Causal Inference and Data Fusion in Econometrics

Paul Hünermund[†] and Elias Bareinboim[‡]

[†]*Copenhagen Business School, Kilevej 14A, Frederiksberg, 2000, DK*
E-mail: `phu.si@cbs.dk`
[‡]*Columbia University, 500 W 120th Street, New York, NY 10027, USA*
E-mail: `eb@cs.columbia.edu`

Summary   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 and flexible 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 perspective on causality introduced by Haavelmo (1943) and the graph-theoretic approach proposed by Pearl (1995), the artificial intelligence (AI) literature has developed a wide array of techniques for causal inference that allow to leverage information from various imperfect, heterogeneous, and biased data sources (Bareinboim and Pearl, 2016). In this paper, we review 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 learning, in which the above-mentioned problems can be addressed in generality. Second, due to their origin in AI, they come together with sound, efficient, and complete (to be formally defined) algorithmic criteria for automation of the corresponding identification task. And third, because of the nonparametric description of structural models that graph-theoretic approaches build on, they combine the analytical rigor of structural econometrics with the flexibility of the potential outcomes framework, and thus offer a valuable complement to these two literature streams.

Keywords: *Causal Inference, Directed Acyclic Graphs, Causal Diagrams, Artificial Intelligence, Data Fusion*

1. INTRODUCTION

Obtaining causal knowledge by uncovering quantitative relationships in statistical data is arguably one of the most important goals of econometrics since the beginning of the discipline. Policy-makers, legislators, and managers need to be able to forecast the likely impact of their actions in order to make informed decisions. Phillip G. Wright's (1928) seminal contribution on instrumental variable estimation, using the theory of path coefficients developed by his son Sewall Wright (1921, 1923), for example, was motivated by the desire to understand the effect of tariffs on the production and import of agricultural products. In the postwar period, interest in the topic of causal inference initially experienced a decline in attention (Hoover, 2004), but 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, 2020) 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. Following 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 represents the outcome of interest, Z a treatment under study, W other socioeconomic variables, and U unobserved exogenous background factors.[2] Since W enters in both equations of the system, it creates a correlation between Z and Y that is not the result of a causal impact. Therefore, to predict how Y reacts to induced changes in Z, the causal mechanism that naturally determines Z needs to be replaced to avoid non-causal (spurious) sources of variation. In this particular example, the values that Z attains must be decoupled 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 of the intervention can be traced via equation (1.2') to pin down Z's causal effect on Y.

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 another variable X. In mathematical notation, this can be written as $Q = P(Y = y|do(X = x))$, where $do(X = x)$ denotes the replacement of whatever mechanisms were there for X, say f_x, by 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 often do not have available. In many economic settings, experiments can be difficult to implement, due to cost, technical, or ethical considerations. Likewise, exactly knowing the structural mechanisms that truly govern the data generating process is hard in the social sciences, where often only qual-

---

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

[2] We follow the usual notation of denoting random variables by uppercase and their realized values by lowercase letters.
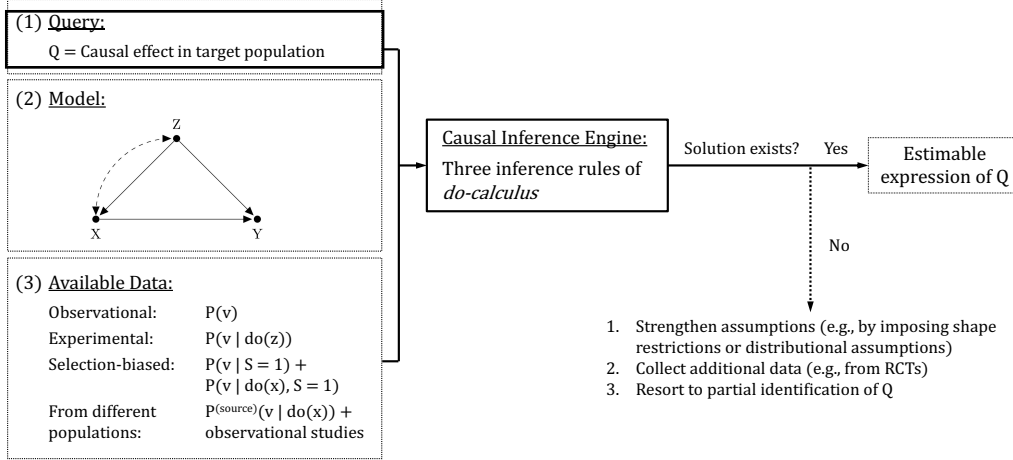
Figure 1: Schematic illustration of the data fusion process.

itative knowledge about causal relationships is available.[3] This means that the counterfactual distribution $Q = P(y|do(x))$ will, in general, not be immediately estimable. In practice, 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 machinery that warrants such kinds of syntactic transformations is called *do-calculus* (Pearl, 1995).

Do-calculus can be seen as a causal inference engine that takes three inputs:

1. A causal quantity Q, which is the query the researcher wants to answer;
2. A model G that encodes the qualitative understanding about the structural dependencies between the economic variables under study; and
3. A collection of datasets $P(v|\cdot)$ that are available to the analyst, including observational, experimental, selection-biased samples, from different populations, and so on.

Building on these inputs, do-calculus consists of three inference rules for transforming probabilistic sentences involving do-expressions into equivalent expressions. The inferential goal is then to re-express the causal quantity Q (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 from 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 fully general, formal treatment of

---

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

nonparametric causal inference (i.e., relying solely on exclusion and independence restrictions encoded in the graph) in recursive models. 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, 2006; Huang and Valtorta, 2006; Bareinboim and Pearl, 2012c; Lee et al., 2019; Correa and Bareinboim, 2020b; Correa et al., 2021). 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).[4]

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.[5] 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, observational distribution $P(y|x)$ – a distinction that has been popularized through the mantra *"correlation does not imply causation"*. 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, 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 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 fulfil certain requirements (e.g., market wages are only observable if individuals are employed; Heckman, 1979). Selection-biased data aggravate the identification problem as $P(y|do(x))$

---

[4]These tools are implemented in the free software `www.causalfusion.net` and available to researchers.

[5]There is a growing interest in the graph-theoretic approach to causal inference in economics. Early examples constitute Adams et al. (2003), Neuberg (2003), White and Chalak (2006), and Frölich (2008), and more recent ones include Heckman and Pinto (2015) (examined in detail in Pearl, 2013) and Imbens (2020). However, these papers do not cover the newer advances that were made in the causal AI literature in the last decade, which is the primary focus of this paper.

needs to be transformed into an expression solely comprised of probabilities from a non-random sample of the population (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 subjects in a laboratory setting), but needs to be brought to productive use in other domains 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 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 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 licenses this kind of information transfer and reconciliation (Bareinboim and Pearl, 2016). In the remainder of the paper, we will describe such a causal modelling 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.[6] 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*

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

---

[6]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.

   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. 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.[7]

To emphasize the interpretation of f_i's as stimulus-response relationships, and in contrast to the standard notation in econometrics, the artificial intelligence 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 at 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 contexts, however, precise knowledge of the functional relationships, f_i, and the distribution of the background factors, $P(u)$, 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., 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 way of encoding a parsimonious set of assumptions of the system sufficient 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.[8] An example is given in Figure 2a, which refers to the following

---

[7] 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 a conditional mean function.

[8] As it is common in graph theory, 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 $B \rightarrow D$, A is an ancestor of E, since $A \rightarrow D \rightarrow E$, and E is a child of D, since $D \rightarrow E$.

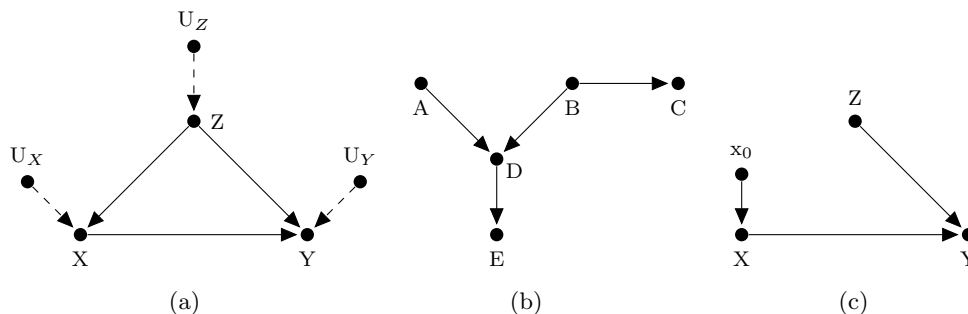**Figure 2**: Examples of directed acyclic graphs representing structural causal models.

underlying structural causal model

$$Z \leftarrow f_z(U_z),$$
$$X \leftarrow f_x(Z, U_x), \qquad\qquad (2.1)$$
$$Y \leftarrow f_y(X, Z, U_y).$$

Note that Z appears as an argument in the structural function of X, f_x. Accordingly, 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 at a time. 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). I.e., the arc $X \dashleftarrow\dashrightarrow Y$ serves as a shortcut notation for $X \leftarrow U \rightarrow Y$, where the set of common causes U is unobservable to the analyst.

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 refer to as *recursive*, is of central importance in causal inference, because they describe 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).[9] It is important to note, however, that the axioms of structural counterfactuals in SCMs (Pearl, 2009, ch. 7) also hold in nonrecursive models, as discussed in Halpern (2000). Graphical causal model with directed cycles, which can incorporate feedback loops and equilibrium behavior, are an active area of research. The interested reader is referred to Spirtes et al. (2001, ch. 12), Pearl (2009, ch. 3.6), and Bongers et al. (2021).

Working with the graphical representation of an SCM entails a deliberate choice by the analyst to refrain from distributional and functional form assumptions, since the shape

---

[9]Incidentally, the potential outcomes framework in the econometric treatment effects literature also interprets the link between treatment and outcome as a stimulus-response relationship and therefore implicitly maintains the assumption of acyclicty (Heckman and Vytlacil, 2007). To witness the discussion about the causal interpretation of individual functional relationships in recursive versus nonrecursive models in the early econometrics literature, see Haavelmo (1943), Bentzel and Wold (1946; as cited in Wold, 1981), Bentzel and Hansen (1954), Strotz and Wold (1960), Wold (1960), and Basman (1963).

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.[10] Shape restrictions on functions (such as separability, monotonicity, or differentiability) and distributional assumptions might sometimes be licensed by economic theory (Heckman and Vytlacil, 2007; Matzkin, 2007, 2013). In case 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 distributional and functional form assumptions once the explanatory power of nonparametric identification 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.[11]

One key property of DAGs is that they are falsifiable through testable implications over the observed distributions, including conditional independence relationships between variables in the model.[12] 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: (a) p contains at least one arrow-emitting node that is in Z, or (b) 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 as $X \perp\!\!\!\perp Y | Z$.[13]*

Conditional independence licensed by d-separation (d stands for "directional") holds for any distribution $P(v)$ over the variables in the model that is compatible with the causal assumptions encoded in the graph. Remarkably, this is true *regardless* of the parametrization of the arrows. An example is given in Figure 2b, where the path $A \rightarrow D \leftarrow B \rightarrow C$ 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 $\{\emptyset\}$ as well. D acts as a so-called *collider* node in this path, because of two arrows pointing into it. Therefore, according to the second condition of Definition 2.2, the path between A and C is blocked without any conditioning. Conversely, when conditioned on, a collider would open up a path that has

---

[10]Note that "nonparametric" in the artificial intelligence literature refers to the absence of assumptions involving error distributions *as well as* constraints over the form of the structural functions in the SCM. Instead, the shared features assumed to be available across structural systems are topological, that is, exclusion and independence restrictions are encoded in the causal graph. This difference in terminology should be kept in mind for what follows.

[11]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 Kumor et al. (2020).

[12]Historically, DAGs were first introduced to the AI literature in the early 1980s 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).

[13]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*: $A \rightarrow B \rightarrow C$.

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.[14] 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

$$A \perp\!\!\!\perp B; \qquad A \perp\!\!\!\perp C; \qquad A \perp\!\!\!\perp E|D; \qquad B \perp\!\!\!\perp E|D;$$
$$C \perp\!\!\!\perp D|B; \qquad C \perp\!\!\!\perp E|D; \qquad C \perp\!\!\!\perp E|B. \tag{2.2}$$

These independence relations can be checked using statistical hypothesis testing, and if rejected, the hypothesized model could be refuted and/or revised. 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 and criticism.

Conditional independence assumptions are one of the main building blocks 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 be easily automated (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.[15] As alluded to earlier, this action is denoted by a mathematical operator

---

[14]Anand et al. (2023) prove d-separation as well as further identification results for so-called *cluster causal diagrams* (C-DAGs), which allow for the partial specification of relationships among variables based on limited prior knowledge.

[15]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.). For a more technical discussion on the semantics and inevitability of the assumptions encoded in such models, please refer to (Bareinboim et al., 2022).

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 two probability distributions associated with M_{x_0} and M_{x_1} (i.e., the system under the intervention $do(X = x_1)$) captures the variations in outcome Y that is the result of a causal impact of $\Delta x = x_1 - x_0$. A randomized control trial closely follows this idea. Experimentation ties the value of a variable to the outcome of a coin flip (or randomization device), 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, Definition 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

$$
\begin{aligned}
Y_{x_0} &\leftarrow f(x_0, Z, U_y), \tag{2.5} \\
Y_{x_1} &\leftarrow f(x_1, Z, U_y), \tag{2.6}
\end{aligned}
$$

which immediately follow from M_{x_0} and M_{x_1}, respectively. In other words, counterfactuals are *derived* from first principles in SCMs and the corresponding causal mechanisms, 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.[16]

DEFINITION 2.3. *(Observational identifiability; Pearl, 2000) Let $Q(M)$ be any computable quantity of a model M. Q is identifiable (ID, 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 encoded in G, we have*

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

This definition requires that for any two (unobserved) SCMs M_1 and M_2, if their induced

---

[16]See Matzkin (2007) and Lewbel (2019) for related definitions of identifiability used in econometrics; see also footnote 10.

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 derived from it, such as risk ratios, odds ratios, quantile effects, etc.) can be computed as[17]

$$E[Y|do(X = x_1)] - E[Y|do(X = x_0)] = \sum_y y[P(y|do(x_1)) - P(y|do(x_0))]. \qquad (2.8)$$

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"*. 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 – backdoor and frontdoor adjustment – 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 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 labour 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 relatively higher adverse health effects, which ultimately reduce life-time earnings. Finally, E represents a set of socioeconomic variables that influence both the probability to graduate from college as well as individuals' future earning

---

[17]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.
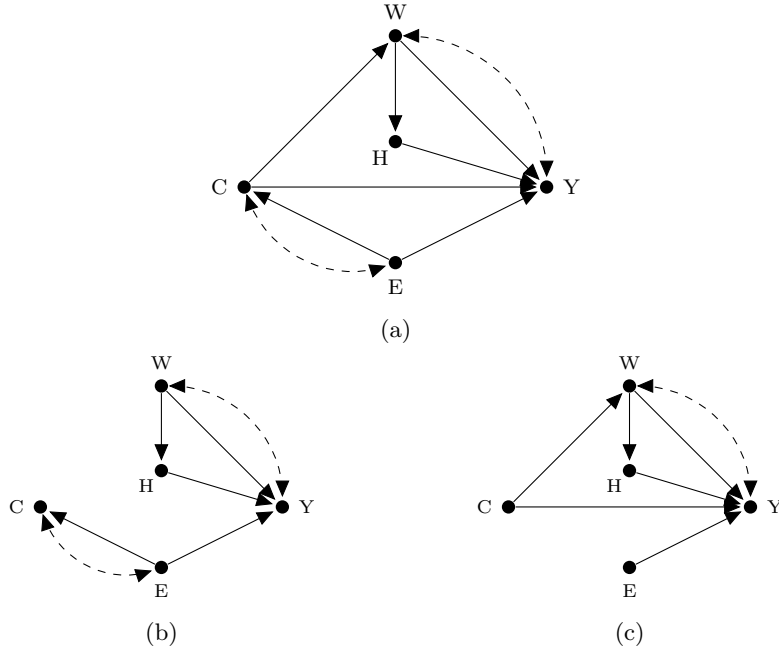
Figure 3: A graphical causal model G for estimating the college wage premium (together with the mutilated graphs $G_{\underline{C}}$ (b) and $G_{\overline{C}}$ (c) used in the do-calculus derivations in Section 3.3).

potentials. Dashed bidirected arrows depict unmeasured common causes that lead to a dependence between the background factors U associated with 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 directed acyclic graph 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 the modified graph $G_{\underline{C}}$ for the college wage premium example, where the edges $C \to Y$ 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., they "enter through the backdoor") create an association between X and Y that is not due to any causal influence exerted by X.[18] By adjusting for variables along these paths, this spurious association can be eliminated such that only the causal effect of X on Y remains.

In the particular example of Figure 3a, the set $Z = \{E\}$ satisfies the backdoor criterion

---

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

and is thus an admissible adjustment set.[19] 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 $C \rightarrow W \leftarrow\!\!-\!\!-\!\!-\!\!\rightarrow Y$, and thus would open, or unblock, this path when conditioned on. As a consequence, adjusting for W would inject bias in this example, creating a non-causal (spurious) correlation between C and Y, and would therefore be a serious mistake.

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

THEOREM 3.1. *(Backdoor adjustment criterion) If a set of variables Z 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*

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

Practically speaking, estimation can be carried out by propensity score matching (Rosenbaum and Rubin, 1983; Heckman et al., 1998), inverse probability weighting (Horvitz and Thompson, 1952; Robins, 1999), deep neural networks (Shi et al., 2019), or weighted empirical risk minimization (Jung et al., 2020), among other efficient estimation methods. If the cardinality of Z is high, regularization techniques such as the double machine learning framework by Chernozhukov et al. (2018) can be applied, including for arbitrary graph structures (Jung et al., 2021).

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

THEOREM 3.2. *(Counterfactual interpretation of backdoor; Pearl, 2009) If a set of variables Z satisfies the backdoor criterion 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 on identifying biasing paths and admissible sets (Definition 3.1), 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 4, which presents a model of input additionality of R&D subsidies. The model stipulates that research grants ($RDSUB$) are assigned based on the quality of projects with which firms apply ($PQUAL$). In addition, young (AGE) as well as small and medium-sized enterprises (defined as having fewer than 250 employees in the EU, EMP) often receive preferential treatment under many policy regimes (Howell, 2017; Hünermund and Czarnitzki, 2019). Other important covariates in this context are whether firms are exporters (EXP, Peters et al., 2022), the degree to which they are

---

[19]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.

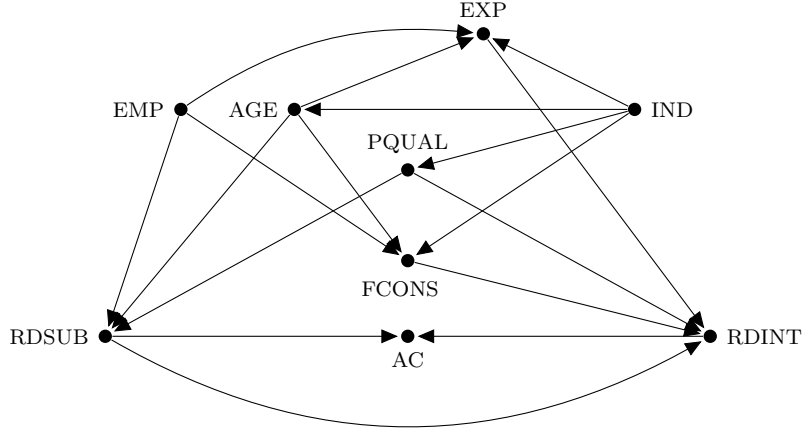**Figure 4**: A graphical model of R&D subsidy additionality. AC: absorptive capacity, AGE: firm age, EMP: number of employees, EXP: exporter, $FCONS$: financial constraints, IND: industry, $PQUAL$: project quality, $RDINT$: R&D intensity, $RDSUB$: R&D subsidy.

financially constrained ($FCONS$, Hottenrott and Peters, 2012), their level of absorptive capacity (AC, Cohen and Levinthal, 1989), and the industry they are operating in (IND). The outcome of interest is whether R&D subsidies are able to raise firms' R&D intensity ($RDINT$), measured as R&D expenditures in percentage of sales.

The two smallest (minimum) admissible adjustment sets for identifying the effect of $RDSUB$ on $RDINT$ in Figure 4 are given by

$$\{EXP, FCONS, PQUAL\} \quad \text{and} \quad \{AGE, EMP, PQUAL\}. \tag{3.3}$$

This example illustrates that it is neither necessary nor sufficient to adjust for all covariates in a model. The analyst could, for example, decide that financial constraints at the firm-level are too difficult to measure and instead proceed with the second admissible adjustment set.

At the same time, it would be a mistake to condition on the node AC. Absorptive capacity is the result of firms' R&D investments (Cohen and Levinthal, 1990, p. 141). Furthermore, there is evidence that R&D subsidies are able to increase absorptive capacity through a process that is commonly referred to as behavioral additionality (Clarysse et al., 2009). Taken together, these two assumptions turn AC into a collider, which, according to the d-separation criterion (Definition 2.2), would open up the path between $RDSUB$ and $RDINT$ and lead to bias if conditioned on. The largest admissible adjustment set is thus given by $\{AGE, EMP, EXP, FCONS, IND, PQUAL\}$, which could be preferred over a smaller set for reasons of estimation efficiency (White and Lu, 2011; Cinelli et al., 2022). These intricacies are hard to spot without the use of a causal diagram. We therefore regard DAGs as a useful complement to reduced-form approaches that mainly rely on verbal theorizing. More formalized approaches can likewise benefit from the visualization and automation properties of causal diagrams, which we will further describe below.

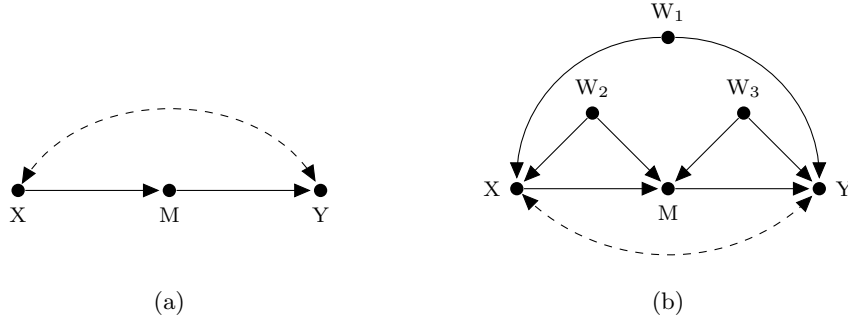*3.2. Frontdoor adjustment in the presence of unmeasured confounders*

(a) (b)

Figure 5: The frontdoor criterion.

Identification via backdoor adjustment requires that all backdoor paths can be blocked by a set of observed nodes, which might not always be feasible. In situations where no set of observables is backdoor-admissible, another (somewhat less familiar to economists) identification strategy might be applicable. Figure 5a presents an example in which adjusting for a set of observable variables is not sufficient to close all backdoor paths between X and Y. For any possible adjustment set, there are unobserved confounders remaining in the graph, represented by the bidirected arc $X \leftdashedarrow Y$. At the same time, the entire effect of X is assumed to be mediated by another observed variable M and there are no unobserved confounders connecting M with X and Y. These assumptions may be plausible in settings in which a test result M provides a noisy signal for the presence of a characteristic X. For example, congenital anomalies are routinely tested via ultrasound screenings during pregnancy. However, these screenings exhibit a positive type-1 and type-2 error rate (Debost-Legrand et al., 2014). If diagnostic missclassifications occur randomly, or are related to observables such as maternal BMI, confounding at the mediator can be controlled for. Full mediation in this case implies that congenital anomalies will only affect the likelihood of an outcome Y, such as the decision to terminate the pregnancy, if they are detected during an ultrasound screening.[20]

In this setting, the causal effect $P(y|do(x))$ is identifiable with the help of the following criterion (generalizing Theorem 2 in Pearl, 1995).

DEFINITION 3.2. *(Conditional frontdoor criterion) A set of variables M is said to satisfy the conditional frontdoor criterion (frontdoor, for short) relative to a triplet (X, Y, W) if: (a) M intercepts all directed paths from X to Y, (b) there is no unblocked backdoor path from X to M given W, and (c) all the backdoor paths from M to Y are blocked by $\{X, W\}$.*

THEOREM 3.3. *(Conditional frontdoor adjustment) If a set of variables M satisfies the*

---

[20]In case other prenatal testing methods than ultrasound are applied that could lead to the detection of congenital anomalies, they should be included as additional mediators in the model. By contrast, if congenital anomalies affect other unobservable markers that lead to pregnancy termination, the assumption of full mediation would be violated due to a direct (from the perspective of the analyst) effect $X \to Y$.

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

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

Frontdoor adjustment amounts to a sequential application of the backdoor criterion. In Figure 5b, first, the effect of X on M can be identified by adjusting for W_2. Second, the backdoor path $M \leftarrow X \leftarrow\text{-}\text{-}\text{-}\text{-}\rightarrow Y$, which remains open after adjusting for W_1 and W_3, can be blocked by conditioning on X, to identify the effect of M on Y. 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 in the presence of direct unobserved confounding between treatment and outcome (i.e., 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.[21]

3.3. Causal calculus and the algorithmatization of identification strategies

The backdoor and frontdoor criteria offer simple graphical identification rules that are easy to check in the causal diagram. However, while definitely important, they only represent a very 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 that 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.

RULE 1. *(Insertion/deletion of observations)*
$$P(y|do(x), z, w) = P(y|do(x), w) \qquad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}.$$
(3.5)

RULE 2. *(Action/observation exchange)*
$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \qquad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}.$$
(3.6)

RULE 3. *(Insertion/deletion of actions)*
$$P(y|do(x), do(z), w) = P(y|do(x), w) \qquad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\overline{Z(W)}}},$$
(3.7)

---

[21] Glynn and Kashin (2017) present an interesting application of the frontdoor criterion (FDC) for evaluating the effect of the National Job Training Partnership Act program (Heckman et al., 1997) on earnings by complementing the FDC with a difference-in-differences-type identification approach that tackles potential bias stemming from unobserved confounders between M and Y. Bellemare et al. (2022) apply the FDC to study how ride sharing affects tipping behavior on popular ride-hailing apps such as Uber and Lyft.

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 $G_{\overline{X}}$, i.e., for any interventional distribution $do(X)$. 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 passively observed $Z = z$. 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 conditions a manipulation of Z does not affect the probability of Y. Roughly speaking, 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.[22]

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 provides the basis for consistent estimation of Q from nonexperimental data. In the following, we demonstrate this process by showing a step-by-step do-calculus derivation for the causal effect of C on Y in the college wage premium example in Figure 3a. Notice that there are two backdoor paths in Figure 3a, which can both be blocked by E. By the law of total probability, $P(y|do(c))$ can be written as

$$P(y|do(c)) = \sum_e P(y|do(c), e)P(e|do(c)). \tag{3.8}$$

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

$$P(y|do(c), e) = P(y|c, e). \tag{3.9}$$

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

$$P(e|do(c)) = P(e). \tag{3.10}$$

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 – in principle – be estimated from nonexperimental data.

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, 2006; Huang and Valtorta, 2006; Bareinboim and Pearl, 2012a; Lee et al., 2019) from a combination of observational and experimental data. Soundness assures that an answer returned by do-calculus is correct. Completeness means that do-calculus is guaranteed to return a solution for the identification problem, whenever such a solution exists. 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 only consisting of 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 making assumptions about the distribution of the background factors U. In fact, this result can also be seen algorithmically,

---

[22]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 provided with the proofs in Pearl (1995).

which allows one to fully automate the often tedious task of transforming causal effect queries into do-free expressions. This way, the identification of causal effects becomes a straightforward exercise that can be solved with the help of a computer (Tian and Pearl, 2002a).[23]

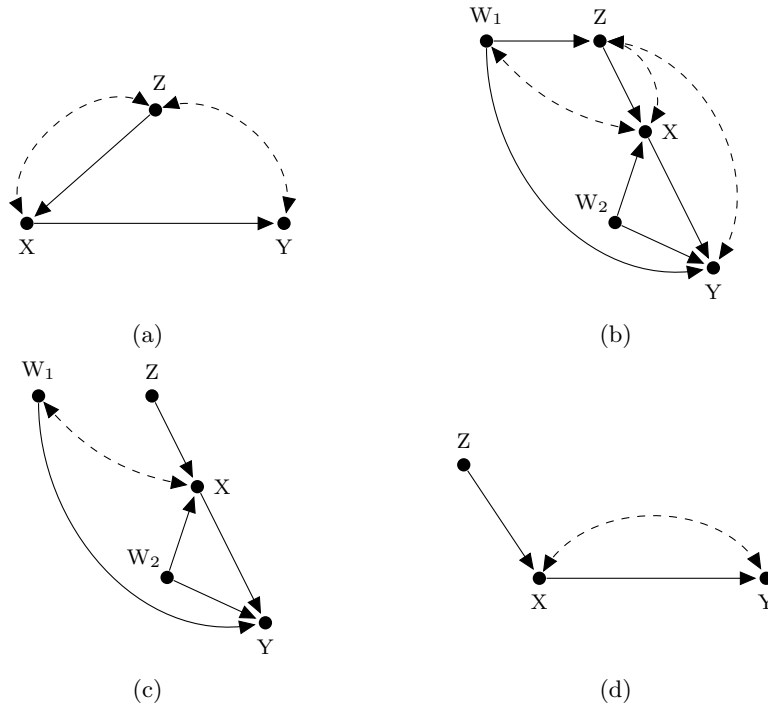### 3.4. Identification by surrogate experiments

Figure 6: Identification problems involving 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. Such surrogate experiments are sometimes referred to as *"encouragement designs"* in economics (Duflo et al., 2008).

Consider the situation in Figure 6a, in which X represents participation in a financial support program that allows job seekers to reimburse private expenses incurred for travel and materials during the application process. Y measures the employment status of an individual six months after the finish of the program. Finally, Z is an indicator for

---

[23]More recently, for effects that are provably not point identifiable, very general machinery for partial identification has been developed that can be applied from any causal diagram and any arbitrary combination of observational and experimental distributions (Zhang et al., 2022).

whether an individual is aware of the existence of the program. For simplicity, we assume that the financial incentives provided by the program are sufficiently attractive, such that awareness is the only parent node of X in the diagram (but the relationship is not deterministic). However, awareness is itself influenced by unobservables that create an unblocked backdoor path $X \leftarrow Z \leftarrow\!\!--\!\!\rightarrow Y$. Unfortunately, adjusting for Z is not sufficient, as this would open up the path $X \leftarrow\!\!--\!\!\rightarrow Z \leftarrow\!\!--\!\!\rightarrow Y$ on which Z is a collider. Thus, $P(y|do(x))$ is not identifiable via backdoor adjustment in this example. Program participation cannot be forced, which rules out a direct manipulation of X. Nonetheless, if the analyst is able to manipulate awareness Z instead, e.g., by sending information about the program only to a randomly selected group of individuals, identification becomes possible.[24] We will illustrate such a strategy involving auxiliary experiments on ancestor nodes of a treatment X in the following.

To make the theory of surrogate experiments in causal diagrams more concrete, Figure 6b 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 $X \leftarrow\!\!--\!\!\rightarrow Z \leftarrow\!\!--\!\!\rightarrow Y$, 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 refinement of the identification problem formulated earlier (see Definition 2.3).

DEFINITION 3.3. *(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 z-identifiable (zID, 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 3.4. *(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 zID 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; Lee et al., 2019, Theorem 3); i.e., it reaches a solution to the *zID* 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.

---

[24]Perfect manipulation of Z implies that individuals cannot learn about the existence of the program via different channels and that there is no communication between individuals, which may or may not be plausible given the context.

THEOREM 3.5. *(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 zID in G if one of the following conditions hold:

  (a) Q is identifiable in G; or
  (b) There exists $Z' \subseteq Z$ such that the following conditions hold,

      (i) X intercepts all directed paths from Z' to Y, and
      (ii) Q is identifiable in $G_{\overline{Z'}}$.

Condition *(a)* is the base case for when standard identifiability is reached. Whenever
this is not the case, condition *(b:i)* requires that all directed paths from Z to Y are
blocked by X. This means that Z has no direct 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 *(b:ii)* 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 6b 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 3.5 can be
applied. Note that all the directed paths from Z to Y are blocked by X, which satisfies
condition *(b:i)*. It is also the case that in the graph $G_{\overline{Z}}$ (see Figure 6c), the set $\{W_1, W_2\}$
is backdoor admissible (by Theorem 3.1), which in turn satisfies condition *(b:ii)*. After
all, the effect $P(Y = y|do(X = x))$ is identifiable and given by the expression

$$\sum_{w_1, w_2} P(Y = y|do(Z = z), X = x, w_1, w_2) P(w_1, w_2|do(Z = z)). \qquad (3.11)$$

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 of the latest results).

$\mathcal{Z}$-identification exploits experimental variation in a surrogate variable that causally
affects the treatment of interest. It thus bears close resemblance to instrumental variable
(IV) estimation (Wright, 1928). The two are not equivalent though. Take the canonical
IV setting with an exogenous instrument depicted in Figure 6d. In contrast to Figure
6a, there is an unobserved confounder directly connecting treatment and outcome. As a
result, $P(y|do(x))$ is not *zID* in this graph, because the bidirected arc between X and Y
violates condition *(b:ii)* of Theorem 3.5.[25]

The fact that $P(y|do(x))$ remains unidentifiable in Figure 6d 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, 1997). Introducing
additional functional form restrictions, such as monotonicity or linearity, would allow one
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 *zID*, the entire post-
interventional distribution, including the average treatment effect, is computable from

---

[25]Theorem 3.5 is only sufficient, but not necessary. Nonetheless, z-identification can be proved to be
impossible for the graph in Figure 6d, following the general treatment developed in Lee et al. (2019).

data. Moreover, the solution concepts provided by Theorems 3.4 and 3.5 are applicable for arbitrary graphs, beyond specific settings such as Figure 6d.[26] Therefore, we consider z-identification to be 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. Beyond 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 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 larger than zero, which is captured by the selection indicator s_i. Economically, this expresses the idea that individuals will choose to remain unemployed if the market wage they are able to attain (determined by the vector of socioeconomic 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 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. (2020) point out another illustrative case.[27] They critique studies which attempt to estimate the extent of racial bias in policing using administrative data (Fryer, 2019). Problematic in this context is that individuals only appear in such records if police officers decide 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 accounting for the sampling selection mechanism. We will realize this goal by adding a new special variable called S to the graph. This variable 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

---

[26]There exist more refined strategies to identify effects beyond this graph and IVs, including Brito and Pearl (2002); Chen et al. (2016, 2017); Kumor et al. (2019, 2020).

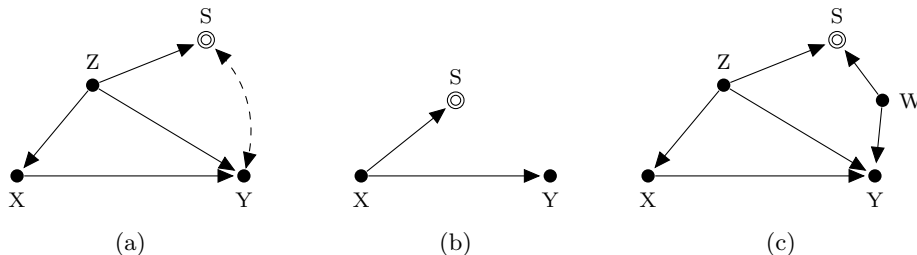[27]See Durlauf and Heckman (2020) for a similar argument.

Figure 7: Examples of selection diagrams.

probabilities, we will add an arrow from these variables to S, which will constitute the specification of the selection mechanism (Bareinboim and Pearl, 2012b).[28] Figure 7a depicts a DAG for the female labor supply example that has been augmented by such a selection node; the resulting graph is denoted by G_S. An individual's socioeconomic 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 to preserve the fully nonparametric nature of causal graphs also in this task. It refrains from introducing functional form assumptions (such as monotonicity or joint normality) related to the selection-propensity score $P(s_i|pa_i)$, as well as *a priori* assuming ignorability of the selection mechanism, which are the approaches most commonly taken in econometrics (Angrist, 1997.) Nevertheless, even with such a limited set of assumptions as a starting point, several positive results for the recoverability of causal effects from selection bias can be derived.

4.1. Recoverability of conditional distributions

As a first step to make progress, 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

---

[28]We will consider the case here where the sample selection nodes are only allowed to have incoming arrows, but will not emit arrows themselves.

strategy for also recovering do-expressions from selection bias (Bareinboim and Tian, 2015).

COROLLARY 4.1. *(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 7b 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 a causal effect 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 labour supply model of Figure 7a.

Selection-biased data complicate identification in observational studies because confounding and selection need to be addressed simultaneously. An example is given by the graph in Figure 8a. Consider the case of a group of entrepreneurs who are looking to crowdfund their business ideas. A researcher is interested in the effect of campaign success X on venture growth Y. Idea quality is captured by Q, which affects the quality of the crowdfunding campaign, C (possibly multivalued), that is presented to potential investors on a digital platform. To increase their chances of getting funded, some entrepreneurs take part in a training workshop. This decision, W, is not necessarily random, which is reflected by the bidirected dashed arc between X and W in the graph. In this example, a problem of selection arises, because the sample that is available to the researcher contains disporportionally many workshop participants, for whom contact addresses were most easily obtainable ($W \to S$). Finally, Z denotes a set of other confounders that the researcher wishes to control for.[29]

Without selection bias, the researcher would have the choice between two backdoor admissible adjustment sets: $\{Z, W, C\}$ and $\{Z, Q\}$. However, with preferential selection into the sample, recoverability can only be obtained with the latter. That is, because in the adjustment formula (Equation 3.1), the prior distribution of the adjustment set needs to be recovered as well, and $\{Z, Q\}$ is the only conditioning set that is marginally d-separated from S (readers are encouraged to check). Thus, following the strategy dictated by Corollary 4.1, the estimable backdoor adjustment expression in this example is

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

4.2. A general solution for recovering from selection bias

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 4.1 with respect

---

[29]Figure 8a encodes the assumption that the effect of campaign quality on venture growth is negligible, e.g., because the research question concerns equity crowdfunding in a business-to-business market. Entrepreneurs might additionally use crowdfunding campaigns as an advertising opportunity for their products. Insofar as this gives rise to a direct effect $C \to Y$, recoverability should be determined with this added assumption.
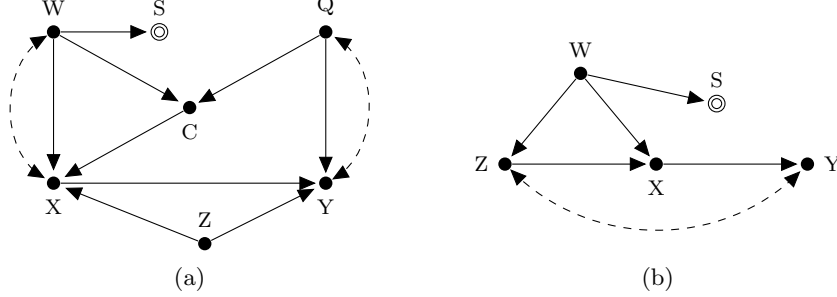
Figure 8: More challenging examples for recovering from selection bias.

to do-expressions. This is exemplified by the graph in Figure 8b. Due to unobserved confounders between Z and Y, and the fact that Z is a collider on the path $X \leftarrow W \rightarrow Z \leftarrow\!\text{-}\!\text{-}\!\text{-}\!\rightarrow Y$, identification via the backdoor criterion would require to adjust for both Z and W, which will close all the backdoor paths. However, $\{Z, W\}$ is not d-separable from S (W has a direct arrow to S), and an attempt to apply Corollary 4.1 will thus fail. Nevertheless, and perhaps surprisingly, $P(y|do(x))$ can still be recovered in Figure 8b with the help of do-calculus using a slightly more sophisticated approach.[30] 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

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

Finally, since $(Z \perp\!\!\!\perp X|W, S)$ 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 data are observational and always contain $S = 1$, given that the samples were selected preferentially. Taken together, this ensures recoverability of the target interventional distribution.

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

---

[30]The following do-calculus derivations are shown in more detail, with corresponding subgraphs depicted alongside, in Appendix A.1.

4.3. 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 important socioeconomic variables can often be measured without bias, e.g., from population-level statistics. Interestingly, this is the case in the original Heckman selection model, where number of children, household assets, husband's wage, labour market experience, and education are observed both for working and non-working women (Heckman, 1976). To illustrate how this can facilitate recoverability, we revisit the example from Figure 7a, but now assume that the common parent node of wages Y and the selection node S is observable as W (see Figure 7c, 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 3.1), 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 4.1 is not applicable. However, if in addition to the selection-biased 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 4.1. *(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:*

1. *Z^+ blocks all backdoor paths from X to Y in G_S;*
2. *X and Z^+ block all paths between Z^- and Y in G_S, namely, $(Z^- \perp\!\!\!\perp Y|X, Z^+)$;*
3. *X and Z block all paths between S and Y in G_S, namely, $(Y \perp\!\!\!\perp S|X, Z)$; and*
4. *Z and $Z \cup \{X, Y\}$ are measured in the unbiased and biased studies, respectively.*

The following theorem can then be proved.

THEOREM 4.2. *(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. (2019b) 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 domains is a fundamental problem in causal inference. Experiments are usually conducted in different contexts than those in which the lessons drawn from them are supposed to be applied. 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 artificial intelligence literature, while social scientists usually refer to it as *"external validity"* (Shadish et al., 2002; Bareinboim and Pearl, 2013b; Pearl and Bareinboim, 2014). 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 a 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 academic achievements, at least in the short-run. Interestingly, however, while treatment effects on mathematics 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. By contrast, baseline skill levels in mathematics did not differ significantly. The remedial

education program, which targeted only the most basic competencies in the curriculum, was therefore equally effective for mathematics skills.
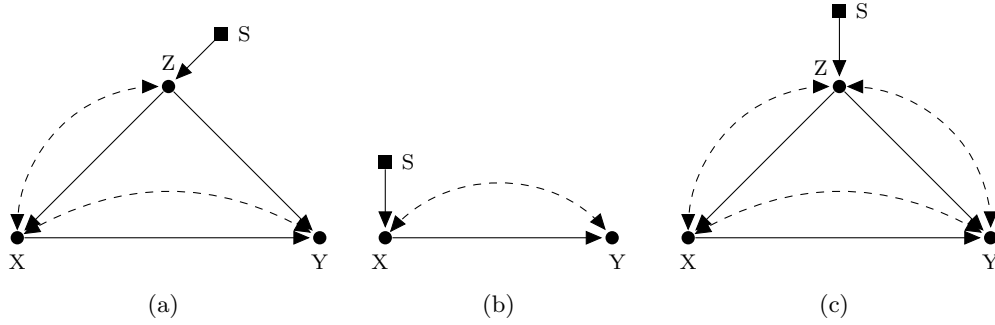

Figure 9: Examples of selection diagrams for the transportability task.

The graph in Figure 9a 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^*). However, we are aware 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 domain knowledge ex-ante 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 sources of structural discrepancies across domains that threaten transportability. Switching between two populations Π and Π^* is then captured by conditioning on different values of S.[31] Next, we define the joint graphical representation of the corresponding structural models in the source and target populations, which is required to establish 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: (a) every edge in G is also an edge in D; and (b) D contains an extra edge $S_i \to V_i$ 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 nodes in the graph, to express the notion that the two populations are maximally structurally heterogeneous (i.e., there are no 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,

---

[31]For clarity, S-nodes invoked for transportability are depicted by squares (■), in order to distinguish them from the selection bias case. Also note that now S is emitting arrows, whereas selection nodes indicating preferential inclusion into the sample only receive arrows.

which allows to transport experimental results obtained in a source Π to another target domain Π^*, where only passive observation is possible.[32]

THEOREM 5.1. *(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 $(Y \perp\!\!\!\perp S|Z, X)_{D_{\overline{X}}}$.*

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 5.3. *(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).*

Syntactically, 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). Semantically, this separation of an S-node indicates that the target distribution is insensitive to the structural disparities represented in the selection diagram, and, therefore, the effects are invariant across populations. Looking at the selection diagram in Figure 9a, 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.

$$P^*(y|do(x)) = P(y|do(x), s), \tag{5.1}$$

$$= \sum_z P(y|do(x), z, s)P(z|do(x), s), \tag{5.2}$$

$$= \sum_z P(y|do(x), z)P(z|s), \tag{5.3}$$

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

The first equation follows from the definition that distributions in the target domain Π^* are denoted by conditioning on S. The second line follows from conditioning and summing over Z. The third line is derived by using the s-admissibility of Z and recognizing the

---

[32]Following Definition 5.2, both domains Π and Π^* 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) developed a complete algorithm for this task. We will not further pursue this topic in what follows and refer the interested reader to the respective papers.

fact that X is a child of Z and, therefore, exerts no causal influence on Z (formally, rule 3 of do-calculus applies). The last line is then just a restatement.

As long as Figure 9a 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 for Mumbai is required. This result is stated in its full generality in the following corollary.

COROLLARY 5.1. *(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 5.1 that any S variable that points into X can be ignored. The causal effect $P(y|do(x))$ is thus directly transportable in Figure 9b. 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 9c provides a cautionary tale in that regard. Apart from the unobserved confounder between Z and Y, it is identical to Figure 9a. Here, however, s-admissibility is violated because conditioning on Z would open up the path $S \to Z \leftarrow\!\!\dashrightarrow Y$. It can be shown that transporting $P(y|do(x))$ is impossible in this selection diagram. The example thus illustrates how 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) is well known in the econometrics literature (Hotz et al., 2005; Dehejia et al., 2021; Andrews and Oster, 2019). Most commonly, approaches in this area build on 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, as the example in Figure 9c demonstrates. 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 5.1.

THEOREM 5.2. *(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 10a. 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
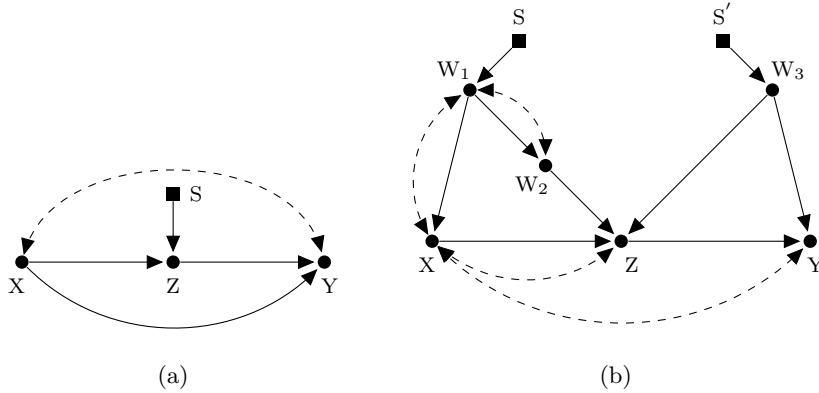
Figure 10: Examples of more challenging transportability tasks, including S-nodes on post-treatment variables.

occur for RCTs in development economics, where the success of a policy is partly dependent on the level of care with which a program is implemented. Duflo et al. (2008) discuss the problem that pilot trials often employ particularly highly qualified program officials. This is difficult to replicate once the program is supposed to be scaled up, which threatens the generalizability of these pilot studies.[33]

Gordon et al. (2019) 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 cannot easily be controlled by 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 10a, 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

---

[33]Similarly, Banerjee et al. (2017) discuss how market equilibrium effects can be an obstacle for the generalizability of pilot studies. A large, nationwide experiment may have an effect on wages and prices of nontradable goods such as land, which is likely to be negligible in smaller RCTs. These intermediate variables might be important for the overall outcome of a program and could thus lead to different expected results in a small versus a larger study population.

$P^*(z|do(x))$ 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 10a, this can be achieved by setting $P^*(z|do(x)) = P^*(z|x)$ (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 5.2 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 10b, in which the correct transport formula is given by

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

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-identifiability. 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 5.2, the rules of do-calculus can be used to transfer causal knowledge from surrogate experiments in the following way.

THEOREM 5.3. *(Bareinboim and Pearl, 2013a, 2014) 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 and only 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 5.3 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 diagram D and a list of variables that were manipulated 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 that allow to combine empirical knowledge from several source domains. Dehejia et al. (2021) 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. Naïve pooling of results 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.[34]

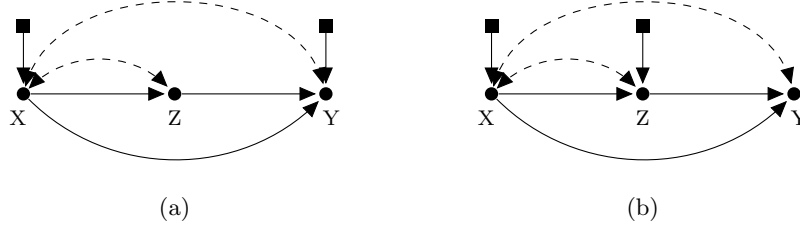(a)   (b)

Figure 11: Selection diagrams representing two heterogeneous source domains.

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 11, in which panel (a) depicts the selection diagram that corresponds to source domain π_a, while panel (b) refers to π_b. Square nodes indicate where discrepancies between the target domain π^* and the source domains arise.[35] 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 11 is a simple extension of a graph that was presented earlier (see Figure 10a). In contrast to before, the unobserved confounder between X and Z (denoted by the dashed

---

[34]Meta-transportability is related to the idea of "data combination" presented e.g. 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.

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

bidirected arc $X \leftarrow\!-\!-\!-\!\rightarrow Z$), which was added to the diagram, now renders individual transportability impossible.[36] Interestingly though, μ-transportability is feasible by combining information from both source domains. To see this, note that the post-intervention distribution in the target domain π^* can be written as

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

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

where the second line follows from rule 2 of do-calculus, since $(Z \perp\!\!\!\perp Y|X)$ in $D_{\overline{X}\underline{Z}}$.[37] 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 $D_{\overline{X}}^{(a)}$. And $P^*(y|do(x),do(z))$ is directly transportable from π^b, since $(S \perp\!\!\!\perp Y)$ in $D_{\overline{X,Z}}^{(b)}$. 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

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

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 in Card et al. (2010), Dehejia et al. (2021), and Meager (2019). A drawback of standard meta-analytical approaches is, however, that they do not incorporate knowledge about domain heterogeneities related to causal mechanisms and background factors. Instead, they attempt to "average out" differences across populations.[38] By contrast, the transportability techniques we have presented make it transparent how discrepancies between study results arise and how they can nonetheless be leveraged to identify a

---

[36]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), $X \leftarrow\!-\!-\!-\!\rightarrow Z$ prevents to set $P^*(z|do(x)) = P^*(z|x)$, which was instrumental for establishing transportability following equation (5.8).

[37]These do-calculus derivations are shown in detail, with corresponding subgraphs depicted next to it, in Appendix A.2.

[38]To the extent that these studies consider domain heterogeneity, this is done in a purely statistical fashion, without explicitly modelling structural differences across populations (Dehejia et al., 2021; Meager, 2019). This leaves open the question whether domains are actually structurally sufficiently similar for transportability to be feasible.

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 techniques thus bear the potential to save on discipline-wide data collection costs and to render causal inference a truly collective endeavor.[39]

6. CONCLUSION

From the end of the 1980s onwards, the artificial intelligence field 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 (Sloman, 2005). A human-like AI, therefore, needs to possess an internal representation of causality in order to mimic human behaviour and communicate with us in a meaningful way (Pearl and Mackenzie, 2018; Bareinboim et al., 2022). 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 various different data collection and knowledge contexts (Bareinboim and Pearl, 2016; Bareinboim et al., 2022).

We are convinced that the causal inference and data fusion techniques we have discussed in this paper also have a lot to offer to econometricians. Until today, the possibilities to completely automate the identification task, which is a necessary ingredient for causal machine learning, remain largely unexplored in econometrics. 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.[40] Moreover, graphical representations of structural causal models do not require the learner – whether artificial or human – to impose any dis-

---

[39]Other contributions to transportability theory have been made by Correa et al. (2019a), 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, unifying several of the techniques discussed in this section. Furthermore, these results have been extended to cover stochastic interventions, where the *sigma calculus*, a generalization of the do-calculus, has been introduced (Correa and Bareinboim, 2020a), and used to solve stochastic-transportability (Correa and Bareinboim, 2020b). More recently, these results were generalized for the case of transporting nested counterfactual from an arbitrary combination of observational and experimental distributions (Correa et al., 2022).

[40]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 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 AI field (Spirtes et al., 2001; Pearl, 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, 2022).

tributional or functional form restrictions on the underlying causal mechanisms under study. The approach remains fully nonparametric and thus naturally incorporates, e.g., treatment effect heterogeneity. At the same time, crucial identification conditions, such as conditional independence, are derived from the properties of the underlying structural model, rather than being assumed to hold *a priori*. Causal graphical models thus combine the flexibility and accessibility of potential outcomes with the analytical rigor of structural econometrics (Rust, 1987; Keane and Wolpin, 1997; Heckman and Vytlacil, 2007). These properties are of great value for applied empirical work. Economists should therefore feel encouraged to engage in a productive exchange with AI researchers for mutual benefit from the numerous useful tools for causal inference developed in both disciplines.

ACKNOWLEDGEMENTS

REFERENCES

Adams, P., M. D. Hurd, D. McFadden, A. Merrill, and T. Ribeiro (2003). Healthy, wealthy, and wise? tests for direct causal paths between health and socioeconomic status. *Journal of Econometrics 112*, 3–56.

Anand, T. V., A. H. Ribeiro, J. Tian, and E. Bareinboim (2023). Effect identification in cluster causal diagrams. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*.

Andrews, I. and E. Oster (2019). A simple approximation for evaluating external validity bias. *Economics Letters 178*, 58–62.

Angrist, J. D. (1997). Conditional independence in sample selection models. *Economics Letters 54*, 103–12.

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

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*, 1171–76.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives 31*(4), 73–102.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007). Remedying education: Evidence from two randomized experiments in india. *Quartely Journal of Economics 122*, 1235–64.

Bareinboim, E., J. D. Correa, D. Ibeling, and T. Icard (2022). On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556.

Bareinboim, E., S. Lee, V. Honavar, and J. Pearl (2013). Transportability from multiple environments with limited experiments. In *Advances in Neural Information Processing Systems (NeurIPS 2013)*.

Bareinboim, E. and J. Pearl (2012a). Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pp. 113–120.

Bareinboim, E. and J. Pearl (2012b). Controlling selection bias in causal inference. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, pp. 100–8.

Bareinboim, E. and J. Pearl (2012c). Transportability of causal effects: Completeness results. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*.

Bareinboim, E. and J. Pearl (2013a). Causal transportability with limited experiments. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013)*.

Bareinboim, E. and J. Pearl (2013b). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference 1*, 107–34.

Bareinboim, E. and J. Pearl (2013c). Meta-transportability of causal effects: A formal approach. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*.

Bareinboim, E. and J. Pearl (2014). Transportability from multiple environments with limited experiments: Completeness results. In *Advances of Neural Information Processing Systems (NeurIPS 2014)*.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences 113*, 7345–52.

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

Bareinboim, E., J. Tian, and J. Pearl (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*.

Basman, R. L. (1963). The causal interpretation of non-triangular systems of economic relations. *Econometrica 31*, 439–48.

Bellemare, M. F., J. R. Bloom, and N. Wexler (2022). The paper of how: Estimating treatment effects using the front-door criterion. Working Paper.

Bentzel, R. and B. Hansen (1954). On recursiveness and interdependency in economic modeks. *Review of Economic Studies 22*, 153–68.

Bentzel, R. and H. Wold (1946). On statistical demand analysis from the viewpoint of simulatneous equations. *Skandivavisk Aktuarietidskrift 29*, 95–114.

Bongers, S., P. Forré, J. Peters, and J. M. Mooij (2021). Foundations of sturctural causal models with cycles and latent variables. *The Annals of Statistics 49*(5), 2885–2915.

Brito, C. and J. Pearl (2002). Generalized instrumental variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pp. 85–93.

Card, D., J. Kluve, and A. Weber (2010). Active labour market policy evaluations: A meta-analysis. *The Economic Journal 120*, 452–77.

Cartwright, N. (2007). *Hunting Causes and Using Them*. Cambridge, UK: 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*, pp. 1156–1164.

Chen, B., J. Pearl, and E. Bareinboim (2016). Incorporating knowledge into structural equation models using auxiliary variables. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 3577–3583.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*, 1–68.

Cinelli, C., A. Forney, and J. Pearl (2022). A crash course in good and bad controls. *Sociological Methods and Research*. forthcoming.

Clarysse, B., M. Wright, and P. Mustar (2009). Behavioural additionality of r&d subsidies: A learning perspective. *Research Policy 38*(10), 1517–1533.

Cohen, W. M. and D. A. Levinthal (1989). Innovation and learning: The two faces of r & d. *The Economic Journal 99*(397), 569–596.

Cohen, W. M. and D. A. Levinthal (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly 35*(1), 128–152.

Correa, J. and E. Bareinboim (2020a). A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*.

Correa, J. and E. Bareinboim (2020b). General transportability of soft interventions: Completeness results. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Correa, J., S. Lee, and E. Bareinboim (2021). Nested counterfactual identification from arbitrary surrogate experiments. In *Advances in Neural Information Processing Systems*, Volume 34.

Correa, J. D. and E. Bareinboim (2019). From statistical transportability to estimating the effect of stochastic interventions. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*.

Correa, J. D., S. Lee, and E. Bareinboim (2022). Counterfactual transportability: A formal approach. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*.

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 (AAAI 2018)*.

Correa, J. D., J. Tian, and E. Bareinboim (2019a). Adjustment criteria for generalizing experimental findings. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*.

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

Debost-Legrand, A., H. Laurichesse-Delmas, C. Francannet, I. Perthus, D. Lémery, D. Gallot, and F. Vendittelli (2014). False positive morphologic diagnoses at the anomaly scan: marginal or real problem, a population-based cohort study. *BMC Pregnancy and Childbirth 14*(112).

Dehejia, R., C. Pop-Eleches, and C. Samii (2021). From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics 39*, 217–43.

Duflo, E., R. Glennerster, and M. Kremer (2008). Using randomization in development economics research: A toolkit. In *Handbook of Development Economics*, Volume 4, Chapter 61. Elsevier.

Durlauf, S. N. and J. J. Heckman (2020). An empirical analysis of racial differences in police use of force: A comment. *Journal of Political Economy 128*.

Frölich, M. (2008). Parametric and nonparametric regression in the presence of endogenous control variables. *International Statistical Review 76*, 214–227.

Fryer, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy 127*.

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 (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science 38*, 193–364.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica 11*, 1–12.

Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research 12*, 317–37.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement 5*, 475–92.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–61.

Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*, 261–94.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies 64*, 605–54.

Heckman, J. J. and R. Pinto (2015). Causal analysis after Haavelmo. *Econometric Theory 31*, 115–51.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part 1: Causal models, structural models and econometric policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *The Hanbook of Econometrics*, Volume 6B, pp. 4779–874. Elsevier.

Hoover, K. D. (2004). Lost causes. *Journal of the History of Economic Thought 26*, 149–64.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–85.

Hottenrott, H. and B. Peters (2012, November). Innovative Capability and Financing Constraints for Innovation: More Money, More Innovation. *The Review of Economics and Statistics, 94*(4), 1126–1142.

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*, 241–70.

Howell, S. T. (2017). Financing innovation: Evidence from R&D grants. *American Economic Review 107*(4), 1136–1164.

Huang, Y. and M. Valtorta (2006). Pearl's calculus of interventions is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2006)*.

Hünermund, P. and D. Czarnitzki (2019). Estimating the causal effect of R&D subsidies in a pan-European program. *Research Policy 48*(1), 115–124.

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*, 356–69.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86*, 4–29.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature 58*(4), 1129–79.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*, 467–75.

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

Jaber, A., A. Ribeiro, J. Zhang, and E. Bareinboim (2022). Causal identification under markov equivalence: Calculus, algorithm, and completeness. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems.*

Jaber, A., J. Zhang, and E. Bareinboim (2018a). Causal identification under Markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence.*

Jaber, A., J. Zhang, and E. Bareinboim (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).*

Jaber, A., J. Zhang, and E. Bareinboim (2019). Causal identification under Markov equivalence: Completeness results. In *Proceedings of the 36th International Conference on Machine Learning.*

Jung, Y., J. Tian, and E. Bareinboim (2020). Learning causal effects via weighted empirical risk minimization. In *Advances in Neural Information Processing Systems*, pp. 12697–12709.

Jung, Y., J. Tian, and E. Bareinboim (2021). Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence.*

Keane, M. P. and K. I. Wolpin (1997). The career decisions of young men. *Journal of Political Economy 105*, 473–522.

Knox, D., W. Lowe, and J. Mummolo (2020). The bias is built in: How administrative records mask racially biased policing. *American Political Science Review 114*, 619–37.

Kumor, D., B. Chen, and E. Bareinboim (2019). Efficient identification in linear structural causal models with instrumental cutsets. In *Advances in Neural Information Processing Systems 32*, pp. 12477–12486.

Kumor, D., C. Cinelli, and E. Bareinboim (2020). Efficient identification in linear structural causal models with auxiliary cutsets. In *Proceedings of the 37th International Conference on Machine Learning.*

Lee, S., J. D. Correa, and E. Bareinboim (2019). General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI 2019).*

Lee, S., J. D. Correa, and E. Bareinboim (2020). General transportability – synthesizing observations and experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020).*

Levitt, S. D. and J. Porter (2000). Sample selection in the estimation of air bag and seat belt effectiveness. *Review of Economics and Statistics 83*, 603–15.

Lewbel, A. (2019). The identification zoo: Meanings of indetification in econometrics. *Journal of Economic Literature 57*(4), 835–903.

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

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. New York, NY: Springer.

Matzkin, R. L. (2007). Nonparametric identification. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, North Holland. Elsevier.

Matzkin, R. L. (2013). Nonparametric identification in structural economic models. *Annual Review of Economics 5*, 457–86.

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.

Nakamura, E. and J. Steinsson (2018). Identification in macroeconomics. *Journal of Economic Persepctives 32*(3), 59–86.

Neuberg, L. G. (2003). Review of 'causality: Models, reasoning, and inference'. *Econometric Theory 19*, 675–85.

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.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*, 669–709.

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

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

Pearl, J. (2013). Reflections on Heckman and Pinto's 'causal analysis after Haavelmo'. Technical Report R-420, UCLA.

Pearl, J. (2015a). Generalizing experimental findings. *Journal of Causal Inference 3*, 259–66.

Pearl, J. (2015b). Trygve haavelmo and the emergence of causal calculus. *Econometric Theory 31*, 152–9.

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 (AAAI 2011)*.

Pearl, J. and E. Bareinboim (2014). External validity: From do-calculus to transportability across populations. *Statistical Science 29*, 579–95.

Pearl, J. and D. Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. New York, NY: 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. *Journal of Machine Learning Research 18*, 8132–93.

Peters, B., M. J. Roberts, and V. A. Vuong (2022). Firm r&d investment and export market exposure. *Research Policy 51*(105228), 104601.

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. In J. J. Heckman and E. E. Leamer (Eds.), *The Handbook of Econometrics*, Volume 6B, pp. 5470–547. Elsevier.

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

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.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica 55*, 999–1033.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont: Wadsworth, Cengage Learning.

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 (NeurIPS 2019)*.

Shpitser, I. and J. Pearl (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. In *Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*.

Sloman, S. (2005). *Causal Models: How People Think about the World and Its Alternatives*. Oxford, UK: Oxford University Press.

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

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*, 417–27.

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 (UAI 2011)*.

Tian, J. and J. Pearl (2002a). A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI 2002)*.

Tian, J. and J. Pearl (2002b). A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI 2002)*.

Verma, T. and J. Pearl (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence (UAI1988)*.

White, H. and K. Chalak (2006). A unified framework for defining and identifying causal effects. *UCSD Department of Economics Discussion Paper*.

White, H. and X. Lu (2011). Causal diagrams for treatment effect estimation with application to efficient covariate selection. *Review of Economics and Statistics 93*, 1453–1559.

Wold, H. (1981). Review and current outlook. In *The Fix-point Approach to Interdependent Systems*, Amsterdam, Netherlands, pp. 1–36. North-Holland Publishing Company.

Wold, H. O. A. (1960). A generalization of causal chain models (part iii of a triptych on causal chain systems). *Econometrica 28*, 443–63.

Woodward, J. (2003). *Making Things Happen.* Oxford, UK: Oxford University Press.

Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils.* New York, NY: MacMillan.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research 20*, 557–85.

Wright, S. (1923). The theory of path coefficients. *Genetics 8*, 239–55.

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

Zhang, J., J. Tian, and E. Bareinboim (2022). Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*.

APPENDIX

This appendix shows 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.1. Selection bias example (Section 4, Figure 8b)
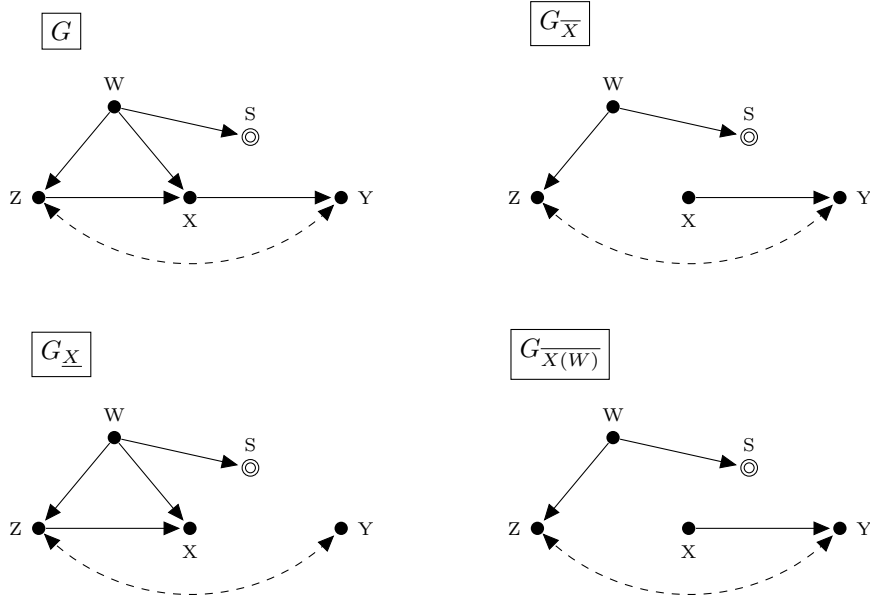
Consider the causal effect of X on Y in graph G. In graph $G_{\overline{X}}$, Z is a collider on the path connecting S and W with Y. Therefore, $(S, W \perp\!\!\!\perp Y)_{G_{\overline{X}}}$, and by the first rule of

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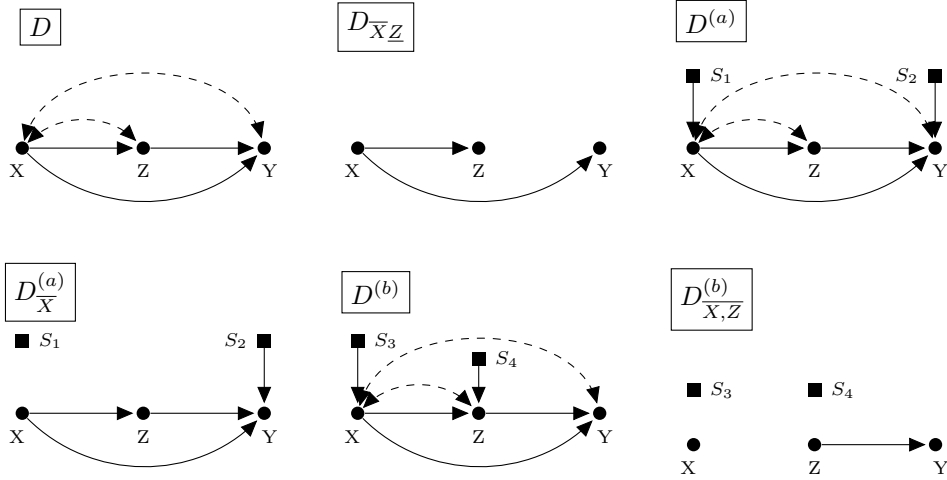
$$P(y|do(x)) = \sum_z P(y|x, z, w, S = 1)P(z|do(x), w, S = 1).$$

Finally, notice that there are no open paths between X and Z in $G_{\overline{X(W)}}$ (which is equivalent to $G_{\overline{X}}$, as X is not an ancestor of W). Thus, since $(Z \perp\!\!\!\perp X|W, S)_{G_{\overline{X(W)}}}$ (this independence holds both conditional and unconditional), rule 3 of do-calculus can be applied to the second term, such that

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

$\square$

A.2. M-Transportability example (Section 5.2, Figure 11)

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

$$P^*(y|do(x)) = \sum_z P^*(y|do(x), z)P^*(z|do(x)),$$

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

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 $D^{(a)}_{\underline{X}}$, 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 $D^{(b)}_{\underline{X,Z}}$, $P^*(y|do(x), do(z))$ 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

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

$\square$