A Course in the Calculus of Interventions

TBD

PRELIMINARY DRAFT —DO NOT CIRCULATE

May 9, 2023

Abstract

Empirical scientists often face the difficult task of drawing causal inferences from multiple, heterogeneous data sets, marred by confounding, sampling selection, missing data, and cross-population biases. To address these challenges, researchers must determine whether the available data, along with some domain knowledge about the data generating process (encoded as assumptions in a causal model), is sufficient to uniquely identify the target of inference—also known as an "identification problem." Recent developments in causal graphical models provide a general, non-parametric framework for representing, understanding, and resolving such problems. Central to these developments is a calculus of interventions—the *do*-calculus—a set of three simple syntactic rules that facilitate the manipulation of probabilistic statements involving interventions. This paper demonstrates the use of graphical models and the *do*-calculus through a series of illustrative examples covering many of the typical identification problems faced by empirical scientists, such as: (i) the identification of effects of single or multiple (sequential) interventions from observational data; (ii) the identification of these effects from a combination of multiple data sources, including observational and experimental data; (iii) recovering from selection bias and missing data; and, (iv) generalizing causal effects across domains. By making this content accessible to non-specialists, we hope to make it easier for researchers to understand and apply these tools in their own work.

1 Introduction

(Motivation here.)

2 Notation & Preliminaries

We assume that readers are familiar with the basic notions of probability theory, causal inference, and causal diagrams (in particular, path-blocking and d-separation). Many introductions to those topics have been published elsewhere.[1] Here we briefly set notation and review basic results needed to follow *do*-calculus derivations.

[1] Readers can find introductions to causal diagrams in a number of applied fields, such as sociology (Elwert, 2013; Morgan and Winship, 2015), economics (Hünermund and Bareinboim, 2019; Cunningham, 2021), psychology (Rohrer, 2018), epidemiology (Greenland et al., 1999; Hernán and Robins, 2020) and statistics (Pearl et al., 2009, 2016). For a quick review of probability theory, we recommend Pearl (2009, Section 1.1) and Pearl et al. (2016, Section 1.3). Book-length introduction to probability theory can be found in Ross (2010) and Blitzstein and Hwang (2015).

Probability Theory. We denote random variables by capital letters (e.g, X) and their realized values by lowercase letters (e.g, $X = x$). We use the abbreviation $P(y, x)$ to denote the probability of the event $P(Y = y, X = x)$. Most derivations rely on well-known results of probability theory, such as the *chain rule*,

$$P(y, x, z, w) = P(y|x, z, w)P(x|z, w)P(z|w)P(w);$$

Bayes' rule,

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)};$$

and, *the law of total probability*:

$$P(y) = \sum_{x} P(y|x)P(x).$$

Throughout the paper, and without loss of generality, we assume that random variables are discrete. Readers accustomed to continuous random variables should replace probability mass functions with probability density functions, and summations with integrals, where appropriate. Here we limit our analysis to probability distributions that are strictly positive, that is, $P(v) > 0$, for all possible events $V = v$.

Causal Diagrams. Causal diagrams provide a parsimonious, qualitative representation of the process that generates the data. In a causal diagram, capital letters (e.g, X) represent random variables and arrows, such as $X \to Y$, represent a (possible) direct causal effect of X on Y. All causal diagrams we are going to consider do not contain cycles, and are thus directed acyclic graphs (DAGs). Bidirected arrows, such as $X \leftrightarrow Y$, denote latent common causes of X and Y, and can be interpreted as shorthand notation for $X \leftarrow U \to Y$, where U is unobserved. Traditionally, causal diagrams are *non-parametric*, meaning that no assumptions are made regarding the functional form of the causal relationships, nor regarding the distribution of the random variables. In the spirit of this course, and slightly departing from the traditional graphical models literature, here we heavily use *causal terminology* to denote the relationship of variables in a causal diagram (e.g, cause, direct cause, effect, direct effect, common cause, common effect, mediator, etc)[2].

Path-blocking and d-separation. Crucial for this course is understanding the three main patterns of association that form the building blocks of a DAG, and when these are blocked or opened.

1. *Mediators* are patterns of the form $X \to Z \to Y$, meaning that X *causally* affects Y through the mediator Z. Conditioning on the mediator Z *blocks* (closes) this flow of association.

2. *Common causes* are patterns of the form $X \leftarrow Z \to Y$, meaning X and Y share a common cause (a confounder) Z. This induces a *non-causal* (or spurious) association between the two variables. Conditioning on the common cause Z *blocks* this flow of association.

3. *Common effects*, also known as *colliders*, are patterns of the form $X \to Z \leftarrow Y$, meaning that both X and Y share a common effect Z. In contrast with the two previous patterns, a common effect does not induce association between X and Y. However, conditioning on the common effect Z opens a *non-causal* (or spurious) association between the two variables.

[2]In the graphical models literature, the most common way to describe such relationships is using *kinship terminology*, such as parents, children, descendants, ancestors, etc. Here, to build causal intuition, we will mostly use causal terminology.

Finally, controlling for the effects of a variable is equivalent to "partially" controlling for that variable. Any arbitrary path from X to Y (consisting of a sequence of mediators, common causes, or common effects) will be blocked conditional on Z if, and only if, the conditioning set contains a common cause or mediator along the path, or if there is common effect in the path, the conditioning set does *not* contain that common effect, nor any of its subsequent effects. We say that Z d-separates X from Y in the graph G if Z blocks all paths from X to Y, written as $(Y \perp\!\!\!\perp X|Z)_G$; d-separation implies that Y and X are conditionally independent given Z (Pearl, 2009).

3 A calculus of interventions

In order to derive a calculus of interventions, we first need to mathematically define what an intervention is. One simple and appealing way to model interventions is to treat them as variables within the system (Pearl, 1994).

3.1 Interventions as variables

Let us introduce a new binary variable I_x, denoting the regime in which X was generated. For our purposes, this variable takes two values, $I_x = \emptyset$ and $I_x = do(x)$. The value $I_x = \emptyset$ denotes the traditional, "observational regime," where no intervention is performed. This is the usual type of data collected in observational studies. The value $I_x = do(x)$ denotes the "interventional regime," in which we set the value of X to x *by external force*. This would be the data we would collect, if we could perform *controlled experimentation* on X.

In this formulation, our target of inference—the **query**—is the distribution of Y when we experimentally set X to x, namely,

$$\mathcal{Q} = P(y|I_x = do(x)).$$

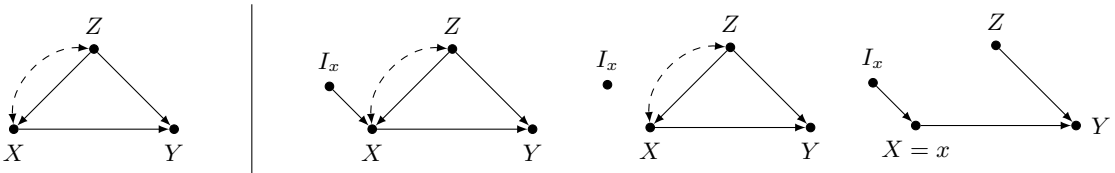
However, in the absence of experiments, we do not have any direct measurements of $P(y|I_x = do(x))$; instead, we only have passive observations of Y, X, and Z, under the no-intervention regime. Formally, then, the **data** available to us is,

$$\mathcal{D} = P(y, x, z|I_x = \emptyset).$$

Clearly, there is a gap between the data which is available to us and our target of inference. Without further assumptions, these distributions can be arbitrarily different. In other words, if the action $do(x)$ is allowed to arbitrarily change the system, then clearly nothing can be inferred from passive observations alone. We thus must specify how our intervention modifies the original system, and, most importantly, what remains *invariant across regimes*.

There are two key assumptions about our intervention $do(x)$ which will allow us to connect it with the observational regime: (i) first, we assume it is *effective*, meaning that when $I_x = do(x)$ this indeed implies that $X = x$, as desired; (ii) second, we assume that our intervention is *local*, meaning that only the local mechanism by which X takes its values is overruled by our actions—all other mechanisms remain the same.

These assumptions are easier to explain via an example. Consider the causal diagram of Figure 1a. In this model, X causes Y, and Z causes both X and Y. Additionally, there is a latent confounder between X and Z, represented by the dashed bi-directed arrow. Now to represent the action $do(x)$ in the system, we add the node I_x to G, as shown in Figure 1b. Note that I_x is only directly connected to X, in accordance with our assumption of locality. How do the different values of the regime indicator I_x change the model? When $I_x = \emptyset$, the system is unperturbed and we just have



(a) Model 1: G (b) $G' = G \cup \{I_x \to X\}$ (c) $G'_{I_X=\emptyset}$ (d) $G'_{I_x=do(x)}$

Figure 1: Interventions as Variables

the same graph as before in Figure 1c. When $I_x = do(x)$, the variable X does not "listen" to its original parent Z, and it is set by the external intervention to $X = x$, *regardless of the value of* Z, as shown in Figure 1d.

With this simple graphical representation, we have now determined what changes and what remains invariant after the intervention. Questions about invariance of probabilities under different regimes are now reduced to simple questions of graphical separation between the variable of interest and the intervention node. We have thus postulated our **model**, $\mathcal{M} = G$, and we can ask whether our query can be computed from the available data, given the assumptions encoded in the diagram. This is known as an identification problem.

It turns out that, in this example, the interventional distribution $P(y|I_x = do(x))$ can indeed be computed from passive observations $P(y, x, z|I_x = \emptyset)$, and the derivation goes as follows:

$$
\begin{aligned}
P(y|I_x = do(x)) &= \sum_z P(y|z, I_x = do(x))P(z|I_x = do(x)) && \text{Law of total probability} \\
&= \sum_z P(y|z, I_x = do(x))P(z|I_x = \emptyset) && (Z \perp\!\!\!\perp I_x)_{G'} \\
&= \sum_z P(y|x, z, I_x = do(x))P(z|I_x = \emptyset) && I_x = do(x) \implies X = x \\
&= \sum_z P(y|x, z, I_x = \emptyset)P(z|I_x = \emptyset) && (Y \perp\!\!\!\perp I_x|X, Z)_{G'}
\end{aligned}
$$

Let us go through the derivation step-by-step. Our goal is remove $I_x = do(x)$ from our query and obtain a final formula in which all elements are expressed in terms of $I_x = \emptyset$, namely, the data available to us. First note that, to remove the regime indicator from the expression, we need to condition on X, as there is a direct path from I_x to Y. Second, note that the effect of X on Y is confounded and thus we cannot naively use the conditional distribution of Y given X to estimate the causal effect of X on Y. This can be seen by the fact that $(Y \not\perp\!\!\!\perp I_x|X)_{G'}$— the paths $I_x \to X \leftarrow Z \to Y$ and $I_x \to X \leftrightarrow Z \to Y$ are open when we condition on X. Thus $P(y|x, I_x = \emptyset) \neq P(y|I_x = do(x))$.

In order to make progress, the natural thing to do is to stratify our query by the levels of the confounder Z, in an attempt to block the back-door paths from X to Y. This leads to the first step of the derivation, $P(y|I_x = do(x)) = \sum_z P(y|z, I_x = do(x))P(z|I_x = do(x))$. No causal assumptions are invoked in this step, this is simply probability theory. Now we can tackle each of these terms separately. First, since X has no effect on Z, intervening on X does not change the probability of Z. This assumption is represented graphically by the fact that $(I_x \perp\!\!\!\perp Z)_{G'}$, resulting in $P(z|I_x = do(x)) = P(z|I_x = \emptyset)$. Finally, adjusting for Z blocks all backdoor paths between X and Y, and thus, for those individuals with the same value of $Z = z$, "seeing" x is equal to "doing" x. Mathematically, this translates into the final two steps of the derivation, one invoking the effectiveness of our intervention $I_x = do(x) \implies X = x$, and the other unconfoundedness given

Z, $(Y \perp\!\!\!\perp I_x | X, Z)_{G'}$. This gives us $P(y|z, I_x = do(x)) = P(y|x, z, I_x = \emptyset)$. Combining these two results yields our final formula, which is known by many names across different disciplines, such as the "backdoor" formula, "adjustment" formula, "g-formula," or "standardization" (Rosenbaum and Rubin, 1983; Robins, 1986; Pearl, 1995, 2009; Imbens and Rubin, 2015; Shpitser et al., 2012; Perkovic et al., 2018).

In general, the backdoor formula holds whenever we can find a set of variables Z that (i) blocks all confounding paths from X to Y and (ii) does not block any of the relevant causal paths from X to Y—see Pearl and Mackenzie (2018, Chapter 4) and Cinelli et al. (2022) for a didactic exposition of identification via covariate adjustment. As we will see in the next sections, the backdoor formula plays a key role in other settings—thus, now that we have learned it, we can make use of it whenever we see this graphical pattern again. But before moving forward, let us take a step back and try to extract the essential rules for manipulating probabilistic expressions involving interventions, which are already present in this simple example.

Simplifying the notation. Instead of decorating all expressions with regime indicators, as above, we can simply distinguish observational from interventional distributions by including or omitting the do-operator. Hereafter we will write, for instance, $P(y|z, do(x))$ to denote $P(y|z, I_x = do(x))$. The absence of $do(\cdot)$ operators for certain variables simply means that these variables are in their natural (observational) regime.

3.2 The rules of *do*-calculus

Recall that a causal diagram encodes essentially two assumptions: (i) the *absence* of a directed arrow $V_i \to V_j$ means that variable V_i does not directly cause V_j; and, (ii) the *absence* of a bi-directed arrow $V_i \leftrightarrow V_j$ means that there are no direct unobserved common causes of V_i and V_j. Note how these two assumptions were leveraged in our previous derivation of the backdoor formula. Given our assumption that X does not cause Z, we concluded that $P(z|do(x)) = P(z)$; and given our assumption that Z blocks all confounding paths from X to Y, we concluded that $P(y|do(x), z) = P(y|x, z)$. This suggests two key symbolic operations that should govern the manipulation of interventional probabilities: (i) ignoring the effect of actions, by dropping a $do(\cdot)$ operator, whenever these actions do not change the relevant probabilities of interest; and, (ii) exchanging action (say, $do(x)$) with an observation ($X = x$), whenever the effect of X in the relevant probabilities is not confounded. Our goal in this section is to generalize these types of operations, and construct an axiomatic system that allows manipulating $do(\cdot)$ expressions.

Consider an arbitrary disjoint set of variables Y, Z, X, W. Suppose that X is held fixed by intervention. As we have seen in Figure 1d, $I_x = do(x)$ implies no other variable is a cause X, and thus all arrows from such variables pointing to X are to be removed. To simplify our graphical representation, we denote such interventional graph by $G_{\overline{X}}$, i.e, the graph G with all arrows pointing to X removed (and omitting the intervention node I_x. Moreover, if the node X is held fixed, all paths emanating from X are naturally blocked. Thus, in all separation statements, we will always condition on X. With this in mind, we can now derive our three rules for manipulating interventional distributions.

Rule 1 (Ignoring Observations):

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

Rule 2 (Action/Observation Exchange):

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

Rule 3 (Ignoring Actions):

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

Rule 1 follows from d-separation, the only difference being that here X is held fixed by intervention. The validity of Rule 2 follows immediately from:

$$P(y|I_x = do(x), I_z = do(z), w) = P(y|I_x = do(x), I_z = do(z), x, z, w) = P(y|I_x = do(x), I_z = \emptyset, x, z, w)$$

Where the first equality follows from effectiveness, and the second equality is licensed by the graphical separation of the regime indicator conditional on X, W and Z. The rule is essentially saying that, when X is held fixed by intervention, if W blocks all backdoor paths from Z to Y, we can replace the action $I_z = do(z)$ with the observation of z in the passive regime $I_z = \emptyset$. Our final rule generalizes conditions for ignoring the effects of an action, and it follows immediately from:

$$P(y|I_x = do(x), I_z = do(z), w) = P(y|I_x = do(x), I_z = do(z), x, w) = P(y|I_x = do(x), I_z = \emptyset, x, w)$$

Where the first equality follows from effectiveness, and the second equality is licensed by the graphical separation of the regime indicator conditional on X and W. These three rules are collective known as the rules of *do*-calculus (Pearl, 1993, 1994).

Checking separation without intervention nodes. An equivalent way to express the three rules of *do*-calculus is to check for the separation conditions directly on the original causal diagram, without decorating the graph with intervention nodes:

Rule 1 (Ignoring Observations):

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

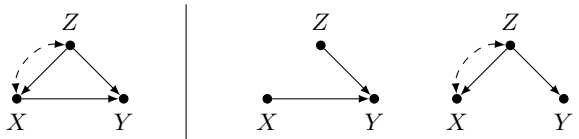
Rule 2 (Action/Observation Exchange):

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

Rule 3 (Ignoring Actions):

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

Where here the subscript $\underline{Z}$ means that all outgoing arrows from Z are removed from the graph. The equivalence of these conditions can be directly seen from the following facts. For Rule 2, recall that all directed paths from I_z to Y goes through Z. Since Z is conditioned on, all these paths are blocked, and the only possibly open paths from I_z to Y are paths with arrows pointing to Z. Thus, when checking for the separation of I_z and Y, we can simply remove all arrows outgoing from Z, and check whether there are any open backdoor paths from Z to Y, leading to the condition $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$. For Rule 3, one could imagine that checking for the separation of I_z and Y is equivalent to simply checking for the separation of Z and Y, after removing edges incoming into Z. This is almost correct, except that, when the conditioning set W contains effects of Z, this partially opens paths of the form $I_z \to Z \leftarrow \cdots \to Y$, and thus for such cases, these arrows should not be removed. This realization leads to the condition $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X\,Z(W)}}}$, where $Z(W)$ is the subset of nodes in Z for which we are not conditioning on any its effects.



(a) G (b) $G_{\overline{X}}$ (c) $G_{\underline{X}}$

Figure 2: Auxiliary DAGs for the *do*-calculus Derivation

Backdoor revisited. Going back to our previous derivation of Model 1, the derivation appealing directly to the rules as articulated above would proceed as follows:

$$\begin{aligned} P(y|do(x)) &= \sum_z P(y|do(x),z)P(z|do(x)) && \text{Law of total probability} \\ &= \sum_z P(y|do(x),z)P(z) && \text{Rule 3: } (Z \perp\!\!\!\perp X)_{G_{\overline{X}}} \\ &= \sum_z P(y|x,z)P(z) && \text{Rule 2: } (Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}} \end{aligned}$$

4 Illustrative Examples

We are now ready to practice the use of graphical models and the *do*-calculus through a series of illustrative examples. We will work, step-by-step, through the derivations of many of typical identification problems faced by empirical scientists, such as: (i) the identification of effects of single or multiple (sequential) interventions from observational data; (ii) the identification of these effects from a combination of multiple data sources, including observational and experimental data; (iii) recovering from selection bias and missing data; and, (iv) generalizing causal effects across domains.

All these problems are characterized by the answer to three main questions:

1. **Query:** What do we want to know?
2. **Data:** What data do we have?
3. **Model:** What do we already know?

Our task is to decide whether the data we have, along with our model assumptions, is sufficient to answer the query of interest. *(explain further)*

We defer the auxiliary graphs that license each step of the derivation to the appendix.

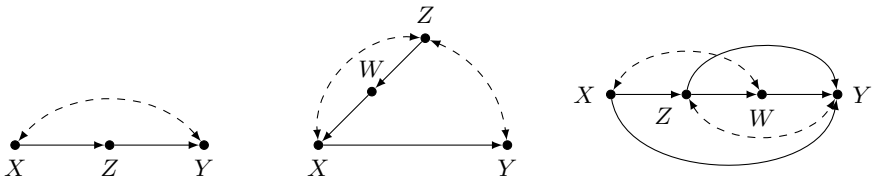
4.1 Single Interventions

We start with what is perhaps the most common task in causal inference studies: the identification of the causal effect of a treatment X on an outcome Y from observational data. Figure 3 provides the illustrative models for this section. Formally, in the following examples,

- Our **Query** is $Q = P(y|do(x))$, i.e, we want to know the effect of intervening on X on the distribution of Y ("what do we want to know?");
- The **Data** available to us is $\mathcal{D} = P(y, x, z, w)$, that is, we have *non-experimental* data of the joint distribution of the observed variables ("what data do we have?"); and,

- Our **Model** is given by the qualitative assumptions encoded by a causal diagram, $\mathcal{M} = G$, representing assumptions of no unobserved confounding (absence of bidirected edges) or no direct effects (absence of directed edges) between certain variables ("what do we already know?").

Our derivation strategy will be that of *query conversion*, we will manipulate the query $Q = P(y|do(x))$ into an equivalent expression that is "do-free," i.e, written solely in terms of the available data $P(y, x, z, w)$. Each step of the query conversion process should be licensed by the assumptions in our causal model, encoded by G.



(a) Model 2 (b) Model 3 (c) Model 4

Figure 3: Illustrative causal diagrams for the identification of the effect of single interventions. In all examples, our **Query** is $Q = P(y|do(x))$, the available **Data** is $\mathcal{D} = P(y, x, z, w)$, and the **Model** is specified by a causal diagram, $\mathcal{M} = G$.

Model 2. Although probably one of the most well-known identification results, backdoor adjustment is not the only way to identify a causal effect. Our first example that goes beyond simple adjustment is Figure 3a. Here the causal effect of X and Y is directly confounded by latent variables, as represented by the bidirected edge $X \leftrightarrow Y$. Since we do not observe these latent variables, there is no way to block the backdoor path, and, therefore, identification by simple adjustment is impossible. However, the causal effect is still identified, as demonstrated below:

$$
\begin{aligned}
P(y|do(x)) &= \sum_z P(y|do(x), z)P(z|do(x)) && \text{Law of total probability} \\
&= \sum_z P(y|do(x), do(z))P(z|x) && \text{Rule 2: } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\underline{Z}}} \text{ and } (Z \perp\!\!\!\perp X)_{G_{\underline{X}}} \\
&= \sum_z P(y|do(z))P(z|x) && \text{Rule 3: } (Y \perp\!\!\!\perp X|Z)_{G_{\overline{X}\overline{Z}}} \\
&= \sum_z P(z|x) \sum_{x'} P(y|z, x')P(x') && \text{Backdoor adj. with } X
\end{aligned}
$$

As before, we start again by applying the law of total probability, stratifying the query by the *mediator* Z: $P(y|do(x)) = \sum_z P(y|do(x), z)P(z|do(x))$. We now have to remove the *do* operator from two quantities, $P(z|do(x))$ and $P(y|do(x), z)$. The first term is easy, since the effect of X on Z is not confounded, and thus we immediately have $P(z|do(x)) = P(z|x)$ (Rule 2). Now all that is left to us is to handle the second term $P(y|do(x), z)$. First note that, when X is held fixed by intervention, there are no backdoor paths between Z and Y, and we can write $P(y|do(x), z) = P(y|do(x), do(z))$ (Rule 2). Here one might surmise that we made our task harder, since we now have two *do* operators to remove! However, note that Z completely mediates the effect of X on Y; thus, if Z is held fixed by intervention, manipulating X does not change the probability of Y, allowing us to drop $do(x)$ from the expression, obtaining the equality $P(y|do(x), do(z)) = P(y|do(z))$ (Rule 3). Our task has

thus simplified to that of identifying the effect of Z on Y, and we already know how to do that! Since X blocks all confounding paths from Z to Y, we can simply use backdoor adjustment, obtaining $P(y|do(z)) = \sum_{x'} P(y|z,x')P(x')$, as in the last step of the derivation[3]. This final expression is known as the "front-door" formula, and it was first derived in Pearl (1993, 1995). The name "front-door" contrasts this identification strategy with "backdoor" adjustment, since here we identify the causal effect not by adjusting for confounders, but by leveraging its mediating paths. Historically, the front-door model was perhaps the first known *non-parametric* identification result that goes beyond simple covariate adjustment.

Model 3. After learning about front-door and backdoor adjustments, readers may wonder if these exhaust all possible non-parametric identification strategies. While this may be true for simple models consisting of only three variables, it is not true in general. Consider the model shown in Figure 3b. First note that there are no observed mediators, therefore front-door adjustment is immediately ruled out. As for backdoor adjustment, we have an open backdoor path $X \leftarrow W \leftarrow Z \leftrightarrow Y$. While adjusting for Z blocks this confounding path, it introduces another source of bias, by opening the colliding path $X \leftrightarrow Z \leftrightarrow Y$. Conditioning on W, a descendant of Z, has the same issue. Therefore, the causal effect of X on Y cannot be identified neither via front-door nor backdoor adjustments. However, it is still identified, albeit with a different formula:

$$
\begin{aligned}
P(y|do(x)) &= P(y|do(x), do(w)) && \text{Rule 3: } (Y \perp\!\!\!\perp W|X)_{G_{\overline{X}\overline{W}}} \\
&= P(y|x, do(w)) && \text{Rule 2: } (Y \perp\!\!\!\perp X)_{G_{\overline{W}\underline{X}}} \\
&= \frac{P(y, x|do(w))}{P(x|do(w))} && \text{Def. of conditional probability} \\
&= \frac{\sum_z P(y, x|w, z)P(z)}{\sum_z P(x|w, z)P(z)} && \text{Backdoor adj. with } Z
\end{aligned}
$$

Intuitively, the strategy works as follows: intervening on W effectively blocks the confounding paths from X to Y, and we can express the causal effect of X as the conditional effect of W on Y; this conditional effect, by its turn, can be identified via a ratio of backdoor adjustments, controlling for Z. Now in detail. First note that W only affects Y through X; hence, when we intervene on X, manipulating W does not change the probability of Y, allowing us to add $do(w)$ to the expression (Rule 3). Moreover, when W is held fixed *by intervention* (not by conditioning), we block all confounding paths between X and Y, and we obtain $P(y|do(x), do(w)) = P(y|x, do(w))$ (Rule 2). Finally, we apply the definition of conditional probability to rewrite the expression as a ratio, $P(y|x, do(w)) = \frac{P(y,x|do(w))}{P(x|do(w))}$, and note that Z blocks all backdoor paths from W to $\{X, Y\}$—we can thus apply backdoor adjustment both to the numerator and to the denominator. This is the simplest model known to us that cannot be identified neither with front-door or backdoor adjustment, and it has been called the "new napkin" problem in Pearl and Mackenzie (2018), or the "trapdoor" model in Helske et al. (2021). Beyond the unusual identification formula, it also has an interesting peculiarity not shared by the previous two models. Note that, like Models 1 and 2, Model 3 does not entail any conditional independencies; however, and in contrast with the other two models, it does have a testable implication known as a "Verma Constraint" (see Verma and Pearl (1991), Tian and Pearl (2002b), Evans (2012, 2018), Shpitser and Pearl (2008), and Richardson et al. (2017)). This testable implication can be seen in the final formula for the causal effect of X on Y, which is *functionally independent* of the value $W = w$.

[3] We use x' to distinguish the value of X in the summation from the value $X = x$ of the term $P(z|x)$.

Model 4. The front-door model requires the treatment X to have no causal paths to the outcome Y other than through the mediator Z. Similarly, it requires the absence of unmeasured confounders between the mediator Z and the outcome Y. These conditions are necessary—if we add a direct path $X \to Y$, or a bidirected arrow $Z \leftrightarrow Y$ to the diagram of Figure 3b, identification is no longer possible. What can we do then, if such paths exist? An alternative route to identify the effect of X on Y is to find a second mediator W on the path from Z to Y, such that: (i) it intercepts the latent confounders of X and Y; (ii) it is not directly caused by X; and, (iii) it is not itself confounded with Y. This motivates the final example of this section, given by Model 4 in Figure 3c. In the presence of such variable, identification is restored:

$$
\begin{aligned}
P(y|do(x)) &= \sum_{z,w} P(y|do(x),z,w)P(z|do(x))P(w|do(x),z) && \text{Law of total probability} \\
&= \sum_{z,w} P(y|x,z,w)P(z|x)P(w|do(x),z) && \text{Rule 2: } (Y \perp\!\!\!\perp X|\{Z,W\})_{G_{\underline{X}}} \text{ and } (Z \perp\!\!\!\perp X)_{G_{\underline{X}}} \\
&= \sum_{z,w} P(y|x,z,w)P(z|x)\sum_{x'} P(w|z,x')P(x') && \text{Same steps as Model 2}
\end{aligned}
$$

The derivation is simple, and uses similar strategies we have learned so far. We again start by applying the law of total probability, decomposing our query into three terms. The first two terms are easy, and can be solved by a simple application of Rule 2: (i) conditional on Z and W, there are no open confounding paths from X to Y, and thus $P(y|do(x),z,w) = P(y|x,z,w)$; (ii) similarly, X is not confounded with Z, therefore $P(z|do(x)) = P(z|x)$. Finally, we can handle the last term, $P(w|do(x),z)$, following the same steps used in the front-door model: we know that $P(w|do(x),z) = P(w|do(z))$, and to identify $P(w|do(z))$ it suffices to adjust for x, namely, $P(w|do(z)) = \sum_{x'} P(w|z,x')P(x')$.

Some Useful Dodges

There are some very simple and useful dodges that can be used to decide the identifiability of an interventional query from observational data. These dodges allow us to completely bypass algebraic derivations with a simple inspection of the graph. For example, if there exists a bidirected edge from the treatment to any of its immediate effects (in the causal pathway to the outcome), this already rules out (non-parametric) identification from observational data[4]. In such cases, the graph instantly warns the researcher that any attempt of derivation is futile—to make progress, further assumptions or new measurements are needed. Moreover, if there is no bidirected path (i.e, a path consisting entirely of bidirected edges) from the treatment to its immediate effects, then the causal effect *is* identifiable, no matter how complicated the graph structure (again, we need only to consider those variables in the causal pathway to the outcome). With this last dodge, for instance, we can immediately verify the identifiability of $P(y|do(x))$ in Models 1, 2 and 4—but the dodge is not enough to derive the result of Model 3. See Tian and Pearl (2002a) and Pearl (2009, Chapter 3) for further discussion.

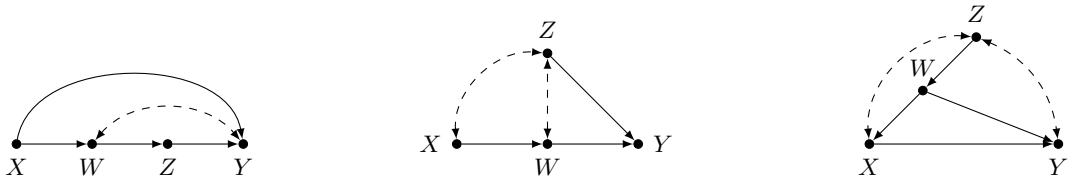
4.2 Multiple and Sequential Interventions

We now move to the identification of multiple and sequential interventions. In the next set of examples, provided in Figure 4, the target of analysis is not just the causal effect of a single random variable X on Y, but the causal effect of a *joint* intervention on *both* X *and* Z on Y. Formally,

[4]See examples of non-identifiable models where this dodge applies in Figures 5b and 5c of Section 4.3.

- Our **Query** is now $\mathcal{Q} = P(y|do(x), do(z))$;
- The **Data** available to us is still the same, namely, $\mathcal{D} = P(y, x, z, w)$; and,
- Our **Model** is given by a causal diagram, $\mathcal{M} = G$.

As before, our strategy will be that of *query conversion*; we will manipulate the target query $\mathcal{Q} = P(y|do(x), do(z))$ into an equivalent expression that can be expressed solely in terms of the available data $P(y, x, z, w)$.



(a) Model 5 (b) Model 6 (c) Model 7

Figure 4: Illustrative causal diagrams for the identification of the effect of multiple interventions. In all examples, our **Query** is $\mathcal{Q} = P(y|do(x), do(z))$, the available **Data** is $\mathcal{D} = P(y, x, z, w)$, and the **Model** is specified by a causal diagram, $\mathcal{M} = G$.

As we will see, some important lessons emerge from multiple or sequential interventions. First, more often than not, simple covariate adjustment is not sufficient for identification—this happens because we may need to adjust for mediators to block confounding paths, while still preserving the causal effect that flows through those mediators. Another curious, and perhaps under-appreciated lesson is that it may be *easier* to identify the joint effect of X and Z than the effect of X alone.

Model 5. Our first example represents a common practical problem of "time-varying" treatments in randomized control trials (Robins and Wasserman, 1997; Pearl and Mackenzie, 2018). One possible story for this model goes as follows. At period 1, a treatment X (say, a drug) is randomized to patients. Next, at period 2, a second dose of the treatment (denoted by Z) is administered. Here, however, the dosage of the drug depends on the patient's current health conditions W. Note that the patient's health condition is not randomized, thus there could be unobserved confounders between W and the final health outcome Y (represented by the bidirected arrow $W \leftrightarrow Y$). Our goal is to identify the effect of a *treatment plan* (i.e, the effect of two doses), given by the joint intervention $P(y|do(x), do(z))$. While X is not confounded with Y, there is an open confounding path between Z and Y, namely, $Z \leftarrow W \leftrightarrow Y$. We thus need to adjust for W to block this confounding path; however, naively doing so opens the colliding path $X \rightarrow W \leftrightarrow Y$. Thus, simple covariate adjustment yields biased estimates. This difficulty has led to several papers by Wermuth and Cox (Wermuth and Cox, 2008, 2014; Cox and Wermuth, 2015), who analyzed the problem under the rubric of "indirect confounding." The *do*-calculus derivation, nevertheless, takes only three lines,

$$
\begin{aligned}
P(y|do(x), do(z)) &= \sum_w P(y|do(x), do(z), w) P(w|do(x), do(z)) && \text{Law of total probability} \\
&= \sum_w P(y|do(x), do(z), w) P(w|do(x)) && \text{Rule 3: } (W \perp\!\!\!\perp Z|X)_{G_{\overline{XZ}}} \\
&= \sum_w P(y|x, z, w) P(w|x) && \text{Rule 2: } (Y \perp\!\!\!\perp \{X, Z\}|W)_{G_{\underline{ZX}}} \text{ and } (W \perp\!\!\!\perp X)_{G_{\underline{X}}}
\end{aligned}
$$

In words, we apply the law of total probability to the original query, effectively decomposing our query into two sub-queries: $P(y|do(x), do(z)) = \sum_w P(y|do(x), do(z), w)P(w|do(x), do(z))$. Noting that manipulating Z has no effect on the probability of W, we can drop $do(z)$ from the second term, $P(w|do(x), do(z)) = P(w|do(x))$ (Rule 3). We can now simply apply Rule 2 to both terms; (i) as a randomized trial, the effect of X on W is not confounded, and $P(w|do(x)) = P(w|x)$. Similarly, conditioning on W blocks all backdoor paths from $\{X, Z\}$ to Y, and we have $P(y|do(x), do(z), w) = P(y|x, z, w)$. Note that the final expression is not a standard regression adjustment for W, as it requires averaging over the conditional distribution $P(w|x)$ instead of averaging over the marginal distribution $P(w)$.

Model 6. Consider now Figure 4b. Contrary to the previous model, here the effect of X on Y is *not* identifiable. We have an open backdoor path $X \leftrightarrow Z \rightarrow Y$, and while adjusting for Z blocks this path, it opens the colliding path $X \leftrightarrow Z \leftrightarrow W \rightarrow Y$. Front-door adjustment is prevented for a similar reason. In fact, it can be shown that there is no sequence of *do*-calculus operations that can express $P(y|do(x))$ as a function of the observed data. Despite that, the *joint effect* of X and Z is still identifiable! The derivation largely parallels the previous one,

$$\begin{aligned} P(y|do(x), do(z)) &= \sum_w P(y|do(x), do(z), w)P(w|do(x), do(z)) && \text{Law of total probability} \\ &= \sum_w P(y|do(z), w)P(w|do(x)) && \text{Rule 3: } (Y \perp\!\!\!\perp X|\{Z, W\})_{G_{\overline{Z}}} \text{ and } (W \perp\!\!\!\perp Z|X)_{G_{\overline{XZ}}} \\ &= \sum_w P(y|z, w)P(w|x) && \text{Rule 2: } (Y \perp\!\!\!\perp Z|W)_{G_{\underline{Z}}} \text{ and } (W \perp\!\!\!\perp X)_{G_{\underline{X}}} \end{aligned}$$

The law of total probability give us $P(y|do(x), do(z)) = \sum_w P(y|do(x), do(z), w)P(w|do(x), do(z))$. Since Z does not cause W we have $P(w|do(x), do(z)) = P(w|do(x))$ (Rule 3). Similarly, Z and W blocks all causal and non-causal paths from X to Y, thus $P(y|do(x), do(z), w) = P(y|do(z), w)$ (Rule 3). Finally, noting that both terms are unconfounded concludes the derivation. The causal diagram of Model 6 also has testable implications: Y should be independent of W given Z and X, otherwise, this implies our assumptions are false.

Model 7. We now examine another case in which a joint intervention is easier to identify than a single one. Consider a variation of Model 3, given by Figure 4c (Model 7). The only difference between Model 7 and Model 3 is the addition of the directed edge $W \rightarrow Y$. This modification, however, is sufficient to prevent the identification of $P(y|do(x))$. Nevertheless, the joint intervention $P(y|do(x), do(z))$ is still identifiable. To witness, as usual, start with the law of total probability:

$$P(y|do(x), do(z)) = \sum_w \underbrace{P(y|do(x), do(z), w)}_{(2)} \underbrace{P(w|do(x), do(z))}_{(1)} \qquad \text{Law of total probability}$$

Term (1) of this expression is easy to handle, since X does not cause W and the effect of Z on W is not confounded:

$$\begin{aligned} (1) : P(w|do(x), do(z)) &= P(w|do(z)) && \text{Rule 3: } (W \perp\!\!\!\perp X \mid Z)_{G_{\overline{XZ}}} \\ &= P(w|z) && \text{Rule 2: } (W \perp\!\!\!\perp Z)_{G_{\underline{Z}}} \end{aligned}$$

The second term of the expression can then be reduced to $P(y|x, do(w))$ in the following way:

$$
\begin{aligned}
(2) : P(y|do(x), do(z), w) &= P(y|do(x), do(z), do(w)) && \text{Rule 2: } (Y \perp\!\!\!\perp W \mid \{X, Z\})_{G_{\overline{XZ}\underline{W}}} \\
&= P(y|do(x), do(w)) && \text{Rule 3: } (Y \perp\!\!\!\perp Z \mid \{X, W\})_{G_{\overline{WXZ}}} \\
&= P(y|x, do(w)) && \text{Rule 2: } (Y \perp\!\!\!\perp X \mid W)_{G_{\overline{W}\underline{X}}}
\end{aligned}
$$

In words, fixing X and Z by intervention blocks all backdoor paths from W to Y and we can thus replace w with $do(w)$; next, intervening on Z does not affect Y whenever X and W are hold fixed by intervention, allowing us to drop $do(z)$; and, finally, holding w fixed by intervention blocks all confounding paths from X to Y, and we can replace $do(x)$ with x. The identification of $P(y|x, do(w))$, then, uses the same steps as those of Model 3, resulting in:

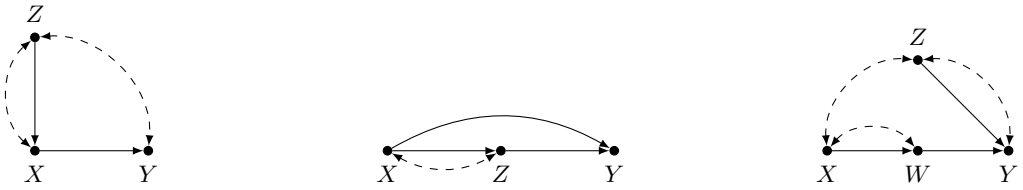
$$
P(y|do(x), do(z)) = \sum_{w} \frac{\sum_{z'} P(y, x|w, z')P(z')}{\sum_{z'} P(x|w, z')P(z')} P(w|z).
$$

4.3 Surrogate Interventions and Surrogate Outcomes

In the previous two sections we discussed the identification of causal effects from observational data. A natural question then, arises—do we need the *do*-calculus, when experiments are available? Perhaps surprisingly, the answer is yes. In applied settings, even when experimental data is available to the analyst, often there is a mismatch between the data and the target query, thus formal causal analysis is still needed. For instance, we may want to measure the effect of a treatment X (e.g, cholesterol) on an outcome Y (e.g, heart disease), but intervening directly on X is not feasible. Instead, we can only run experiments by manipulating Z (e.g, the diet of the individual). Moreover, even when we can randomize the treatment of interest X (e.g, vaccine), the primary outcome of interest Y (e.g, immunity) may not be immediately measurable during the period of the trial, and only a preliminary endpoint Z (e.g, antibodies) is available. Finally, policy makers may need to estimate the effect of a joint intervention on X and Z, but only trials where either X or Z were randomized separately are available. These problems are known as "surrogate interventions" or "surrogate outcomes" (Bareinboim and Pearl, 2012; Tikka and Karvanen, 2018; Lee et al., 2020), and they motivate our next batch of examples, shown in Figure 5. Here,

- Our **Query** remains the same as in the previous two sections. We want to identify the effect of interventions, such as the marginal effect X on Y, $\mathcal{Q} = P(y|do(x))$, or the joint effect of X and Z on Y, $\mathcal{Q} = P(y|do(x), do(z))$;
- However, the **Data** available to us is now *richer*. In each of these problems, although we do not have experimental data that directly answers the query, we have a *combination* of experimental and observational data, such as $\mathcal{D} = \{P(y, x, z), P(z|do(x))\}$, which may be combined to answer the query. Finally,
- Our **Model** is given by a causal diagram, $\mathcal{M} = G$.

As before, our derivation strategy is that of *query conversion*: we will manipulate the target query into an equivalent expression that is written solely in terms of the available data. The only difference from the previous section is that, since now we may have access to experimental data (e.g, RCT in which we manipulate Z), we do not need to remove all $do(\cdot)$ operators—we need only to make sure that all terms containing "do's" in the final expression match the experimental data that is available to us (e.g, $do(z)$).



(a) Model 8 (b) Model 9 (c) Model 10

Figure 5: Illustrative causal diagrams with examples of how to combine data from different experimental conditions: in (a) our goal is to estimate the causal effect of X on Y, but we can only run experiments in Z (surrogate intervention); in (b) our goal is to estimate the causal effect of X on Y, but we only measure Z during the trial (surrogate outcome); in (c) our goal is to estimate the *joint* effect of Z and X on Y, but we only have data from separate trials, in which either X or Z were manipulated independently; in (d) our goal is to estimate the *joint* effect of Z and X on Y, but we can only manipulate W.

Model 8. In this example we are interested in the causal effect of X (e.g, body fat) on Y (e.g, survival). However, X is something that cannot be randomized directly, either due technical or due to ethical concerns, and we can only run experiments in which Z is randomized (e.g, diet or exercise). Formally, our query is $\mathcal{Q} = P(y|do(x))$ and the data available to us is $\mathcal{D} = P(y, x|do(z))$. Can we identify the effect of X by randomizing Z instead of X? Surprisingly, under the assumptions of Model 8, the answer is positive. The derivation takes just two steps of *do*-calculus:

$$\begin{aligned} P(y|do(x)) &= P(y|do(x), do(z)) && \text{Rule 3: } (Y \perp\!\!\!\perp Z \mid X)_{G_{\overline{X}\overline{Z}}} \\ &= P(y|x, do(z)) && \text{Rule 2: } (Y \perp\!\!\!\perp X)_{G_{\overline{Z}\underline{X}}} \end{aligned}$$

That is, we first note that X mediates all the effects of Z on Y, thus once we hold X fixed (by intervention) manipulating Z has no effect on the probability of Y, resulting in $P(y|do(x)) = P(y|do(x), do(z))$ (Rule 3). Now we simply note that, when we intervene on Z, there are no back-door paths from X to Y, giving us our final expression $P(y|do(x), do(z)) = P(y|x, do(z))$ (Rule 2). Since $P(y|x, do(z))$ is written solely in terms of the data available to us (experimental data in which we manipulate Z) we are done, and we conclude that we can measure the effect of X on Y, by performing a surrogate intervention on Z. Note the assumptions that Z does not cause Y directly, and that Z fully intercepts the confounding paths from X to Y, are necessary: if any of them are violated, the surrogate intervention is not enough to identify the effect of X on Y.

Model 9. Here we are interested in the effect of X (e.g, vaccine), and we can actually perform a trial in which we randomize X. However, our primary outcome of interest Y (e.g, immunity) cannot be measured during the period of the trial—instead, we can only measure a "surrogate endpoint" Z (e.g, antibodies). Finally, suppose non-experimental data of the association between X, Z and Y is available. Formally, our query is $\mathcal{Q} = P(y|do(x))$ and the data available to us is $\mathcal{D} = \{P(y, x, z), P(z|do(x))\}$. Under the assumptions of Figure 5b, identification using the surrogate outcome Z is indeed possible, and given by

$$\begin{aligned} P(y|do(x)) &= \sum_w P(y|z, do(x))P(z|do(x))) && \text{Law of total probability} \\ &= \sum_z P(y|x, z)P(z|do(x)) && \text{Rule 2: } (Y \perp\!\!\!\perp X \mid Z)_{G_{\underline{X}}} \end{aligned}$$

The derivation again uses only one step of do-calculus, and simply notes that the conditional effect of X on Y, given Z, is unconfounded (Rule 2). Note how all terms of the final expression can be computed from the available data. Starting from the second term, $P(z|do(x))$, this is given by the randomized control trial, in which we randomize the treatment X and measure the surrogate endpoint Z. As for the first term, $P(y|x,z)$, this can be obtained with non-experimental data.

Model 10. Now imagine we have separate trial data for two drugs, X and Z, but our goal is to understand the effect of a treatment plan that prescribes both *simultaneously* (a joint intervention). We may be interested on the joint effect, because, for example, the drugs may interact with each other in unexpected, harmful ways. Mathematically, our query is then the joint effect $P(y|do(x),do(z))$ and data available to us is $\mathcal{D} = \{P(y,w|do(x)), P(y,w|do(z))\}$. In Model 10, identification is possible, and indeed the two trials can be combined to estimate our query:

$$\begin{aligned} P(y|do(x),do(z)) &= \sum_w P(y|w,do(x),do(z))P(w|do(x),do(z)) && \text{Law of total probability} \\ &= \sum_w P(y|w,do(z))P(w|do(x)) && \text{Rule 3: } (Y \perp\!\!\!\perp X \mid W,Z)_{G_{\overline{Z}}} \text{ and } (W \perp\!\!\!\perp Z \mid X)_{G_{\overline{ZX}}} \end{aligned}$$

As usual, we start the derivation with law of total probability. The second line, then, applies Rule 3 twice: (i) intervening on X does not modify the conditional effect of Z on Y, given W; and, (ii) Z does not affect W. Note there is only one $do(\cdot)$ operator in each term of the final expression, meaning that they can be estimated from separate trials, in which only either Z or X are randomized.

4.4 Recovering from Selection Bias

Up until now we have considered cases in which the data available to the researcher, be it experimental or observational, came from the target population of interest. Often, however, especially with randomized trials, the study sample *is not* a representative sample of the general population. This problem is usually known as selection bias (Elwert and Winship, 2014; Bareinboim and Tian, 2015; Bareinboim and Pearl, 2016; Correa et al., 2019)[5]. When, and how, can we recover the causal effect in the general population, in the presence of selection-biased data?

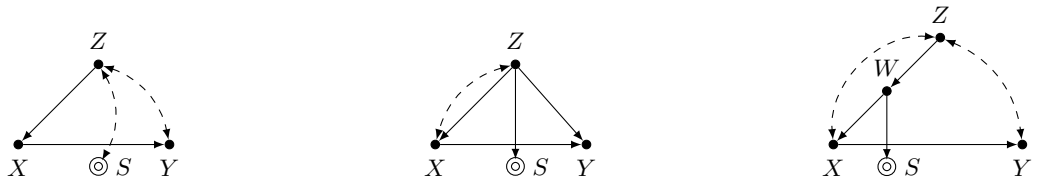
As it is usually the case in causal inference, in order to answer that question, we need to *model the selection mechanism.* In keeping with the spirit of the previous sections, we do so non-parametrically, by simply providing a qualitative description of the determinants of inclusion of units in the study sample. To encode the selection mechanism we introduce a new variable S called a *selection node.* S is a binary variable such that

$$S = \begin{cases} s, & \text{if the unit is selected to the study sample;} \\ s', & \text{otherwise.} \end{cases}$$

Once we have a selection node, the causal diagram should now include an arrow from any variable V to S, whenever we believe V affects the probability that a unit would be enlisted in the study. As usual, unobserved common causes of S and other variables in the system are encoded as bidirected edges. With this in mind, we move to the next set models, provided in Figure 6. Here:

[5]Some economists use the term "selection bias" to denote confounding bias, in the sense of "preferential selection to treatment" (Angrist and Pischke, 2009, 2014). Here we use selection bias to denote preferential selection of units into the study sample.

- The **Query** is the causal effect of X on Y in the *general population*, $\mathcal{Q} = P(y|do(x))$, that is, the causal effect *without conditioning on* $S = s$;
- The **Data** available to us, however, is the selection-biased data, conditioned on $S = s$, i.e, $\mathcal{D} = P(y, x, z, w|s)$. Sometimes, we may also have *auxiliary demographic data* from the general population for some variables, such as $P(z)$. And,
- The **Model** is still our causal diagram $\mathcal{G}$, but now enriched with a selection node S, describing the determinants of the selection process.



(a) Model 11 (b) Model 12 (c) Model 13

Figure 6: Illustrative causal diagrams where recovering from selection bias is possible. In all examples, our **Query** is $\mathcal{Q} = P(y|do(x))$. The available **Data** varies in each example, and it consists of a mix of observational data from the study sample $\mathcal{D} = P(y, x, z|s)$, and some (or none) auxiliary, demographic data, from the general population. The **Model** is given by the causal diagram $\mathcal{M} = G$.

Model 11. Our query is the causal effect of X on Y in the general population, $\mathcal{Q} = P(y|do(x))$, and available to us is non-experimental data from the study sample, $\mathcal{D} = P(y, x, z|s)$. No demographic data from the general population was collected. Note the causal effect is identifiable in the general population via backdoor adjustment, namely, $P(y|do(x)) = \sum_z P(y|x, z)P(z)$. This expression, however, is not directly estimable from the available data, since it is not conditioned on the selection indicator S. Moreover, note that $Y \not\perp\!\!\!\perp S|\{X, Z\}$ and $Z \not\perp\!\!\!\perp S$. Thus, the graph does note directly license the inclusion of S in each of the terms that composes our query, and one might be tempted to believe that there is no way around it. Yet, a causal look at the problem, however, tells us otherwise

$$\begin{aligned} P(y|do(x)) &= P(y|do(x), s) && \text{Rule 1: } (Y \perp\!\!\!\perp S|X)_{G_{\overline{X}}} \\ &= \sum_z P(y|x, z, s)P(z|s) && \text{Backdoor adj. with } Z \end{aligned}$$

The key step of the derivation is the first line (Rule 1), revealing that the causal effect of X on Y is *the same both* in the study sample and general population, *despite* the biased selection process. This happens because the study participants do not differ from the general population in any potential effect modifiers (i.e, causes of Y that are not fully mediated by the treatment X). Once we have this realization, it suffices then to identify the causal effect in the study sample, and for that we can simply apply backdoor adjustment using Z.

Model 12. Often it will not be possible to recover the causal effect in the general population without auxiliary demographic data. Model 12 is one such example. Here, identification of $P(y|do(x))$ is not possible from selection-biased data alone, because the selection process depends on Z, a potential effect modifier of the effect of X on Y. If, however, we enrich our data collection to include

measurements of Z in the target population, then identification is possible:

$$\begin{aligned} P(y|do(x)) &= \sum_z P(y|x,z)P(z) && \text{Backdoor adj. with } Z \\ &= \sum_z P(y|x,z,s)P(z) && \text{Rule 1: } (Y \perp\!\!\!\perp S|X,Z)_G \end{aligned}$$

The derivation starts by recognizing that the causal effect is identifiable using backdoor adjustment, then noting that the conditional distribution of Y given X and Z is the same both in the study sample and the target population. We then simply re-weight the z-specific effect with $P(z)$. Note the final expression only contains terms that are estimable from the available data, $\mathcal{D} = \{P(y,x,z|s), P(z)\}$. Specifically, the first term is estimable from the study sample, whereas the second term is given by demographic data.

A different look at Model 12. There is another way of solving Model 12. First note that Z d-separates all remaining nodes from the selection node, that is, $\{X,Y\} \perp\!\!\!\perp S|Z$. We can thus recover the full joint distribution of the general population,

$$\begin{aligned} P(y,x,z) &= P(y,x|z)P(z) && \text{Chain rule} \\ &= P(y,x|z,s)P(z). && \text{Rule 1: } (\{X,Y\} \perp\!\!\!\perp S|Z)_G \end{aligned}$$

That is, we can recover $P(y,x,z)$ by simply re-weighting the study sample with $\frac{P(z)}{P(z|s)}$. Once $P(y,x,z)$ is recovered, we can proceed as if we had access to unbiased population data, and estimate the causal effect as usual, via backdoor adjustment. This strategy can facilitate derivations in many examples. For instance, imagine instead of $Z \to S$ we now have a classical case of what is known as "case-control" bias, with an arrow $Y \to S$. Now we have $\{X,Z\} \perp\!\!\!\perp S|Y$, and thus, by similar reasoning, it suffices to have demographic data on $P(y)$ to recover the population distribution, as well as the causal effect of interest. In fact, if we include a selection node $Z \to S$ (or $Y \to S$) in any of the previous models 1 to 7, it trivially follows from the above considerations that the causal effect is recoverable given external data on Z (or Y). In the sequel, we revisit the "napkin" problem of Model 3 to illustrate how both strategies used in Model 12 can again be fruitfully deployed.

Model 13. Model 13 is similar to Model 3, except that we have selection-biased data, and the selection mechanism depends on W. Note that the causal effect in the general population, $P(y|do(x))$, cannot be recovered without external data. Now we show how to recover the causal effect using demographic data, either on W or on Z. Beginning with W, note that $\{Y,X,Z\} \perp\!\!\!\perp S|W$. Thus, as in the previous derivation, we can recover the joint distribution of the general population with $P(y,x,z,w) = P(y,x,z|w,s)P(w)$. After that, we can proceed with identification as usual, obtaining

$$P(y|do(x)) = \frac{\sum_z P(y,x|w,z)P(z)}{\sum_z P(x|w,z)P(z)}$$

where each term above is given by the recovered distribution. As to using external data on Z, we note that $\{Y,X\} \perp\!\!\!\perp S|Z,W$. Thus, $P(y,x|w,z) = P(y,x|w,z,s)$, and we may directly rewrite the identifying expression as:

$$P(y|do(x)) = \frac{\sum_z P(y,x|w,z,s)P(z)}{\sum_z P(x|w,z,s)P(z)}.$$

4.5 Recovering from Missing Data

Missing data is a pervasive issue in data analysis. If the data is missing because subjects with missing values are different from those with observed values, this may affect our ability to extract the true causal relations from the data. In fact, the problem of selection bias we studied in the last section can be seen as a special case of the general problem of missing data—i.e, when an individual is not included in the study sample, i.e, $S = s'$, all variables for that individual are missing. Here, instead, we now allow each variable to have its own unique missingness mechanism.

Mathematically, for each variable X in our causal diagram, we define two extra special variables. First, we encode the missingness process of X using a binary indicator R_x, such that

$$R_x = \begin{cases} r'_x, & \text{if the observation for } X \text{ is } \textit{not} \text{ missing;} \\ r_x, & \text{otherwise.} \end{cases}$$

Moreover, since X may not be fully observed, we need to introduce a *proxy* variable X^*, denoting the actual observations available to the analyst, defined as

$$X^* = \begin{cases} X, & \text{if } R_x = r'_x; \\ \text{missing}, & \text{if } R_x = r_x. \end{cases}$$

In words, the proxy variable X^* equals X whenever X is observed, and it is missing otherwise. In principle, all variables should have their missingness indicators. However, whenever a variable is fully observed, we simply omit its corresponding missingness indicator from the DAG. Similarly, depicting proxy variables can be often redundant. When that is the case, we omit them from the DAG.

Selection Bias

X	Y	S
1	0	1
0	1	1
1	0	1
0	1	1
NA	NA	0
NA	NA	0
NA	NA	0
NA	NA	0

Missing Data

X*	Y*	R_x	R_y
NA	NA	1	1
NA	NA	1	1
1	NA	0	1
0	NA	0	1
NA	0	1	0
NA	1	1	0
1	0	0	0
1	1	0	0

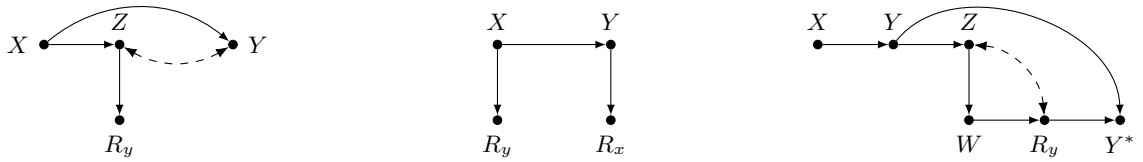
Table 1: Selection Bias vs. Missing Data

Table 1 depicts the distinction between selection bias and missing data visually. In the case of selection bias, the full row of an individual is affected by the selection process, and thus all variables for that individual are missing. On the other hand, in the general case of missing data, each variable is allowed to have its own missingness mechanism, and thus for any individual some variables may or may not be missing simultaneously.

With this in mind, in the next batch of examples:

- The **Query** is $\mathcal{Q} = P(y|do(x))$, the causal effect of X on Y;

- The **Data** are the actual measurements (i.e, the proxy variables) and their corresponding missingness indicators, $\mathcal{D} = P(y^*, x^*, z^*, R_x, R_y, R_z)$.
- The **Model** is a causal diagram $\mathcal{G}$ enriched with missingness indicators, describing the determinants of missingness for each variable. These graphs are sometimes called *missingness* graphs, or m-graphs (Mohan and Pearl, 2021).



(a) Model 14 (b) Model 15 (c) Model 16

Figure 7: Illustrative causal diagrams where recovering from missing data is possible. In all examples, our **Query** is $\mathcal{Q} = P(y|do(x))$. The available **Data** is $\mathcal{D} = P(y^*, x^*, z^*, R_x, R_y, R_z)$. The **Model** is given by the causal diagram (missingness graph) $\mathcal{M} = G$. Proxy variables omitted for clarity. The omission of a missingness indicator for a variable encodes the assumption that the respective variable is fully observed.

Model 14. This model illustrates bias due to attrition in a randomized control trial, namely, subjects dropping out from the trial before the final outcome is observed (Breskin et al., 2018; Cinelli and Pearl, 2018). We have complete data on the treatment X (e.g, vaccine assignment) and on a side effect Z (e.g injection site pain); however, we only have partial data on the outcome Y (e.g, disease), because participants with side-effects were more likely to drop out of the study (thus the arrow $Z \to R_y$). The bidirected arrow $Z \leftrightarrow Y$ denotes latent confounders of the side-effect Z and the outcome Y. Our goal is to estimate $P(y|do(x))$, despite the bias due to missing data. The derivation goes as follows,

$$
\begin{aligned}
P(y|do(x)) &= P(y|x) && \text{Rule 2: } (Y \perp\!\!\!\perp X)_{G_{\underline{X}}} \\
&= \sum_z P(y|x,z)P(z|x) && \text{Law of total probability} \\
&= \sum_z P(y|x,z,r'_y)P(z|x) && \text{Rule 1: } (Y \perp\!\!\!\perp R_y|X,Z)_G
\end{aligned}
$$

The first equality is licensed by randomization (null backdoor condition), the second by the law of total probability, and the last step by d-separation. All components of the final expression can be estimated from the available data; the first factor is estimable from the units we have outcome data, $R_y = r'_y$, whereas the second term is estimable from all units entering the trial.

Model 15. This model illustrates a more complex missingness process where not only X can cause the missingness of Y but also Y can cause the missingness of X. Note that, since X and Y are not confounded, as long as we can recover $P(x, y)$, we can identify $P(y|do(x)) = P(y|x)$. By the chain rule, it is easy to see that if we can identify the probability of non-missingness $P(r'_x, r'_y|x, y)$, we can recover the joint distribution $P(y, x)$:

$$P(x,y) = \frac{P(x,y,r'_x,r'_y)}{P(r'_x,r'_y|x,y)} = \frac{P(x^*,y^*,r'_x,r'_y)}{P(r'_x,r'_y|x,y)} \qquad \text{Chain rule}$$

Where in the second equality we use the fact that r'_x implies $x = x^*$ and r'_y implies $y = y^*$. We can now identify $P(r'_x, r'_y|x, y)$ by applying the chain rule and d-separation:

$$\begin{aligned} P(r'_x,r'_y|x,y) &= P(r'_x|y,x,r'_y)P(r'_y|x,y) && \text{Chain rule} \\ &= P(r'_x|y^*,r'_y)P(r'_y|x^*,r'_x) && \text{Rule 1: } (R_y \perp\!\!\!\perp \{Y,R_x\}|X)_G \text{ and } (R_x \perp\!\!\!\perp X|Y)_G \end{aligned}$$

Combining both results, we thus recover the joint distribution of X and Y:

$$P(x,y) = \frac{P(x^*,y^*,r'_x,r'_y)}{P(r'_x|y^*,r'_y)P(r'_y|x^*,r'_x)} \tag{1}$$

Note how all terms in the final expression involve non-missing data. The causal effect $P(y|do(x)) = P(y|x)$ can thus be computed from the recovered distribution.

Model 16. For our final missing-data example, it is useful to depict the proxy variable explicitly. The derivation then proceeds as follows:

$$\begin{aligned} P(y|do(x)) &= P(y|x) && \text{Rule 2: } (Y \perp\!\!\!\perp X)_{G_{\underline{X}}} \\ &= P(y|x,do(w)) && \text{Rule 3: } (Y \perp\!\!\!\perp W|X)_{G_{\overline{W}}} \\ &= P(y^*|x,do(w),r'_y) && \text{Rule 1: } (Y \perp\!\!\!\perp R_y|W)_{G_{\overline{W}}} \\ &= \frac{\sum_z P(y^*,r'_y|x,w,z)P(z|x)}{\sum_z P(r'_y|x,w,z)P(z|x)} && \text{Same as Model 3.} \end{aligned}$$

The first line follows from the randomization of X. The next line simply notes that W does not affect Y. Moreover, if W is held fixed by intervention, Y is independent of its missingness indicator R_y, allowing us to keep our focus only on the cases with complete data. Now all that is left to us is to identify the effect of W on Y^*, conditional on X. Note, however, that this structure is similar to that of Model 3, and the same steps applied there can be employed here.

4.6 Generalizing Causal Effects Across Domains

(Still working on it.)

We extend our causal diagram with "selection nodes" (S) indicating structural discrepancies between populations. Formally, switching between populations is represented by conditioning on different values of S. For instance, consider a source population Π and a target population Π^*. In this setting, the selection node S takes two values,

$$S = \begin{cases} s, & \text{if the data comes from the source population } \Pi; \\ s^*, & \text{if the data comes from the target population } \Pi^*. \end{cases}$$

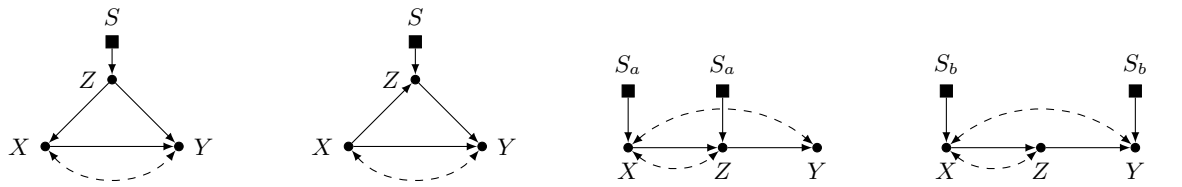
Concretely, for instance, $P(y|do(x), s)$ denotes the experimental distribution of Y in the source domain Π while $P(y|do(x), s^*)$ represents the experimental distribution of Y in the target domain Π^*. The selection node simply act as a "switcher," and accounts, non-parametrically, for any discrepancy

between the two populations. The presence of an edge $S \to V$ in a causal diagram means the *local* mechanism that assigns values to variable V may be different between populations. Conversely, the *absence* of an edge $S \to V$ represents the assumption that the *local* mechanism that assigns values to V is the same in both populations (Pearl, 1995, 2009; Bareinboim and Pearl, 2016). For clarity, and given their special status, selection nodes (S) are represented by square nodes (■).

In the following set of models (Figure 8):

- The **Query** is $\mathcal{Q} = P(y|do(x), s^*)$, that is the effect of X on Y in the target population Π^*;
- The **Data** available to us is a combination of observational data from the target domain Π^*, and experimental data from the source domain Π, such as $\mathcal{D} = \{P(y, x, z|s^*), P(y, z|do(x))\}$;
- The **Model** is a causal diagram $\mathcal{G}$, enriched with selection nodes S, describing discrepancies across domains.

Symbolically, our task is to remove the conditioning on $S = s^*$ on any $do(\cdot)$ expression, since we do not have experimental data on the target domain Π^*. Graphically, we will check for separation of the source of discrepancy (S) from key variables in the terms that describe out target quantity.



(a) Model 17 (b) Model 18 (c) Model 19: Source A (d) Model 19: Source B

Figure 8

References

Angrist, J. and Pischke, J.-S. (2009). *Mostly harmless econometrics: an empiricists guide*. Princeton: Princeton University Press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Bareinboim, E. and Pearl, J. (2012). Causal inference by surrogate experiments: z-identifiability. *Uncertainty in Artificial Intelligence*.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Bareinboim, E. and Tian, J. (2015). Recovering causal effects from selection bias. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3475–3481.

Blitzstein, J. K. and Hwang, J. (2015). *Introduction to probability*. Crc Press Boca Raton, FL.

Breskin, A., Cole, S. R., and Hudgens, M. G. (2018). A practical example demonstrating the utility of single-world intervention graphs. *Epidemiology*, 29(3):e20–e21.

Cinelli, C., Forney, A., and Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, 00491241221099552.

Cinelli, C. and Pearl, J. (2018). On the utility of causal diagrams in modeling attrition: a practical example. Technical report, Cognitive Systems Laboratory, UCLA.

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

Cox, D. R. and Wermuth, N. (2015). Design and interpretation of studies: Relevant concepts from the past and some extensions. *Observational studies*, 1(1):165–170.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

Elwert, F. (2013). Graphical causal models. In *Handbook of causal analysis for social research*, pages 245–273. Springer.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*, 40:31–53.

Evans, R. J. (2012). Graphical methods for inequality constraints in marginalized dags. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE.

Evans, R. J. (2018). Margins of discrete bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656.

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

Helske, J., Tikka, S., and Karvanen, J. (2021). Estimation of causal effects with small data in the presence of trapdoor variables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3):1030–1051.

Hernán, M. and Robins, J. (2020). *Causal inference: What if*. Boca Raton: Chapman & Hill/CRC.

Hünermund, P. and Bareinboim, E. (2019). Causal inference and data-fusion in econometrics. *arXiv preprint arXiv:1912.09104*.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Lee, S., Correa, J. D., and Bareinboim, E. (2020). General identifiability with arbitrary surrogate experiments. In *Uncertainty in artificial intelligence*, pages 389–398. PMLR.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. Forthcoming, *Journal of the American Statistical Association*, 0(0):1–16.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Pearl, J. (1993). Mediating instrumental variables. Technical report, Cognitive Systems Laboratory, UCLA.

Pearl, J. (1994). A probabilistic calculus of actions. In *Uncertainty Proceedings 1994*, pages 454–462. Elsevier.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: a primer*. John Wiley & Sons.

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

Perkovic, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2017). Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Robins, J. M. and Wasserman, L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 409–420.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1):27–42.

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

Ross, S. (2010). *A first course in probability*. Pearson.

Shpitser, I. and Pearl, J. (2008). *Dormant independence*. AAAI Press.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2012). On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*.

Tian, J. and Pearl, J. (2002a). A general identification condition for causal effects. Technical report, Cognitive Systems Laboratory, UCLA.

Tian, J. and Pearl, J. (2002b). On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 519–527. Morgan Kaufmann Publishers Inc.

Tikka, S. and Karvanen, J. (2018). Surrogate outcomes and transportability. *arXiv preprint arXiv:1806.07172*.

Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. *Uncertainty Artificial Intelligence, Elsevier, Amsterdam, The Netherlands*, pages 255–270.

Wermuth, N. and Cox, D. R. (2008). Distortion of effects caused by indirect confounding. *Biometrika*, 95(1):17–33.

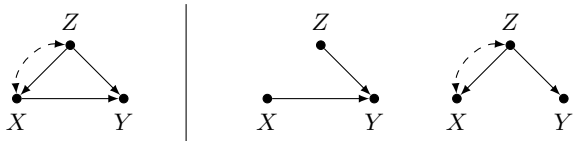
Wermuth, N. and Cox, D. R. (2014). Graphical markov models: overview. *arXiv preprint arXiv:1407.7783*.

Appendix A do-calculus derivations

A.1 Single Interventions

Model 1:

Auxiliary DAGs for the Main Text

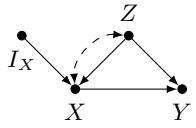


(a) G (b) $G_{\overline{X}}$ (c) $G_{\underline{X}}$

Figure 9: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= \sum_z P(y|do(x), z)P(z|do(x)) && \text{Law of total probability} \\
&= \sum_z P(y|do(x), z)P(z) && \text{Rule 3: } (Z \perp\!\!\!\perp I_X)_G \\
&= \sum_z P(y|x, z)P(z) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X|\{Z, X\})_G
\end{aligned}
$$

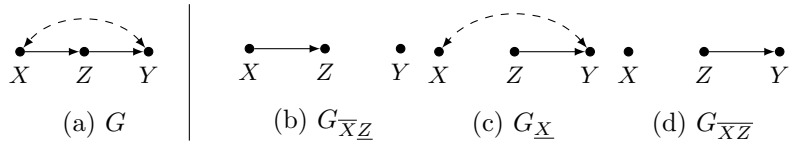


(a) $G \cup I_X$

Figure 10: Auxiliary DAGs for the Intervention Variable Formulation

Model 2:

Auxiliary DAGs for the Main Text

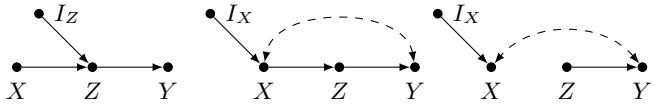


(a) G (b) $G_{\overline{X}\underline{Z}}$ (c) $G_{\underline{X}}$ (d) $G_{\overline{XZ}}$

Figure 11: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= \sum_z P(y|do(x), z)P(z|do(x)) && \text{Law of total probability} \\
&= \sum_z P(y|do(x), do(z))P(z|x) && \text{Rule 2: } (Y \perp\!\!\!\perp I_Z|\{X,Z\})_{G_{\overline{X}}} \text{ and } (Z \perp\!\!\!\perp I_X|X)_G \\
&= \sum_z P(y|do(z))P(z|x) && \text{Rule 3: } (Y \perp\!\!\!\perp I_X|Z)_{G_{\overline{Z}}} \\
&= \sum_z P(z|x) \sum_{x'} P(y|z, x')P(x') && \text{Backdoor adj. with } X
\end{aligned}
$$

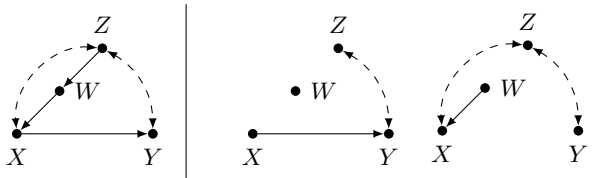


(a) $G_{\overline{X}} \cup I_Z$ (b) $G \cup I_X$ (c) $G_{\overline{Z}} \cup I_X$

Figure 12: Auxiliary DAGs for the Intervention Variable Formulation

Model 3:

Auxiliary DAGs for the Main Text

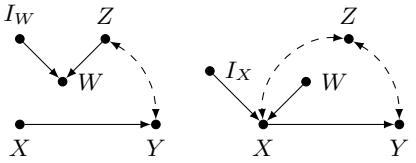


(a) G (b) $G_{\overline{XW}}$ (c) $G_{\overline{W}\underline{X}}$

Figure 13: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= P(y|do(x), do(w)) && \text{Rule 3: } (Y \perp\!\!\!\perp I_W|X)_{G_{\overline{X}}} \\
&= P(y|x, do(w)) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X|X)_{G_{\overline{W}}} \\
&= \frac{P(y, x|do(w))}{P(x|do(w))} && \text{Def. of conditional probability} \\
&= \frac{\sum_z P(y, x|w, z)P(z)}{\sum_z P(x|w, z)P(z)} && \text{Backdoor adj. with } Z
\end{aligned}
$$

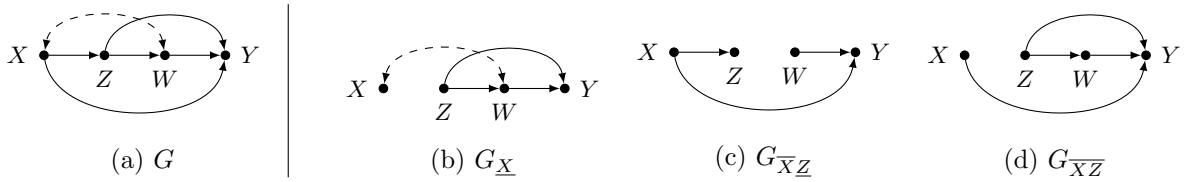


(a) $G_{\overline{X}} \cup I_W$ (b) $G_{\overline{W}} \cup I_X$

Figure 14: Auxiliary DAGs for the Intervention Variable Formulation

Model 4:

Auxiliary DAGs for the Main Text

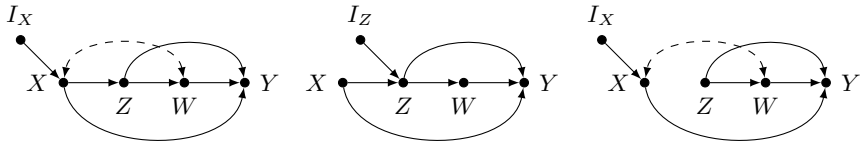


(a) G (b) $G_{\underline{X}}$ (c) $G_{\overline{X}\underline{Z}}$ (d) $G_{\overline{XZ}}$

Figure 15: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= \sum_{z,w} P(y|do(x),z,w)P(z|do(x))P(w|do(x),z) && \text{Law of total probability} \\
&= \sum_{z,w} P(y|x,z,w)P(z|x)P(w|do(x),z) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X|\{Z,W,X\})_G \text{ and } (Z \perp\!\!\!\perp I_X|X)_G \\
&= \sum_{z,w} P(z|x)P(y|x,z,w)P(w|do(x),do(z)) && \text{Rule 2: } (W \perp\!\!\!\perp I_Z \mid \{X,Z\})_{G_{\overline{X}}} \\
&= \sum_{z,w} P(z|x)P(y|x,z,w)P(w|do(z)) && \text{Rule 3: } (W \perp\!\!\!\perp I_X \mid Z)_{G_{\overline{Z}}} \\
&= \sum_{z,w} P(y|x,z,w)P(z|x)\sum_{x'} P(w|z,x')P(x') && \text{Backdoor adj. with } X
\end{aligned}
$$



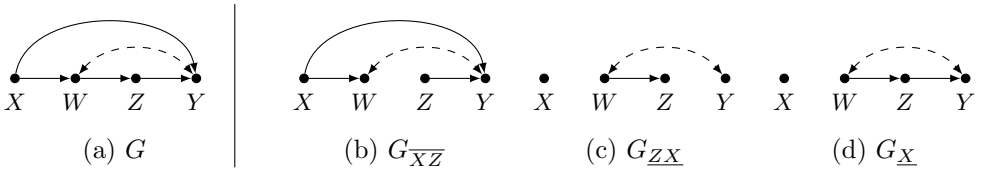
(a) $G \cup I_X$ (b) $G_{\overline{X}} \cup I_Z$ (c) $G_{\overline{Z}} \cup I_X$

Figure 16: Auxiliary DAGs for the Intervention Variable Formulation

A.2 Multiple and Sequential Interventions

Model 5:

Auxiliary DAGs for the Main Text

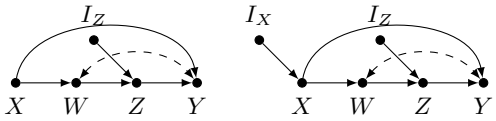


(a) G (b) $G_{\overline{XZ}}$ (c) $G_{\underline{ZX}}$ (d) $G_{\underline{X}}$

Figure 17: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x),do(z)) &= \sum_w P(y|do(x),do(z),w)P(w|do(x),do(z)) && \text{Law of total probability} \\
&= \sum_w P(y|do(x),do(z),w)P(w|do(x)) && \text{Rule 3: } (W \perp\!\!\!\perp I_Z|X)_{G_{\overline{X}}} \\
&= \sum_w P(y|x,z,w)P(w|x) && \text{Rule 2: } (Y \perp\!\!\!\perp \{I_X,I_Z\}|\{W,X,Z\})_G \text{ and } (W \perp\!\!\!\perp I_X|X)_G
\end{aligned}
$$

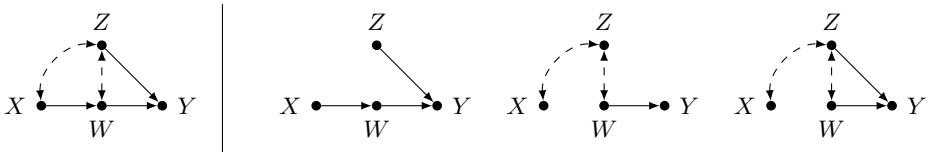


(a) $G_{\overline{X}} \cup I_Z$ (b) $G \cup \{I_X, I_Z\}$

Figure 18: Auxiliary DAGs for the Intervention Variable Formulation

Model 6:

Auxiliary DAGs for the Main Text

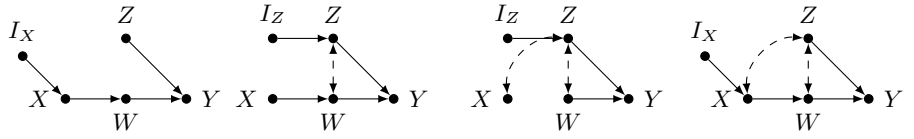


(a) G (b) $G_{\overline{Z}}$ or $G_{\overline{XZ}}$ (c) $G_{\underline{ZX}}$ (d) $G_{\underline{X}}$

Figure 19: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x),do(z)) &= \sum_w P(y|do(x),do(z),w)P(w|do(x),do(z)) && \text{Law of total probability} \\
&= \sum_w P(y|do(z),w)P(w|do(x)) && \text{Rule 3: } (Y \perp\!\!\!\perp I_X|\{Z,W\})_{G_{\overline{Z}}} \text{ and } (W \perp\!\!\!\perp I_Z|X)_{G_{\overline{X}}} \\
&= \sum_w P(y|z,w)P(w|x) && \text{Rule 2: } (Y \perp\!\!\!\perp I_Z|\{W,Z\})_G \text{ and } (W \perp\!\!\!\perp I_X|X)_G
\end{aligned}
$$

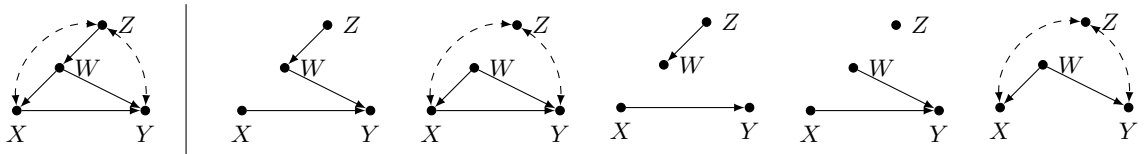


(a) $G_{\overline{Z}} \cup I_X$ (b) $G_{\overline{X}} \cup I_Z$ (c) $G_{\underline{X}} \cup I_Z$ (d) $G \cup I_X$

Figure 20: Auxiliary DAGs for the Intervention Variable Formulation

Model 7:

Auxiliary DAGs for the Main Text

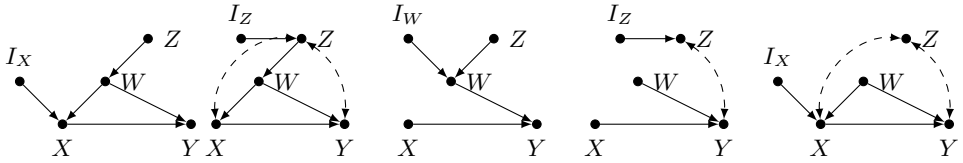


(a) G (b) $G_{\overline{XZ}}$ (c) $G_{\underline{Z}}$ (d) $G_{\overline{XZ}\underline{W}}$ (e) $G_{\overline{WXZ}}$ (f) $G_{\overline{W}\underline{X}}$

Figure 21: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x),do(z)) &= \sum_w P(y|do(x),do(z),w)P(w|do(x,z)) && \text{Law of total probability} \\
&= \sum_w P(y|do(x),do(z),w)P(w|do(z)) && \text{Rule 3: } (W \perp\!\!\!\perp I_X \mid Z)_{G_{\overline{Z}}} \\
&= \sum_w P(y|do(x),do(z),w)P(w|z) && \text{Rule 2: } (W \perp\!\!\!\perp I_Z|Z)_G \\
&= \sum_w P(y|do(x),do(z),do(w))P(w|z) && \text{Rule 2: } (Y \perp\!\!\!\perp I_W \mid \{X,Z,W\})_{G_{\overline{XZ}}} \\
&= \sum_w P(y|do(w),do(x))P(w|z) && \text{Rule 3: } (Y \perp\!\!\!\perp I_Z \mid X,W)_{G_{\overline{WX}}} \\
&= \sum_w P(y|do(w),x)P(w|z) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X \mid \{W,X\})_{G_{\overline{W}}} \\
&= \sum_w \frac{P(y,x|do(w))}{P(x|do(w))}P(w|z) && \text{Conditional Probability} \\
&= \sum_w \frac{\sum_{z'} P(y,x|w,z')P(z')}{\sum_{z'} P(x|w,z')P(z')}P(w|z) && \text{Backdoor adj. with } X
\end{aligned}
$$



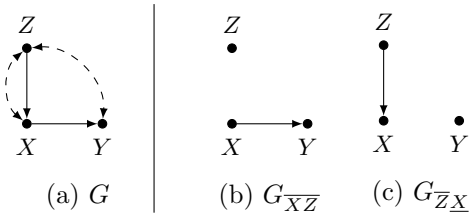
(a) $G_{\overline{Z}} \cup I_X$ (b) $G \cup I_Z$ (c) $G_{\overline{XZ}} \cup I_W$ (d) $G_{\overline{WX}} \cup I_Z$ (e) $G_{\overline{W}} \cup I_X$

Figure 22: Auxiliary DAGs for the Intervention Variable Formulation

A.3 Surrogate Interventions and Surrogate Outcomes

Model 8:

Auxiliary DAGs for the Main Text

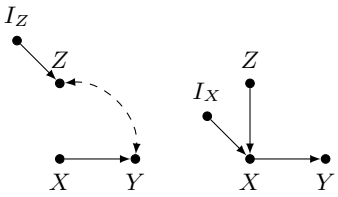


(a) G (b) $G_{\overline{XZ}}$ (c) $G_{\overline{Z}\underline{X}}$

Figure 23: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= P(y|do(x), do(z)) && \text{Rule 3: } (Y \perp\!\!\!\perp I_Z \mid X)_{G_{\overline{X}}} \\
&= P(y|x, do(z)) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X|X)_{G_{\overline{Z}}}
\end{aligned}
$$



(a) $G_{\overline{X}} \cup I_Z$ (b) $G_{\overline{Z}} \cup I_X$

Figure 24: Auxiliary DAGs for the Intervention Variable Formulation

Model 9:

Auxiliary DAGs for the Main Text

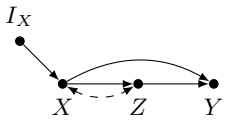
(a) G (b) $G_{\underline{X}}$

Figure 25: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$P(y|do(x)) = \sum_w P(y|z, do(x))P(z|do(x)) \qquad \text{Law of total probability}$$

$$= \sum_z P(y|x, z)P(z|do(x)) \qquad \text{Rule 2: } (Y \perp\!\!\!\perp I_X \mid \{Z, X\})_G$$

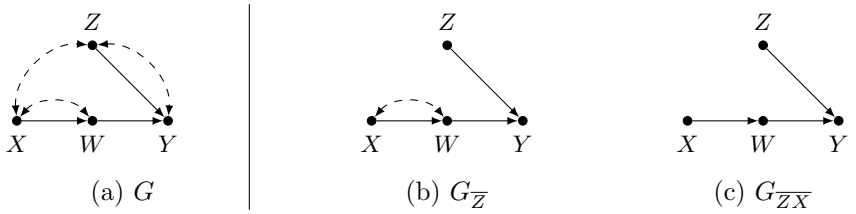


(a) $G \cup I_X$

Figure 26: Auxiliary DAGs for the Intervention Variable Formulation

Model 10:

Auxiliary DAGs for the Main Text



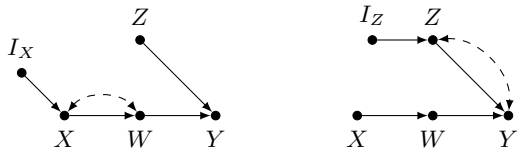
(a) G (b) $G_{\overline{Z}}$ (c) $G_{\overline{ZX}}$

Figure 27: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$P(y|do(x),do(z)) = \sum_w P(y|w,do(x),do(z))P(w|do(x),do(z)) \qquad \text{Law of total probability}$$

$$= \sum_w P(y|w,do(z))P(w|do(x)) \qquad \text{Rule 3: } (Y \perp\!\!\!\perp I_X \mid W,Z)_{G_{\overline{Z}}} \text{ and } (W \perp\!\!\!\perp I_Z \mid X)_{G_{\overline{X}}}$$



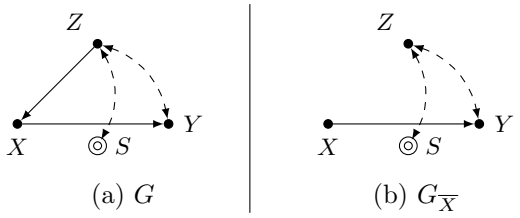
(a) $G_{\overline{Z}} \cup I_X$ (b) $G_{\overline{X}} \cup I_Z$

Figure 28: Auxiliary DAGs for the Intervention Variable Formulation

A.4 Recovering from Selection bias

Model 11:

Auxiliary DAGs for the Main Text



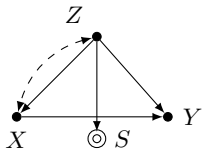
(a) G (b) $G_{\overline{X}}$

Figure 29: Auxiliary DAGs for the *do*-calculus Derivation

$$
\begin{aligned}
P(y|do(x)) &= P(y|do(x), s) && \text{Rule 1: } (Y \perp\!\!\!\perp S|X)_{G_{\overline{X}}} \\
&= \sum_z P(y|x, z, s)P(z|s) && \text{Backdoor adj. with } Z
\end{aligned}
$$

Model 12:

Auxiliary DAGs for the Main Text



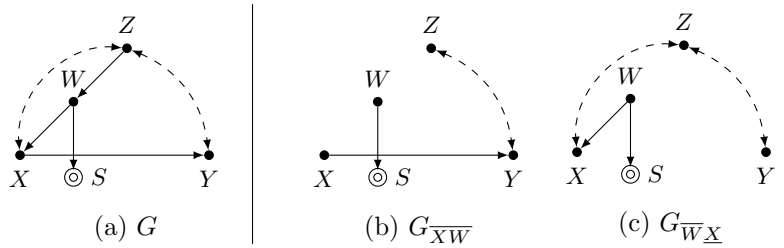
(a) G

Figure 30: Auxiliary DAGs for *do*-calculus Derivation

$$
\begin{aligned}
P(y|do(x)) &= \sum_z P(y|x, z)P(z) && \text{Backdoor adj. with } Z \\
&= \sum_z P(y|x, z, s)P(z) && \text{Rule 1: } (Y \perp\!\!\!\perp S|X, Z)_G
\end{aligned}
$$

Model 13:

Auxiliary DAGs for the Main Text

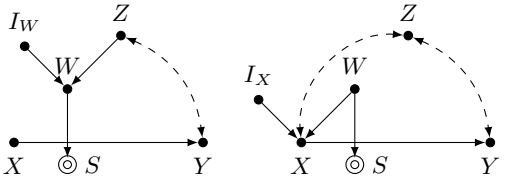


(a) G (b) $G_{\overline{XW}}$ (c) $G_{\overline{W}\underline{X}}$

Figure 31: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= P(y|do(x), do(w)) && \text{Rule 3: } (Y \perp\!\!\!\perp I_W|X)_{G_{\overline{X}}} \\
&= P(y|x, do(w)) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X|X)_{G_{\overline{W}}} \\
&= \frac{P(y, x|do(w))}{P(x|do(w))} && \text{Def. of conditional probability} \\
&= \frac{\sum_z P(y, x|w, z)P(z)}{\sum_z P(x|w, z)P(z)} && \text{Backdoor adj. with } Z \\
&= \frac{\sum_z P(y, x|w, z, s)P(z)}{\sum_z P(x|w, z, s)P(z)} && \text{Rule 1: } (\{Y, X\} \perp\!\!\!\perp S|\{W, Z\})_G
\end{aligned}
$$



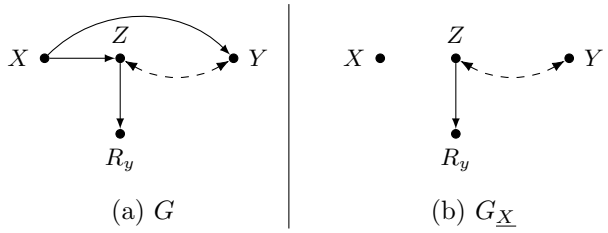
(a) $G_{\overline{X}} \cup I_W$ (b) $G_{\overline{W}} \cup I_X$

Figure 32: Auxiliary DAGs for the Intervention Variable Formulation

A.5 Recovering from Missing Data

Model 14:

Auxiliary DAGs for the Main Text

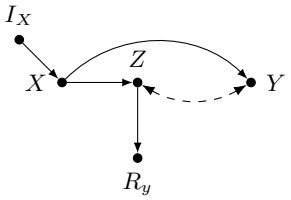


(a) G (b) $G_{\underline{X}}$

Figure 33: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= P(y|x) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X|X)_G \\
&= \sum_z P(y|x,z)P(z|x) && \text{Law of total probability} \\
&= \sum_z P(y|x,z,r'_y)P(z|x) && \text{Rule 1: } (Y \perp\!\!\!\perp R_y|X,Z)_G
\end{aligned}
$$

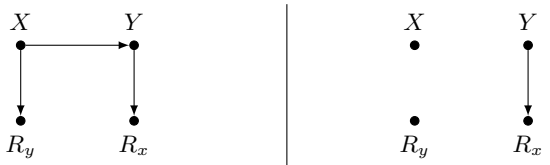


(a) $G \cup I_X$

Figure 34: Auxiliary DAGs for the Intervention Variable Formulation

Model 15:

Auxiliary DAGs for the Main Text

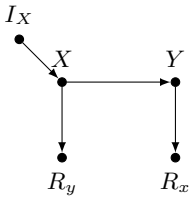


(a) G (b) $G_{\underline{X}}$

Figure 35: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= P(y|x) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X|X)_G \\
&= P(y^*|x, r'_y) && \text{Rule 1: } (Y \perp\!\!\!\perp R_y|X)_G \\
&= \frac{P(y^*, x, r'_y)}{\sum_{y^*} P(y^*, x, r'_y)} && \text{Definition of conditional probability} \\
&= \frac{P(x|y^*, r'_y)P(y^*, r'_y)}{\sum_{y^*} P(x|y^*, r'_y)P(y^*, r'_y)} && \text{Chain rule} \\
&= \frac{P(x^*|y^*, r'_y, r'_x)P(y^*, r'_y)}{\sum_{y^*} P(x^*|y^*, r'_y, r'_x)P(y^*, r'_y)} && \text{Rule 1: } (X \perp\!\!\!\perp R_x|Y)_G
\end{aligned}
$$

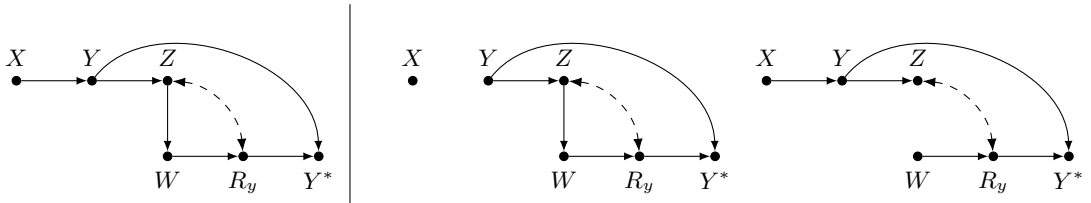


(a) $G \cup I_X$

Figure 36: Auxiliary DAGs for the Intervention Variable Formulation

Model 16:

Auxiliary DAGs for the Main Text

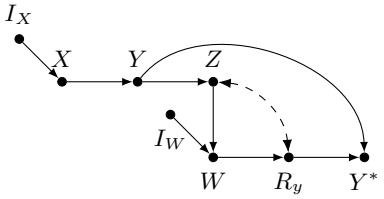


(a) G (b) $G_{\underline{X}}$ (c) $G_{\overline{W}}$

Figure 37: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x)) &= P(y|x) && \text{Rule 2: } (Y \perp\!\!\!\perp I_X \mid X)_G \\
&= P(y|x, do(w)) && \text{Rule 3: } (Y \perp\!\!\!\perp I_W|X)_G \\
&= P(y^*|x, do(w), r'_y) && \text{Rule 1: } (Y \perp\!\!\!\perp R_y|W)_{G_{\overline{W}}} \\
&= \frac{\sum_z P(y^*, r'_y|x, w, z)P(z|x)}{\sum_z P(r'_y|x, w, z)P(z|x)} && \text{New napkin model with backdoor adj. with } Z
\end{aligned} \tag{2}
$$



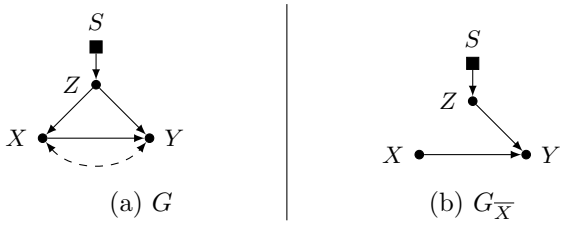
(a) $G \cup \{I_X, I_W\}$

Figure 38: Auxiliary DAGs for the Intervention Variable Formulation

A.6 Generalizing causal Effects Across Domains

Model 17:

Auxiliary DAGs for the Main Text

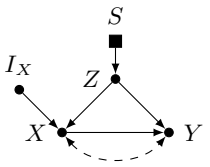


(a) G (b) $G_{\overline{X}}$

Figure 39: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x), s^*) &= \sum_z P(y|do(x), z, s^*) P(z|do(x), s^*) && \text{Law of total Probability} \\
&= \sum_z P(y|do(x), z, s^*) P(z|s^*) && \text{Rule 3: } (Z \perp\!\!\!\perp I_X \mid S)_G \\
&= \sum_z P(y|do(x), z, s) P(z|s^*) && \text{Rule 1: } (Y \perp\!\!\!\perp S \mid X, Z)_{G_{\overline{X}}}
\end{aligned}
$$

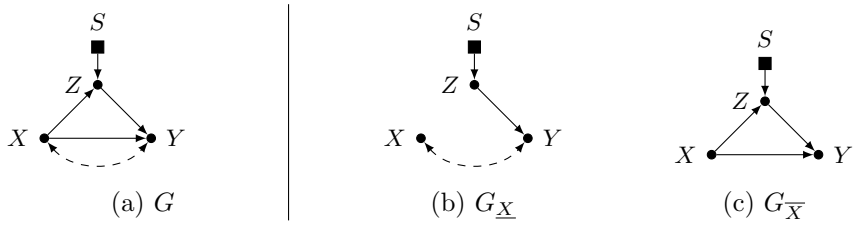


(a) $G \cup I_X$

Figure 40: Auxiliary DAGs for the Intervention Variable Formulation

Model 18:

Auxiliary DAGs for the Main Text

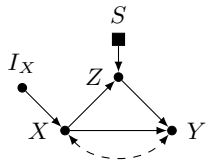


(a) G (b) $G_{\underline{X}}$ (c) $G_{\overline{X}}$

Figure 41: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x), s^*) &= \sum_z P(y|do(x), z, s^*)P(z|do(x), s^*) && \text{Law of total Probability} \\
&= \sum_z P(y|do(x), z, s^*)P(z|x, s^*) && \text{Rule 2: } (Z \perp\!\!\!\perp I_X \mid \{S, X\})_G \\
&= \sum_z P(y|do(x), z, s)P(z|x, s^*) && \text{Rule 1: } (Y \perp\!\!\!\perp S \mid X, Z)_{G_{\overline{X}}}
\end{aligned}
$$

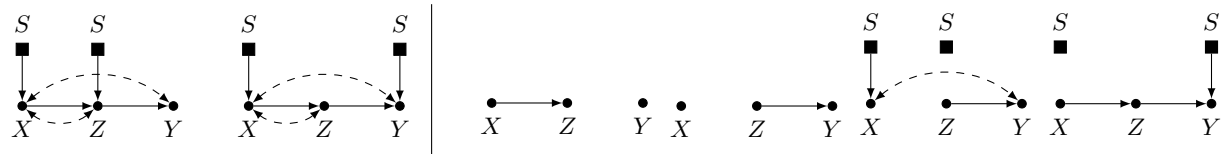


(a) $G \cup I_X$

Figure 42: Auxiliary DAGs for the Intervention Variable Formulation

Model 19:

Auxiliary DAGs for the Main Text

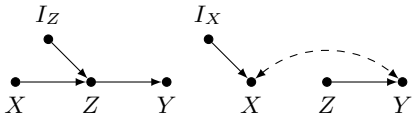


(a) G^A: Source A (b) G^B: Source B (c) $G_{\overline{X}\underline{Z}}$ (d) $G_{\overline{XZ}}$ (e) $G^A_{\overline{Z}}$ (f) $G^B_{\overline{X}}$

Figure 43: Auxiliary DAGs for the *do*-calculus Derivation

Intervention Variable Formulation

$$
\begin{aligned}
P(y|do(x), s^*) &= \sum_z P(y|do(x), z, s^*) P(z|do(x), s^*) && \text{Law of total probability} \\
&= \sum_z P(y|do(x), do(z), s^*) P(z|do(x), s^*) && \text{Rule 2: } (Y \perp\!\!\!\perp I_Z \mid \{X, Z\})_{G_{\overline{X}}} \\
&= \sum_z P(y|do(z), s^*) P(z|do(x), s^*) && \text{Rule 3: } (Y \perp\!\!\!\perp I_X \mid Z)_{G_{\overline{Z}}} \\
&= \sum_z P(y|do(z), s_a) P(z|do(x), s_b) && \text{Rule 1: } (Y \perp\!\!\!\perp S = s_a \mid Z)_{G^A_{\overline{Z}}} \text{ and } (Z \perp\!\!\!\perp S = s_b \mid X)_{G^B_{\overline{X}}}
\end{aligned}
$$



(a) $G_{\overline{X}} \cup I_Z$ (b) $G_{\overline{Z}} \cup I_X$

Figure 44: Auxiliary DAGs for the Intervention Variable Formulation