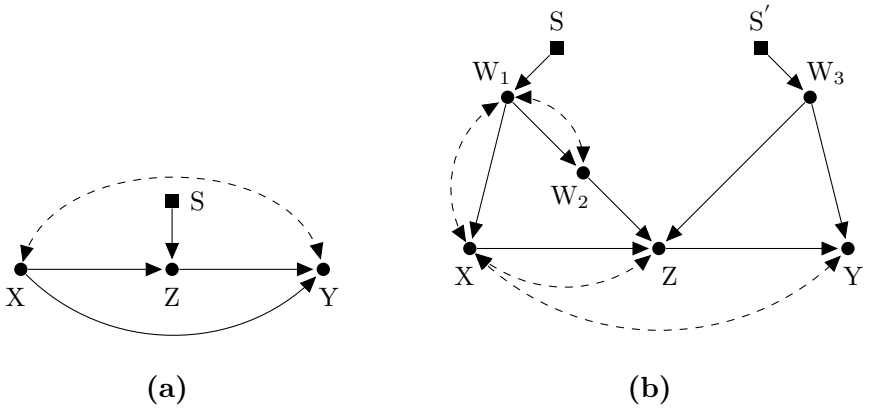


Figure 9: *(a) $P^*(y|do(x))$ is transportable even though S points into a post-treatment variable. (b) A more complex graph in which transportability can be decided algorithmically by the criteria developed in Bareinboim and Pearl (2013b).*

Theorem 5.6. *(Pearl and Bareinboim, 2011) Let D be the selection diagram characterizing two populations, Π and Π^*, and S as set of selection variables in D. The relation $R = P^*(y|do(x))$ is transportable from Π to Π^* if the expression $P(y|do(x), s)$ is reducible, using the rules of do-calculus, to an expression in which S appears only as a conditioning variable in do-free terms.*

One such class of models is given when domains differ due to variables that are themselves causally affected by the treatment, as in Figure 9a. Here, the effect of X on Y is partly transmitted by Z, and domains differ either according to the distribution of background factors U_Z or the mechanism f_Z that determines Z. Such a situation can occur, for example, in development programs, where the success of a policy is partly dependent on the level of care with which it is implemented. Duflo et al. (2008) discuss the problem that pilot trials often employ particularly highly qualified program officials, which is difficult to replicate once the program is supposed to be scaled up and thus threatens the generalizability of these pilot studies.

Gordon et al. (2018) provide a similar example from an entirely different context. The effectiveness of advertising campaigns on social media platforms depends on how frequently clients are exposed to the ads. Exposure thus acts as a mediator for the effect of advertising on an outcome of interest, e.g., the click-through rate. And since exposure is determined by user behavior, it is difficult to control for the advertiser. If a social media company running advertising experiments wants

to transport results obtained on a desktop version of the platform to users with mobile devices, it will need to take into account that exposure might differ across domains, e.g., due to differences in user demographics.

If post-treatment variables, such as in Figure 9a, are s-admissible, the causal effect of X can be transported as

$$P^*(y|do(x)) = P(y|do(x), s) \tag{5.6}$$

$$= \sum_z P(y|do(x), z, s)P(z|do(x), s) \tag{5.7}$$

$$= \sum_z P(y|do(x), z)P^*(z|do(x)), \tag{5.8}$$

where the last line follows from s-admissibility (Pearl and Bareinboim, 2014). Given equation (5.8), we can see that transportability of $P^*(y|do(x))$ then requires to transform $P^*(z|do(x))$ into a do-free expression, since by definition no manipulation can be carried out in the target domain. Recognizing that X and Z are unconfounded in Figure 9a, this can be achieved by $P^*(z|do(x)) = P^*(z|x)$ (formally, rule 2 of do-calculus applies).

The resulting transport formula, when domains differ according to post-treatment variables, is different from the simple expression in equation (5.5). It prescribes to reweight the z-specific effects by the conditional (instead of the uncoditional) distribution of Z in the target population:

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z|x). \tag{5.9}$$

Theorem 5.6 was proven to be a necessary and sufficient criterion for transporting causal effect estimates across domains by Bareinboim and Pearl (2012c). However, it is only procedural in nature and, therefore, does not specify the sequence of do-calculus steps that need to be taken to arrive at the desired expression. In order to fill this gap, Bareinboim and Pearl (2013b) develop a complete algorithmic solution for carrying out the transformation. The benefits of solving the transportability problem algorithmically become particularly apparent for more

complex graphs, such as in Figure 9b, in which the correct transport formula is:

$$P^*(y|do(x)) = \sum_{z,w_2,w_3} P(y|do(x), z, w_2, w_3)P(z|do(x), w_2, w_3)P^*(w_2, w_3). \quad (5.10)$$

Note also that this expression does not contain W_1. Applying the transportability algorithm thus helps to decide which measurements are required for transportability and thereby allows to economize on data collection efforts in the target domain.

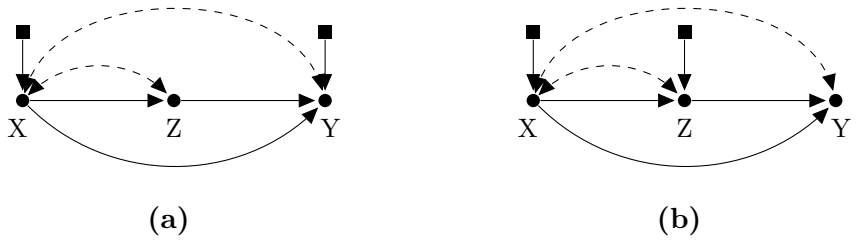
5.1. Transportability with surrogate experiments

Bareinboim and Pearl (2013a) combine the idea of transportability with the previously introduced concept of z-identification, to develop a theory they call *z-transportability*. Owing to this extension, it becomes possible to not only transfer causal knowledge obtained from direct randomized control trials, but also from the encouragement designs, discussed in Section 3.4, that rely on surrogate experiments. Researchers are thus given the flexibility to learn from knowledge across domains even in cases when direct manipulation of a treatment would be prohibitively costly, both in the target and in the source domain.

Remarkably, z-transportability is a distinct problem and reduces neither to ordinary transportability nor to z-identifiabilitty. Bareinboim and Pearl (2013a) demonstrate this fact by presenting examples of causal queries which are $z\mathcal{ID}$ in the source domain Π, but that may or may not be z-transportable. Analogous to Theorem 5.6, the rules of do-calculus can be used to transfer causal knowledge from surrogate experiments in the following way.

Theorem 5.7. *(Bareinboim and Pearl, 2013a) Let D be the selection diagram characterizing two populations, Π and Π^*, and S be the set of selection variables in D. The relation $R = P^*(y|do(x))$ is z-transportable from Π to Π^* in D if the expression $P(y|do(x), s)$ is reducible, using the rules of do-calculus, to an expression in which all do-operators apply to subsets of Z, and the S-variables are separated from these do-operators.*

Again, Theorem 5.7 provides no indication of the sequence of do-calculus steps that need to be taken in order to establish z-transportability. To this end, Bareinboim and Pearl (2013a) develop a complete algorithm, which takes the selection



(a) **(b)**

Figure 10: *Selection diagrams relative to two heterogeneous source domains π_a and π_b. Square nodes indicate discrepancies between the source and target domains. Meta-transportability entails to combine causal knowledge from both π_a and π_b to arrive at an estimate for $P^*(y|do(x))$ in the target domain.*

diagram D, and information on the variable that has been intervened on in the source domain as inputs, and then returns a transport formula expression whenever such an expression exists.

5.2. Combining causal knowledge from several heterogeneous source domains

Transportability techniques are particularly valuable in situations where it is possible to combine empirical knowledge from several source domains. Dehejia et al. (2015) consider the case of a policy-maker who is faced with the decision to either learn about a desired treatment effect from extrapolation of an existing experimental evidence base, or to commission a costly new experiment. The challenge in this situation is that previous experiments have possibly been conducted in very different contexts than the one of interest, and underlying populations might be quite heterogeneous. Naive pooling of results, for example, is thus likely to fail. Based on the approaches presented in the previous sections, Bareinboim and Pearl (2013c) introduce the concept of *meta*-transportability (or μ-transportability, for short), which provides a principled solution to this problem.[28]

Let $\mathcal{D} = \{D_1, \ldots, D_n\}$ be a collection of selection diagrams relative to source domains $\Pi = \{\pi_1, \ldots, \pi_n\}$. An example is given by Figure 10, in which panel (a) depicts the selection diagram that corresponds to source domain π_a, while panel (b)

[28] Meta-transportability is related to the ideas concerning "data combination" presented in Ridder and Moffitt (2007). In this case, however, the goal is to combine causal knowledge from several heterogeneous populations that share at least some causal mechanisms.

refers to π_b. Square nodes indicate where discrepancies between the target domain π^* and the source domains arise.[29] In line with Definition 5.2, these discrepancies can occur due to differences in causal mechanisms as well as background factors related to the the variables that square nodes point into.

Figure 10 is a simple extension of a graph that was presented earlier (see Figure 9a). In contrast to before, the unobserved confounder between X and Z (denoted by the dashed bidirected arc $X \dashleftarrow\dashrightarrow Z$), which was added to the diagram, now renders individual transportability impossible.[30] Interestingly, however, by combining information from both source domains, μ-transportability is feasible. To see this, note that the post-intervention distribution in the target domain π^* can be written as:

$$P^*(y|do(x)) = \sum_z P^*(y|do(x), z)P^*(z|do(x)), \tag{5.11}$$

$$= \sum_z P^*(y|do(x), do(z))P^*(z|do(x)), \tag{5.12}$$

where the second line follows from rule 2 of do-calculus, since $(Z \perp\!\!\!\perp Y|X)$ in $D_{\overline{X}\underline{Z}}$.[31] Using this representation, each component can be shown to be individually transportable from one of the source domains. $P^*(z|do(x))$ is directly transportable from π^a, because $(S \perp\!\!\!\perp Z)$ in $D^{(a)}_{\overline{X}}$. And $P^*(y|do(x), do(z))$ is directly transportable from π^b, since $(S \perp\!\!\!\perp Y)$ in $D^{(b)}_{\overline{X,Z}}$. The individual components of equation (5.12) can therefore be written as $P^*(z|do(x)) = P^{(a)}(z|do(x))$ and $P^*(y|do(x), do(z)) = P^{(b)}(y|do(x), do(z))$. This leads to the final transport formula:

$$P^*(y|do(x)) = \sum_z P^{(b)}(y|do(x), do(z))P^{(a)}(z|do(x)). \tag{5.13}$$

[29] The causal diagram for the target domain is accordingly obtained by deleting all square nodes from the selection diagrams.

[30] The algorithm by Bareinboim and Pearl (2013b) would exit without returning a transport formula expression for both selection diagrams. Intuitively, in panel (a), transportability is prohibited by the selection node pointing directly into Y. In (b), $X \dashleftarrow\dashrightarrow Z$ prevents to set $P^*(z|do(x)) = P^*(z|x)$, which was instrumental for establishing transportability following equation (5.8).

[31] These do-calculus derivations are shown in detail, with corresponding subgraphs depicted alongside, in Appendix A.2.3

In addition to demonstrating that multiple pairwise transportability is not a necessary condition for μ-transportability, the example illustrates the superior inferential power obtained by combining multiple datasets over each individual dataset alone.

Bareinboim and Pearl (2013c) develop a complete algorithmic solution for deciding about μ-transportability. The approach is further extended by Bareinboim et al. (2013) who combine μ-transportability with z-transportability, to allow for combining causal knowledge from multiple heterogeneous sources when only surrogate experiments on a subset Z of variables in $\mathcal{D}$ are possible. This latter task is called *mz-transportability* and can be automated by an algorithm that was proved to be complete by Bareinboim and Pearl (2014).

In recent years, meta-analyses, which synthesize the results of several studies on a specific subject, are becoming increasingly important. Examples from economics can be found, inter alia, in Card et al. (2010), Dehejia et al. (2015), and Meager (2019). A drawback of standard meta-analytical approaches is, however, that they do not incorporate knowledge about domain heterogeneity in terms of causal mechanisms and background factors. Instead, they attempt to "average out" differences across populations.[32] By contrast, the transportability techniques we have presented make it transparent how discrepancies in study results arise and how they can nonetheless be leveraged to identify a target query of interest in a principled and efficient manner. Moreover, they discipline the analyst to think carefully about the assumptions and shared mechanisms that allow extrapolation across domains to actually take place.

Transportability theory thereby enables the research community to devise an effective strategy for leveraging the entire evidence base that exists related to a specific problem. Causal knowledge obtained by an individual experiment does not need to, and should not, be regarded in isolation. Rather, it contributes to a larger body of empirical work that can be recombined to tackle entirely new policy problems, which were unimagined at the time of the original study. In combination with undergoing efforts to make more data sets openly available, transportability

[32] To the extent that these studies consider domain heterogeneity, this is done in a purely statistical fashion, without explicitly modeling structural differences across populations (Dehejia et al., 2015; Meager, 2019). This leaves open the question whether domains are actually structurally sufficiently similar such that transportability of study results can be ensured.

techniques thus bear the potential to save on discipline-wide data collection costs and to render causal inference a truly collective endeavor.[33]

6. CONCLUSION

From the end of the 1980s onwards, the artificial intelligence literature has developed an increased interest in causal inference (Pearl, 1988, 2009; Bareinboim and Pearl, 2016; Pearl and Mackenzie, 2018). Causation is a fundamental concept in human thinking and structures the way in which we interact with our environment (Woodward, 2003; Mumford and Anjum, 2013). A human-like AI, therefore, needs to possess an internal representation of causality in order to mimic human behavior and communicate with us in a meaningful way (Pearl and Mackenzie, 2018). Tremendous progress over the last three decades has led to the development of a powerful causal inference engine, which puts an artificial learner into the position to acquire and combine causal knowledge from many diverse sources in its surroundings. In particular, several important contributions to the literature in recent years have made this engine more robust, general, and practical, by expanding its applicability to the various different data collection and knowledge contexts we have discussed in this paper.

We are convinced that these causal inference tools originating from AI have also a great deal to offer to econometricians. Until today, the possibilities to completely automatize the identification task, which is a necessary ingredient for causal machine learning, still remain largely unexplored in econometric practice. The applications of do-calculus we have discussed only require the analyst to provide a model of the economic context under study and a description of the available data, the rest can be handled automatically by an algorithm.[34]

[33] Other recent contributions to transportability theory have been made by Correa and Bareinboim (2019), who develop adjustment criteria for generalizing experimental findings in the presence of selection bias (see Section 4) and Lee et al. (2020), who present a general treatment of transportability theory, which is able to unify several of the techniques that have been discussed in this section.

[34] Up to a certain extent, directed acyclic graphs can also be learned from observational data. Respective techniques rely on the testable implications of DAGs that were discussed in Section 2 in order to find an equivalence class of models that is compatible with the d-separation relations in the data. The interested reader is referred to the literature on "causal structure learning" and "causal discovery" in the artificial intelligence field (Spirtes et al., 2001; Pearl,

Moreover, graphical representations of structural causal models do not require the learner – whether artificial or human – to impose any distributional or functional-form restrictions on the underlying causal mechanisms under study. The approach remains fully nonparametric, a characteristic it shares with the potential outcomes framework (Imbens and Rubin, 2015; Imbens, 2019). At the same time, however, crucial identification assumptions, such as ignorability, are derived from the properties of the underlying structural model, rather than being assumed to hold in a coarse, a priori way. Causal graphical models thus combine the accessibility and flexibility of potential outcomes with the preciseness and analytical rigor of structural econometrics (Heckman and Vytlacil, 2007; Matzkin, 2013; Lewbel, 2019). The balance graphical approaches strike between these two currently competing econometric streams is of great value for applied empirical work. Economists should therefore feel encouraged to engage in a productive exchange with AI researchers in order to mutually benefit from the numerous useful tools for causal inference developed in both disciplines.

REFERENCES

Andrews, I. and E. Oster (2018): "Weighting for External Validity," NBER Working Paper No. 23826.

Angrist, J. D. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic Review*, 80, 313–336.

——— (1997): "Conditional independence in sample selection models," *Economics Letters*, 54, 103–112.

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Balke, A. and J. Pearl (1995): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

2009; Peters et al., 2017). Automation of the identification task in these settings has also gained traction recently (Zhang, 2006; Perkovic et al., 2017; Jaber et al., 2018b,a, 2019).

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007): "Remedying Education: Evidence from Two Randomized Experiments in India," *The Quartely Journal of Economics*, 122, 1235–1264.

Bareinboim, E., S. Lee, V. Honavar, and J. Pearl (2013): "Transportability from multiple environments with limited experiments," in *Advances in Neural Information Processing Systems*, ed. by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, vol. 26, 136–144.

Bareinboim, E. and J. Pearl (2012a): "Causal Inference by Surrogate Experiments: z-identifiability," in *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 113–120.

——— (2012b): "Controlling Selection Bias in Causal Inference," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, 100–108.

——— (2012c): "Transportability of Causal Effects: Completeness Results," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

——— (2013a): "Causal Transportability with Limited Experiments," in *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 95–101.

——— (2013b): "A general algorithm for deciding transportability of experimental results," *Journal of Causal Inference*, 1, 107–134.

——— (2013c): "Meta-Transportability of Causal Effects: A Formal Approach," in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Scottsdale, AZ, vol. 31.

——— (2014): "Transportability from Multiple Environments with Limited Experiments: Completeness Results," in *Advances of Neural Information Processing Systems*, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, vol. 27, 280–288.

——— (2016): "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, 113, 7345–7352.

Bareinboim, E. and J. Tian (2015): "Recovering Causal Effects from Selection Bias," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ed. by S. Koenig and B. Bonet, Association for the Advancement of Artificial Intelligence, Palo Alto, CA: AAAI Press.

Bareinboim, E., J. Tian, and J. Pearl (2014): "Recovering from Selection Bias in Causal and Statistical Inference," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Basman, R. L. (1963): "The Causal Interpretation of Non-Triangular Systems of Economic Relations," *Econometrica*, 31, 439–448.

Bentzel, R. and B. Hansen (1954): "On Recursiveness and Interdependency in Economic Modeks," *The Review of Economic Studies*, 22, 153–168.

Bentzel, R. and H. Wold (1946): "On Statistical Demand Analysis from the Viewpoint of Simulatneous Equations," *Skandivavisk Aktuarietidskrift*, 29, 95–114.

Card, D., J. Kluve, and A. Weber (2010): "Active Labour Market Policy Evaluations: A Meta-Analysis," *The Economic Journal*, 120, 452–477.

Cartwright, N. (2007): *Hunting Causes and Using Them*, Cambridge University Press.

Chen, B., D. Kumor, and E. Bareinboim (2017): "Identification and Model Testing in Linear Structural Equation Models using Auxiliary Variables," in *Proceedings of the 34th International Conference on Machine Learning*, ed. by D. Precup and Y. W. Teh, PMLR, vol. 70 of *Proceedings of Machine Learning Research*, 757–766.

Correa, J. D. and E. Bareinboim (2019): "From Statistical Transportability to Estimating the Effect of Stochastic Interventions," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*.

Correa, J. D., J. Tian, and E. Bareinboim (2018): "Generalized Adjustment Under Confounding and Selection Biases," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, LA: AAAI Press, vol. 32, 6335–6342.

——— (2019): "Identification of Causal Effects in the Presence of Selection Bias," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Dehejia, R., C. Pop-Eleches, and C. Samii (2015): "From Local to Global: External Validity in a Fertility Natural Experiment," NBER Working Paper No. 21459.

Duflo, E., R. Glennerster, and M. Kremer (2008): "Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics*, Elsevier, vol. 4, chap. 61.

Duflo, E. and E. Saez (2003): "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," *Quarterly Journal of Economics*, 118, 815–842.

Frisch, R. (1933): "Editor's Note," *Econometrica*, 1, 1–4.

Fryer, R. G. (2018): "An Empirical Analysis of Racial Differences in Police Use of Force," .

Glynn, A. N. and K. Kashin (2017): "Front-Door Difference-in-Differences Estimators," *American Journal of Political Science*, 61, 989–1002.

Gordon, B. R., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2018): "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," .

Haavelmo, T. (1943): "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11, 1–12.

Halpern, J. Y. (2000): "Axiomatizing Causal Reasoning," *J. Artif. Int. Res.*, 12, 317–337.

Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *The Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies*, 65, 261–294.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 64, 605–654.

Heckman, J. J. and R. Pinto (2013): "Causal Analysis after Haavelmo," *Econometric Theory*, 31, 115–151.

Heckman, J. J. and E. J. Vytlacil (2007): "Econometric Evaluation of Social Programs, Part 1: Causal Models, Structural Models and Econometric Policy Evaluation," in *Hanbook of Econometrics*, Elsevier B.V., vol. 6B.

Hoover, K. D. (2004): "Lost Causes," *Journal of the History of Economic Thought*, 26.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005): "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics*, 125, 241–270.

Huang, Y. and M. Valtorta (2006): "Pearl's Calculus of Interventions Is Complete," in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI2006)*.

Ihlanfeldt, K. R. and J. Martinez-Vazquez (1986): "Alternative Value Estimates of Owner-Occupied Housing: Evidence on Sample Selection Bias and Systematic Errors," *Journal of Urban Economics*, 20, 356–369.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

——— (2014): "Instrumental Variables: An Econometrician's Perspective," *Statistical Science*, 29, 323–358.

——— (2019): "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics," ArXiv:1907.07271.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

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

Jaber, A., J. Zhang, and E. Bareinboim (2018a): "Causal Identification under Markov Equivalence," in *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, UAI'18, 978–987.

——— (2018b): "A Graphical Criterion for Effect Identification in Equivalence Classes of Causal Diagrams," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, 5024–5030.

——— (2019): "Causal Identification under Markov Equivalence: Completeness Results," in *Proceedings of the 36th International Conference on Machine Learning*, ICML'19.

Knox, D., W. Lowe, and J. Mummolo (2019): "The Bias Is Built In: How Administrative Records Mask Racially Biased Policing," .

Kowalski, A. E. (2018): "How to examine External Validity Within an Experiment," NBER Working Paper 24834.

Lee, S., J. D. Correa, and E. Bareinboim (2019): "General Identifiability with Arbitrary Surrogate Experiments," in *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, 144.

——— (2020): "General Transportability – Synthesizing Observations and Experiments from Heterogeneous Domains," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

Levitt, S. D. and J. Porter (2000): "Sample Selection in the estimation of air bag and seat belt effectiveness," *The Review of Economics and Statistics*, 83, 603–615.

Lewbel, A. (2019): "The Identification Zoo: Meanings of Indetification in Econometrics," *Journal of Economic Literature*, 57, 835–903.

List, J. A. (2011): "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off," *Journal of Economic Perspectives*, 25, 3–16.

Maddala, G. S. (1986): *Limited-Dependent and Qualitative Variables in Econometrics*, Econometric Society Monographs.

Manski, C. F. (1990): "Nonparametric bounds on treatment effects," *American Economic Review, Papers and Proceedings*, 80, 319–323.

——— (2003): *Partial Identification of Probability Distributions*, New York: Springer.

Matzkin, R. L. (2007): "Nonparametric Identification," in *Handbook of Econometrics*, vol. 6B.

——— (2013): "Nonparametric Identification in Structural Economic Models," *Annual Review of Economics*, 5, 457–486.

Meager, R. (2019): "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," *American Economic Journal: Applied Economcis*, 11, 57–91.

Morgan, M. S. (1991): "The Stamping Out of Process Analysis in Econometrics," in *Appraising Economic Theories: Studies in the Methodology of Research Programs*, ed. by N. D. Marchi and M. Blaug, Aldershot, UK: Edward Elgar, 237–265.

Mumford, S. and R. L. Anjum (2013): *Causation: A Very Short Introduction*, Great Clarendon Street, Oxford, OX2DP, United Kingdom: Oxford University Press.

Nakamura, E. and J. Steinsson (2018): "Identification in Macroeconomics," *Journal of Economic Persepctives*, 32, 59–86.

Neyman, J. (1923): "Sur les applications de la thar des probabilities aux experiences agraricales: Essay des principle," English translation of excerpts (1990) by D. Dabrowska and T. Speed in *Statistical Science*, 5:463-472.

Page, S. E. (1999): "Computational models from A to Z," *Complexity*, 5, 35–41.

Pearl, J. (1988): *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA: Morgan Kaufmann.

——— (1995): "Causal diagrams for empirical research," *Biometrika*, 82, 669–709.

——— (2000): *Causality: Models, Reasoning, and Inference*, New York, United States, NY: Cambridge University Press, 1st ed.

——— (2009): *Causality: Models, Reasoning, and Inference*, New York, United States, NY: Cambridge University Press, 2nd ed.

——— (2013): "Reflections on Heckman and Pinto's 'Causal analysis after Haavelmo'," Tech. Rep. R-420, Univiversity of California, Los Angeles.

——— (2015a): "Generalizing Experimental Findings," *Journal of Causal Inference*, 3, 259–266.

——— (2015b): "Trygve Haavelmo and the Emergence of Causal Calculus," *Econometric Theory*, 31, 152–179.

Pearl, J. and E. Bareinboim (2011): "Transportability of Causal and Statistical Relations: A Formal Approach," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press, 247–254.

——— (2014): "External Validity: From Do-Calculus to Transportability Across Populations," *Statistical Science*, 29, 579–595.

Pearl, J. and D. Mackenzie (2018): *The Book of Why: The New Science of Cause and Effect*, New York: Basic Books.

Perkovic, E., J. Textor, M. Kalisch, and M. H. Maathuis (2017): "Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs," *The Journal of Machine Learning Research*, 18, 8132–8193.

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

Ridder, G. and R. Moffitt (2007): *The Econometrics of Data Combination*, Elsevier B.V., vol. 6B, chap. 75, 5470–5547.

Robins, J. M. (1999): "Testing and estimation of of directed effects be reparameterizing directed acyclic with structural nested models," in *Computation, Causation, and Discovery*, ed. by C. N. Glymour and G. F. Cooper, Cambridge, MA: AAAI/MIT Press, 349–405.

Rosenbaum, P. R. and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Shi, C., D. Blei, and V. Veitch (2019): "Adapting Neural Networks for the Estimation of Treatment Effects," in *Advances in Neural Information Processing Systems 32*, ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Curran Associates, Inc., –.

Shpitser, I. and J. Pearl (2006a): "Identification of Conditional Interventional Distributions," in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI2006)*.

——— (2006b): "Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models," in *Twenty-First National Conference on Artificial Intelligence*.

Spirtes, P., C. Glymour, and R. Scheines (2001): *Causation, Prediction, and Search*, Cambride, MA: The MIT Press, 2nd ed.

Strotz, R. H. and H. O. A. Wold (1960): "Recursive vs. Nonrecursive Systems: An Attempt At Synthesis (Part I of a Triptych on Causal Chain Systems)," *Econometrica*, 28, 417–427.

Textor, J., J. Hardt, and S. Knüppel (2011): "DAGitty: A Graphical Tool for Analyzing Diagrams," *Epidemiology*, 5, 745.

Textor, J. and M. Liśkiewicz (2011): "Adjustment Criteria in Causal Diagrams: An Algorithmic Perspective," in *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, AUAI press, 681–688.

Tian, J. and J. Pearl (2002a): "A general identification condition for causal effects," in *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press/The MIT Press, 567–573.

——— (2002b): "A general identification condition for causal effects," in *Aaai/iaai*, 567–573.

Verma, T. and J. Pearl (1988): "Causal networks: Semantics and expressiveness," in *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, Mountain View, CA, 352–359.

Wold, H. (1954): "Causality and Econometrics," *Econometrica*, 22, 162–177.

——— (1981): *The Fix-point Approach to Interdependent Systems*, North-Holland Publishing Company, chap. The Fix-point Approach to Interdependent Systems: Review and Current Outlook, 1–36.

Wold, H. O. A. (1960): "A Generalization of Causal Chain Models (Part III of a Triptych on Causal Chain Systems)," *Econometrica*, 28, 443–463.

Woodward, J. (2003): *Making Things Happen*, Oxford Studies in Philosophy of Science, Oxford University Press.

Zellner, A. (1979): "Causality and econometrics," *Carnegie-Rochester Conference Series on Public Policy*, 10, 9–54.

Zhang, J. (2006): "Causal inference and reasoning in causally insufficient systems," Ph.D. thesis, Carnegie Mellon University.

APPENDIX

A.1. Causality in recursive and interdependent systems

In this paper, attention is restricted to a class of models that can be described by directed acyclic graphs, in which the rules of do-calculus apply. The requirement of acyclicity gives rise to what economists commonly denote as *recursive* systems (Wold, 1954; Pearl, 2009, p. 231). Yet, many standard models in economics, such as the canonical supply and demand relationship, as well as game theoretic models, are nonrecursive or *interdependent*. In the aftermath of Haavelmo's celebrated paper on simultaneous equation models (Haavelmo, 1943), an intensive discussion about the conceptual interpretation of recursive versus interdependent models emerged in the econometrics literature (see Morgan, 1991, for an excellent historical account). The debate was particularly motivated by practical concerns of estimation, as Haavelmo demonstrated for the first time in full clarity that the method of least squares does not lead to unbiased parameter estimates in interdependent simultaneous equation models.[35] However, it also touched on the causal interpretation of interdependent models and the adequacy of cyclic causal relationships as a representation of economic processes. One central argument, most notably formulated in Bentzel and Hansen (1954) and Strotz and Wold (1960), was that individual equations in an interdependent model do not have a causal interpretation *"in the sense of a stimulus-response relationship"* (Strotz and Wold, 1960, p. 417).[36] Instead, interdependent systems with equilibrium conditions are regarded as *"shortcut"* descriptions (Wold, 1960; Imbens, 2014) of the underlying dynamic behavioral processes.[37]

In this context, Strotz and Wold (1960) discuss the example of the cobweb

[35] As a matter of fact, Haavelmo never made a distinction between recursive and interdependent models in his 1943 paper. Starting from an interdependent simultaneous equation model, he demonstrated that OLS is biased in this context. Later, Bentzel and Wold (1946; as cited in Wold, 1981) were able to show that least squares estimation is indeed appropriate if the system is recursive.

[36] More than two decades later, Maddala (1986, p. 111) presented a similar point of view in his influential textbook.

[37] Herman Wold coined the term *causal chain* for the latter. Bentzel and Hansen (1954) point out that interdependency can also be the result of an aggregation of variables measured at an inappropriate frequency, even if the underlying data generating process is fully recursive.

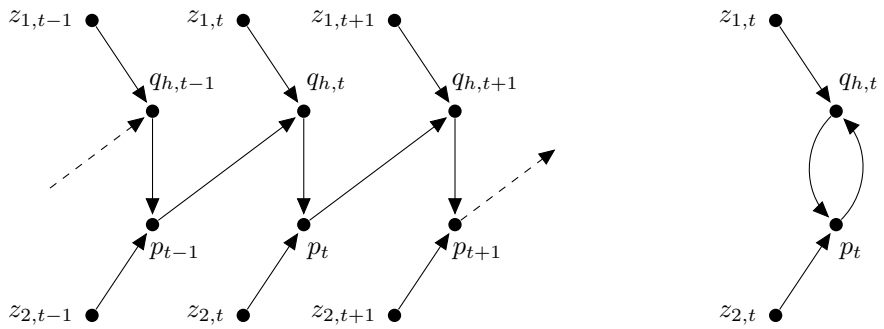


Figure 11: *(a) Dynamic, recursive model of a market for crops. (b) Nonrecursive model of the same market after imposing an equilibrium constraint.*

model, a particular form of a dynamic supply and demand system, based on Jan Tinbergen's microeconometric work in the 1920s (see Morgan, 1991).

$$q_{h,t} \leftarrow \gamma + \delta p_{t-1} + \nu z_{1,t} + u_{1,t}, \tag{A.1}$$
$$p_t \leftarrow \alpha - \beta q_{h,t} + \varepsilon z_{2,t} + u_{2,t}. \tag{A.2}$$

This model is recursive. The first equation determines the quantity of a particular crop harvested at time t, based on the crop's price p_{t-1} in the previous period. The second equation describes crop demand and pins down prices in t, depending on current supply. Moreover, the model incorporates exogenous supply and demand shifters z_1 and z_2. By imposing an equilibrium assumption on the system, such that prices are required to remain constant over time

$$p_{t-1} = p_t, \tag{A.3}$$

the model becomes interdependent, as price and quantity now affect each other simultaneously in the same period.

$$q_{h,t} \leftarrow \gamma + \delta p_t + \nu z_{1,t} + u_{1,t} \tag{A.4}$$
$$p_t \leftarrow \alpha - \beta q_{h,t} + \varepsilon z_{2,t} + u_{2,t} \tag{A.5}$$

Figure 11 illustrates the step from the fully dynamic model to a nonrecursive equilibrium model graphically. Note, however, that the equilibrium assumption

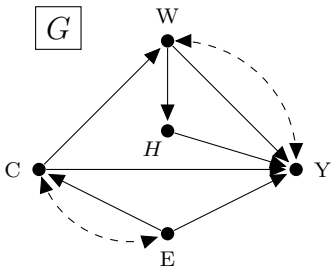
(A.3) carries no behavioral interpretation and may or may not describe the data adequately. Likewise, the individual equations of the interdependent system do not represent autonomous causal relationships in the stimulus-response sense, since the endogenous variables are determined jointly by all equations in the system (Matzkin, 2013; Heckman and Pinto, 2013). Thus, it would not be possible, for example, to directly use p_t in equation (A.4) to bring about a desired change in $q_{h,t}$.

This discussion does not imply – as these authors have stated repeatedly – that equilibrium models cannot be useful for learning about individual causal parameters (Strotz and Wold, 1960, p. 426), nor that a causal interpretation cannot be given to a nonrecursive model as a whole (Bentzel and Hansen, 1954; Basman, 1963; Zellner, 1979). However, if individual functions of an economic model are supposed to be interpreted as stimulus-response relationships, cyclic patterns need to be excluded. Otherwise, stimuli would be permitted to be causes of themselves, which would violate the notion of asymmetry usually attached to them (Woodward, 2003; Cartwright, 2007). Incidentally, the potential outcomes framework in the econometric treatment effects literature also interprets the link between a treatment and an outcome as a stimulus-response relationship and therefore implicitly maintains the assumption of acyclicty (Heckman and Vytlacil, 2007).

A.2. Do-calculus derivations

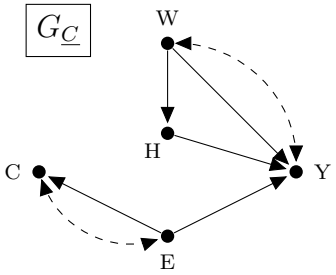
In this section, we show step-by-step solutions for the do-calculus derivations discussed in the main text. For illustration purposes, subgraphs used in the respective steps are placed alongside.

A.2.1. College wage premium example (Section 3.1, Figure 3a)



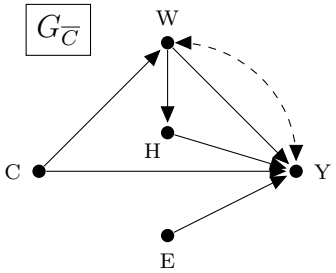
Consider the causal effect of C on Y in graph G. There are two backdoor paths in G, which can both be blocked by E. Conditioning and summing over all values of E yields:

$$P(y|do(c)) = \sum_e P(y|do(c), e)P(e|do(c)).$$



By rule 2 of do-calculus, since $(Y \perp\!\!\!\perp C|E)$ in subgraph $G_{\underline{C}}$, it holds that:

$$P(y|do(c), e) = P(y|c, e).$$



In $G_{\overline{C}}$, E is d-separated from C, because Y is a collider on every path connecting them. Thus, $(E \perp\!\!\!\perp C)_{G_{\overline{C}}}$, and by rule 3 of do-calculus:

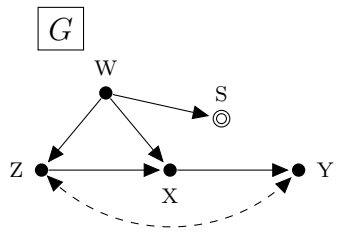
$$P(e|do(c)) = P(e).$$

Combining these two expressions yields:

$$P(y|do(c)) = \sum_e P(y|c, e)P(e).$$

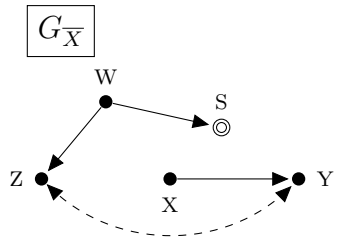
The right-hand-side expression is do-free and can therefore be estimated from observational data.

A.2.2. Selection bias example (Section 4, Figure 7a)



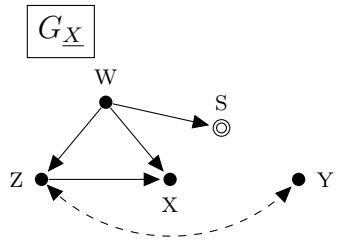
Consider the causal effect of X on Y in graph G. In graph $G_{\overline{X}}$, Z is a collider on the path connecting S and W with Y. Therefore, $(S, W \perp\!\!\!\perp Y)_{G_{\overline{X}}}$, and by the first rule of do-calculus it holds that:

$$P(y|do(x)) = P(y|do(x), w, S = 1),$$
$$= \sum_z P(y|do(x), z, w, S = 1)P(z|do(x), w, S = 1).$$



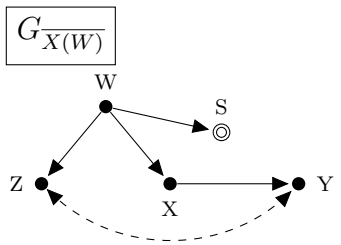
Moreover, because $(Y \perp\!\!\!\perp X|W, Z, S)$ in $G_{\underline{X}}$, rule 2 of do-calculus applies to the first factor, which leads to:

$$P(y|do(x)) = \sum_z P(y|x, z, w, S = 1)P(z|do(x), w, S = 1).$$

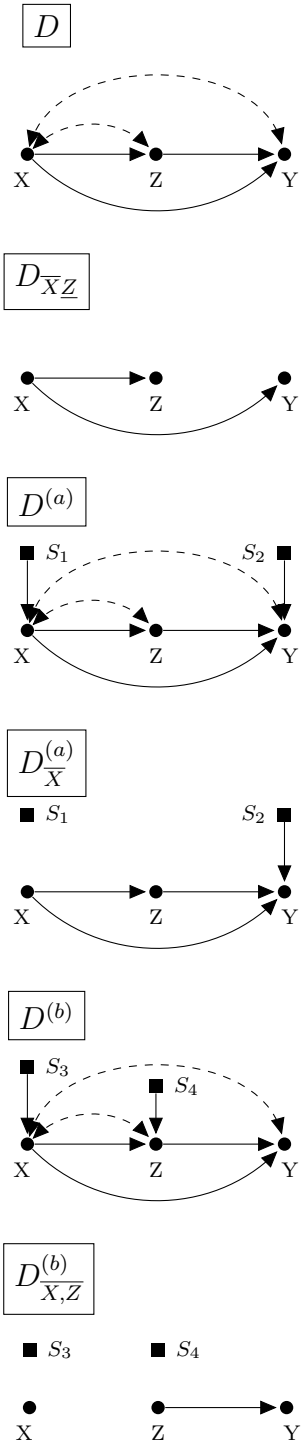


Finally, notice that W blocks any path between X and Z conditional on $S = 1$ in $G_{\overline{X(W)}}$. Thus, since $(Z \perp\!\!\!\perp X|W)_{G_{\overline{X(W)}}}$, rule 3 of do-calculus applies to the second term, such that:

$$P(y|do(x)) = \sum_z P(y|x, z, w, S = 1)P(z|w, S = 1).$$



A.2.3. M-Transportability example (Section 5.2, Figure 10)



Consider the causal effect of X on Y in graph D, in target domain π^*:

$$P^*(y|do(x)).$$

Note that X d-separates Z and Y in $D_{\overline{X}\underline{Z}}$. Thus, since $(Z \perp\!\!\!\perp Y|X)_{D_{\overline{X}\underline{Z}}}$, it follows from rule 2 of do-calculus that:

$$\begin{aligned} P^*(y|do(x)) &= \sum_z P^*(y|do(x), z)P^*(z|do(x)), \\ &= \sum_z P^*(y|do(x), do(z))P^*(z|do(x)). \end{aligned}$$

Let the selection diagrams for the two source domains π^a and π^b be given by D_a and D_b, respectively.
Note that $(S_1, S_2 \perp\!\!\!\perp Z)$ in $D^{(a)}_{\overline{X}}$, therefore, $P^*(z|do(x))$ is directly transportable from π^a as:

$$P^*(z|do(x)) = P^{(a)}(z|do(x)).$$

Furthermore, since $(S_3, S_4 \perp\!\!\!\perp Y)$ in $D^{(b)}_{\overline{X,Z}}$, $P^*(y|do(x), do(z))$ is directly transportable from π^b:

$$P^*(y|do(x), do(z)) = P^{(b)}(y|do(x), do(z)).$$

Combining the two expressions leads to the final transport formula:

$$P^*(y|do(x)) = \sum_z P^{(b)}(y|do(x), do(z))P^{(a)}(z|do(x)).$$