Revisiting sample selection as a threat to internal validity: A new analytical tool, lessons, and examples [DRAFT]

Adam Rohde[*], Chad Hazlett[†]

October 2022

Abstract

Researchers often seek to estimate the effect of a treatment on an outcome within a sample that has been drawn in some selective way from a larger population. Such selective sampling not only changes the population about which we make inferences, but can bias our estimate of the causal effect for the units in the sample, thus threatening the "internal validity" of the estimate Campbell (1957). Further, it is not possible to know what the causal effect would be in any other population of eventual interest if we cannot first obtain an unbiased estimate in the observed sample—a result we formalize below. That selective sampling can threaten even internal validity has long been known, and over the decades different research traditions have offered guidelines for assessing the threats to internal validity posed by sample selection. We employ formal graphical tools for causal reasoning to more fully and rigorously characterize the (i) the settings in which selective sampling does and does not bias the "internal effect estimate", and (ii) the conditions under which this bias can theoretically be corrected, and how to do so. These results are collectively conveyed through a graphical criterion that investigators can apply in their circumstances to examine the threats to bias and opportunities for correction given a graphical causal model. A number of common lessons emerge, including that many forms of selection, including selection processes influenced by the treatment or a mediator, are not always problematic. That said, the central lesson is that many complications may arise, requiring the researcher to to use these tools to examine how selection processes bias the result or can be corrected under specific causal structures the user cannot reject as plausible.

1   Introduction

In applied quantitative research, we typically estimate quantities of interest using samples of data drawn in non-random ways from the population of eventual scientific or policy interest. Where we are interested in learning the causal effect of one thing (D) on another (Y), this can create two types of problems. One problem involves *external validity* (Campbell, 1957): how might the effect of interest look when averaged over some population of interest, as compared to its average in our sample? This is an important question which has achieved considerable recent attention (for a recent review and development, see Egami and Hartman (2022)). While investigators are accustomed to obviating this problem by restricting their inferences to the observed sample, the second and more pernicious problem is that selective sampling can threaten even *internal validity*, biasing our estimate of the treatment effect as defined as an average over the sample. This has long been known. For example, Berk (1983) states, that under selective sampling, "Both internal and external validity are implicated. There is no escape by limiting one's causal conclusions to the population from which the nonrandom sample was drawn (or even the sample itself)."

---

[*]Department of Statistics, UCLA. adamrohde@ucla.edu

[†]Associate Professor, Departments of Statistics & Political Science, UCLA. chazlett@ucla.edu

Notwithstanding increased recent attention to external validity through a rapidly advancing literature on "transportation" and "generalization" (e.g., Pearl and Bareinboim (2011); Pearl (2015b); Bareinboim and Pearl (2016); Correa et al. (2018, 2019); Egami and Hartman (2021)), internal validity remains a key concern for investigators for two reasons. First, internal validity may reasonably be of sufficient scientific interest in many settings. For example, we may study a causal mechanism that is thought to be approximately the same in most or all individuals or populations thereof. Or, we may design a study so that the sample in hand is already representative of a sufficiently interesting target group, e.g. those who would be eligible for a new policy or therapeutic intervention. Second, even where ultimate interest lies in a claim of generalization to some broader or different population, we formally show how internal validity is always required for (and easier to achieve than) external validity: identification of internally valid causal effects requires a subset of the causal assumptions required for identification of causal effect estimates that generalize from the sample.[1]

What detailed knowledge must investigators have to avoid or remedy these biases? What types of selection processes are problematic and what types are not? Can the internal validity of randomized experiments be threatened by sample selection? Where there are threats, under what conditions can they be remedied and how? Methodologists in a variety of disciplines have long sought to address these perennial questions, proposing guidelines and methodological fixes since at least Campbell (1957), with well-known later contributions in Greenland (1977), Heckman (1979), Berk (1983), Hernán et al. (2004), Wooldridge (2010), and Elwert and Winship (2014), among others. However, these approaches do not amount to a comprehensive and rigorous treatment of the problem, nor have they resulted in a complete approach that a researcher can apply to reveal any problem or solution that exists for any causal structure they deem plausible in their setting. As noted by Berk (1983), "while considerable effort has been devoted to documenting sampling biases within traditional survey sample approaches...we are a very long way from a formal theory."

Fortunately, we are now in a position to answer Berk's 1983 call for such a formal theory, and indeed to provide a procedure for answering such questions as they apply to any causal structure. In recent years, researchers have benefited from the growth and formalization of methods that rigorously define causal quantities, characterize when they can or cannot be estimated from the data, and point to solutions for correcting biases, employing the devices of potential outcomes, structural causal models, and graphical causal models.[2] With these tools comes the possibility of more completely and rigorously posing and answering questions about sample selection and internal validity. In particular, we develop a graphical approach to understanding the threats such sample selection can pose for internally valid causal effect estimates. This involves first introducing "internal selection graphs", an extension of standard graphical approaches that visually shows the consequences of sample selection for the relationships between variables. Second, we provide rules for how to use these extended graphs to determine when causal quantities are identifiable under selective

---

[1]While we offer a formalization of these claims, we do so to add rigor to statements that have long been made. For example, Campbell (1957) states that "Internal validity is the prior and indispensable consideration", and Campbell and Stanley (1963) argue that "Internal validity is the basic minimum without which any experiment is uninterpretable..." Shadish et al. (2002) clarify that the primacy of internal validity is specific to "cause-probing research," which is the context of our paper. Additionally, any claim of internal validity entails some conception of the population from which the sample was drawn. Without this, we cannot begin to analyze whether the sample selection mechanism threatens our ability to draw causal inferences for the selected sample.

[2]While we direct readers to texts such as Pearl (1988, 2009); Imbens and Rubin (2015); Hernán and Robins (2020) for a fuller review of these concepts, our goal is to explain our use of these tools as they arise so as to provide a mostly self-contained guide to users.

sampling. These tools aim to provide a wider audience with the ability to analyze how sample selection might threaten internal validity in their applications. Applied to a number of common causal structures, our tools support a few broad findings of note, including: (i) sample selection is not always problematic to internal validity (e.g., post-treatment selection, or confounders of selection and the outcomes (Hernán, 2017) are not biasing on their own), (ii) some causal effects can still be identified when sample selection is based on a mediator,[3] (iii) sample selection can influence the identification of causal effects even when it is not a collider, and (iv) the threats from sample selection for internal validity are not the same as those for external validity, nor are the means of addressing those threats. While communicating those broad conclusions helps to signal the complexity and possibly non-intuitive nature of this problem, our key message echoing Berk, 1983 and Greenland, 2022, is that for any specific application, the details of the causal structure and sample selection mechanism determine whether sample selection threatens internal validity and what might be done about it. Our primary contribution is thus the graphical criterion and tool we provide that enables investigators to reliably perform such diagnostic and prescriptive analyses in their setting.

2   Working example: Racial bias in policing

For concreteness, we employ a single working example throughout the paper.[4] Inspired by Knox et al. (2020), we look to data from police "stops" (an encounter in which a police officer stops and interacts with a civilian, on foot or in a vehicle). The question is then what can be learned from such data about how the *police-perceived* race of the civilian stopped alters the chances that police employ force in that encounter. The emphasis on *police-perceived* race is important for two reasons. First, it reminds us of the possible misperception and conceptual ambiguity regarding the police officer's belief about the civilian, as opposed to how the civilian would identify. Second, it reminds us that we are interested in the question of how the police officer's *belief* regarding the civilian's race might have influenced the outcome.

The key challenge we consider here is that the data are limited to administrative records that are produced only when the police officer stops a civilian, thus making a report, citation, or arrest that appears in the data. Hence, such studies are restricted to a sample of civilian-police encounters that has been selected in a non-random way, as the encounters in which the officers stop the civilian depends greatly on characteristics of the encounter (in particular on characteristics of the officer and the civilian).

To begin addressing this, we must first be willing to contemplate the causal structure of the system in question, meaning that we consider how each variable in this system *could* cause or be caused by others observed variables, or by a web of unobserved variables that influence more than one observable. Here we assume it is possible that police-perceived race influences the ways in which the officers interact with the civilian, both through whether or not the officers make a stop and whether or not the officers use force. Second, police stopping of a civilian is a prerequisite to police use of force; if no stop is made, then officers cannot use force. Third, the police administrative records do not capture all of the factors that influence whether or not officers make a stop and/or use force; that is, there are unobserved common causes of making

---

[3]Mediators are nodes in causal graphs the lie on at least one causal path from the treatment to the outcome. See below for a discussion of causal graphs and paths.

[4]Appendix A offers additional illuminating—and perhaps entertaining—exercises which the reader can use to test and develop their understanding. These include understanding why taller NBA players are worse free-throw shooters, whether imagining applying eyeliner helps one lose weight, and seeing if doing more can feel like less, and more.

a stop and of using force. These possible relationships are represented in the usual causal graphical form (that is, as a directed acyclic graph (DAG)) in Figure 1(a), which mimics Knox et al. (2020) Figure 1.[5] Here D represents an indicator for police–civilian encounters involving a civilian that was perceived to be from a minority ethnic group, S represents an indicator for police–civilian encounters in which the police make a stop, Y is an indicator for police use of force, and U are the unobserved common causes of making a stop and use of force.

Figure 1: Racial Discrimination in the Use of Force by Police

(a) *Proposed directed acyclic graph (DAG)*

(b) *Extended DAG revealing "bridge" formed by selection*
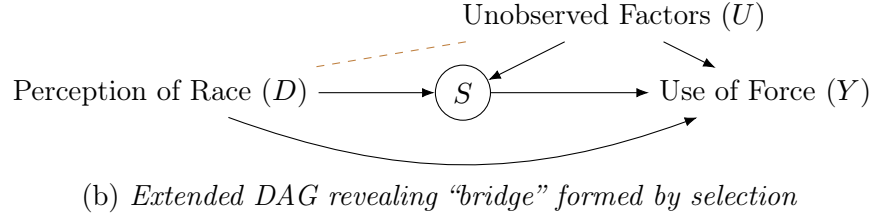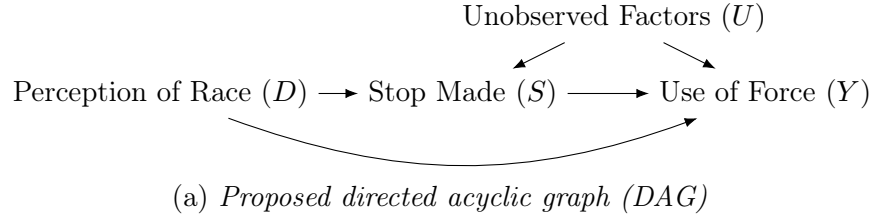
This causal structure immediately points to a number of familiar problems. We use Figure 1(b) to annotate the original DAG in ways that make these problems apparent. First, and central to this project, we are forced to "select" data for which a stop occurred ($S = 1$), and this stop is a mediator between perceived ethnicity (D) and the use of force (Y). This coniditioning on S is represented by the circle around S in Figure 1(b), and it has two immediate consequences. First, because some of the effect of D flows through S to Y, conditioning on S in this way blocks some of the effect we wish to study, leaving only the part of the effect that flowed directly from D to Y to be estimable. Second, by conditioning on $S = 1$, we are conditioning on a "collider" or a common consequence of D and U.[6] Conditioning on a collider (or a descendant[7] of a collider) *can* create purely statistical associations between the parents of the collider. We can represent this purely statistical association with a dashed undirected edge $D\text{-}\text{-}\text{-}U$ on Figure 1 (b). This leads to additional problems. For example, by generating an association between D and U, the pathway $D - U - Y$ now generates an assocation

---

[5]For those unaccustomed to relying on DAGs, we note that it is important that the user include on such a DAG any arrow that could exist, meaning that to leave out an arrow requires a strong argument for why no such arrow exists. The requirement that such a causal structure be assumed at this level of detail may seem like a drawback. However, it asks little more than what is absolutely necessary to draw conclusions about how selection impacts the result and what can be done about it indeed depends on the causal structure at this level of detail. An unwillingness to transparently state what causal structure the researcher believes to be plausible would not free us from the consequences of that structure, only blind us to the possible problems and solutions associated with that structure.

[6]A collider is a node in the graph into which two arrows point: $D \to S \leftarrow U$. See Pearl (2009) for an introduction to causal graphical models and colliders.

[7]We do not consider a node to be a descendant of itself. See Definition B.9 in Appendix B.

between D and Y that is not due to the effect of D on Y. Thus, a comparison of the rates of use of force across values for (perceived) civilian race will be biased for the total effect and direct effect *even for encounters in which a stop was made* (the selected sample). Knox et al. (2020) agree with these conclusions but use a non-graphical approach to reach them.

We note that conditioning on a collider does not *always* create an association among its parents nodes, as there are a number of counterexamples. These counterexamples often impose strong or knife-edge assumptions and cannot easily be defended, but in other cases they are plausible. In Section B.6 of Appendix B, we discuss conditions under marginally independent parents of a collider maintain their independence conditional on the collider for discrete variables and the implausible circumstances under which no association would be created between D and U conditional on $S = 1$ in the present example. In Section B.7 of Appendix B, we show that zero interaction information implies that marginally independent parents of a collider maintain their independence conditional on the collider. Because conditioning on a collider can and often will create an association among the parents, we proceed as though it does to maintain a more conservative analysis.

Finally, an important feature of these scenarios is to be clear and explicit about the causal estimand we are aiming to identify and estimate. In plain languge, Wwe are interested in "the effect of perceived race on use-of-force", specifically among (averaged over) those cases where a stop occurred. This description, however, is ambiguous regarding which counterfactuals we mean to compare. Specifically we could be referring to two different quantities:

- **The direct effect, among those who were stopped.** Once a stop is made, how does perceived race influence the risk that force is used? This considers an individual i for whom a stop was actually made, and compare whether force would have been had they been perceived to be of one race ($r_i = r$) or another ($r_i = r'$).

- **The total effect, among those who were stopped** considers the "the effect of perceived race through triggering a stop that might not have otherwise occurred", as well as the direct effect above that applies "once a stop occurs". Consider an individual i with perceived race $r_i = r$, for whom a stop really did occur ($s_i = 1$). For this individual, had they been perceived to be of a different race ($r_i = r'$), a stop may or may not have still occurred, and force may or may not have been used. We are interested in the outcome (use-of-force) for this individual, had we changed their perceived race, recognizing that doing so might also have changed whether a stop would have occurred. The average total effect among those who were stopped — or more generically the (average) "internal total effect" — is be composed of such counterfactual comparisons for all units i that were, in the real data, actually subject to a stop (without manipulation of their perceived race).[8]

---

[8]Describing these quantities in terms of the "ideal" or "target" randomized trials one would have to conduct can be a useful exercise, in part for the gap it may reveal between causal quantitites and what can straightforwadaly be learned by experimentation. For the "total effect", we must be able to randomly manipulate the perceived race of the civilian at the beginning of the encounter (before the officer decides whether to make a stop). Yet, to limit our inference to the individuals who would have been stopped, we must have a way of recording whether, when the officer perceived the race as they would have without intervention, they would have made a stop. Such an experiment would require a time-travel or memory-wiping mechanism, or the ability to monitor "two worlds" (one with the natural perceived race, one with the counterfactual). For the indirect effect, the manipulation in question comes later: we wait and see if a stop is made, and only would we need to wipe the officer's memory and intervene on their perception of race. Alternatively we could intervene on perceived race prior to the encounter, determine when a stop would have occured for unit i under their naturally perceived race (keeping only individuals i for whom that is the case), and then, when we set race for each such indivudal i to its non-unobserved value, we must simultaneously force a stop to occur, even if it would not have otherwise. Such interventions are clearly infeasible, which

3 Background

Notation and key quantities Let's introduce some notation to clarify the types of casual effects we mean when we say internally valid causal effects and causal quantities. A potential outcome, $Y_d[i]$, is the value that the variable Y would have taken for unit i, if the variable D for unit i had been set, possibly counterfactually, to the value d. (Splawa-Neyman et al., 1990; Rubin, 1974, 1978, 1990) The unit-level causal effect of setting D to d relative to D to d' is $\tau_i = Y_d[i] - Y_{d'}[i]$.

The fundamental problem of causal inference, however, is that we are never able to observe more than one of the potential outcomes for a given unit and so cannot calculate unit level causal effects. (Rubin, 1978; Holland, 1986; Imbens and Rubin, 2015; Westreich et al., 2015) Despite this, these are the building blocks of typical causal inferential targets. When readers see "internally valid causal effects," we suspect that most have in mind something like the sample average treatment effect (SATE), $\frac{1}{N}\sum_{i=1}^{N}\tau_i$, which is the simple average of the unit level effects across the units that are observed in the sample. This is a perfectly good target causal effect and the discussion that follows will apply to this. However, researchers might also be interested in the causal effect for the *sup-population for which the selected sample is representative.* We write this as $\mathbb{E}[\tau_i|S_i = 1]$ and call it the selected-population average treatment effect (SPATE). We will use a binary variable, S, to denote non-random sample selection. $S_i = 1$ can be interpreted as indicating that a unit is included in the observed study sample. It may also be thought of as indicating the subset of the population from which the observed study sample was randomly drawn.[9] Both the SATE and SPATE are different from the population average treatment effect (PATE), $\mathbb{E}[\tau_i]$.

$$\text{SATE} = \frac{1}{N}\sum_{i=1}^{N}\tau_i$$

$$\text{SPATE} = \mathbb{E}[\tau_i|S_i = 1]$$

$$\text{PATE} = \mathbb{E}[\tau_i] = \mathbb{E}[\tau_i|S_i = 1]p(S_i = 1) + \mathbb{E}[\tau_i|S_i = 0]p(S_i = 0)$$

If we have a random sample from the population, $\mathbb{E}[\text{SATE}] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}\tau_i] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\tau_i] = \mathbb{E}[\tau_i] = \text{PATE}$. However, in the more interesting case of a non-random sample from the population, we instead have $\mathbb{E}[\text{SATE}|S_i = 1] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}\tau_i|S_i = 1] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\tau_i|S_i = 1] = \mathbb{E}[\tau_i|S_i = 1] = \text{SPATE} \neq \text{PATE}$. In the non-random sampling case, the SATE is unbiased for the SPATE, not the PATE, since the sample at hand can be thought of as a representative sample of the sub-population indicated by $S = 1$.

This brings us to a more formal definition of internal validity: An estimation strategy is said to be "internally valid" if it can unbiasedly or consistently estimate the SPATE. In what follows, we do not always differentiate between units eligible to be in the selected sample from those specifically in the sample in

---

gives some researchers pause, whereas others are satistfied to make such comparisons regardless as they can be clearly defined by their counterfactuals and/or by the stuctural causal model and DAG they reference.

[9]In a setting like our running example, police-civilian encounters in the study sample are selected based on whether or not the police made a stop, which we represent with $S_i = 1$. The specific set of encounters that we observe and are able to analyze may not contain all encounters for which police made stops, however. Perhaps we only have a sample of such encounters - a random sample or one based on geography, time, or other constraints. If we are willing to assume that these other constraints are not material to the causal relationships between perceived race and use of force, then we might aim to estimate a causal effect for all police-civilian encounters in which a stop was made. The fact that we only have a sample of all such encounters does not pose a problem for causal identification of causal effects for the sub-population of encounters in which a stop was made. However, only having a sample that contains encounters in which a stop was made may threaten our ability to identify such a causal effect.

hand. Obtaining a valid estimate of a causal effect for the specific sample, we can then generalize this to the subpopulation. So going forward, we often refer to just the units in the sample at hand, even if our target is really the subpopulation. Other causal effects (i.e., not just means of τ_i) might be of interest as well, but in what follows, we will focus on distributional primitives that allow for the identification of any causal effect.

Identification under conditional ignorability. Any internally valid causal effect can be written using $p(Y_d|S = 1)$. For example, $\mathbb{E}[Y_d - Y_{d'}|S = 1] = \sum y \times p(Y_d = y|S = 1) - \sum y \times p(Y_{d'} = y|S = 1)$. Such quantities can be identified under sufficiently strong assumptions, including that covariates Z are observed such that in each stratum of Z in the selected population, treatment is ignorable, i.e. $Y_d \perp\!\!\!\perp D|Z, S = 1$.[10]

$$
\begin{aligned}
p(Y_d|S = 1) &= \sum_z p(Y_d|Z = z, S = 1)p(Z = z|S = 1) && \text{by law of total probability} \\
&= \sum_z p(Y_d|D = d, Z = z, S = 1)p(Z = z|S = 1) && \text{by } Y_d \perp\!\!\!\perp D|Z, S = 1 \\
&= \sum_z p(Y|D = d, Z = z, S = 1)p(Z = z|S = 1) && \text{by consistency}
\end{aligned}
$$

Structural causal models and DAGs. The key practical concern is: how can we know if D and Y_d are in fact independent conditional on Z in the selected sub-population? It is simple to write a conditional ignorability statement of this kind, and even to understand a hypothetical randomization scheme that would justify such a choice. With observed data, however, supporting or defending such an assumption turns on detailed claims about the underlying data generating process. To convincingly defend an assumption of conditional ignorability is to convincingly argue that the data were generated from one of a class of "structural causal models" that would produce such ignorability. Specifically, such a model must make claims as to how the treatment and outcome causally relate to each other and relevant observed and unobserved covariates. The key assumptions in these models often regard which variables *do not* influence other variables. Such causal relationships can often be non-parametrically encoded in a structural causal model (SCM) which can be represented graphically as a directed acyclic graph (DAG; and extensions thereof). See Pearl (2009) and Appendix B for details on SCMs and DAGs. See Figure 2 for example DAGs. DAGs allow us to visualize dependencies and independencies between variables. On a given DAG, a set of rules known as graphical criteria can then be used to determine if ignorability holds under the sample selection mechanism implied by that DAG. In the following section, we develop such a graphical criterion.

4   Proposal

Our set of tools for analyzing how sample selection can threaten internal validity begins with a method for displaying the effects of sample selection graphically. Next, we present a formal graphical criterion for determining whether internally valid causal effects are identifiable. Our proposed approach is similar to those presented in Pearl (1995); Shpitser et al. (2010); Daniel et al. (2012); Correa et al. (2018), but emphasizes the impact of sample selection for internal validity regardless of the role that sample selction takes in the

---

[10]Here, we ignore problems of measurement bias, missingness, interference, and other wrinkles that are important in practice. These can also be represented graphically.

causal graph. We conclude this section by drawing connections with the generalizability results in Correa et al. (2018) and with a discussion of sample selection based on mediators in more detail. It is important to emphasize that we should have some population in mind from which the sample was selected. This will allow us to attempt to non-parametrically model the sample selection process.[11] As discussed above, we will use a binary variable, S, to denote sample selection.

4.1 Showing selection: internal selection graphs

We will now detail our simple graphical approach to determining whether conditional ignorability of the form $Y_d \perp\!\!\!\perp D | Z, S = 1$ holds. The key is to graphically represent the ways in which sample selection alters the relationships in the selected sample. We do this by defining internal selection graphs, which visually extend traditional causal graphs to represent all the ways that sample selection can change relationships between variables.

Definition 1 (Internal Selection Graph, G_S^+). Let G be the DAG induced by a SCM.
1. Create G_S by adding an appropriately connected binary selection node, S.
2. Draw a circle around S to clearly indicate that we must limit our analysis to $S = 1$.
3. Add to G_S any node which is a parent of the treatment or a parent of a descendant of the treatment. (U_S, the background factors contributing to selection, can be excluded.)
4. Add a dashed undirected edge between all variables between which S is a collider or an ancestor of S is a collider. We will call these dashed, undirected edges *bridges*.

Call the resulting graph an *internal selection graph, G_S^+*.

(This definition is similar to the "modified extended diagram" in Daniel et al. (2012).)

Let us dwell briefly on why we've chosen to denote sample selection with a separate binary selection node, rather than conditioning on some variable already in the causal model. Consider a simple two node DAG, $D \rightarrow Y$, in which both variables are continuous and there is sample selection on the outcome. Suppose this represents the effect of education (D) on income (Y) in a simplified setting in which we assume no common causes of education and income. (Elwert and Winship, 2014; Hausman and Wise, 1977) If we have a sample that is filtered to units with a particular *range* of (low) incomes, then sample selection, S, is not identical to income. There is still some variation in income, despite sample selection. As such, we should represent sample selection as a child of income: $Y \rightarrow S$. In other cases, like our running example related to the the effect of racial discrimination on police use of force, selection may be equivalent to a binary variable in the causal model (in our example sample selection is equivalent to police making a stop). In this case, we should represent the binary variable as selection. There are also many settings in which sample selection is caused by more than one variable and this representation can simplify things while also showing how such variables can become related by selection. In cases when the binary selection variable is not equivalent to another variable in the causal graph, selection does not eliminate all variation in that other variable. This means that paths running through that variable are not blocked by sample selection. Only paths that run directly through

---

[11]While having a population in mind is useful, "There is also the problem of infinite regress. Even if one has a random sample from a defined population, that population is almost certainly a nonrandom subset from a more general population. ... In principle, therefore, there exists an almost infinite regress for any dataset in which at some point sample selection bias becomes a potential problem." Berk (1983)

variables selection is equivalent to will be blocked. We are not the first to use a binary selection variable in this way. See Bareinboim et al. (2014); Correa et al. (2018); Egami and Hartman (2021), among others. We echo Greenland (2022) in the sentiment that "realistic causal diagrams should always have a selection (sampling) indicator node S ... as a part of the data-generating process."[12]

Figure 2: Examples of DAGs and Internal Selection Graphs

(a.i.) DAG   (a.ii.) Internal Selection Graph   (b.i.) DAG   (b.ii.) Internal Selection Graph

(c.i.) DAG   (c.ii.) Internal Selection Graph
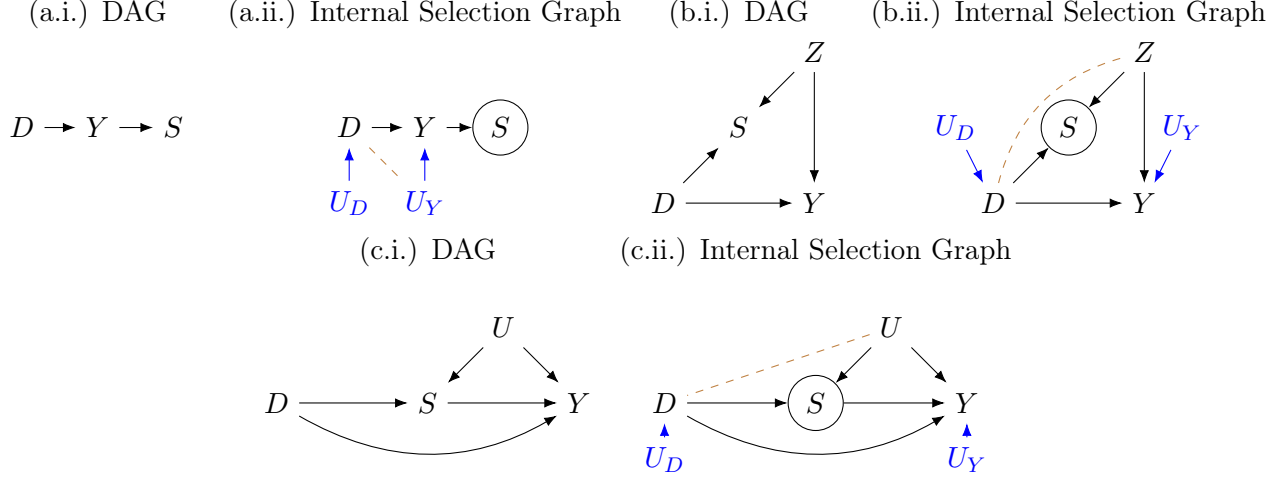
The key features of internal selection graphs[13][14][15] are the inclusion of an encircled sample selection node, specific background variables,[16] and bridges that capture the statistical associations that result from sample selection. Bridges are created as a result of conditioning on a collider; this is also referred to as "collider stratification".[17] These additions ensure sample selection and the changes it requires for identification are visualized in the graph and can be analyzed easily. See Figure 2 and Tables 1 - 4 for examples. Figure 2(c.ii.) contains the internal selection graph for our working example. At first glance, it appears that the types of

---

[12]This sentiment is not new, though the graphical form may be. Berk (1983) states "When considering whether potential sample selection bias is likely to be realized, the initial step is to formulate a theoretical model of the selection process. One needs a theory of selection. Without a theory, it is difficult to draw even preliminary inferences about the nature of the problem and impossible to choose how best to implement sample selection corrections."

[13]Including U_S would lead to the direct parents of S to be associated with each other through U_S. But the direct parents of S will already be associated with each other due to conditioning on selection itself. The associations between U_S and any direct parents of S are otherwise immaterial to ignorability, making the inclusion of U_S unnecessary.

[14]Bridges are simply graphical representations of the purely statistical relationships that arise as a result of conditioning on a collider. Here, we are forced to filter to $S = 1$; so when S is a collider, we are conditioning on a collider.

[15]The value of this sort of graph for evaluating sample selection can be seen in Elwert and Winship (2014); Schneider (2020), papers that informally explore various types of selection bias in sociology and economic history. These papers, without stating a formal approach for doing so, add bridge-like undirected edges to the graphs they use to illustrate issues related to sample selection, but do not formally discuss how these non-causal edges can be incorporated into attempts to identify causal quantities, as we do in this paper. Greenland et al. (1999) also discusses an approach in which undirected edges are added to the graph.

[16]Step three of the internal selection graph definition requires that we include additional nodes in our causal graph. This means including background noise nodes for the treatment and all descendants of the treatment. As a result, these nodes are entirely determined by their parents represented in the graph.

[17]Conditioning on a collider allows information to flow between the direct parents of the collider. This is the opposite of what happens when we condition on other types of nodes, in which case conditioning "blocks" the flow of information along a path. We can state these relationships as follows. Two sets of nodes, D, Y, in a graph G are said to be *d-separated* by a third set, Z, if every path from any node $D_0 \in D$ to any node in $Y_0 \in Y$ is blocked. A path is blocked by Z if either [1] some W is a collider on the path between D, Y and $W \notin Z$ and the descendants of W are not in Z or [2] W is not a collider on the path but $W \in Z$ (see e.g. Pearl, 2009, Chapter 1 for details).

sub-paths in internal selection graphs has expanded. We might wonder if the usual chains, forks, and colliders are joined by additional sub paths containing bridges. But the new additions are just built up from the old. The possible sub-paths incldue: chains $(A \to B \to C)$, forks $(A \leftarrow B \to C)$, colliders $(A \to B \leftarrow C)$, bridge chains $(A \cdots B \to C$ or $A \cdots B \leftarrow C)$ and double bridges $(A \cdots B \cdots C)$. A double bridge might result from something like in Figure 3(a). A bridge chain might result from something like Figure 3(b,c,d,e). So colliders are still defined only with respect to directed edges. Bridges cannot create colliders, since they are really just graphical representations of purely statistical relationships created by sample selection.

Figure 3: New sub-paths are built from the old.
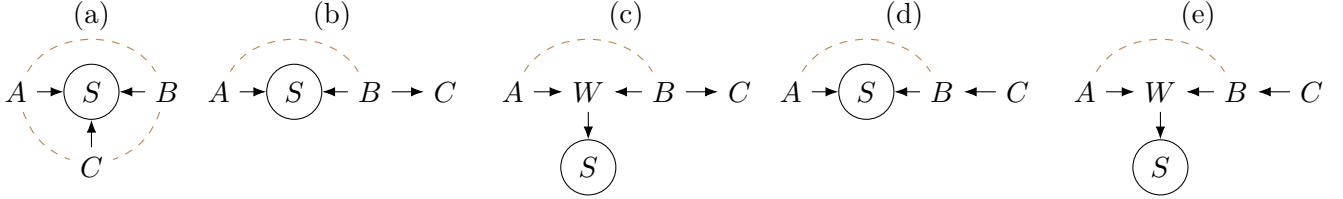
Next, we will differentiate between a few types of paths. Generalized paths are any sequence of nodes and edges (directed edges and/or bridges) where each node appears only once (e.g., $D \cdots Z \to Y$, $D \to Y$, $D \to S \leftarrow Z$, $U_D \to D \to Y$). Causal paths are any generalized path where all edges between the nodes are directed and point in the same direction (e.g., $D \to Y$, $U_D \to D \to Y$). Generalized non-causal paths are any generalized path that isn't a causal path (e.g., $D \cdots Z \to Y$).[18]

Consider again the two node example of education and income where the sample has been selected to include only low income individuals. Figure 2(a) captures this. We do not assume that education explains the entirety of the variation in income. There is also the U_Y background term. Since S is a descendant of Y and Y is a collider between D and U_Y, there is a purely statistical association created between D and U_Y. Therefore, there is a generalized non-causal path from $D \cdots U_Y \to Y$ that will confound estimates of the effect of education on income, despite the fact that we assumed there were no common causes between education and income in this simplified example. This is the simplest example of the need to include such background terms in the analysis of how sample selection alters relationships between variables in the sample. Let us also consider again our running example of racial discrimination's effect on police use of force for civilian-police encounters in which the police make a stop. This is described in Figure 2(c). We see that the internal selection graph is essentially identical to the graph in Figure 1(b) but with two additional background terms. In this example, the background terms do not play a role, but as we've just seen they can play an important role in many cases. They are included here since internal selection graphs are constructed so that all the ways that sample selection can alter relationships in the sample are captured for any causal graph and sample selection mechanism.

---

[18]Following the above discussion, d-separation is defined in the same way for these paths as for regular paths, since colliders are defined in the same way. See Appendix B, in particular, Corollary B.5 and Definition B.10 for details.

4.2   Internal Selection Adjustment Criterion

How can we use internal selection graphs to determine whether conditional ignorability holds more generally? Using the intuition of ruling out "alternative explanations" for why the treatment and outcome covary, we want to leave the causal paths between the treatment and the outcome that we hope to study untouched while removing any other systematic relationships between the treatment and outcome that could provide an alternate explanation. That is, first, we don't want to conditioning on variables that are on causal paths we are interested in, since this would block some of the effect we want to study. Second, we want to block any open generalized non-causal paths between the treatment and outcome. These paths are not those we want to study and confound the causal paths that we do want to study. Third, we don't want to open any previously closed generalized non-causal paths between treatment and outcome as a result of our covariate adjustment. (Pearl, 2009) Sample selection may block paths whether we would like them blocked or not and may open previously closed paths. We formalize these intuitions as a set of rules that fold in the effects of sample selection; we refer to them as the internal selection adjustment criterion.

Definition 2 (Internal Selection Adjustment Criterion (ISAC)). A set of nodes Z in G_S^+ satisfies the internal selection adjustment criterion relative to D (treatment) and Y (outcome) if

1. No element of Z lies on or is a descendant of a node that lies on a causal path originating from D and arriving at Y. Note that an element of Z could be a descendant of D itself, if it is not on a causal path from D to Y. Note also that elements of Z should not be on, or descendants of nodes on, causal paths even if S is also on the causal path.

2. Z blocks every generalized non-causal path between D and Y that does not pass through S. Note that generalized non-causal paths passing through M, when S is a descendant of mediator M, on which M is an ancestor of Y also do not need to be blocked, assuming the previous condition is not violated. Further, M should not be a member of Z from the previous condition.

An important component of this criterion is that we do not need to worry about blocking paths on which S appears. If S is a collider, we have already added bridges that circumvent S itself. If S is not a collider but is on a path, it blocks the path.[19] This also shows us that sample selection can alter relationships in the sample, even when the sample selection node is not a collider. As in the case when the sample selection node is a collider, we also need to take care to account for how it alters relationships when it is not a collider.

Our criterion builds on powerful, well-known, existing criteria like those presented in Pearl (1995); Shpitser et al. (2010); Daniel et al. (2012); Correa et al. (2018). Using these building blocks, ISAC focuses us on what is required to attain internal validity under non-random sample selection by relying on internal selection graphs and allowing us to determine whether causal effects are identifiable under non-random sample selection, regardless of the role that sample selection plays in the casual graph. See Section 5 for a more detailed discussion. The following results use our internal selection adjustment criterion to show how to identify internal causal quantities in the presence of sample selection and confounding, whether selection is post-treatment or

---

[19]When the sample selection node is on a generalized non-causal path but is not a collider, we do not need to consider blocking this path with some additional node Z. Sample selection will already block this path. Further, the associations created when sample selection is a collider are captured by the bridges that circumvent S. So we do not need to consider any path that passes through S. When selection is the descendant of a mediator M, we will be working with potential outcomes for which we intervene on M; see below. This means that paths running through M on which M is an ancestor of Y will be blocked and since we already condition on a descendant of M, parents of M will already be associated due to selection, when M is a collider.

not. Recall that mediators are nodes that lie on at least one causal path from the treatment to the outcome. Mediators are discussed further below.

Internal Validity Results. If a set of nodes Z in G_S^+ satisfies ISAC relative to D (treatment) and Y (outcome) and

- S *is not a mediator or descendant of a mediator between D and Y,* then
    - (ignorability) $Y_d \perp\!\!\!\perp D | Z, S = 1$.
    - (identification) We can identify $p(Y_d | S = 1) = \sum_z p(Y | D = d, Z = z, S = 1) p(Z = z | S = 1)$.
- S *is a mediator between D and Y (but S is not also a descendant of another mediator),* then
    - (ignorability) $Y_{d,S=1} \perp\!\!\!\perp D | Z, S = 1$.
    - (identification) We can identify $p(Y_{d,S=1} | S = 1) = \sum_z p(Y | D = d, Z = z, S = 1) p(Z = z | S = 1)$.
    - (non-ignorability) For any set of observables, W, $Y_d \not\perp\!\!\!\perp D | W, S = 1$ and $Y_{d,S_{d'}} \not\perp\!\!\!\perp D | W, S = 1$.
- S *is a descendant of an observed mediator, M, between D and Y (but S is not also a mediator itself),* then
    - (ignorability) $Y_{d,m} \perp\!\!\!\perp D | Z, S = 1$ and $Y_{d,m} \perp\!\!\!\perp M | D, Z, S = 1$, where $M = m$ is observed.
    - (identification) We can identify
      $p(Y_{d,m} | S = 1) = \sum_z p(Y | D = d, M = m, Z = z, S = 1) p(Z = z | S = 1)$.
    - (non-ignorability) For any set of observables, W, $Y_d \not\perp\!\!\!\perp D | W, S = 1$ and $Y_{d,M_{d'}} \not\perp\!\!\!\perp D | W, S = 1$.

These results are proved in Appendix B in Theorems B.1, B.2, and B.3. Note that all of these identification results equate internal causal quantities with expressions that are estimable from the selected sample alone.[20] The first bullet can be used to identify "total effects"; the second and third bullets can be used to identify "direct effects." The above results show that sample selection can often be seen as an omitted variable problem. Heckman (1979) discusses "the bias that results from using nonrandomly selected samples to estimate behavioral relationships as an ordinary specification bias that arises because of a missing data problem." The idea is essentially that, cast in the right light, sample selection can be thought of as an omitted variable or misspecification problem. Heckman's discussion was in the context of a parametric framework and he proposed a correction procedure in this context. Our results show that something similar is true when we take the graphical, non-parametric view on sample selection as a threat to internal validity. When the threat that sample selection poses to the internal validity of causal effect estimates of D (treatment) on Y (outcome) can be overcome through adjustment on some covariates, Z, the problem of sample selection for internal validity can be viewed as an omitted variable problem. Future work will explore omitted variable based sensitivity analysis for sample selection based on this remark. Before turning to a discussion of direct effects, we first consider some connections to genereralizability.

4.3 Connections to generalizability

As previously discussed, researchers are often concerned with how a causal effect might look averaged over some population of interest, as opposed to averaged over the sample at hand. When the sample is drawn

---

[20]Estimation strategies that would apply to a covariate adjustment or conditional ignorability identification strategy will also apply when we properly account for sample selection. In Section B.5 of Appendix B, we present a discussion of IPW estimation of $\mathbb{E}[Y_d | S = 1]$, $\mathbb{E}[Y_{d,S=1} | S = 1]$, and $\mathbb{E}[Y_{d,m} | S = 1]$. This is meant to illustrate one example of how estimation might proceed for internal causal effects. We also provide this discussion to demonstrate that estimation of "internal controlled direct effects" is also straightforward.

from the target population in some (random or non-random) way, we may hope to use information on the causal effect available from the sample to generalize to a statement about the causal effect in the population.[21] Generalizability in this sense is a form of external validity. We will show that internal validity is more permissive than generalization, by which we mean the causal assumptions required to identify (and the observed data required to estimate) internally valid causal effects are a subset of those required for causal effect estimates that generalize from the sample. In doing so, we formalize what Campbell and Stanley (1963) first claimed with respect to experiments for approaches using covariate adjustment: "Internal validity is the basic minimum without which any experiment is uninterpretable..."[22][23] We now show a similar result using graphical criteria and limit our discussion to causal quantities containing potential outcomes of the form Y_d.

Definition 3 (Generalization Criterion (GC)). This definition is a translation of Definition 8 from Correa et al. (2018). A set of nodes Z in G_S^+ satisfies the generalization criterion relative to D (treatment) and Y (outcome) if

- Z satisfies ISAC relative to D and Y and
- $Z_{\text{Ext}} \subset Z$ blocks all causal and generalized non-causal paths between Y and S in G_S^+ other than those that end in a causal path from D to Y.

Generalization Results. If a set of nodes Z in G_S^+ satisfies GC relative to D (treatment) and Y (outcome) and *S is not a mediator or descendant of a mediator between D and Y*, then

- (ignorability) $Y_d \perp\!\!\!\perp D | Z, S = 1$ and $Y_d \perp\!\!\!\perp S | Z_{\text{Ext}}$.
- (identification) We can identify $p(Y_d) = \sum_z p(Y|D = d, Z = z, S = 1) p(Z_{\text{Int}} = z_{\text{Int}} | Z_{\text{Ext}} = z_{\text{Ext}}, S = 1) p(Z_{\text{Ext}} = z_{\text{Ext}})$, where $Z_{\text{Int}} = Z - Z_{\text{Ext}}$.

This translates Definition 7 and Theorem 1 in Correa et al. (2018) to use potential outcomes.

Remark 1. *The causal assumptions invoked by ISAC to identify $p(Y_d|S = 1)$ are a subset of those invoked by GC to identify $p(Y_d)$ using covariate adjustment.*

Further, we can estimate $p(Y_d|S = 1)$ with observed data from the selected sample for Y, D, Z alone. On the other hand, $p(Y_d)$ can be estimated with observed data from the full population on Z_{Ext} in addition to data from the selected sample for Y, D, Z. In this way, we have formalized that generalization "costs more" in

---

[21]In this section, we discussion connections between our results and recent results addressing generalization (Correa et al., 2018). This is a distinct but related to the problem of transportability. See Bareinboim and Pearl (2016) for a discussion of the differences between generalization and transportability. We are considering a single DAG. In many interesting observational settings, the same DAG might not hold across all settings of interest. For example, there may be a reason why, at the particular hospital we have data from, the DAG we are evaluating holds and further that we can sustain related ignorability statements or satisfy graphical criteria related to this DAG, but the DAG might be different at other hospitals. Often we have to be identification opportunists, looking for someplace that the treatment was assigned in some way that is conducive to identification, and that story (DAG) may not hold elsewhere. You might be able to extend to a larger population in which the same DAG holds, but not to a population in which the DAG does not hold. When considering external validity, you have to have a specific target population in mind; and here we're talking about external validity for the population in which the same DAG holds; i.e., generalization not transportation.

[22]Shadish et al. (2002) clarify that the special role of internal validity is specific to "cause-probing research", as we are discussing here, but not all forms of research.

[23]An alternative version of this idea can be stated as follows. If observed covariates and data limited to the selected sample ($S = 1$) are insufficient identify $p(Y_d|S = 1)$, then they will also be insufficient to identify $p(Y_d) = p(Y_d|S = 1)p(S = 1) + p(Y_d|S = 0)p(S = 0)$, since $p(Y_d)$ extends beyond the selected sample which is all we have information about. [I'm not sure this is a logical deduction.] We show this within the context of graphical criteria in this section.

terms of both causal assumptions and observed data. These generalization results are proved in Appendix B; see Theorem B.4. Additional connections are also shown in Lemma B.24.[24]

Let's explore some examples and discuss how identifiability can differ for internal and external validity. First, let's consider our running example of racial discrimination in policing in which we are interested in the effect of perceived civilian race on police use of force in which we have data only on civilian-police encounters in which the police made a stop. The internal selection graph for this can be found in Figure 2(c.ii.). We immediately see that there is no hope for generalizing: since S is a direct cause of Y there is no hope of satisfying GC, regardless of the choice of Z_{Ext}. Further, we only have access to data for police-civilian encounters in which the police made a stop, so we do not have any data for the full population. But do we satisfy the internal selection adjustment criterion? Well, as we saw before, no we don't. This is because of the generalized non-causal path from $D \dashrightarrow U \to Y$, which violates ISAC. So no total or direct effect is identifiable. We will discuss this example more in the next section.

Next let's consider Figure 4(a), which we will just consider in the abstract. Here there is a common cause of sample selection and the outcome, but the two are not directly associated and sample selection is not a collider. Let's consider internal validity first. Suppose we choose not to condition on any covariates. We easily see that we satisfy ISAC. This means that we are indeed able to identify internal causal quantities without any covariate adjustment. We also only need data from the selected sample to estimate effects for the selected sample. What about identifying external causal quantities? We see that letting $Z = Z_{\text{Ext}}$ means that we satisfy GC. So we can identify external causal quantities. However, since we need to adjust for Z, we could only estimate generalized effects so long as we have data for Z for the full population. Finally, Figure 4(b) is similar to Figure 4(a), but it turns out that we can identify internal causal quantities adjusting for W or Z or both. Estimation would only require data on one of these for the selected sample. However, to identify external causal quantities we must adjust for Z (or Z and W). Estimation would require data on Z for the full population, data for the selected sample alone is insufficient, as is data only on W, even if it is for the entire population.

Figure 4:   Examples for comparison of ISAC and GC.
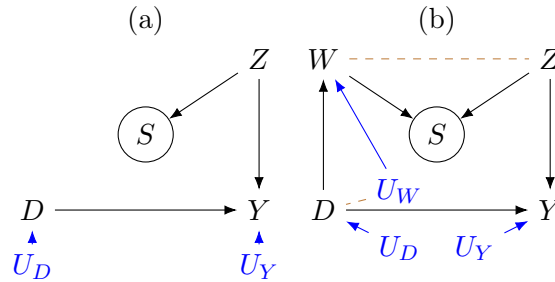
4.4   Sample Selection based on a mediator

In this section, we return to potential outcomes of the the form $Y_{d,S=1}$ and $Y_{d,m}$, which allow us to consider particular types of direct effects called "controlled direct effects." Above we saw that, when sample selection

---

[24]See Correa et al. (2018) for an IPW estimator for generalization.

plays the role of a mediator or the descendant of a mediator and we satisfy ISAC, we can identify $p(Y_{d,S=1}|S=1)$ or $p(Y_{d,m}|S=1)$, respectively. Identifying these internal causal quantities allows us to identify what we call "internal controlled direct effects," which are defined below. But what are these strange looking effects?

Definition 4 (Internal Controlled Direct Effects (ICDEs)). Define $\mathbb{E}[Y_{D=d,S=1} - Y_{D=d',S=1}|S=1]$ and $\mathbb{E}[Y_{D=d,M=m} - Y_{D=d',M=m}|S=1]$ to be the internal controlled direct effect when selection is a mediator between D and Y and when selection is a descendant of a mediator, M, between D and Y, respectively.

In mediation analysis, we have a treatment D, a mediator M, and an outcome Y, in addition to other relevant covariates. There are two possible paths along which the treatment might effect the outcome. First is the familiar direct path: $D \to Y$. Second is the indirect path: $D \to M \to Y$. This set up follows the mediation discussion from Baron and Kenny (1986). For our purposes, we consider the settings in which the sample selection node is itself a mediator between D and Y[25] or is a descendant of a mediator between D and Y. There are a variety of causal effects to consider when considering mediation, including total effects, controlled direct effects, natural direct effects, and natural indirect effects (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2011; VanderWeele and Vansteelandt, 2009; Richiardi et al., 2013).

In both of our settings, a causal path between D and Y is blocked or partially blocked as a result of sample selection. This means that total effects, like $\mathbb{E}[Y_d - Y_{d'}|S=1]$, which capture all causal paths along which the treatment D can effect the outcome Y, are not identifiable. The indirect effect of D on Y that runs along the path on which S lies or is a descendant will also not be identifiable. Only the causal paths $D \to Y$ that do not relate to S remain unaltered. Are we able to identify the direct effects that run along these unaltered paths?

In our present discussion, we are limited to the study sample. So the type of causal effects that we are able to identify when sample selection is a mediator or descendant of a mediator will still be for the selected sample alone. We are therefore interested in direct effects for the selected sample. ICDEs are just this sort of effect: direct effects averaged over the units in the selected sample. ICDEs are distinct from CDEs for the entire population and are also distinct from total effects for the selected sample. Instead, ICDEs compare setting $D = d$ with setting $D = d'$, while also setting M to m (or S to 1) for both versions of treatment interventions, *for units in the selected study sample.*

To understand ICDEs better, let's return to our working example. First, we note that total effects cannot be identified since sample selection (police making a stop) is a mediator blocking one of the causal paths between the treatment (police perception of race) and outcome (police use of force). But we might consider an ICDE as a possible effect of interest. The ICDE is the difference in police use of force between

- a setting in which we intervene to force police to perceive each civilian as being from the *minority* racial group and intervene to force police to stop each civilian
- a setting in which we intervene to force police to perceive each civilian as being from the *majority* racial group and intervene to force police to stop each civilian

where we average this difference in use of force only over the police-civilian encounters in which police actually did make a stop. A key to understanding this effect is to consider that, for some encounters in which a stop was made in reality, under a different perceived civilian race the officer would not have made a stop. ICDEs

---

[25]Selection can be a a mediator only when it is equivalent to a substantive binary variable in the causal graph. Our working example is an example of this.

evaluate what would have happened if we intervened in these cases to ensure that the officer still made a stop, since we only have data on encounters in which stops were actually made and so can only make estimates about cases in which stops were made. In this example, the path $D \cdots U \to Y$ violates ISAC and confounds the ICDE. That is, sample selection creates a purely statistical association between perceptions of race and use of force that we cannot untangle from the direct effect, even for police-civilian encounters we observe.

5 Discussion

5.1 Connections to existing work

Numerous literatures relate closely to this problem, yet we argue that the tools proposed here fill a gap in the toolkit available to researchers to fully examine the potential threats due to sample selection and to illuminate possible solutions. Sample selection can arise at various points in the a study: during study entry (e.g., from non-participation or participation that is not representative of the population) or the data gathering process (e.g., only gathering data on some segment of the population), between study entry and analysis (e.g., loss to follow-up), or even during analysis as a result of conditioning or subsetting. Montgomery et al. (2018) illustrate how sample selection can threaten not only observational studies but also experiments. Sample selection and the associated bias goes by many different names in various fields: sample truncation bias, non-response bias, attrition bias, ascertainment bias, Heckman selection bias (Heckman, 1979), selection on the treatment, selection on the outcome, Berkson's bias, homophily bias, survival bias, m-bias, differential loss to follow up, volunteer bias, self-selection bias, healthy worker bias, and others. See Hernán et al. (2004); Elwert and Winship (2014); Schneider (2020) for informal overviews of the various forms that sample selection can take and real difficulties that researchers grapple with. Different research traditions have proposed informal guidelines for determining when sample selection threatens internal validity. Most notably, Campbell, Stanley, and their co-authors (Campbell, 1957; Campbell and Stanley, 1963; Cook and Campbell, 1979; Shadish et al., 2002) introduce the language of validity[26] to discuss conceptually how bias can arise in the design and implementation of studies and how these challenges can be overcome. Pearl and others (Pearl, 1995, 2009, 2014; Shpitser et al., 2010; Spirtes et al., 2000) advocate causal graphs and structural causal models that subsume the potential outcomes approach (Rubin, 1974, 1978, 1990) to causal inquiry.[27] Existing graphical approaches have not, thus far, gone further than informal suggestions as to how to address the threat sample selection poses to internal validity.

There are graphical approaches focused on identification of causal effects (Pearl, 1995; Shpitser et al., 2010) that do not directly discuss sample selection. There are many recently-developed approaches focused on generalizability and transportability (Bareinboim and Pearl, 2012, 2016; Correa and Bareinboim, 2017; Correa et al., 2018, 2019; Bareinboim et al., 2014; Lesko et al., 2017; Pearl, 2015b; Pearl and Bareinboim, 2011, 2014, 2019; Hartman et al., 2015; Egami and Hartman, 2021). There are graphical approaches focused on generalizing conditional causal effects from complete cases when there is missing data (Daniel et al., 2012), generalizing in the face of missing data and sample selection (Saadati and Tian, 2019), and generalizing

---

[26]Shadish et al. (2002) describe "internal validity" as concerning whether the covariation between the treatment and outcome results from a causal relationship for the study sample.

[27]Matthay and Glymour (2020) take the very useful step of explicitly connecting the graphical approach to the Campbell tradition.

from missing data alone (Mohan and Pearl, 2014, 2021).[28] Didelez et al. (2010) focus on outcome dependent sampling, on the causal odds ratio, and mostly on generalization; they also present some results on testing for the presence of conditional causal effects.

Our criterion builds on these powerful, existing graphical criteria, but focuses on and facilitates easy analysis of internal validity under non-random sample selection by visualizing teh effects of sample selection and illustrating whether identification of causal effects is possible regardless of the role that the sample selection node plays in the causal graph. Let us review some of the key existing graphical criteria. First, ISAC is similar to the back-door criterion of Pearl (1995). Understanding that sample selection is a form of conditioning that could be contemplated on a DAG, simply including S in the adjustment set of the back-door criterion would violate the back-door criterion when S is post-treatment, since the back-door criterion does not allow post-treatment conditioning. Thus, we would violate the back-door criterion in cases where we can make causal progress. Further, Pearl (2009) Section 11.3 discusses how background noise nodes that are often left off of DAGs can become important under certain types of conditioning. We ensure that, when relevant, such variables are included on internal selection graphs so that all effects of sample selection are represented graphically. Second, ISAC is also similar to the adjustment criterion of Shpitser et al. (2010). However, simply including S in the adjustment set of the adjustment criterion would violate the adjustment criterion when S is a mediator or a descendant of a mediator. This is because the adjustment criterion disallows conditioning on nodes that appear on causal paths or that are descendants of nodes on causal paths. Thus, we would violate the adjustment criterion in cases where we can make causal progress. Third, ISAC is similar to the generalized back-door criterion of Daniel et al. (2012), which is focused on generalization from complete cases when there is missing data and not internal validity. Moreover, the generalized back-door criterion does not allow for any post-treatment adjustment and, in Daniel et al. (2012), is shown to only to identify Z-conditional causal quantities. Finally, ISAC is similar to the generalized adjustment criterion of Correa et al. (2018), but ISAC is less restrictive since it focuses on internal rather than external validity. Specifically, ISAC requires a subset of the causal assumptions required by the generalized adjustment criterion.

Despite existing approaches not focusing on sample selection and internal validity, sample selection is recognized as a threat to internal validity that is fundamentally different from common cause confounding (Hernán et al., 2004; Hernán and Robins, 2020; Infante-Rivard and Cusson, 2018; Matthay and Glymour, 2020; Smith, 2020; Elwert and Winship, 2014; Schneider, 2020). Other authors discuss sample selection in the context of external validity (Arah, 2019; Flanders and Ye, 2019; Thompson and Arah, 2014) and mention the threat to internal validity posed by sample selection (Berk, 1983; Tripepi et al., 2010; Cuddeback et al., 2004; Hernán and Robins, 2006; Larzelere et al., 2004; Smith and VanderWeele, 2019; Westreich et al., 2018). The celebrated approach presented in Heckman (1979) makes parametric assumptions and requires data for the entire population to estimate a model for the probability of selection. Hernán (2017) shows how sample selection can present different problems for generalization and external validity than it does for internal validity; we provide clarity on this in the general.

It is possible to conduct a more formal and rigorous treatment of sample selection's threats to internal validity that determines when causal effect estimates in the given sample have been biased by the sample selection processes, when the correct estimate can be recovered, and how. We propose a formal graphical

---

[28]Missing data can be seen as a generalization of sample selection. (Saadati and Tian, 2019; Westreich, 2012; Howe et al., 2015)

criteria (a set of rules) for determining when internally-valid causal effects are identifiable, and provide tools enabling researchers to easily determine how sample selection may bias their estimates of causal effects across their study sample, as well as what might be able to be done to correct any bias. To our knowledge no such graphical approach exists. While some existing approaches hint at how we might formally approach sample selection for internal validity, none focus on it in similar measure to confounding.[29] We advocate that it should be standard practice to consider the sample selection mechanism in the context of the causal model under investigation, even when attempting to estimate a causal effect in the sample at hand alone. No causal model should be without a sample selection node, since, in practice, we are analyzing a particular sample of data which can be thought of as *a representative sample of some subset* of a larger population. This sample selection mechanism can have wide-ranging implications for causal inference. Our discussion encompasses post-treatment selection, selection as a mediator, selection as a descendant of a mediator, and all other major roles that selection can play in the structure of causal models. We also provide numerous examples and discuss key lessons. We hope that these allow researchers to easily determine how sample selection may bias their estimates of sample treatment effects and learn what might be able to be done to correct any bias.

5.2   Library of Examples

In this section, we consider more examples (in Tables 1 - 4) and see what lessons researchers might take from everything we've seen. We start with examples where covariate adjustment turns out not to be necessary. In (1.a.), sample selection is a confounder of the treatment-outcome relationship. In (1.d.), sample selection is indirectly associated with the outcome through a common cause. In (1.e.), sample selection is indirectly associated with the treatment through a common cause. In (1.f.), treatment is a direct cause of sample selection. (1.g.) is the setting where sample selection is not related to the treatment or outcome directly or indirectly. For the examples in which selection is related directly or indirectly with the outcome, we might not be able to generalize from the sample to the population and in the best case would still need data external to the selected sample on the covariates that separate selection from the outcome. In all of the examples in Table 1, the adjustment set can be the empty set and satisfy ISAC. No generalized non-causal paths between D and Y, that do not pass through S, exist and we do not condition on variables along causal paths. While these settings are desirable in terms of not having to do any work to get valid identification, researchers should not simply assume they are in this setting. Only principled argument in accordance with knowledge of the selection mechanism can justify this. We should also recognize the value of a formal approach to evaluating ignorability. All the assumptions and information relevant to this determination are clearly captured in the internal selection graph; and we only need to check the conditions of ISAC to see if identification is possible.

Next, consider examples where identification of internal causal quantities is possible but only with covariate adjustment. These can be found in Table 2. (2.b.) has post-treatment selection in addition to sample selection being indirectly associated with the outcome through a common cause. This would create a purely statistical association between the treatment and outcome in the selected sample that confounds causal effects. The fact that we are forced by sample selection to condition on a post-treatment variable presents us with no problem. ISAC provides a straightforward approach to determining that ignorability is possible by conditioning on Z.

---

[29]For example, the typical three node confounder graph (see Figure **??**) is incomplete for internal validity since it does not address sample selection in any way.

All generalized non-causal paths between D and Y, that do not pass through S, are blocked by Z. Of course, if Z is not observed, we may need to resort to other options.

If some unobserved variable(s) being measured would allow us to satisfy the conditions of ISAC and therefore identify causal quantities, we might consider a few solutions. First, when possible, the best option is to gather measurements of the variable(s) for the sample at hand. These can then be used to block the necessary generalized non-causal paths. Second, we might consider gathering measurements of proxies of the variable. To do so, the proxies must be included in the internal selection graph, connected appropriately, and not open any new generalized non-causal paths when conditioned on. The proxies must also be sufficiently associated with the variable of interest so as to reduce bias. We do not explore this option further here, but what qualifies this sufficiency needs to be carefully considered. Third, we might consider using sensitivity analyses to explore how different assumptions about the unobserved variable(s) relates to the treatment and outcome change the estimate that does not use the unobserved variable. We plan to explore such sensitivity analyses in subsequent work. Finally, we also discuss possible tests of "causal nulls" in Footnote ?? that might be possible when other approaches are not.

Sadly, we will not always be able to identify internal causal quantities, whether or not we observe all variables. Consider the examples in Table 3. At least one generalized non-causal paths from selection cannot be blocked by any covariate adjustment set. Therefore, ISAC cannot be satisfied. For some of these examples, the generalized non-causal path $D \cdots U_Y \to Y$ is created because sample selection is a descendant of the outcome and the outcome is a collider. This can be clearly seen only in the internal selection graphs in which U_Y is included and the path $D \cdots U_Y$ is drawn. In the original causal graphs, these important elements are not included. Many other situations will also not allow us to identify internal causal quantities when covariates that might block generalized non-causal paths are not observed. We should also note the similarity between the graphs (2.c.) and (3.e.) as well as (2.d.) and (3.c.). In these, the direction of only one edge is changed but the ability to identify is completely different.

Finally, consider the setting where sample selection is a mediator or a descendant of a mediator; Table 4. (4.d.) provides an interesting case, where selection is a mediator and the central node of an M-shape DAG. ISAC still provides a clear approach to identifying ICDEs by adjusting for Z. All generalized non-causal paths between D and Y, that do not pass through S, are blocked by Z.

5.3   Lessons

There are several important lessons to note briefly here. First, researchers need to be careful about the details of the causal graph that they are studying. Including a sample selection node in every causal model and careful consideration of the sample selection mechanisms is required to determine the treat that sample selection poses to internal validity and what, if anything, might be able to be done. Second, there is potential for users to intentionally or unintentionally favor one graph over another very similar graph in order to show that some causal quantity is identifiable. These are difficult but inherent problems in causal study and good-faith efforts to do credible causal inference should spend ample time defending the specific causal model being analyzed. Let's consider what other lessons we can take away from what we've seen.

- Informal applications of other adjustment criteria to identification of internal causal quantities in light of sample selection can be misleading and should be avoided. Standard causal graphs often omit important

background variables and require the user to remember that certain paths are open. These difficulties increase with the complexity of the causal graph.

- Sample selection can influence identification of internal causal quantities, even when it is not a collider. Whether selection is a collider, confounder, mediator, or indirectly related to variables of interest, ISAC provides clear guidance on identification.
- When sample selection manifests as attrition or some other post-treatment type of selection, randomization of treatment assignment does not automatically ameliorate problems of sample selection even for internal validity. Therefore, the discussion here is not exclusively for observational studies.
- Sample selection does not always present a problem. An application of ISAC is the best way to be sure.
- Selection on the outcome is usually a problem for identification of internal causal quantities.[30] However, association of the outcome and selection does not automatically present a problem. When a third variable causes both and the two are only indirectly related, there may be no problem.
- Post-treatment selection is typically not a problem on its own.[31]
- Indirect association between selection and the outcome or selection and the treatment are typically not problems on their own but can be when they appear together.
- When selection is a mediator or descendant of a mediator, causal quantities can still be identified.
- The specific causal quantities that can be identified, and whether identification of them is possible, depends on the causal model and which variables are observed.[32]

These lessons can also provide useful insights for the role that sample selection plays in identification strategies other than simple covariate adjustment, like instrumental variables.[33] While such strategies often require some additional assumptions, they typically also involve conditional ignorability assumptions and/or can be guided by graphical analysis. The internal validity of these strategies can also be threatened by sample selection in the ways similar to those we've seen, in which case the tools provided here can provide clarity. For example, instrumental variables approaches involve ignorability of the instrument with respect to potential outcomes. Sample selection can threaten whether such ignorability holds in a variety of ways. See Table 5 for examples. As with simple covariate adjustment, the tools and lessons provided in this paper can be useful for other identification strategies to analyze the effects of sample selection.

In conclusion, the target of causal studies is often a causal effect averaged across the study sample alone. The ability to estimate such an effect without bias is called internal validity. While there are many existing covariate adjustment criteria, none focus on sample selection for the purposes of internal validity. We've presented a simple graphical framework for dealing with sample selection that allows us to reliably attain internal validity for causal effects for the selected study sample. This framework allows sample selection to play any role in the original causal graph (including post-treatment selection, selection as a mediator, selection

---

[30]Though it can be allowed in certain circumstances, like case control studies in which the causal odds ratio is the target. See Daniel et al. (2012) section 4.5 "Missingness driven only by outcome." This and examples like (1.h.) highlight that rules of thumb like "selection on the outcome is biasing" might not correctly characterize all scenarios. Hence, while such general lessons can be useful guidance, it is important to consider the specifics of each application and use formal tools and systematic analysis.

[31]See the next paragraph for a brief discussion of sample selection in instrumental variables, in which case post-treatment selection is typically a problem. Again, this demonstrates how simple rules of thumb like "post-treatment selection is not biasing" can be misleading and the specifics of each application should be considered and analyzed using formal tools.

[32]Also see Appendix section B.8 for a discussion of how one might deal with unknown edge directions.

[33]See Angrist et al. (1996); Hernán and Robins (2006) for discussion of instrumental variables. See Swanson et al. (2015); Canan et al. (2017); Swanson (2019); Hughes et al. (2019); Elwert and Segarra (2022) for discussions of sample selection and instrumental variables.

as a descendant of a mediator, selection as a confounder, selection on the outcome, etc.) and provides clear guidance on which causal quantities are identifiable and what is required for identification through covariate adjustment. We've also seen many examples and have discussed key lessons that, we hope, will prove useful for researchers in a variety of settings, as they attempt to obtain internally valid estimates of causal effects for their study sample.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Arah, O. A. (2019). Analyzing selection bias for credible causal inference. *Epidemiology*, 30(4):517–520.

Balke, A. and Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, page 46–54, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Balke, A. and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI'94, page 230–237. AAAI Press.

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In Lawrence, N. D. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 100–108, La Palma, Canary Islands. PMLR.

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., Tian, J., and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*, 6(51):1173–82.

Berk, R. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, 48(3):386–398.

Campbell, D. T. (1957). Factors relevant to validity of experiments in social settings. *Psychological Bulletin*, 54(4):297–312.

Campbell, D. T. and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago.

Canan, C., Lesko, C., and Lau, B. (2017). Instrumental Variable Analyses and Selection Bias. *Epidemiology*, 28(3):396–398.

Clifford, S., Sheagley, G., and Piston, S. (2021). Increasing precision without altering treatment effects: Repeated measures designs in survey experiments. *American Political Science Review*, 115(3):1048–1065.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin.

Correa, J. and Bareinboim, E. (2017). Causal effect identification by adjustment under confounding and selection biases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Correa, J., Tian, J., and Bareinboim, E. (2018). Generalized adjustment under confounding and selection biases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Correa, J., Tian, J., and Bareinboim, E. (2019). Adjustment criteria for generalizing experimental findings. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1361–1369. PMLR.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Cuddeback, G., Wilson, E., Orme, J. G., and Combs-Orme, T. (2004). Detecting and statistically correcting sample selection bias. *Journal of Social Service Research*, 30(3):19–33.

Daniel, R. M., Kenward, M. G., Cousens, S. N., and De Stavola, B. L. (2012). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256.

Didelez, V., Kreiner, S., and Keiding, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387.

Egami, N. and Hartman, E. (2021). Covariate selection for generalizing experimental results: Application to a large-scale development program in uganda*. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(4):1524–1548.

Egami, N. and Hartman, E. (2022). Elements of external validity: Framework, design, and analysis. *American Political Science Review*, page 1–19.

Elwert, F. and Segarra, E. (2022). *Instrumental Variables with Treatment-Induced Selection: Exact Bias Results*, page 575–592. Association for Computing Machinery, New York, NY, USA, 1 edition.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40(1):31–53. PMID: 30111904.

Flanders, W. D. and Ye, D. (2019). Limits for the magnitude of m-bias and certain other types of structural selection bias. *Epidemiology*, 30(4):501–508.

Ghassami, A. and Kiyavash, N. (2017). Interaction information for causal inference: The case of directed triangle. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1326–1330.

Greenland, S. (1977). Response and follow-up bias in cohort studies. *American Journal of Epidemiology*, (3):184–187.

Greenland, S. (2022). *The Causal Foundations of Applied Probability and Statistics*, page 605–624. Association for Computing Machinery, New York, NY, USA, 1 edition.

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

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):757–778.

Hausman, J. A. and Wise, D. A. (1977). Social experimentation, truncated distributions, and efficient estimation. *Econometrica*, 45(4):919–938.

Hazlett, C. (2020). Angry or weary? how violence impacts attitudes toward peace among darfurian refugees. *Journal of Conflict Resolution*, 64(5):844–870.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Helin, K. (2011). Free throw shooting: It's not about where you're from, it's about height. https://nba.nbcsports.com/2011/03/31/free-throw-shooting-it%E2%80%99s-not-about-where-you%E2%80%99re-from-it%E2%80%99s-about-height/. [Online; posted 31-March-2011].

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586.

Hernán, M., Hernández-Díaz, S., and Robins, J. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

Hernán, M. and Robins, J. (2020). *Causal Inference: What If.* Chapman & Hall/CRC, Boca Raton.

Hernán, M. A. (2017). Invited Commentary: Selection Bias Without Colliders. *American Journal of Epidemiology*, 185(11):1048–1050.

Hernán, M. A. and Robins, J. M. (2006). Instruments for Causal Inference: An Epidemiologist's Dream? *Epidemiology*, 17(4):360–372.

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

Howe, C. J., Cain, L. E., and Hogan, J. W. (2015). Are all biases missing data problems? *Current epidemiology reports*, 2(3):162–171.

Hughes, R. A., Davies, N. M., Davey Smith, G., and Tilling, K. (2019). Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis. *Epidemiology*, 30(3):350–357.

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

Infante-Rivard, C. and Cusson, A. (2018). Reflection on modern methods: selection bias—a review of recent developments. *International Journal of Epidemiology*, 47(5):1714–1722.

Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637.

Krippendorff, K. (2009). Information of interactions in complex systems. *International Journal of General Systems*, 38(6):669–680.

Larzelere, R. E., Kuhn, B. R., and Johnson, B. (2004). The intervention selection bias: An underrecognized confound in intervention research. *Psychological Bulletin*, 130(2):289–303.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., and Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, 28(4):553–561.

Maathuis, M. H. and Colombo, D. (2015). A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060 – 1088.

Matthay, E. C. and Glymour, M. M. (2020). A graphical catalog of threats to validity. *Epidemiology*, 31(3):376–384.

McGill, W. (1954). Multivariate information transmission. *Psychometrika*, (19):97–116.

McMahan, I. (2017). Hacking the free throw: the science behind the most practiced shot in sports. https://www.theguardian.com/sport/2017/nov/22/free-throws-foul-shots-science-of-sports. [Online; posted 22-November-2017].

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Mohan, K. and Pearl, J. (2014). Graphical models for recovering probabilistic and causal queries from missing data. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037.

Montgomery, J. M., Nyhan, B., and Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775.

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

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

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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

Pearl, J. (2014). The deductive approach to causal inference. *Journal of Causal Inference*, 2(2):115–129.

Pearl, J. (2015a). Conditioning on post-treatment variables. *Journal of Causal Inference*, 3(1):131–137.

Pearl, J. (2015b). Generalizing experimental findings. *Journal of Causal Inference*, 3(2):259–266.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, page 247–254.

Pearl, J. and Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579 – 595.

Pearl, J. and Bareinboim, E. (2019). Note on "generalizability of study results". *Epidemiology*, 30(2):186–188.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2017). Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *J. Mach. Learn. Res.*, 18(1):8132–8193.

Richardson, T. and Robins, J. (2013a). Single world intervention graphs: a primer. *Working Paper, University of Washington, Seattle.*

Richardson, T. and Robins, J. (2013b). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Working Paper, Center for Statistics and the Social Sciences, University of Washington, Seattle*, (128).

Richiardi, L., Bellocco, R., and Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, 42(5):1511–1519.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34 – 58.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279–292.

Saadati, M. and Tian, J. (2019). Adjustment criteria for recovering causal effects from missing data.

Schneider, E. B. (2020). Collider bias in economic history research. *Explorations in Economic History*, 78:101356.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton-Mifflin, Boston.

Shahar, D. J. and Shahar, E. (2017). A theorem at the core of colliding bias. *The International Journal of Biostatistics*, 13(1):20160055.

Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, page 352–359, Arlington, Virginia, USA. AUAI Press.

Shpitser, I., VanderWeele, T., and Robins, J. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010, pages 527–536. AUAI Press.

Smith, L. (2020). Selection mechanisms and their consequences: Understanding and addressing selection bias. *Current Epidemiology Reports*, 7:179–189.

Smith, L. H. and VanderWeele, T. J. (2019). Bounding bias due to selection. *Epidemiology*, 30(4):509–516.

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

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465 – 472.

Swanson, S. A. (2019). A Practical Guide to Selection Bias in Instrumental Variable Analyses. *Epidemiology*, 30(3):345–349.

Swanson, S. A., Robins, J. M., Miller, M., and Hernán, M. A. (2015). Selecting on Treatment: A Pervasive Form of Bias in Instrumental Variable Analyses. *American Journal of Epidemiology*, 181(3):191–197.

Thompson, C. A. and Arah, O. A. (2014). Selection bias modeling using observed data augmented with imputed record-level probabilities. *Annals of Epidemiology*, 24(10):747–753.

Tripepi, G., Jager, K. J., Dekker, F. W., and Zoccali, C. (2010). Selection bias and information bias in clinical research. *Nephron Clinical Practice*, 115:94–99.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26.

VanderWeele, T. J. (2011). Controlled direct and mediated effects: definition, identification and bounds. *Scandinavian journal of statistics, theory and applications*, 38(3):551–563.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468.

Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology*, 23(1):159–164.

Westreich, D., Edwards, J. K., Cole, S. R., Platt, R. W., Mumford, S. L., and Schisterman, E. F. (2015). Imputation approaches for potential outcomes in causal inference. *International Journal of Epidemiology*, 44(5):1731–1737.

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., and Stuart, E. A. (2018). Target Validity and the Hierarchy of Study Designs. *American Journal of Epidemiology*, 188(2):438–443.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Zhang, J. (2008). Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(47):1437–1474.

Zhou, H. and Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4):493–504.

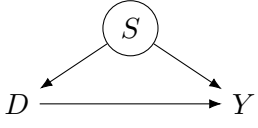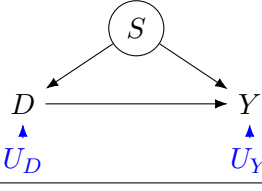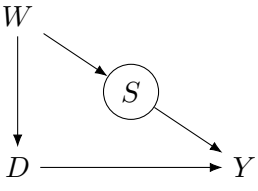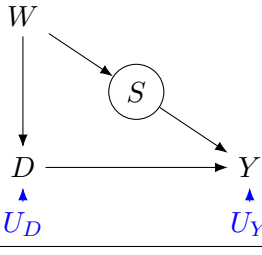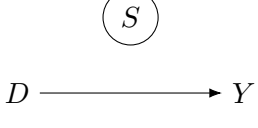Table 1: **Identification Possible without Covariate Adjustment**
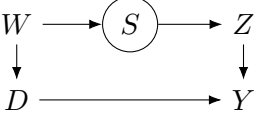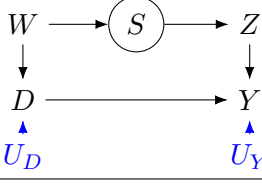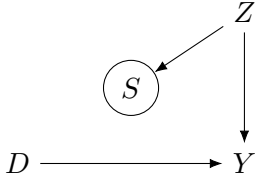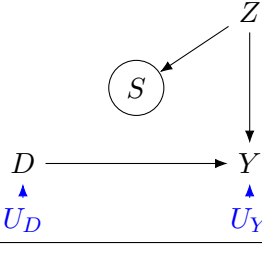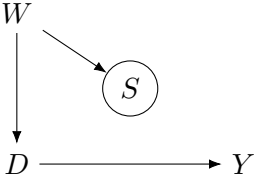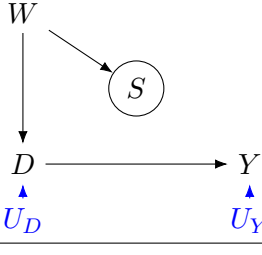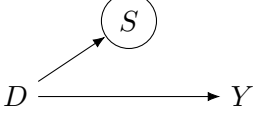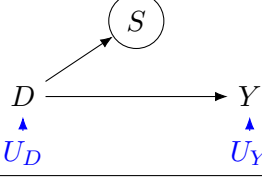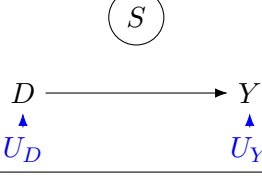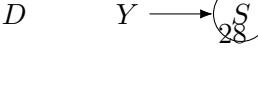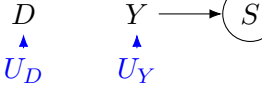
Relations with Selection	Causal Graph	Internal Selection Graph	Explanation
Selection as Confounder or Child of Confounder	(1.a.)		Let the adjustment set be $\{\emptyset\}$. ISAC is satisfied. No generalized non-causal paths between D and Y, that do not pass through S, exist.
	(1.b.)		
	(1.c.)		
Indirect Association with Selection	(1.d.)		
	(1.e.)		
Post-Treatment Selection	(1.f.)		
Selection Unrelated to Treatment and Outcome	(1.g.)		
Selection on Outcome; Treatment Does not Cause Outcome	(1.h.)		

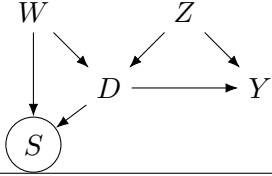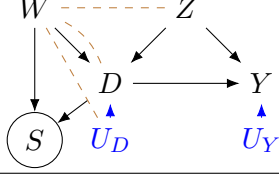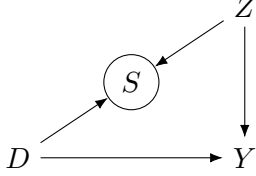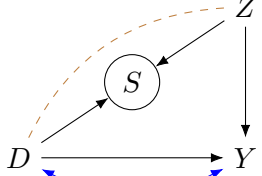Table 2: **Identification Possible with Covariate Adjustment**
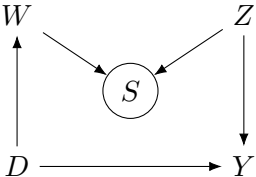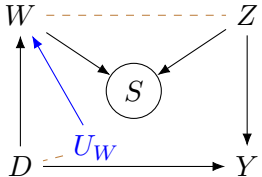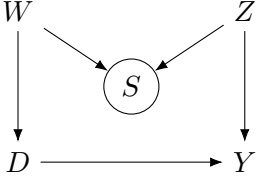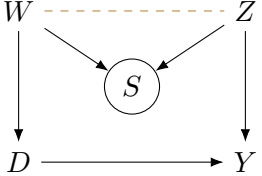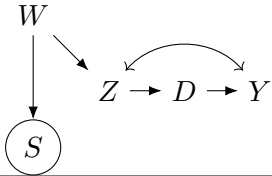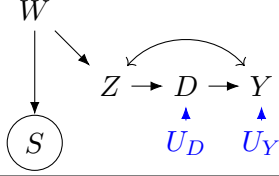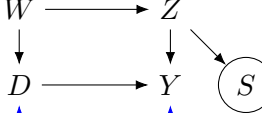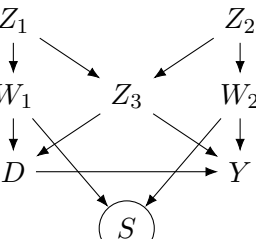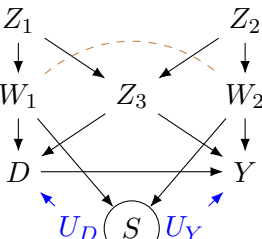
Relations with Selection	Causal Graph	Internal Selection Graph	Explanation
Post-Treatment Selection	**(2.a.)** W, Z, $D \to Y$, S	W, Z, $D \to Y$, S, U_D, U_Y	Let the adjustment set be $\{Z\}$. ISAC is satisfied. All generalized non-causal paths between D and Y, that do not pass through S, are blocked by Z. Note that for (2.c.), (2.d.), and (2.f), the adjustment set could also be $\{W\}$.
	(2.b.) Z, S, $D \to Y$	Z, S, $D \to Y$, U_D, U_Y	
	(2.c.) W, Z, S, $D \to Y$	W, Z, S, U_W, $D \to Y$, U_D, U_Y	
Indirect Association with Selection	**(2.d.)** W, Z, S, $D \to Y$	W, Z, S, $D \to Y$, U_D, U_Y	
	(2.e.) W, S, $Z \to D \to Y$	W, S, $Z \to D \to Y$, U_D, U_Y	
	(2.f.) $W \to Z$, $D \to Y$, S	$W \to Z$, $D \to Y$, S, U_D, U_Y	
	(2.g.) Z_1, Z_2, W_1, Z_3, W_2, $D \to Y$, S	Z_1, Z_2, W_1, Z_3, W_2, $D \to Y$, U_D, S, U_Y	ISAC is satisfied by $\{Z_3, W_1\}$ or $\{Z_3, W_2\}$.

29

Table 3: **Identification Not Possible with Covariate Adjustment**

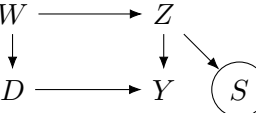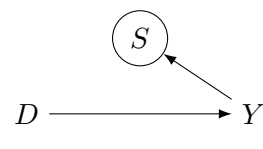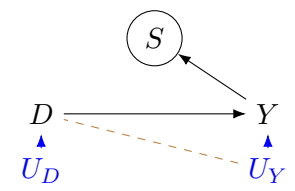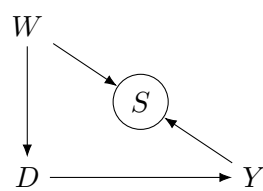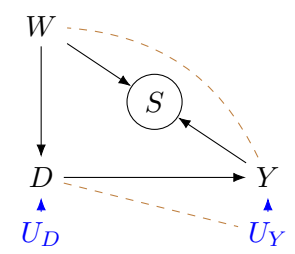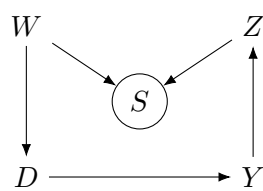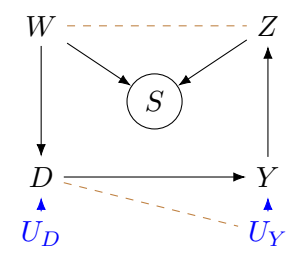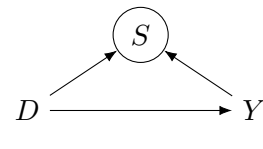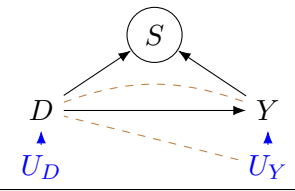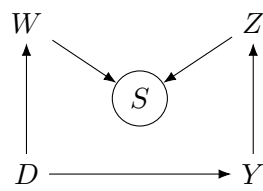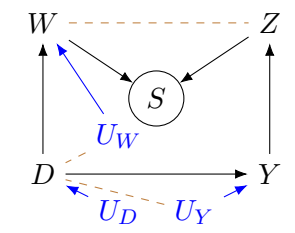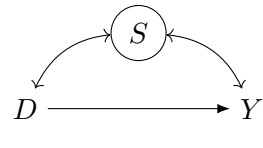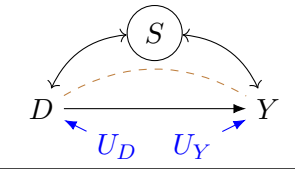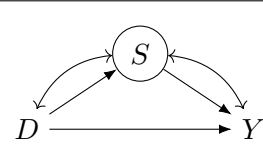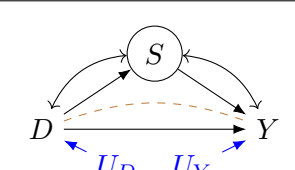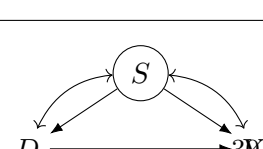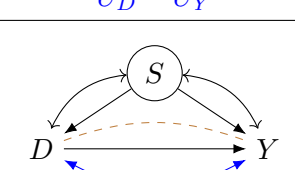| Relations with Selection	Causal Graph	Internal Selection Graph	Explanation
Selection on Outcome	**(3.a.)**		The generalized non-causal path $D \cdots U_Y \to Y$ cannot be blocked by any covariate adjustment set. Therefore, ISAC cannot be satisfied.
	(3.b.)		
	(3.c.)		
Selection on Outcome and Post-Treatment Selection (Berkson's Bias)	**(3.d.)**		
	(3.e.)		
Indirect Association with Selection	**(3.f.)**		The generalized non-causal path $D \cdots Y$ cannot be blocked by any covariate adjustment set because the common direct causes are unobserved. Therefore, ISAC cannot be satisfied. **(3.f.)** is like **(2.d.)**, but where the common direct causes are unobserved.
	(3.g.)		
	(3.h.)		

31

Table 4: **Selection as Mediator or Descendant of Mediator**
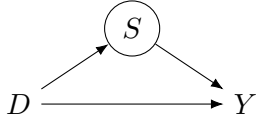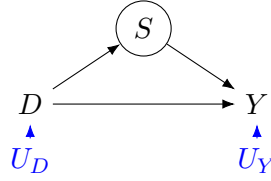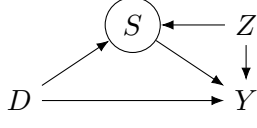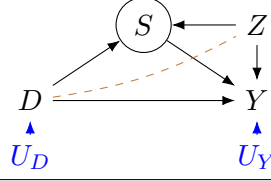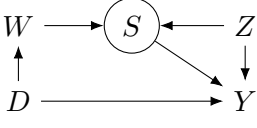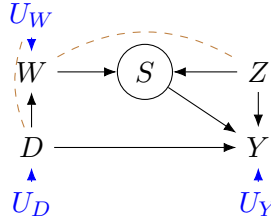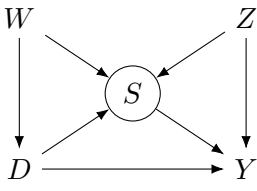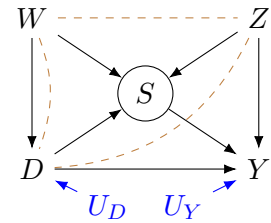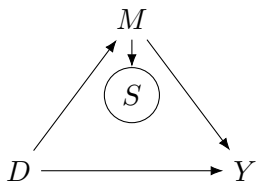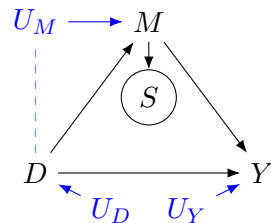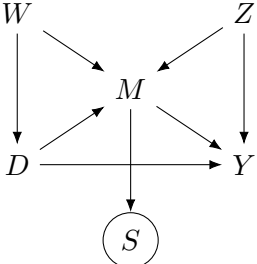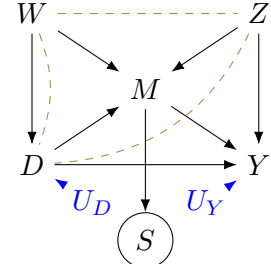
	Causal Graph	Internal Selection Graph	Explanation
(4.a.)			Let the adjustment set be $\{\emptyset\}$. ISAC is satisfied. No generalized non-causal paths between D and Y, that do not pass through S, exist.
(4.b.)			Let the adjustment set be $\{Z\}$. ISAC is satisfied. All generalized non-causal paths between D and Y, that do not pass through S, are blocked by Z.
(4.c.)			
(4.d.)			
(4.e.)			Let the adjustment set be $\{\emptyset\}$. ISAC is satisfied. The only generalized non-causal path between D and Y is one on which M, the mediator of which S is a descendant, is an ancestor of Y.
(4.f.)			Let the adjustment set be $\{Z\}$. ISAC is satisfied. All generalized non-causal paths between D and Y, that do not pass through S, are blocked by Z. And the only other generalized non-causal path between D and Y is one on which M, the mediator of which S is a descendant, is an ancestor of Y. This path contains U_M and is not shown here for clarity.

Table 5:   **Instrumental Variables and Sample Selection**

Causal Graph	Internal Selection Graph	Explanation
(5.a.)		**Future work will address this.** But for now we can see that internal selection graphs can aid in the analysis of whether instrumental variables is possible.
(5.b.)		
(5.c.)		
(5.d.)		
(5.e.)		
(5.f.)		

A   Exercises

Here we consider additional substantive examples that highlight different ways internal validity can relate to sample selection.

A.1   Exercise: Selection on Treatment and Outcome with a Mediator

In this exercise, we look at a piece of common knowledge to any basketball fan: taller players tend to have lower free throw percentages in the NBA. (McMahan, 2017; Helin, 2011) We show that this is indeed the case looking at NBA player-season free-throw percentages from 1950 - 2017.[34] A simple scatter plot of free throw percentage and height can be found in Figure 5. Running a simple linear regression, we find that there is a negative relationship between free throw percentage and height; see Table 6. We won't get bogged down by the model specification or estimation details here, it's easy enough to see the downward trend in the plot.

Figure 5:   Free Throws and Height
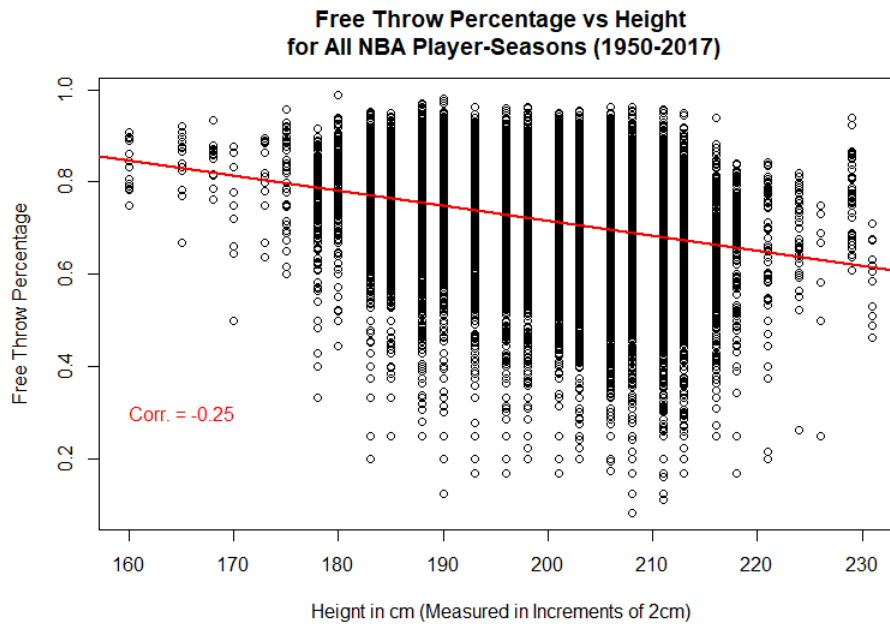
Table 6:   Free Throws and Height

Estimated Effect	SE	CI Low	CI High	DF
-0.0032803	8.34e-05	-0.0034438	-0.0031168	22995

But what might be causing this relationship? Is it really true that being taller causes you to make fewer free throws? We doubt it; we suspect there might be a slightly positive direct effect of height on shooting, since taller people are just closer to the basket, if there is a direct effect. So what might be going on? One

---

[34]The data come from https://www.basketball-reference.com/ downloaded by the authors from Kaggle: https://www.kaggle.com/datasets/drgilermo/nba-players-stats.

issue might be that players that are taller typically end up playing positions that require shooting primarily from ranges that are closer than the free throw line. So taller players might not practice shooting from the free throw distance as much as shorter players do. Another possibility is that there is a form of sample selection bias sometimes called ascertainment bias. (Elwert and Winship, 2014; Schneider, 2020) That is, it is likely the case that shooting ability has a positive influence on making it into the NBA and also that height has a positive influence on making it into the NBA. And so, even if there is no direct causal relationship between height and shooting ability (or even a slightly positive one), the fact that we are looking at a sample limited to players that made it to the NBA can create purely statistical relationship between height and shooting percentage.

Figure 6: DAG and internal selection graph for height (D) and free throw percentage (Y), where M might represent a mediator like players' position and S represent an indicator for players in the NBA

(a) DAG   (b) Internal Selection Graph
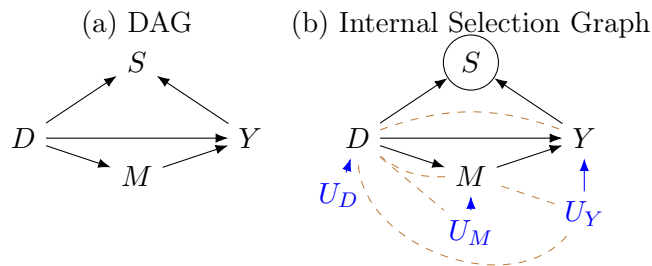
The internal selection graph for this setting might be captured by something like Figure 6. We see that there are generalized non-causal paths between height and shooting percentage. If the sample selection mechanism is strong enough, it could be enough to reverse the sign of the estimated effect. However, in this example, it actually might be reasonable to think that a negative effect mediated by position outweighs any slightly positive direct effect. Though, as the internal selection graph makes clear, even the mediator has purely statistical relationships with the treatment and outcome due to sample selection and paths like $D \cdots U_y \to Y$ and $D \cdots Y$ threaten the internal validity of an analysis of mediation.

A.2   Exercise: Selection on Treatment

In this section we look at two exercises that exhibit selection on the treatment variable in such a way that validity is threatened and bias results. We consider two online survey experiments from Zhou and Fishbach (2016). The authors aim to show that online experiments are often subject to high levels of attrition that can be biasing. The two specific experiments we discuss here were actually designed to induce bias resulting from attrition. Their subject matter is not very serious but they provide intuitive illustrations of how sample selection can bias treatment effect estimates. Both of the experiments have internal selection graphs of the form in Figure 7.

Figure 7: Zhou and Fishbach (2016) Exercise Graphs

(a) DAG   (b) Internal Selection Graph
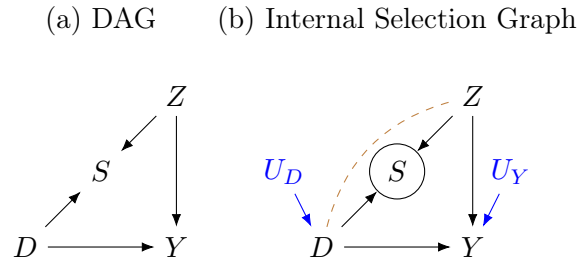
A.2.1 "Can doing more feel like less?"

Zhou and Fishbach (2016) "predicted that an experiment that assigns participants to recall many versus few happy events would result in a biased sample consisting of mainly happy people in the many-events condition, because happy events come to mind easily for these people, whereas the less-happy people in this condition would have to quit this difficult task. As a result of this experimental attrition, recalling many happy events could feel easier than recalling fewer happy events." The authors then conduct an experiment of this form using Amazon Mechanical Turk. Here D is the randomly assigned treatment which consisted of either being assigned to recall 4 or 12 happy events from the last year. Y is the outcome and is a measure of how difficult the treatment task was to complete on a 7 point scale ($1 = $ not difficult; $7 = $ extremely difficult). Z is a latent variable that captures each participants happiness. S is attrition from the study. The authors describe the attrition statistics as follows: "A total of 196 MTurk workers consented to take part in this experiment... Ninety-four of these participants dropped out of the survey once they learned what their first task (i.e., the experimental manipulation) entailed.... The dropout rate in the many condition, 69% (69/100), is significantly higher than in the few condition, 26% (25/96)..." The participants that dropped out did not complete the treatment task or rate the difficulty. The mean difficulty of recall rating for the group asked to recall 12 happy events was 2.74. The mean difficulty of recall rating for the group asked to recall 4 happy events was 3.97. This is despite the fact that, "[a]ll else being equal, recalling 12 happy events from the past year requires more effort than recalling four such events." So clearly the naive treatment effect estimate is biased even for the sample of individuals that did not drop out.

The authors discuss that sample selection introduces a confound as the task is easier for happier people. This would unbalance the treatment groups. The graphical tools we have developed make this mechanism completely clear. Which treatment group you are randomly assigned to influences your decision whether to drop out or not, since one treatment is more demanding than the other and some individuals might not want to put in the extra effort. How happy you are generally influences both your decision to drop out or not, since being happier makes the task less demanding, and your rating of how difficult the task was, for the same reason. Since we are forced to condition on the selection node and this is a collider between the treatment assignment and happiness, a purely statistical relationship is created between the random treatment assignment and happiness. This purely statistical relationship in turn creates a non-causal path between the treatment and the outcome, biasing the effect estimates. This can all easily be seen from the internal selection graph and we can see that ISAC is violated since we cannot block the path $D \dashrightarrow Z \to Y$. This is because

happiness is unobserved. If this were observed, then we could satisfy ISAC. So we see that another study that measures general happiness in some reliable way might be able to reach an unbiased treatment effect estimate.

A.2.2 "Can imagining applying eyeliner help one lose weight?"

Zhou and Fishbach (2016) "predicted that an experiment that assigns participants to imagine applying eyeliner (vs. applying aftershave cream) would end up with a sample that is disproportionally female. As a result, participants assigned to imagine applying eyeliner would report weighing less than those assigned to imagine applying aftershave." Obviously, the true treatment effect should be zero, since considering eyeliner or aftershave will not reduce one's weight. The authors then conduct an experiment of this sort using Amazon Mechanical Turk. Here D is the randomly assigned treatment which consisted of either being assigned to describe how applying versus not applying eyeliner would make one feel differently or being assigned to describe how applying versus not applying aftershave cream would make one feel differently. Y is the outcome and is the participants' self-reported weight in pounds. Z is an indicator variable that captures each participants gender. S is attrition from the study. The authors describe the attrition statistics as follows: "A total of 144 MTurk workers consented to take part in this experiment... Forty-one of these participants dropped out of the survey once they learned what their first task (i.e., the experimental manipulation) entailed.... The dropout rates were comparable across the two conditions: 32.4% (24/74) in the eyeliner condition and 24.3% (17/70) in the aftershave cream condition ..." The participants that dropped out did not complete the treatment task or rate the difficulty. The mean weight for the group asked to discuss eyeliner was 159 pounds. The mean weight for the group asked to discuss aftershave cream was 182 pounds. This is despite the fact that, discussing these products has no effect on weight. So clearly the naive treatment effect estimate is biased even for the sample of individuals that did not drop out. The authors observed that the eyeliner group did indeed have more females than the aftershave group.

Table 7: Zhou and Fishbach (2016) Exercise - Average Weight by Treatment

Treatment	Average Weight	N
Eyeliner	159.02	49
Aftershave	182.08	53

The authors discuss that sample selection introduces a confound as "imagining applying eyeliner would be difficult or even aversive for average adult males, inducing them to quit." This would unbalance the treatment groups. The graphical tools we have developed make this mechanism completely clear. Which treatment group you are randomly assigned to influences your decision whether to drop out or not, since one treatment is more demanding than the other depending on your gender. So your gender and the your treatment assignment both influence your decision to drop out or not. Gender also influences weight on average, given that "females generally weigh less than males." Since we are forced to condition on the selection node and this is a collider between the treatment assignment and gender, a purely statistical relationship is created between the random treatment assignment and gender. This purely statistical relationship in turn creates a non-causal path between the treatment and the outcome, biasing the effect estimates. This can all easily be seen from the internal selection graph and we can see that ISAC is violated since we cannot block the path $D \cdots Z \to Y$.

If gender is observed, and our causal graph is accurate, then we would be able to satisfy ISAC. Since the authors did observe participant gender, they could have actually adjusted for gender to get unbiased treatment effect estimates. We take this step here. Taking simple averages of weight by treatment and gender, we see that there is actually a difference in average weight between females who received the two treatments; something similar is true for males, but less so.

Table 8: Zhou and Fishbach (2016) Exercise - Average Weight by Treatment and Gender

Treatment	Gender	Average Weight	N
Eyeliner	Male	182.34	29
Aftershave	Male	191.76	37
Eyeliner	Female	125.20	20
Aftershave	Female	159.69	16

We might wonder whether some other variables that might predict weight would help reduce some of the variability in weights that could be leading differences in weight, whether these differences are due purely to sampling error or to some systematic causal relationship driving the differences, like gender does. We consider adjusting for (in a simple linear model) individuals' birth year, English as first language, race, education, income, total duration of survey, and location information. All of these are pre-treatment variables and we do not believe adjusting for any of these variables will induce or amplify bias. It is possible that some are also related to attrition, like gender is. So we believe adjusting for these will improve precision and reduce bias.[35] We fit a linear model with these covariates that does not adjust for gender and one that does. We find that, while the model without gender still shows a positive effect on weight for individuals who discussed aftershave. The model that adjusts for gender also shows a positive effect, but the 95% confidence interval includes zero. We might conclude that our simple causal model above is slightly inaccurate, since the covariates might need to be incorporated; further, some additional bias might be working through some unobserved variable. Though, it is also possible that the difference we see is just from sampling error. Note that we are able to make these conclusions because we know the treatment effect should be zero. We will not generally have such information with which to critique our causal model.

Table 9: Zhou and Fishbach (2016) Exercise - Linear Model Treatment Effect Estimates

	Estimated Effect	SE	CI Low	CI High
Not Adjusting for Gender	20.344	9.798	0.88	39.80
Adjusting for Gender	13.937	8.705	-3.35	31.23

A.3   Exercise: Selection on a Mediator

Similar mechanisms and analyses from our working example apply for a range of fairness and discrimination questions. The examples described by Figure 8 are based on examples discussed in Mitchell et al. (2021).

---

[35]We assert this and the conclusions in this paragraph to keep our demonstration simple. Additional consideration might reveal more about how adjusting for these variables alters causal effect estimates.

In the lending example, researchers might be attempting to understand how immigrant status relates to loan repayment or simply predicting repayment. Often the data used in such an investigation would only include individuals who actually received loans, as they have the potential to repay. In the pre-trial release example, we might want to understand how perceptions of race relate to appearance at an appointed court date or simply predicting appearance. Again, the data in such investigations are typically be limited to individuals who were actually released, as they are able to appear or not. In these examples, we again see purely statistical relationships between the protected or sensitive group statuses and outcomes that can bias the results of the investigation.

Figure 8: **Discrimination in Lending**: D indicates immigrant status, S indicates receiving a loan, Y indicates repayment, and U represents unobserved factors. **Discrimination in Pre-Trial Release**: D indicates perception of race, S indicates pre-trial release, Y indicates appearance for court date, and U represents unobserved factors.
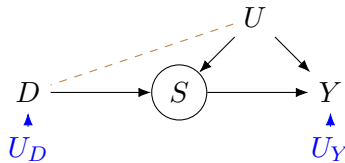
A.4   Exercise: Selection on Pre-Treatment and Post-Treatment Outcomes

In this section, we consider a second exercise that draws from one of the online survey experiments in Clifford et al. (2021). The authors conduct six such experiments in which the outcome is measured both before and after treatment. The paper aims to show that measuring and adjusting for pre-treatment outcomes can increase precision of treatment effect estimates. We leverage this data to simulate how sample selection based on pre- vs post-treatment outcomes leads to very different conclusions for internal validity. While we focus on only one of the experiments from this paper in this section, we consider four more of the experiments in the Appendix.

We focus on the experiment related to education spending information. In this experiment "a random half of respondents were informed of the average annual per pupil spending on public schools. All respondents were then asked whether taxes to support public schools should be increased or decreased on five-point scale." The outcome question was also asked before treatment. We emphasize that the treatment was randomly assigned. The goal of this experiment is to see whether providing information about average annual per pupil spending on public schools causes a change in attitudes about taxes used for education. In the version of the experiment from Clifford et al. (2021), there are no obvious concerns about sample selection. We do not worry about whether or not this is true; rather, we impose sample selection of two forms in a simulation aimed at illustrating how different sample selection mechanisms can change the validity of causal effect estimates. Both of the sample selection mechanism are selection based on the outcome of interest. In one, we impose sample selection based on the post-treatment outcome (i.e., support for increased taxes after being treated); in the other, we impose sample selection based on the pre-treatment outcome (i.e., support for increased taxes before being treated). These selection mechanisms can be thought of as attrition, perhaps related to

shyness about expressing opinions about cutting spending on education. In the experiment, the authors also measured political party identification and ideology. These covariates are used to increase the precision of effect estimates, as is the pre-treatment outcome. The original DAG and internal selection graphs for this experiment and the simulations based off of it are in Figure 9.

Figure 9: Clifford et al. (2021) Exercise Graphs

(a) DAG

(b) Selection on Post-Treatment Outcome   (c) Selection on Pre-Treatment Outcome
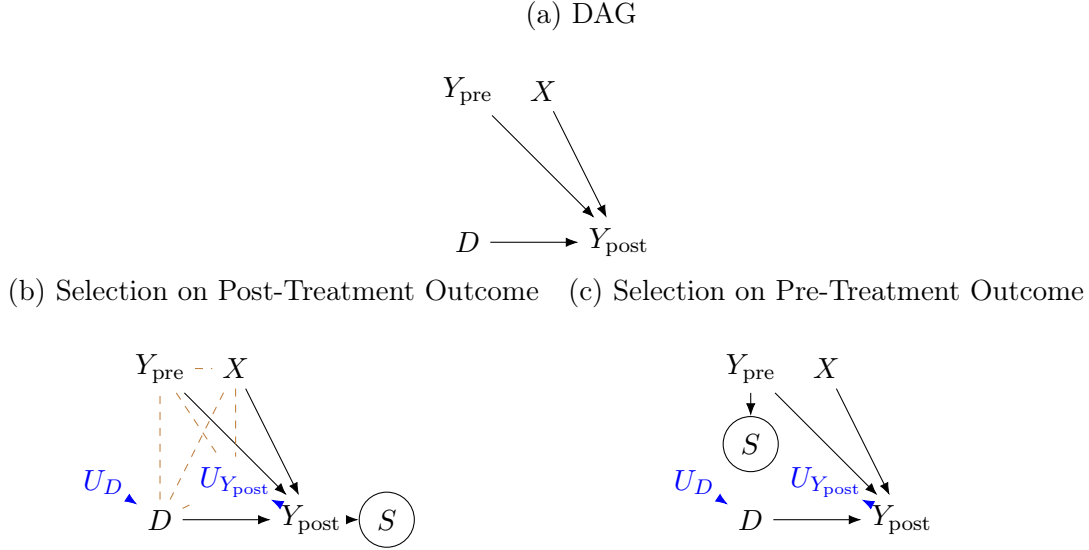
The simulations take the following form:

1. estimate the treatment effect from the experiment adjusting for pre-treatment outcomes, political party identification, and ideology to increase precision
2. fit a simple linear model to the control units
3. use this model to simulate the control potential outcomes for all units
4. simulate treatment potential outcomes as control potential outcomes plus the constant treatment effect we estimated from the experimental data plus mean zero noise
5. calculate a new post-treatment outcome from potential outcomes based on unit treatment status
6. estimate three versions of the treatment effect
   - where there is no sample selection
   - where sample selection is a function of the post-treatment outcome
   - where sample selection is a function of the pre-treatment outcome

The sample selection mechanisms we consider first sorts units by ascending pre- or post-treatment outcome (depending on what which we'll be selecting based on), then by party identification and ideology. We then give units that are sorted to be in the lower 50% of units (i.e., units with lower outcome values) a 20% probability of staying in the sample and units that are sorted to be in the upper 50% of units (i.e., units with higher outcome values) an 80% chance of staying in the sample. This sample selection can be thought of as attrition. The attrition mechanisms are meant to be simple, to lead to similar sample sizes for selection on pre- and post-treatment outcome, and not necessarily to be very realistic; our main goal here is demonstrating the potential for bias. The mean zero noise that we use to simulate the treatment potential outcomes has

a standard deviation that is twice that of the actual post-treatment outcome from the experimental data. Increasing the variance of the noise increases the bias of the estimates with selection on the post-treatment outcome, but has no effect other than on the precision of the estimates with no selection or with selection on the pre-treatment outcome. This is what we should expect based on Figure 9. We want to highlight the possible bias from selection on post-treatment outcomes, so we ensure that the noise is variable enough to see this. Some actual post-treatment outcomes were missing, which lead to some units being excluded from the experimental results. These units are also excluded from the simulations.

Figure 9(b) shows that selection on the post-treatment outcome leads to an unblocked non-causal path that runs $D \cdots U_{Y_{\text{post}}} \to Y_{\text{post}}$, where the relationship between D and U_y is purely statistical and is the result of selection on the post-treatment outcome. A similar relationship is also created between D and Y_{pre} and D and the covariates, X. However, these are observed and we adjust for Y_{pre} and X when estimating effects. So estimates under this type of selection should be biased. Figure 9(c) shows that no spurious relationships are created as a result of sample selection on the pre-treatment outcome and there are no non-causal paths between the treatment and post-treatment outcome. So estimates under this type of selection should not be biased.

The results of the experiment and the simulations are in Table 10. The experimental "true" treatment effect estimate is close to zero and it's confidence interval includes zero. Selection on the post-treatment outcome leads to a positive effect estimate that is far from zero; the confidence interval for this estimate does not include zero. Selection on the pre-treatment outcome leads to an estimate that is close to the experimental effect; the confidence interval for this estimate does include zero. The simulation created a constant treatment for all individuals that is equal to the experimental effect estimate. In this exercise, we end up finding a positive effect when none exists, due to selection on the post-treatment outcome; but there is no such bias resulting from selection on the same variable just measured before treatment. So we have seen that selection based on pre- vs post-treatment outcomes can pose very different threats to the internal validity of effect estimates.

Table 10: Clifford et al. (2021) Exercise Table

	Estimated Effect	SE	CI Low	CI High	Bias	N
"True" Treatment Effect (Experimental Results)	-0.034	0.044	-0.12	0.05	0.00	400
No Selection	0.094	0.132	-0.17	0.35	0.13	400
Selection on Post-Treatment Outcome	0.804	0.176	0.46	1.15	0.84	189
Selection on Pre-Treatment Outcome	0.096	0.183	-0.26	0.46	0.13	204

A.5 Exercise: Selection on Treatment and Unobserved Cause of Outcome

Hazlett (2020) considers the effect of being directly harmed in the conflict in Darfur in early 2000s on attitudes about peace using a survey of individuals in refugee camps. The paper controls for things like village, gender, and other important covariates. While adjustment for these covariates likely reduces non-causal association between the treatment and outcome, being harmed may effect whether someone re-entered the conflict (and hence was not captured in the survey). An individual's pro-peace predisposition (before the conflict) may be a common cause of both whether they re-entered the conflict and their peace attitudes at the time of the survey.

This may cause there to be a generalized non-causal path running from harm to pro-peace predisposition to attitude about peace that could threaten the internal validity of estimated effects. In Figure 10, D is direct harm, Y is attitudes about peace, Z is pro-peace predisposition (before the conflict), S is being in the survey from the refugee camp (i.e., did not re-enter the conflict), and X is observed covariates like village and gender. Hazlett (2020) is able to adjust for the observed covariates, but the path $D \dashrightarrow Z \rightarrow Y$ cannot be blocked since pro-peace predisposition is not observed.

Figure 10: Possible threat to internal validity in Hazlett (2020)

(a) DAG   (b) Internal Selection Graph
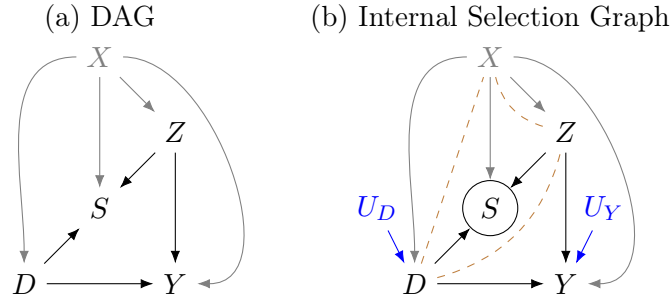
One lesson from this example would be to try to measure pro-peace predisposition, which is essentially a pre-treatment version of the outcome. This might be difficult in practice, however, since this is a pre-treatment variable that would ideally have been measured for individuals before the conflict. Another perspective, which is argued in the paper, is that the process that would drive individuals back into the conflict would "act more powerfully for men of fighting age because in this context, few women or elderly participate directly in the armed opposition groups. If such a process drove the results, we would see the apparent effect most strongly among young men but should see little or no apparent effect among women or the elderly who are far less likely to join the opposition. This is not the case." (Hazlett, 2020) We might then claim that the effect of direct harm on peace attitudes among women and the elderly is perhaps not biased by this sample selection mechanism. But this sample selection mechanism might not allow us to obtain an internally valid effect estimate for fighting age men.

B   Technical Appendix

Here we provide technical details and prove the results found in the main text. First, we briefly discuss why we work with potential outcomes. We then introduce a series of definitions. These are followed by a series of lemmas. Then we state our main results in a set of theorems that follow directly from the lemmas. We conclude with a discussion of IPW estimation of $\mathbb{E}[Y_d|S=1]$, $\mathbb{E}[Y_{d,S=1}|S=1]$, and $\mathbb{E}[Y_{d,m}|S=1]$.

B.1   Sample selection, the do-operator, and potential outcomes

Let us linger on the choice to use potential outcomes as opposed to the do-operator (Pearl, 1995, 2009) in our discussion. We discuss post-treatment selection, selection as a mediator, selection as the child of a mediator, and all other major roles that selection can play in the structure of causal models. In the case of non-post-treatment selection, potential outcomes will typically have the same interpretation as the do-operator. In the case of selection as a mediator and selection as a descendant of a mediator, the causal effects of interest are typically defined using potential outcomes. (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2011; VanderWeele and Vansteelandt, 2009; Richiardi et al., 2013) So here potential outcomes notation is really the natural and traditional choice. In the case of post-treatment selection, where selection is not a mediator or a descendant of a mediator, $p(Y_d|S=1)$ is usually of interest and not $p(Y|do(D=d), S=1) = p(Y_d|S_d=1)$.[36] As Pearl (2015a) states, "By the counterfactual query $Q_c[= p(Y_d|S=1)]$ we mean: Take all units which are currently at level $[S=1]$, and ask what their Y would be had they been exposed to treatment $[D=d]$. This is different from $Q_{do} = [p(Y|do(d), S=1)]$, which means: Expose the whole population to treatment $[D=d]$, take all units which attained level $[S=1]$ (post exposure) and report their Y's." Further, Pearl (2015a) states "Q_{do} is rarely posed as a research question of interest, probably because it lacks immediate causal interpretation. It serves primarily as an auxiliary mathematical object in the service of other research questions. ... I have not seen Q_{do} presented as a target query on its own right." For non-post-treatment selection, $p(Y_d|S=1) = p(Y|do(d), S=1)$, since $S_d = S$. In our context, we will typically be interested in quantities of the type of $p(Y_d|S=1)$, which tell us the distribution of outcomes for units that were selected in reality, had they been exposed to treatment $D = d$.

B.2   Definitions

Definition B.1 (SCM (adapted from Pearl (2009)))**.** *A structural causal model, M, has the following parts*
1. *U is a set of background variables determined by exogenous factors;*
2. *V is a set $\{V_1, V_2, \ldots, V_n\}$ of variables determined by variables in the model;*
3. *F is a set $\{f_1, f_2, \ldots, f_n\}$ of functions that map $f_i : U_i \cup PA_i \to V_i$, where $U_i \subset U$ and $PA_i \subset V \setminus V_i$ and the entire set F forms a mapping from U to V. That is, each f_i assigns a value to V_i that depends on the values of a select set of variables in $V \cup U$ ($v_i = f_i(pa_i, u_i)$), and the entire set F has a unique solution $F(u)$.*
4. *$p(u) = \prod p(u_j)$ is a probability function defined over the domain of U.*

---

[36] $p(Y|do(D=d), S=1) = \frac{p(Y,S=1|do(D=d))}{p(S=1|do(D=d))} = \frac{p(Y_d, S_d=1)}{p(S_d=1)} = p(Y_d|S_d=1)$. (Pearl, 2014) In this case, S_d is the potential selection value when the treatment variable D takes the value that we are investigating in Y_d.

Definition B.2 (Sub-Model (adapted from Pearl (2009))). *Let M be a causal model, D be a set of variables in V, and d a particular realization of D. A submodel M_d of M is the causal model M_d, where F is replaced with F_d, which is formed by deleting the functions for the variables in D and replacing them with constant functions $D = d$.*

Definition B.3 (Potential Outcome (adapted from Pearl (2009))). *Let D and Y be two subsets of variables in V. The counterfactual values of Y when D had been set to d, written Y_d, is the solution for Y of the set of equations F_d, given the realized values of the background variables, U.*

Definition B.4 (Causal Graph (adapted from Shpitser et al. (2010), also see Pearl (1988, 2009))). *A SCM induces a causal graph in the following way. Each variable in the model is represented by a node. A node corresponding to variable V_i has edges pointing to it from every variable whose value is used to determine the value of V_i by the function f_i. Exogenous variables have no edges pointing to them. A causal graph is an I-map (see Definition B.11 below) for $p(v)$.*

Definition B.5 (Path). *A path is a sequence of edges in G where each pair of adjacent edges in the sequence share a node, and each such shared node can occur only once in the path.*

Definition B.6 (Causal Path). *A causal path from D to Y is a path from D to Y on which all edges are directed and point away from D and toward Y.*

Definition B.7 (Proper Causal Path (Shpitser et al., 2010))). *Let D, Y be sets of nodes. A causal path from a node in D to a node in Y is called proper if it does not intersect D except at the end point.*

Definition B.8 (Non-Causal Path). *A non-causal path is a path that is not a causal path.*

Definition B.9 (Parents, Ancestors, and Descendants). *Parents of node X are the nodes in the graph from which an edge points directly to X. An ancestor of X is any node which has a causal path to X. A descendant of X is any node which X has a causal path to.*[37]

Definition B.10 (d-Separation and Blocking (adapted from Pearl (2009))). *Two sets of nodes, D, Y, in a graph G are said to be d-separated by a third set, Z, if every path from any node $D_0 \in D$ to any node in $Y_0 \in Y$ is blocked. A path is blocked by Z if either [1] some W is a collider on the path between D, Y and $W \notin Z$ and the descendants of W are not in Z or [2] W is not a collider on the path but $W \in Z$.*

Definition B.11 (I-map (adapted from Pearl (1988))). *A causal graph G is said to be an I-map of a dependency model M if every d-separation condition displayed in G corresponds to a valid conditional independence relationship in M. That is, for every set of three nodes X, Y, and Z, if Z d-separates X from Y in G, then X is independent of Y given Z.*

Shpitser et al. (2010) discuss a graphical representation called latent projections of causal graphs that contain both directed and bidirected edges. Latent projections allow us to exclude latent variables in convenient ways. Specifically, they include a node for every observed variable. However, two observable nodes A and B are connected by a directed edge only when any and all intervening variables between A and B are latent.

---

[37]We do not consider a node to be a descendant of itself.

Also, A and B are connected by a bidirected edge when there is a path from A to B that is not d-separated that starts with an edge pointing into A and ends with an edge pointing into B and all the nodes on this path are latent other than the end points. As Shpitser et al. (2010) point out, latent projections retain all d-separation statements from the original graph. We will also allow for such latent projections to be used to simplify graphs. For our purposes, we do not allow sample selection to be treated as a latent variable and so it should always be included as a separate node in the graph.

Definition B.12 (Twin-Network (adapted from Shpitser et al. (2010)))**.** *The twin network graph, N, (Balke and Pearl, 1994b,a) displays counterfactual independence among two possible worlds, the pre-intervention world which is represented by the original graph G, and the post-intervention world, which is represented by the graph $G_{\overline{D}}$ (a copy of G with the edges pointing into D deleted and D replaced with $D = d$). The twin network is an I-map for the joint counterfactual distribution $p(v, v_d)$, where V is the set of all observables, and V_d is the set of all observable variables after the intervention $do(D = d)$ was preformed. The observable nodes in these two graphs share the U variables, to signify a common history of these worlds up to the point of divergence due to $do(D = d)$. We add the additional refinement from Shpitser and Pearl (2007) where node copies of all non-descendants of D in G and $G_{\overline{D}}$ are merged in the twin network graph (since such nodes are the same random variable in both the pre and post intervention worlds).*

In our proofs, we will consider causal graphs and twin networks in which each bidirected edge between nodes A and B is replaced with a node U^*_{AB} that is a common cause of the two nodes that were connected with the bidirected edge and points to each of A and B. This replacement does not change d-separations from the original graph. See Figure 11 for a simple example. Correa et al. (2018) make a similar alteration to the causal graphs they consider.

Figure 11:   Twin network with no bidirected edges

(a) DAG with bidirected edge   (b) DAG without bidirected edge

(c) Twin Network without bidirected edge
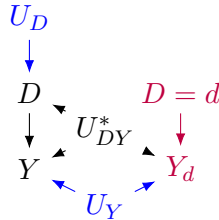
Definition B.13 (Colliders)**.** *A collider is a node in a causal graph into which two (or more) arrow heads point. For nodes A, B, C, let C be a collider between A and B if it appears in the following sub-path of the causal graph: $A \to C \leftarrow B$.*

44

Definition 1 (Internal Selection Graph, G_S^+). Let G be the DAG induced by a SCM.

1. Create G_S by adding an appropriately connected binary selection node, S.
2. Draw a circle around S to clearly indicate that we must limit our analysis to $S = 1$.
3. Add to G_S any node which is a parent of the treatment or a parent of a descendant of the treatment. (U_S, the background factors contributing to selection, can be excluded.)
4. Add a dashed undirected edge between all variables between which S is a collider or an ancestor of S is a collider. We will call these dashed, undirected edges *bridges*.

Call the resulting graph an *internal selection graph*, G_S^+.

(This definition is similar to the "modified extended diagram" in Daniel et al. (2012).)

Definition B.14 (Extended Twin-Network). *An extended twin network, N_S^+, is a twin network, N_S, containing an appropriately connected pre-intervention binary selection node, S, and any corresponding post-intervention versions of it, where we add bridges between all variables between which the pre-intervention S is a collider or an ancestor of pre-intervention S is a collider. (Note that pre and post-intervention versions of S are assumed to have been added to both N_S and N_S^+; we don't use a subscript to indicate this here.) It is easy to see that, like a twin network, an extended twin network displays counterfactual independence among two possible worlds, the pre-intervention world which is represented by the original graph G_S^+, and the post-intervention world, which is represented by the graph $(G_S)_{\overline{D}}$.*

Extended twin networks are useful for the same reason that internal selection graphs are useful. There can be purely statistical relationships between variables in the sample that are not captured in regular twin networks. See Figure 12. As we saw in the main text, bridges do not create colliders, since they are graphical representations of conditioning on sample selection when it is a collider. So bridges do not alter the underlying fully directed graph. Since the addition of bridges does not create any colliders, d-separation and blocking retain their definition in internal selection graphs and extended twin networks. See Lemmas B.4 and B.5 that show how d-separation (using the same definition) in internal selection graphs and extended twin networks corresponds to d-separation in causal graphs and twin networks. As a result, we can then get independence statements by reasoning about internal selection graphs and extended twin networks.

Figure 12: Example of Extended Twin Network
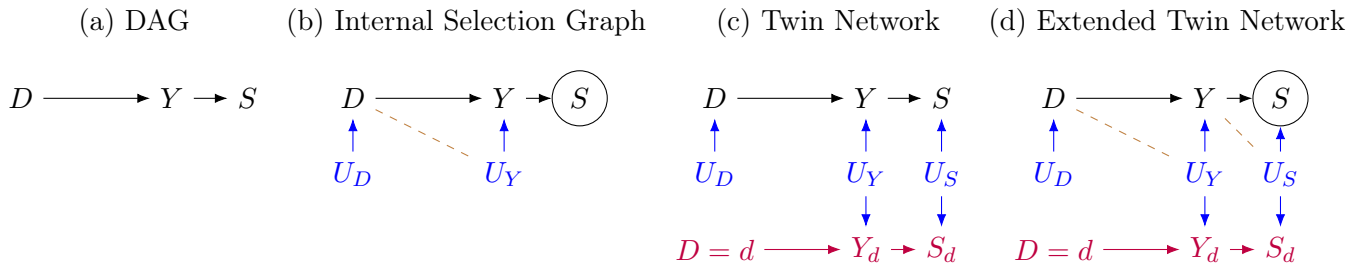
Twin network graphs can become pretty complicated, even when the original causal graph only contains three nodes.[38] This is what makes graphical criteria like the one presented in this paper attractive for

---

[38]Richardson and Robins (2013a,b) also introduce a graphical approach to visualizing how post-intervention world variables

simplifying the analysis that leads to ignorability statements. The internal selection graph maintains only the necessary elements of the extended twin network that allow us to use the internal selection criterion to see when conditional ignorability is possible. We are not advocating that researchers actually work with extended twin networks themselves. We discuss extended twin networks in our proofs only. We advocate using internal selection graphs, which are usually much simpler than twin networks and extended twin networks, and the internal selection adjustment criterion for determining ignorability.

Figure 13: Twin Networks versus Single World Intervention Graphs (SWIGs)

In this example, W can be used to block generalized non-causal paths between D and Y_d; however, only the counterfactual world W_d appears in the "extended SWIG." The actual world W and the counterfactual world W_d both appear in the twin network.

(a) DAG (b) Internal Selection Graph (c) "Extended SWIG" (d) Extended Twin Network
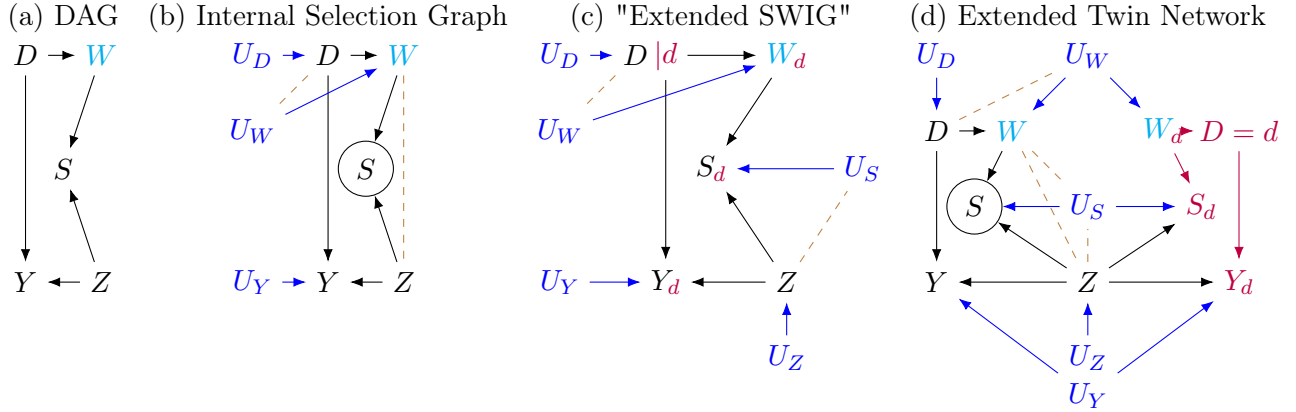
Definition B.15 (Paths and Generalized Non-Causal Paths). *We revise Definition B.5 to state that a path is a sequence of edges in G_S^+ or N_S^+ where each pair of adjacent edges in the sequence share a node, and each such shared node can occur only once in the path, where we allow the edges to be bridges, as well as directed edges. A generalized non-causal path is a path that is not a causal path.*

Definition B.16 (Route (adapted from Shpitser et al. (2010))). *A route from D to Y in a graph, G_S^+ or N_S^+, is a sequence of edges, where each pair of adjacent edges share a node, the unshared node of the first edge is D, and the unshared node of the last edge is Y. (Shared nodes can occur more than once.) A route is d-separated if the same triples are blocked as in the definition of d-separation above. The difference between a route and a path is that paths cannot contain duplicate nodes while routes can. Note that we allow edges to be bridges.*

relate to pre-intervention world variables called Single World Intervention Graphs (SWIGs). SWIGs are often simpler than twin networks. However, they do not provide exactly the same picture of both the pre-intervention world and the post-intervention world that twin networks do; for example, pre-intervention world post-treatment variables do not appear in SWIGs, though they can be used to block generalized non-causal paths between the treatment and potential outcome of interest. See Figure 13. Here, W can be conditioned on to block the open generalized non-causal path between D and Y_d but the pre-intervention world W does not appear in the SWIG, while it does appear in the twin network. The pre-intervention selection node is also missing from the SWIG. Finally, the path $D \to W \cdots Z \to Y_d$ is also missing from the SWIG; we do not want to have W_d touch any bridges since it is not actually a parent of the pre-intervention selection node. As such, we use twin networks for our discussion. This example also demonstrates why the criterion in Daniel et al. (2012) is less general than we might want, since those authors do not allow any adjustment for post-treatment variables. We believe that internal selection graphs and ISAC provide the simplest approach for practitioners to analyze sample selection and internal validity, by only slightly extending the causal graphs that they are used to seeing.

Definition B.17 (Direct Route (adapted from Shpitser et al. (2010))). *Let π be a route from D to Y in G_S^+ or N_S^+. Label each node occurrence in the route π by the number of times the node has already occurred earlier in π. A direct route π^* is a sub-sequence obtained from π inductively as follows:*

- *The first node in π^* is the first node in π with the largest occurrence number.*
- *If the kth shared node in π^* (and the mth node in π) is (X_i, r), and $X_i \neq Y$, let the $k+1$th node in π^* be (X_j, n), where X_j is the $m+1$th node in π, and n is the largest occurrence number of X_j in π.*

Definition 2 (Internal Selection Adjustment Criterion (ISAC)). A set of nodes Z in G_S^+ satisfies the internal selection adjustment criterion relative to D (treatment) and Y (outcome) if

1. No element of Z lies on or is a descendant of a node that lies on a causal path originating from D and arriving at Y. Note that an element of Z could be a descendant of D itself, if it is not on a causal path from D to Y. Note also that elements of Z should not be on, or descendants of nodes on, causal paths even if S is also on the causal path.

2. Z blocks every generalized non-causal path between D and Y that does not pass through S. Note that generalized non-causal paths passing through M, when S is a descendant of mediator M, on which M is an ancestor of Y also do not need to be blocked, assuming the previous condition is not violated. Further, M should not be a member of Z from the previous condition.

Definition 3 (Generalization Criterion (GC)). This definition is a translation of Definition 8 from Correa et al. (2018). A set of nodes Z in G_S^+ satisfies the generalization criterion relative to D (treatment) and Y (outcome) if

- Z satisfies ISAC relative to D and Y and
- $Z_{\text{Ext}} \subset Z$ blocks all causal and generalized non-causal paths between Y and S in G_S^+ other than those that end in a causal path from D to Y.

B.3 Lemmas

Lemma B.1 (adapted from Shpitser et al. (2010); Pearl (1988)). *Let G be a causal graph. Then any model M with a distribution $P(u, v)$ inducing G, if A is d-separated from B by C in G, then A is independent of B given C, which we write $A \perp\!\!\!\perp B | C$ in $P(u, v)$.*

Lemma B.2 (adapted from Shpitser et al. (2010)). *For every route π in G_S^+, the direct route π^* is a path. Moreover, if π is unblocked, then π^* is unblocked.*

B.3.1 Selection unrelated to mediators

Lemma B.3. *If Z satisfies ISAC in G_S^+ relative to D and Y and S is not a mediator or descendant of a mediator between D and Y, then Z d-separates D and Y_d in N_S^+.*

Proof. We very closely follow the structure of the proof of Theorem 4 of Shpitser et al. (2010). We will show the contrapositive: assuming that we are conditioning on Z, an unblocked path from D to Y_d in N_S^+ implies that ISAC is violated in G_S^+ relative to D and Y. We will proceed in the following manner:

1. Discuss the structure of π, an unblocked path from D to Y_d in N_S^+.
2. Discuss how sample selection relates to π.

3. Discuss a procedure for finding the path π^* in G_S^+ that corresponds to π in N_S^+.

4. Discuss possible cases for π^* in G_S^+ and their relation to ISAC.

[1. The structure of π.] We start by assuming that, assuming we are conditioning on Z, there is an unblocked path from D to Y_d in N_S^+. We are going to call this π. We are also able to assume, without loss of generality, that π intersects D only at the starting point of π. What are we able to say about the structure of π across the two halves of N_S^+, the pre-intervention G_S^+ and the post-intervention $(G_S)_{\overline{D}}$? We start by noticing that the elements of Z can only appear on the pre-intervention G_S^+ side of N_S^+. This is because we can only condition on observed variables; we cannot condition on counterfactual variables, which are not observed. This means that we cannot condition on $D = d$ or any of the descendants of $D = d$ in the post-intervention $(G_S)_{\overline{D}}$ side of N_S^+. As such, as soon as π finds it way to the post-intervention side of N_S^+, the remainder of π connecting to Y_d can only contain post-intervention variables, non of which are conditioned on. Moreover, this portion of π in $(G_S)_{\overline{D}}$ can contain only edges pointing toward Y_d. This clarifies that π must be made up of first an unblocked path in the pre-intervention side, G_S^+, that we will label π_1. Next π contains one edge that points from some node in G_S^+ to some node in $(G_S)_{\overline{D}}$, which we label π_2. Recall that we are dealing with graphs in which all bidirected edges have been replaced. We will also see in the next section that π_2 cannot be a bridge. This means that the only type of edge that could connect π_1, which is entirely made up of pre-intervention nodes, to the post-intervention side is a directed edge from the pre-intervention side to the post-intervention side. An edge pointing the other direction would mean that some variables on π_1 are actually post-intervention, a contradiction. Finally, π contains the path we previously discussed, namely, a causal path that contains only descendants of $D = d$ in $(G_S)_{\overline{D}}$ that ends with Y_d. So π is composed of π_1, π_2, π_3. Since N_S^+ is built from G_S^+ and $(G_S)_{\overline{D}}$, π may contain two node "copies" that refer to the same node in G_S^+.

[2. Sample selection and π.] How does sample selection relate to π? Sample selection means we condition on $S = 1$. This is a pre-intervention variable. No post-intervention variable can be an ancestor of the pre-intervention version of S, otherwise we would be considering a post-intervention version of S. So all ancestors of the pre-intervention S are also pre-intervention variables. Therefore, all bridges in N_S^+ appear in the pre-intervention side of the graph, G_S^+, since we've assumed that we've replaced bidirected edges with U^*'s with uni-directional edges that point to the nodes that the bidirected edge had pointed to. Hence, any bridge on π will be in π_1. Since we must condition on the pre-intervention S, any path on which the pre-intervention S appears and is not a collider is blocked and so cannot be π. Also, any path on which the pre-intervention S appears and is a collider (or for which S is a descendant of a collider on the path) will correspond to a generalized non-causal path that is identical to the original path except that the collider is not on the generalized non-causal path and the parents of the collider are connected by a bridge on the generalized non-causal path. If the generalized non-causal path is not blocked then the original path will also not be blocked; if the generalized non-causal path is blocked then so is the original path. Therefore, we can limit our analysis to such generalized non-causal paths. So we consider π that do not contain the pre-interventional S, though π may contain bridges in π_1. Since we have assumed that sample selection is not a mediator or a descendant of a mediator, post-intervention versions of S will not appear on π_3 if they exist at all in N_S^+.

[3. Finding the path π^* in G_S^+ that corresponds to π in N_S^+.] How can we find a path in G_S^+ that corresponds to π? We follow a procedure laid out in Shpitser et al. (2010). First, we create π', a route in G_S^+,

in this way:

1. Start by replacing each instance of a post-intervention variable in π with copy of the same node that appears on the pre-intervention side, G_S^+. We carry along the appropriate occurrence number for each of these replaced nodes.

2. Continue by replacing any instances in which the same variable appears twice in a row with only one copy of that variable. Then reduce the occurrence number of this variable by one and also do this for all the variables that follow.

The portions of π' that were created from π_1 and π_3 (portions of π in N_S^+) will also be unblocked since π_1 and π_3 are unblocked. What about the portion of π created from π_2? This will correspond to a set of three nodes where the center node is the one pointed to by π_2, which we know is a directed edge pointing to some post-intervention node, from the above discussion. The second edge in this triple must be pointing away from the middle node, since all edges in π_3 point toward Y_d, and also must be part of a causal path from D to Y in G_S^+ since the node came from the post-intervention side. But conditioning on nodes on causal paths from D to Y constitutes a violation of ISAC, so we cannot condition on the center node without violating ISAC. Therefore, this last portion of π' is also unblocked, if it exists (it may not if there are no edges in π_3). For example, say that G_S^+ contains $D \to Y$ and $D \to Z \to Y$ and sample selection is not connect to any other node. Then suppose that π is taken to be $D \to Z \leftarrow U_Z \to Z_d \to Y_d$. Here π_1 is $D \to Z \leftarrow U_Z$, π_2 is the edge between U_Z and Z_d, and π_3 is $Z_d \to Y_d$. So π' is $D \to Z \leftarrow U_Z \to Z \to Y$. The node triple in π' that does not correspond to π_1 or π_3 is $U_Z \to Z \to Y$. This is blocked since we condition on Z. However, conditioning on Z is a violation of ISAC since Z lies on a causal path from D to Y in G_S^+. All blocked versions of the node triple in π' that does not correspond to π_1 or π_3 must also violate ISAC for similar reasons. Since the middle node in this node triple is pointed to by π_2, the middle node must be a post-intervention node and so it must lie on a causal path from D to Y in G_S^+, and conditioning on it violates ISAC. Either the node triple is unblocked or it isn't. But, if it isn't, then it could only have resulted from a violation of ISAC. So π' is an unblocked route. By Lemma B.2, π^*, the direct route of π' in G_S^+, is an unblocked path in G_S^+.

[4. Possible cases for π^* in G_S^+ and their relation to ISAC.]

So what are the types of π^* we might see and how do these relate to ISAC?

- The easy case is when π^* is a generalized non-causal path. Here we violate ISAC since Z does not block every generalized non-causal path between D and Y that does not pass through S. Note that any time π contains a bridge, π^* will also contain a bridge and π^* will be a generalized non-causal path.

- The case where π^* is a causal path is somewhat more involved. We again assume that π^* is a proper causal path without loss of generality. The first edge in a π that would create a π^* that is causal would have to be an edge pointing away from some element of D. As we've discussed, π_2 must have been directed in π and pointed to a post-intervention node in $(G_S)_{\overline{D}}$ from a pre-intervention node in G_S^+. The pre-intervention node could not have been a descendant of D, otherwise it would be in the $(G_S)_{\overline{D}}$ part of N_S^+.

  - If there are no node copies that are in both π_1 and π_3 (meaning π_1 has a pre-intervention copy and π_3 has a post-intervention copy of the same node), then π^* cannot be a proper causal path from D to Y in G_S^+. The only way it could be would be for the pre-intervention node to be a descendant of D, a contradiction.

  - If there are node copies that are in both π_1 and π_3, then the only way to reach the pre-intervention

node from D is via a collider unblocked by our conditioning on some element of Z. This would mean that the second node in π (and the second node in π^*) is an ancestor of Z, which violates ISAC.

□

Lemma B.4. *If Z d-separates D and Y_d in N_S^+, then $\{Z, S\}$ d-separates D and Y_d in N_S.*

Proof. We very closely follow the structure of the proof of Lemma 3 in the Web Appendix for Daniel et al. (2012). We start by supposing that the statement that "If Z d-separates D and Y_d in N_S^+, then $\{Z, S\}$ d-separates D and Y_d in N_S." is false. This means that, although all paths from D to Y_d in N_S^+ are blocked by Z, we can find a N_S and a Z for which there is a path, ξ, in N_S from D to Y_d that is not blocked by $\{Z, S\}$. The path ξ is also in N_S^+ since N_S^+ is N_S but with edges added. No edges are removed in extending N_S to N_S^+. If ξ is blocked after conditioning on Z in N_S^+ but is unblocked after conditioning on $\{Z, S\}$ in N_S, then either

- There must be a variable on the path ξ that is not in the set $\{Z, S\}$ but the variable is a in Z. In this way, this variable does not block ξ in N_S but does block ξ in N_S^+. But since Z is in $\{Z, S\}$, this is a contradiction.

- There must be a collider on ξ that satisfies both of the following conditions. The collider is not in Z and does not have descendants in Z. The collider is in $\{Z, S\}$ or has descendants in $\{Z, S\}$. In this way, ξ is blocked in N_S^+ and ξ is not blocked in N_S. Clearly, the collider is S or an ancestor of S. But we can see that ξ is identical to a generalized non-causal path, ξ', in N_S^+ with the exception that the immediate parents of the collider have a bridge between them on ξ' and the collider does not appear on ξ'. If ξ is unblocked in N_S then ξ' must be unblocked in N_S^+. This is because none of the variables along ξ is in $\{Z, S\}$ possibly with the exception of the collider. If they were, then ξ would be blocked. So ξ' must be unblocked in N_S^+, a contradiction.

□

Lemma B.5. *If Z d-separates D and Y in G_S^+, then $\{Z, S\}$ d-separates D and Y in G_S.*

Proof. The same argument as in Lemma B.4 proves this result. □

Lemma B.6. *If $\{Z, S\}$ d-separates D and Y_d in N_S, then $Y_d \perp\!\!\!\perp D | Z, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma B.1. □

Lemma B.7. *If $Y_d \perp\!\!\!\perp D | Z, S = 1$ for every model inducing G_S, then $p(Y_d | S = 1) = \sum_z p(Y | D = d, Z = z, S = 1) p(Z = z | S = 1)$, for every model inducing G_S.*

Proof.

$$
\begin{aligned}
p(Y_d | S = 1) &= \sum_z p(Y_d | Z = z, S = 1) p(Z = z | S = 1) && \text{by law of iterated expectations} \\
&= \sum_z p(Y_d | D = d, Z = z, S = 1) p(Z = z | S = 1) && \text{by } Y_d \perp\!\!\!\perp D | Z, S = 1 \\
&= \sum_z p(Y | D = d, Z = z, S = 1) p(Z = z | S = 1) && \text{by consistency}
\end{aligned}
$$

□

B.3.2 Selection as a mediator

Lemma B.8. *If Z satisfies ISAC in G_S^+ relative to D and Y and S is a mediator between D and Y (but S is not also a descendant of another mediator), then Z d-separates D and $Y_{d,S=1}$ in N_S^+.*

Proof. We rely on the proof of Lemma B.3, but with some caveats for the changes due to S being a mediator between D and Y. See that proof for how the objects in this proof are defined. We consider whether there are new types of open paths π that could connect D to $Y_{d,S=1}$ in N_S^+ and if they exist whether they violate ISAC.

[π **that contain a copy of** S.] In the extended twin network for $Y_{d,S=1}$, any path along which S appears on the post-intervention side of the graph $((G_S)_{\overline{D,S}})$ has been severed since we intervene to set $S = 1$ in addition to setting $D = d$. Again π (the assumed unblocked path from D to $Y_{d,S=1}$ in N_S^+) is constructed from three parts: π_1 (an unblocked path in G_S^+), π_3 (a causal path in $(G_S)_{\overline{D,S}}$ on which every node is a descendant of D and/or S), and π_2 (a single edge connecting π_1 and π_3 in N_S^+). This means that π must land in the post-intervention side on a node that is downstream of $S = 1$ and/or of $D = d$. As in Lemma B.3, due to the construction of N_S^+ we can focus on the generalized non-causal paths that circumvent the pre-interventional S in G_S^+ as candidates for π_1, rather than any such path that passes through the pre-intervention S. Any path on which the pre-intervention S appears where it is not a collider is blocked. So π is not one of these. The paths on the pre-intervention side on which S is a collider correspond to generalized non-causal paths on which S is circumvented and that are blocked whenever the path on which S is a collider is blocked. Due to our additional intervention on $S = 1$, conditioning on the pre-intervention S, and the bridges that we've added, we can conclude that π will not contain either the pre- or post-interventional copy of S.

[π **that do not contain a copy of** S.] Are there any other ways that S as a mediator could add to the cases we need to consider not already covered in Lemma B.3? Conditioning on pre-intervention S could open generalized non-causal paths in the pre-intervention side that contain bridges. But if π contains a bridge, then π^* will also contain a bridge and so it will always be non-causal and hence violate ISAC. (Note that we do not allow elements of Z to appear on causal paths or to be descendants of variables on causal paths, whether or not S is on these paths.) Any remaining unblocked π would be of the types already covered by Lemma B.3 and the same logic from the proof of Lemma B.3 will apply here.

□

Lemma B.9. *If Z d-separates D and $Y_{d,S=1}$ in N_S^+, then $\{Z, S\}$ d-separates D and $Y_{d,S=1}$ in N_S.*

Proof. This proof is similar to that for Lemma B.4. □

Lemma B.10. *If $\{Z, S\}$ d-separates D and $Y_{d,S=1}$ in N_S, then $Y_{d,S=1} \perp\!\!\!\perp D | Z, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma B.1. □

Lemma B.11. *If $Y_{d,S=1} \perp\!\!\!\perp D | Z, S = 1$ for every model inducing G_S, then $p(Y_{D=d,S=1} | S = 1) = \sum_z p(Y | D = d, Z = z, S = 1) p(Z = z | S = 1)$, for every model inducing G_S.*

Proof.

$$p(Y_{d,S=1}|S=1) = \sum_z p(Y_{d,S=1}|Z=z,S=1)p(Z=z|S=1) \qquad \text{by law of iterated expectations}$$

$$= \sum_z p(Y_{d,S=1}|D=d,Z=z,S=1)p(Z=z|S=1) \quad \text{by } Y_{d,S=1} \perp\!\!\!\perp D|Z,S=1$$

$$= \sum_z p(Y|D=d,Z=z,S=1)p(Z=z|S=1) \qquad \text{by consistency}$$

$\square$

Lemma B.12. *If S is a mediator between D and Y, then no set of observed variables, W, can d-separate D and Y_d (or $Y_{d,S_{d'}}$) in N_S^+, unless S is a deterministic function of observed variables. And so no set of variables, W, (along with S) can d-separate D and Y_d (or $Y_{d,S_{d'}}$) in N_S. Hence, the following do not hold for every model inducing G_S: $Y_d \not\perp\!\!\!\perp D|W, S=1$ and $Y_{d,S_{d'}} \not\perp\!\!\!\perp D|W, S=1$.*

Proof. We consider the simplest case in the graphs below. Since we see that we cannot d-separate D and Y_d (or $Y_{d,S_{d'}}$) in N_S^+ in this simplest case, adding more edges and nodes will not change this. The second statement follows from just looking at the twin network contained in the extended twin network here. The last part of the lemma follows from Lemma B.1.

$\square$
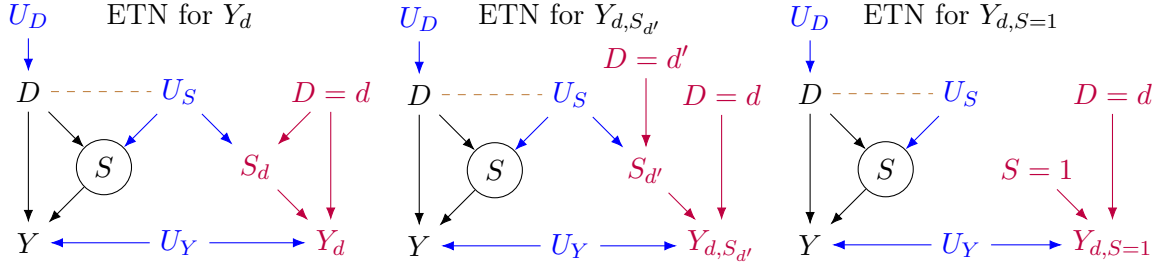
B.3.3 Selection as a descendant of a mediator

Lemma B.13. *If Z satisfies ISAC in G_S^+ relative to D and Y and S is a descendant of a mediator, M, between D and Y (but S is not also a mediator itself), then Z d-separates D and $Y_{d,m}$ in N_S^+.*

Proof. We rely on the proof of Lemma B.3 and Lemma B.8, but with some caveats for the changes due to S being a descendant of a mediator between D and Y. See those proofs for how the quantities in this proof are defined. Note that we are assuming that S is not itself a mediator between D and Y. We consider whether there are new types of open paths π that could connect D to $Y_{d,S=1}$ in N_S^+ and if they exist whether they violate ISAC.

[π **that contain a copy of S or post-intervention M.**] Since S is a descendant of M, a mediator between D and Y in G_S^+, but S is not itself a mediator, in the extended twin network for $Y_{d,m}$, N_S^+, any path on which S or M appear on the post-intervention side of the graph $((G_S)_{\overline{D,M}})$ has been severed since we intervene to set $M=m$ and $D=d$, in addition to the fact that S is not a mediator itself. Again, π (the assumed unblocked path from D to $Y_{d,m}$ in N_S^+) is constructed from three parts: π_1 (an unblocked path in G_S^+), π_3 (a causal path in $(G_S)_{\overline{D,M}}$ on which every node is a descendant of D and/or M), and π_2 (a single

edge connecting π_1 and π_3 in N_S^+). This means that π must land in the post-intervention side on a node that is downstream of $M = m$ and/or of $D = d$. Further, any path containing a post-intervention version of S will not end with $Y_{d,m}$ since S is not a mediator itself. As in Lemmas B.3 and B.8, due to the construction of N_S^+ we can focus on the generalized non-causal paths that circumvent S in G_S^+ as candidates for π_1, rather than any such path that passes through S. So π will not contain either the pre- or post-interventional copy of S. It will also not contain a post-interventional copy of M.

   [π that contain pre-intervention M.] Let's consider the cases in which π might have M on the pre-intervention side. This means that any π in N_S^+ between D and $Y_{d,m}$ that contains M must have M in π_1.

- If M is on a non-causal π^* then we have a violation of ISAC.
- As in Lemma B.3, if M is on a causal π^* then we will also find a violation of ISAC. π_2, again, must have been directed and pointed from a pre-intervention node to a post-intervention node and the pre-intervention node could not have been a descendant of M, otherwise it would be in the $((G_S)_{\overline{D,M}})$ part of N_S^+. M would be on π_1 and the first edge out of M must be an edge pointing away from M, since M only appears on π_1 and π^* is causal. If there are no node copies that are in both π_1 and π_3 (meaning π_1 has a pre-intervention copy and π_3 has a post-intervention copy of the same node), then π^* cannot be a proper causal path from D to Y in G_S^+. The only way it could be would be for the pre-intervention node to be a descendant of M, a contradiction. If there are node copies that are in both π_1 and π_3, then the only way to reach the pre-intervention node from M is via a collider unblocked by our conditioning on some element of Z. This would mean that the second node after M in π (and the second node after M in π^*) is an ancestor of Z, which violates ISAC.

So any π on which M appears corresponds to a π^* that violates ISAC.

Due to the above discussion, any generalized non-causal path on which M is an ancestor of Y in G_S^+ does not need to be blocked by Z. This is because such paths cannot correspond to paths from D to $Y_{d,m}$ in N_S^+ unless we condition on a Z that violates the fist condition of ISAC. This is because these paths point to the pre-interventional Y and the post-interventional version of these paths is severed by our intervention setting $M = m$.

Any remaining unblocked π would be of the types already covered by Lemmas B.3 and B.8 and the same logic from the proofs of Lemmas B.3 and B.8 will apply here. Note that conditioning on M (that is including it in Z) or variables that are on the same causal path that M is on are prohibited by ISAC. $\qquad\square$

Lemma B.14. *If Z d-separates D and $Y_{d,m}$ in N_S^+, then $\{Z, S\}$ d-separates D and $Y_{d,m}$ in N_S.*

Proof. This proof is similar to that for Lemma B.4. $\qquad\square$

Lemma B.15. *If $\{Z, S\}$ d-separates D and $Y_{d,m}$ in N_S, then $Y_{d,m} \perp\!\!\!\perp D | Z, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma B.1. $\qquad\square$

Lemma B.16. *If Z satisfies ISAC in G_S^+ relative to D and Y and S is a descendant of a mediator, M, between D and Y (but S is not also a mediator itself), then $\{Z, D\}$ d-separates M and $Y_{d,m}$ in N_S^+.*

Proof. We will again show the contrapositive: assuming we are conditioning on Z and D, we show that any unblocked path from M to $Y_{d,m}$ in N_S^+, ϕ, means there is an unblocked path from D to $Y_{d,m}$ in N_S^+ (where we only condition on Z), π, which implies that ISAC is violated in G_S^+ relative to D and Y based on Lemma B.13.

We can connect any ϕ that starts with an arrow into M (i.e., $M \leftarrow \ldots Y_{d,m}$ or $M \leftrightarrow \ldots Y_{d,m}$) to $D \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow M$ to create π since M in this case is a collider and we condition on S, a descendant of M; and we cannot condition on any node like W on $D \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow M$ since W would also be on causal paths from D to Y, since M is a mediator, which would be a violation of ISAC. Note that this also holds when M is a direct descendant of D: $D \rightarrow M$ can be connected to ϕ and since M is a collider between D and some variable on ϕ, there is an open π. We can similarly connect any ϕ that starts with a bridge touching M (i.e., $M \cdots \ldots Y_{d,m}$) can be connected to $D \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow M$ to create π since bridges do not create colliders. Any path with an arrow pointing out of M (i.e., $M \rightarrow \ldots$) can only end at $Y_{d,m}$ after traversing a bridge or a collider that has been conditioned on, since M is a pre-intervention node but $Y_{d,m}$ is a descendant of the intervention $M = m$ not of M. In both of these cases, this path can also be connected with $D \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow M$ to create π. The key is that ϕ is an open path from M to $Y_{d,m}$ and so it can be linked to the causal path from D to M (which cannot be blocked without violating ISAC). Any such unblocked path π from D to $Y_{d,m}$ in N_S^+ (where we only condition on Z) violates ISAC following Lemma B.13.

Conditioning on D blocks any open paths between D and $Y_{d,m}$ that could be connected to the causal path from D to M to create an open path between M and $Y_{d,m}$. Could conditioning on D create an open ϕ from M to $Y_{d,m}$ without there also being an open π between D and $Y_{d,m}$? This could only happen if conditioning on D opened a previously closed path, otherwise, no additional paths are opened by conditioning on D (in addition to Z) and we're in the situation above where any ϕ between M and $Y_{d,m}$ can be used to create a π between D and $Y_{d,m}$. Conditioning on D (in addition to Z) would only open paths between M and $Y_{d,m}$ if D is a collider on such paths. This would mean that there would have to be an open path between D and M that ends with an arrow into D as well as an open path between D and $Y_{d,m}$ that ends with an arrow into D. But any open path between D and $Y_{d,m}$ that ends with an arrow into D is an open π that, as we've seen, would violate ISAC. Note that conditioning on M (that is including it in Z) or variables that are on the same causal path that M is on are prohibited by ISAC.

$\square$

Lemma B.17. *If $\{Z, D\}$ d-separates M and $Y_{d,m}$ in N_S^+, then $\{Z, D, S\}$ d-separates M and $Y_{d,m}$ in N_S.*

Proof. This proof is similar to that for Lemma B.4. $\square$

Lemma B.18. *If $\{Z, D, S\}$ d-separates M and $Y_{d,m}$ in N_S, then $Y_{d,m} \perp\!\!\!\perp M | D, Z, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma B.1. $\square$

Lemma B.19. *If $Y_{d,m} \perp\!\!\!\perp D | Z, S = 1$ and $Y_{d,m} \perp\!\!\!\perp M | D, Z, S = 1$ for every model inducing G_S, then $p(Y_{d,m} | S = 1) = \sum_z p(Y | D = d, M = m, Z = z, S = 1) p(Z = z | S = 1)$, for every model inducing G_S.*

Proof.

$$p(Y_{d,m}|S=1) = \sum_z p(Y_{d,m}|Z=z, S=1)p(Z=z|S=1) \qquad \text{by law of iterated expectations}$$

$$= \sum_z p(Y_{d,m}|D=d, Z=z, S=1)p(Z=z|S=1) \qquad \text{by } Y_{d,m} \perp\!\!\!\perp D|Z, S=1$$

$$= \sum_z p(Y_{d,m}|D=d, M=m, Z=z, S=1)p(Z=z|S=1) \quad \text{by } Y_{d,m} \perp\!\!\!\perp M|D, Z, S=1$$

$$= \sum_z p(Y|D=d, M=m, Z=z, S=1)p(Z=z|S=1) \qquad \text{by consistency}$$

$\square$

Lemma B.20. *If S is a descendant of a mediator, M, between D and Y, then no set of observed variables, W, can d-separate D and Y_d (or $Y_{d,M_{d'}}$) in N_S^+, unless S is a deterministic function of observed variables. And so no set of variables, W, (along with S) can d-separate D and Y_d (or $Y_{d,M_{d'}}$) in N_S. Hence, the following do not hold for every model inducing G_S: $Y_d \not\perp\!\!\!\perp D|W, S=1$ and $Y_{d,S_{d'}} \not\perp\!\!\!\perp D|W, S=1$.*

Proof. We consider the simplest case in the graphs below. Since we see that we cannot d-separate D and Y_d (or $Y_{d,M_{d'}}$) in N_S^+ in this simplest case, adding more edges and nodes will not change this. The second statement follows from just looking at the twin network contained in the extended twin network here. The last part of the lemma follows from Lemma B.1.

$\square$

B.3.4   Generalization

Lemma B.21. *If a set of nodes Z in G_S^+ satisfies GC relative to D (treatment) and Y (outcome) and S is not a mediator or descendant of a mediator between D and Y, then Z_{Ext} d-separates S and Y_d in N_S^+.*

Proof. We take a similar approach as in the proof of Lemma B.3. We will show the contrapositive: assuming that we are conditioning on Z, which includes Z_{Ext}, an unblocked path from S to Y_d in N_S^+ implies that GC is violated in G_S^+ relative to D and Y.

- We define π like in Lemma B.3: We start by assuming that, assuming we are conditioning on Z, there is an unblocked path, π, from S to Y_d in N_S^+. Like in Lemma B.3, we can consider π that do not contain the pre-interventional S, though π may contain bridges in π_1. Since we have assumed that sample selection is not a mediator or a descendant of a mediator, post-intervention versions of S will not appear on π_3 if they exist at all in N_S^+.

- As in in Lemma B.3, we can create a π' that is an unblocked route in G_S^+ and π^*, the direct route of π' in G_S^+, is an unblocked path in G_S^+.
- Now, let's consider the types of paths that π^* could be.
  - π^* that do not end in a causal path from D to Y violate of GC.
  - For π^* that do end in a causal paths from D to Y, we want to show that these would imply contradictions for π or violate GC.
    * How can we have D on π^*? This could only result from the pre-intervention side copy of D appearing on π, since no edges point into $D = d$, no bridges can connect to $D = d$, and π must land downstream of $D = d$ on the post-intervention side.
    * So how can the pre-intervention side copy of D appear on π? This could only result from a non-causal or causal path from S to D, where both are on the pre-intervention side. So this non-causal or causal path from S to D is entirely on the pre-intervention side and, therefore, is part of π_1.
    * Then the question becomes: how can we get a causal path from D to Y in π^*, when D is pre-intervention and Y_d is post-intervention on π? From here we can follow the logic in the section bullet of part 4. of the proof of Lemma B.3, which shows us that this either results in a contradiction or violates GC since it violates ISAC.

$\square$

Lemma B.22. *If Z_{Ext} d-separates S and Y_d in N_S^+, then Z_{Ext} d-separates S and Y_d in N_S.*

Proof. We start by supposing that the statement "If Z_{Ext} d-separates S and Y_d in N_S^+, then Z_{Ext} d-separates S and Y_d in N_S." is false. This means that although all paths from S to Y_d in N_S^+ are blocked by Z_{Ext}, we can find a N_S and a Z_{Ext} for which there is a path, ξ, in N_S from S to Y_d that is not blocked by Z_{Ext}. Without loss of generality, we assume that ξ only intersects S at the endpoint. N_S^+ is identical to N_S except for N_S^+ contains bridges created as a result of conditioning on S. Bridges are added to N_S between all variables between which S is a collider or an ancestor of S is a collider resulting in N_S^+. If ξ is unblocked in N_S, then ξ traverses no bridges, since bridges do not appear in N_S. The path ξ is also in N_S^+ and is unblocked since N_S^+ is N_S but with edges added and we are conditioning on the same set of nodes, Z_{Ext}, in both. No edges are removed in extending N_S to N_S^+. In this way, N_S^+ "contains" all of N_S. But an unblocked path between S and Y_d in N_S^+ is a contradiction. $\square$

Lemma B.23. *If Z_{Ext} d-separates S and Y_d in N_S, then $Y_d \perp\!\!\!\perp S | Z_{Ext}$ for every model inducing G_S.*

Proof. This follows from Lemma B.1. $\square$

Lemma B.24. *If $Z_{Ext} \subset Z$ and $Z_{Int} = Z - Z_{Ext}$, $Y_d \perp\!\!\!\perp S | Z_{Ext}$, and $Y_d \perp\!\!\!\perp D | Z, S = 1$ for every model*

inducing G_S, then, for every model inducing G_S,

$$p(Y_d|S = 1) = \sum_z p(Y_d|Z = z, S = 1)p(Z_{Int} = z_{Int}|Z_{Ext} = z_{Ext}, S = 1)p(Z_{Ext} = z_{Ext}|S = 1)$$

$$= \sum_z p(Y|D = d, Z = z, S = 1)p(Z_{Int} = z_{Int}|Z_{Ext} = z_{Ext}, S = 1)p(Z_{Ext} = z_{Ext}|S = 1)$$

$$and \ p(Y_d) = \sum_z p(Y_d|Z = z, S = 1)p(Z_{Int} = z_{Int}|Z_{Ext} = z_{Ext}, S = 1)p(Z_{Ext} = z_{Ext})$$

$$= \sum_z p(Y|D = d, Z = z, S = 1)p(Z_{Int} = z_{Int}|Z_{Ext} = z_{Ext}, S = 1)p(Z_{Ext} = z_{Ext}).$$

Proof.

$$p(Y_d|S = 1) = \sum_z p(Y_d|Z = z, S = 1)p(Z = z|S = 1)$$

$$= \sum_z p(Y_d|D = d, Z = z, S = 1)p(Z = z|S = 1) \text{ by } Y_d \perp\!\!\!\perp D|Z, S = 1$$

$$= \sum_z p(Y|D = d, Z = z, S = 1)p(Z = z|S = 1) \text{ by consistency}$$

$$= \sum_z p(Y|D = d, Z = z, S = 1)p(Z_{\text{Int}} = z_{\text{Int}}|Z_{\text{Ext}} = z_{\text{Ext}}, S = 1)p(Z_{\text{Ext}} = z_{\text{Ext}}|S = 1)$$

$$p(Y_d) = \sum_{z_{\text{Ext}}} p(Y_d|Z_{\text{Ext}} = z_{\text{Ext}})p(Z_{\text{Ext}} = z_{\text{Ext}})$$

$$= \sum_{z_{\text{Ext}}} p(Y_d|Z_{\text{Ext}} = z_{\text{Ext}}, S = 1)p(Z_{\text{Ext}} = z_{\text{Ext}}) \text{ by } Y_d \perp\!\!\!\perp S|Z_{\text{Ext}}$$

$$= \sum_{z_{\text{Ext}}} \left[\sum_{z_{\text{Int}}} p(Y_d, Z_{\text{Int}} = z_{\text{Int}}|Z_{\text{Ext}} = z_{\text{Ext}}, S = 1) \right] p(Z_{\text{Ext}} = z_{\text{Ext}})$$

$$= \sum_{z_{\text{Ext}}} \left[\sum_{z_{\text{Int}}} p(Y_d|Z_{\text{Int}} = z_{\text{Int}}, Z_{\text{Ext}} = z_{\text{Ext}}, S = 1)p(Z_{\text{Int}} = z_{\text{Int}}|Z_{\text{Ext}} = z_{\text{Ext}}, S = 1) \right] p(Z_{\text{Ext}} = z_{\text{Ext}})$$

$$= \sum_z p(Y_d|Z = z, S = 1)p(Z_{\text{Int}} = z_{\text{Int}}|Z_{\text{Ext}} = z_{\text{Ext}}, S = 1)p(Z_{\text{Ext}} = z_{\text{Ext}})$$

$$= \sum_z p(Y_d|D = d, Z = z, S = 1)p(Z_{\text{Int}} = z_{\text{Int}}|Z_{\text{Ext}} = z_{\text{Ext}}, S = 1)p(Z_{\text{Ext}} = z_{\text{Ext}}) \text{ by } Y_d \perp\!\!\!\perp D|Z, S = 1$$

$$= \sum_z p(Y|D = d, Z = z, S = 1)p(Z_{\text{Int}} = z_{\text{Int}}|Z_{\text{Ext}} = z_{\text{Ext}}, S = 1)p(Z_{\text{Ext}} = z_{\text{Ext}}) \text{ by consistency}$$

$\square$

B.4 Theorems

Theorem B.1. *If a set of nodes Z in internal selection graph G_S^+ satisfies ISAC relative to D (treatment) and Y (outcome) and S is not a mediator or descendant of a mediator between D and Y, then, for every model inducing G_S, $Y_d \perp\!\!\!\perp D|Z, S = 1$ and we can then identify $p(Y_d|S = 1) = \sum_z p(Y|d, z, S = 1)p(z|S = 1)$, all of which is estimable from the selected sample alone.*

Proof. Lemmas B.3, B.4, B.6, and B.7 prove the result. $\square$

Theorem B.2. *If a set of nodes Z in internal selection graph G_S^+ satisfies ISAC relative to D (treatment) and Y (outcome) and S is a mediator between D and Y (but S is not also a descendant of another mediator), then, for every model inducing G_S, $Y_{d,S=1} \perp\!\!\!\perp D|Z, S = 1$ and we can then identify $p(Y_{d,S=1}|S = 1) = \sum_z p(Y|d, z, S = 1)p(z|S = 1)$, all of which is estimable from the selected sample alone. Further note that, for any set of observables, W, $Y_d \not\perp\!\!\!\perp D|W, S = 1$ and $Y_{d,S_{d'}} \not\perp\!\!\!\perp D|W, S = 1$, unless S is a deterministic function of observed variables.*

Proof. Lemmas B.8, B.9, B.10, B.11, and B.12 prove the result. □

Theorem B.3. *If a set of nodes Z in internal selection graph G_S^+ satisfies ISAC to D (treatment) and Y (outcome), where S is a descendant of an observed mediator, M, between D and Y (but S is not also a mediator itself), then, for every model inducing G_S, $Y_{d,m} \perp\!\!\!\perp D|Z, S = 1$ and $Y_{d,m} \perp\!\!\!\perp M|D, Z, S = 1$, where $M = m$ is a value observed in the sample. We can then identify, for every model inducing G_S, $p(Y_{d,m}|S = 1) = \sum_z p(Y|d, m, z, S = 1)p(z|S = 1)$, all of which is estimable from the selected sample alone. Further note that, for any set of observables, W, $Y_d \not\perp\!\!\!\perp D|W, S = 1$ and $Y_{d,M_{d'}} \not\perp\!\!\!\perp D|W, S = 1$, unless S is a deterministic function of observed variables.*

Proof. Lemmas B.13, B.14, B.15, B.16, B.17, B.18, B.19, and B.20 prove the result. □

If S is both a mediator and a descendant of a mediator (M) between D and Y, then we might consider potential outcomes of the form $Y_{d,m,S=1}$ (intervening to set $D = d$, $M = m$, and $S = 1$) and something like these theorems should hold. We do not demonstrate this for brevity. The following remark states that we can sometimes think about the threats that sample selection poses for internal validity as an omitted variable problem.

$*$

The following result concerning generalization translates the results in Correa et al. (2018) to use potential outcomes.

Theorem B.4. *If a set of nodes Z in G_S^+ satisfies GC relative to D (treatment) and Y (outcome) and S is not a mediator or descendant of a mediator between D and Y, then $Y_d \perp\!\!\!\perp D|Z, S = 1$ and $Y_d \perp\!\!\!\perp S|Z_{Ext}$. We can identify $p(Y_d) = \sum_z p(Y|D = d, Z = z, S = 1)p(Z_{Int} = z_{Int}|Z_{Ext} = z_{Ext}, S = 1)p(Z_{Ext} = z_{Ext})$.*

Proof. Theorem B.1 and Lemmas B.21, B.22, B.23, and B.24 prove the result. □

Remark 1. *The causal assumptions invoked by ISAC to identify $p(Y_d|S = 1)$ are a subset of those invoked by GC to identify $p(Y_d)$ using covariate adjustment.*

Proof. This follows immediately from Theorems B.1 and B.4. □

B.5 IPW estimation

Following related discussions in Hernán and Robins (2006); VanderWeele (2009); Correa et al. (2018); Hernán and Robins (2020) and elsewhere, we present IPW estimators for $\mathbb{E}[Y_d|S = 1]$, $\mathbb{E}[Y_{d,S=1}|S = 1]$, and $\mathbb{E}[Y_{d,m}|S = 1]$. These are familiar results but tailored to our internal validity for the selected sample focus.

If $Y_d \perp\!\!\!\perp D|Z, S = 1$ (following an application of ISAC) and we have SUTVA, consistency, and positivity, then we can show that, by the law of large numbers, the IPW estimator

$$\hat{\mu}_d = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i \times \mathbb{1}_{D_i=d}}{\hat{p}(D_i=d|Z_i=z,S=1)}$$

is consistent for $\mathbb{E}[Y_d|S=1]$ when the propensity score model $\hat{p}(D_i=d|Z_i=z,S=1)$ is correctly specified:

$$
\begin{aligned}
\mathbb{E}[Y_d|S=1] &= \sum_y y \times p(Y_d=y|S=1) \\
&= \sum_y \sum_z y \times p(Y=y|D=d,Z=z,S=1)p(Z=z|S=1) \quad \text{by Lemma B.7} \\
&= \sum_y \sum_z y \times \frac{p(Y=y,D=d|Z=z,S=1)}{p(D=d|Z=z,S=1)}p(Z=z|S=1) \\
&= \sum_y \sum_z y \times \frac{p(Y=y,D=d,Z=z|S=1)}{p(D=d|Z=z,S=1)} \\
&= \sum_y \sum_z \sum_d y \times \mathbb{1}_{D=d} \times \frac{p(Y=y,D=d,Z=z|S=1)}{p(D=d|Z=z,S=1)} \\
&= \mathbb{E}\left[\frac{Y \times \mathbb{1}_{D=d}}{p(D=d|Z=z,S=1)}\Big|S=1\right] = \mathbb{E}[\hat{\mu}_d|S=1]
\end{aligned}
$$

If $Y_{d,S=1} \per\!\!\!\perp D|Z,S=1$ and we have SUTVA, consistency, and positivity, then we can show that, by the law of large numbers, the IPW estimator $\hat{\mu}_d$ is consistent for $\mathbb{E}[Y_{d,S=1}|S=1]$ when the propensity score model $\hat{p}(D=d|Z=z,S=1)$ is correctly specified:

$$
\begin{aligned}
\mathbb{E}[Y_{d,S=1}|S=1] &= \sum_y y \times p(Y_{d,S=1}=y|S=1) \\
&= \sum_y \sum_z y \times p(Y=y|D=d,Z=z,S=1)p(Z=z|S=1) \quad \text{by Lemma B.11} \\
&= \sum_y \sum_z y \times \frac{p(Y=y,D=d|Z=z,S=1)}{p(D=d|Z=z,S=1)}p(Z=z|S=1) \\
&= \sum_y \sum_z y \times \frac{p(Y=y,D=d,Z=z|S=1)}{p(D=d|Z=z,S=1)} \\
&= \sum_y \sum_z \sum_d y \times \mathbb{1}_{D=d} \times \frac{p(Y=y,D=d,Z=z|S=1)}{p(D=d|Z=z,S=1)} \\
&= \mathbb{E}\left[\frac{Y \times \mathbb{1}_{D=d}}{p(D=d|Z=z,S=1)}\Big|S=1\right] = \mathbb{E}[\hat{\mu}_d|S=1]
\end{aligned}
$$

If $Y_{d,m} \per\!\!\!\perp D|Z,S=1$ and $Y_{d,m} \per\!\!\!\perp M|D,Z,S=1$ and we have SUTVA, consistency, and positivity, then we can show that, by the law of large numbers, the IPW estimator

$$\hat{\mu}_{d,m} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i \times \mathbb{1}_{M_i=m} \times \mathbb{1}_{D_i=d}}{\hat{p}(M_i=m|D_i=d,Z_i=z,S=1)\hat{p}(D_i=d|Z_i=z,S=1)}$$

is consistent for $\mathbb{E}[Y_{d,m}|S=1]$ when the propensity score model $\hat{p}(D_i=d|Z_i=z,S=1)$ and the mediator model $\hat{p}(M_i=m|D_i=d,Z_i=z,S=1)$ are correctly specified:

$$\mathbb{E}[Y_{d,m}|S=1] = \sum_y y \times p(Y_{d,m}=y|S=1)$$

$$= \sum_y \sum_z y \times p(Y=y|D=d,M=m,Z=z,S=1)p(Z=z|S=1) \text{ by Lemma B.19}$$

$$= \sum_y \sum_z y \times \frac{p(Y=y,M=m|D=d,Z=z,S=1)}{p(M=m|D=d,Z=z,S=1)}p(Z=z|S=1)$$

$$= \sum_y \sum_z y \times \frac{p(Y=y,M=m|D=d,Z=z,S=1)}{p(M=m|D=d,Z=z,S=1)}\frac{p(D=d|Z=z,S=1)}{p(D=d|Z=z,S=1)}p(Z=z|S=1)$$

$$= \sum_y \sum_z y \times \frac{p(Y=y,M=m,D=d|Z=z,S=1)}{p(M=m|D=d,Z=z,S=1)p(D=d|Z=z,S=1)}p(Z=z|S=1)$$

$$= \sum_y \sum_z y \times \frac{p(Y=y,M=m,D=d,Z=z|S=1)}{p(M=m|D=d,Z=z,S=1)p(D=d|Z=z,S=1)}$$

$$= \sum_y \sum_z \sum_m \sum_d y \times \mathbb{1}_{M=m} \times \mathbb{1}_{D=d} \times \frac{p(Y=y,M=m,D=d,Z=z|S=1)}{p(M=m|D=d,Z=z,S=1)p(D=d|Z=z,S=1)}$$

$$= \mathbb{E}\left[\frac{Y \times \mathbb{1}_{M=m} \times \mathbb{1}_{D=d}}{p(M=m|D=d,Z=z,S=1)p(D=d|Z=z,S=1)}\middle|S=1\right]$$

$$= \mathbb{E}[\hat{\mu}_{d,m}|S=1]$$

B.6 Conditioning on a collider

Shahar and Shahar (2017) discuss the conditions under which an association is created between the parents of a collider when the collider is conditioned on for discrete variables. They show that, "[i]f [D] and [U] are marginally independent causes of [S], then [D] and [U] are dependent conditional on [$S=1$] if and only if [D] and [U] modify each other's effects on [$S=1$]." These authors define effects on the collider as well as effect modification in terms of probability ratios. Section B.7 of Appendix B further shows that non-zero interaction information is a requirement for dependency to be created between two marginally independent parents of a collider, when the collider is conditioned on. Also relevant to the discussion in the present paper, Shahar and Shahar (2017) show that for marginally independent causes D,U of a binary collider S, if the effects of D and U on S are not null, then D,U modify each other's effects in at least one stratum of S and are dependent in at least one stratum of S.

Using our working example, let us consider what would be required for the parents of the sample selection node to remain independent after stratifying to $S=1$. Simplifying the example inspired by Knox et al. (2020), we consider how the effect of police perception of majority vs minority race ($D \in \{\text{majority}, \text{minority}\}$) on police making a stop or not ($S \in \{0,1\}$) might be modified by an unobserved factor that represents police officer stress levels ($U \in \{\text{high}, \text{low}\}$). Following Shahar and Shahar (2017), we consider the probability ratios in (1):

$$\frac{p(S=1|D=\text{majority},U=\text{high})}{p(S=1|D=\text{minority},U=\text{high})} \overset{?}{=} \frac{p(S=1|D=\text{majority},U=\text{low})}{p(S=1|D=\text{minority},U=\text{low})} \tag{1}$$

The left hand side of (1) is the effect of police perception of race on police making a stop when police

officers are under high stress. The right hand side of (1) is the effect of police perception of race on police making a stop when police officers are under low stress. Police officer stress not modifying the effect of police perception of race on police making a stop would mean that the left and right hand sides of (1) are equal. Under what circumstances would this occur?

The simplest scenario would be when people perceived to be from the majority racial group are never stopped. That is, the numerators on both sides of (1) are zero: $p(S = 1|D = \text{majority}, U = \text{high}) = 0$ and $p(S = 1|D = \text{majority}, U = \text{low}) = 0$. This would make both probability ratios zero, meaning no effect moderation. This is clearly an absurd scenario. The other case would be when the two probability ratios perfectly balance, meaning that there is no effect moderation. This is also not likely; police officer stress levels likely do modify the effect of perceptions of race on making a stop in some way. But if one of these scenarios holds, following Shahar and Shahar (2017)'s result quoted above it can be shown that $p(D = \text{majority}|U = \text{high}, S = 1) = p(D = \text{majority}|S = 1)$. We demonstrate this in full below. That is, police perception of race is independent of police stress levels, despite stratifying to $S = 1$, if one the the two (unrealistic) scenarios holds. Further, if one the the two scenarios holds and so long as $p(S = 1|D = \text{minority}, U = \text{high}) \neq p(S = 1|D = \text{minority}, U = \text{low})$ (i.e., there is an effect from police officer stress), then we would see that the two sides of (2) are not equal, meaning there is effect moderation and hence police perception of race is dependent on police stress levels, stratifying to $S = 0$.

$$\frac{p(S = 0|D = \text{majority}, U = \text{high})}{p(S = 0|D = \text{minority}, U = \text{high})} \overset{?}{=} \frac{p(S = 0|D = \text{majority}, U = \text{low})}{p(S = 0|D = \text{minority}, U = \text{low})} \tag{2}$$

Practitioners would only know that they're in the setting in which stratifying to $S = 1$ does not create association between the parents of S if they have detailed knowledge of the selection mechanism, like the unrealistic scenarios above. In such a situation, even if the parents of the sample selection node are dependent due to selection, the practitioner could use inverse probability of selection weighting to estimate unbiased effects. (Thompson and Arah, 2014) But we emphasize that such knowledge is hard to come by and the assumptions required to fall into such a setting will often be absurd.

We follow Shahar and Shahar (2017) to show that $p(D = \text{majority}|U = \text{high}, S = 1) = p(D = \text{majority}|S = 1)$. In the derivation below, we abbreviate "majority" as "ma" and "minority" as "mi." We rely on the fact that there is no effect moderation to show this. First note that we can write

$$\frac{p(S = 1|D = \text{ma}, U = \text{high})}{p(S = 1|D = \text{mi}, U = \text{high})} = \frac{p(S = 1|D = \text{ma}, U = \text{high})p(U = \text{high}|D = \text{ma})}{p(S = 1|D = \text{mi}, U = \text{high})p(U = \text{high}|D = \text{mi})} = \frac{p(S = 1, U = \text{high}|D = \text{ma})}{p(S = 1, U = \text{high}|D = \text{mi})}$$

Next we show that we can write

$$\frac{p(S = 1|D = \text{ma}, U = \text{high})}{p(S = 1|D = \text{mi}, U = \text{high})}$$
$$= \frac{p(S = 1|D = \text{ma}, U = \text{high})}{p(S = 1|D = \text{mi}, U = \text{high})}\frac{p(S = 1, U = \text{high}|D = \text{mi})}{p(S = 1|D = \text{mi})} + \frac{p(S = 1|D = \text{ma}, U = \text{low})}{p(S = 1|D = \text{mi}, U = \text{low})}\frac{p(S = 1, U = \text{low}|D = \text{mi})}{p(S = 1|D = \text{mi})}$$
$$= \frac{p(S = 1, U = \text{high}|D = \text{ma})}{p(S = 1, U = \text{high}|D = \text{mi})}\frac{p(S = 1, U = \text{high}|D = \text{mi})}{p(S = 1|D = \text{mi})} + \frac{p(S = 1, U = \text{low}|D = \text{ma})}{p(S = 1, U = \text{low}|D = \text{mi})}\frac{p(S = 1, U = \text{low}|D = \text{mi})}{p(S = 1|D = \text{mi})}$$
$$= \frac{p(S = 1, U = \text{high}|D = \text{ma})}{p(S = 1|D = \text{mi})} + \frac{p(S = 1, U = \text{low}|D = \text{ma})}{p(S = 1|D = \text{mi})} = \frac{p(S = 1|D = \text{ma})}{p(S = 1|D = \text{mi})}$$

Finally, we show that

$$p(D = \text{majority}|U = \text{high}, S = 1)$$

$$= \frac{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma}|U = \text{high})}{p(S = 1|U = \text{high})}$$

$$= \frac{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma}|U = \text{high})}{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma}|U = \text{high}) + p(S = 1|D = \text{mi}, U = \text{high})p(D = \text{mi}|U = \text{high})}$$

$$= \frac{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma})}{p(S = 1|D = \text{ma}, U = \text{high})p(D = \text{ma}) + p(S = 1|D = \text{mi}, U = \text{high})p(D = \text{mi})}$$

$$= p(D = \text{ma}) \left[\frac{p(S = 1|D = \text{ma}, U = \text{high})}{p(S = 1|D = \text{ma}, U = \text{high})}p(D = \text{ma}) + \frac{p(S = 1|D = \text{ma}, U = \text{high})}{p(S = 1|D = \text{mi}, U = \text{high})}p(D = \text{mi}) \right]^{-1}$$

$$= p(D = \text{ma}) \left[\frac{p(S = 1|D = \text{ma})}{p(S = 1|D = \text{ma})}p(D = \text{ma}) + \frac{p(S = 1|D = \text{ma})}{p(S = 1|D = \text{mi})}p(D = \text{mi}) \right]^{-1} \quad \text{from above}$$

$$= \frac{p(S = 1|D = \text{ma})p(D = \text{ma})}{p(S = 1|D = \text{ma})p(D = \text{ma}) + p(S = 1|D = \text{mi})p(D = \text{mi})}$$

$$= \frac{p(S = 1|D = \text{ma})p(D = \text{ma})}{p(S = 1)}$$

$$= p(D = \text{majority}|S = 1)$$

B.7 Colliders and interaction information

We draw on discussion of interaction information and colliders from Ghassami and Kiyavash (2017) (also see McGill (1954)) to show that zero interaction information for marginally independent random variables implies continued independence conditional on a collider. Section B.6 of Appendix B further discusses collider stratification in the context of our working example. Suppose we have two random variables X_1, X_2 and variable S that is a collider between X_1, X_2: $X_1 \rightarrow S \leftarrow X_2$. Interaction information is a generalization of mutual information to the case when there are multiple variables. See Cover and Thomas (2006) for an introduction to mutual information. We can write interaction information as (Ghassami and Kiyavash, 2017)

$$\text{MI}(S; X_1; X_2) = \text{MI}(X_1; X_2) - \text{MI}(X_1; X_2|S)$$
$$= \text{MI}(S; X_1) - \text{MI}(S; X_1|X_2)$$
$$= \text{MI}(S; X_2) - \text{MI}(S; X_2|X_1)$$

Using the first expression for interaction information above, we see that $\text{MI}(X_1; X_2|S) = \text{MI}(X_1; X_2) - \text{MI}(S; X_1; X_2)$. It is trivial to see that, $\text{MI}(S; X_1; X_2) = 0$ means that $\text{MI}(X_1; X_2|S) = \text{MI}(X_1; X_2)$. Meaning that zero interaction information implies that there is no change in mutual information between X_1, X_2 when we condition on the collider between them. For marginally independent X_1, X_2, $\text{MI}(X_1; X_2|S) = -\text{MI}(S; X_1; X_2)$ and so if $\text{MI}(S; X_1; X_2) = 0 \implies \text{MI}(X_1; X_2|S) = 0$. Meaning that zero interaction information implies that conditional on the collider, X_1, X_2 remain independent.

Another view of interaction information is possible by writing

$$\text{MI}(S; [X_1, X_2]) = \text{MI}(S; X_2) + \text{MI}(S; X_1|X_2)$$
$$\implies \text{MI}(S; [X_1, X_2]) - \text{MI}(X_1; S) - \text{MI}(X_2; S) = \text{MI}(S; X_2) + \text{MI}(S; X_1|X_2) - \text{MI}(S; X_1) - \text{MI}(S; X_2)$$
$$= -(\text{MI}(S; X_1) - \text{MI}(S; X_1|X_2))$$
$$= -\text{MI}(S; X_1; X_2)$$
$$\implies \text{MI}(X_1; S) + \text{MI}(X_2; S) - \text{MI}(S; [X_1, X_2]) = \text{MI}(S; X_1; X_2)$$

where $\text{MI}(S; [X_1, X_2]) = \text{MI}(S; X_2) + \text{MI}(S; X_1|X_2)$ captures the information that X_1, X_2 jointly share with S. Interaction information would be zero when $\text{MI}(X_1; S) + \text{MI}(X_2; S)$ equals $\text{MI}(S; [X_1, X_2])$. That is, when the information that X_1, X_2 jointly share with S is exactly equal to the the mutual information between X_1 and S added to the mutual information between X_2 and S. This would arise when the information shared between X_1 and S does not overlap with the information shared between X_2 and S.

We include this discussion only to show that conditioning on a collider does not necessarily lead to mutual information (and therefore dependence) between marginally independent parents of the collider or a change in the mutual information between marginally dependent parents of the collider. We've shown that it does not precisely when interaction information is zero.

A few additional notes on interaction information. If X_1 and X_2 are marginally independent causes of S, one might be tempted to think of interaction information as arising through the "interaction" between X_1 and X_2 in determining S, since no shared information existed marginally. However, interaction information can be difficult to interpret and what is meant by this "interaction" is not necessarily what we might expect. See a Krippendorff (2009) for a discussion of these difficulties, including that interaction information can be negative.

The "interaction" is not necessarily something like an interaction term in a linear model. It is easy to show that for Gaussian random variables $X_1 \sim \mathcal{N}(0, 1)$, $X_2 \sim \mathcal{N}(0, 1)$, $\epsilon \sim \mathcal{N}(0, 1)$, and $S = X1 + X2 + \epsilon$, where there is no interaction term $X_1 \times X_2$ in the data generating process for S, there is still non-zero interaction information between X_1, X_2, S. This arises from the simple fact that mutual information for Gaussians has the following relationship with R^2's: $\text{MI}(A; B) = -\frac{1}{2}\log(1 - R_{A,B}^2)$. (Cover and Thomas, 2006) From above, we know that $\text{MI}(S; X_1; X_2) = \text{MI}(X_1; S) + \text{MI}(X_2; S) - \text{MI}(S; [X_1, X_2])$. From the relationship between mutual information and R^2, we have that $\text{MI}(X_1; S) = -\frac{1}{2}\log(1 - R_{X_1,S}^2) = -\frac{1}{2}\log(1 - \frac{1}{3}) \approx 0.203$ and similarly, $\text{MI}(X_2; S) \approx 0.203$. However $\text{MI}(S; [X_1, X_2]) = -\frac{1}{2}\log(1 - R_{S,X_1+X_2}^2) = -\frac{1}{2}\log(1 - \frac{2}{3}) \approx 0.55 \neq 0.406 \approx \text{MI}(X_1; S) + \text{MI}(X_2; S)$. So we see that interaction information is non-zero, this in turn means conditional mutual information is non-zero, i.e. $\text{MI}(X_1; X_2|S) \neq 0$. The partial R^2, $R_{X_1,X_2|S}^2 = \left(\frac{R_{X_1,X_2} - R_{X_1,S}R_{X_2,S}}{\sqrt{1 - R_{X_1,S}^2}\sqrt{1 - R_{X_2,S}^2}}\right)^2 = \frac{R_{X_1,S}^2 R_{X_2,S}^2}{(1 - R_{X_1,S}^2)(1 - R_{X_2,S}^2)} = \frac{(\frac{1}{3})^2}{(1 - \frac{1}{3})^2} = \frac{1}{4}$, is also not zero.

We note two additional interesting cases. First, even if $\text{MI}(X_1; S)$ and $R_{X_1,S}^2$ are zero, we can have non-zero interaction information, conditional mutual information and partial R^2 if the zeroes are due to perfect balancing of the path coefficients (e.g., $X_1 \sim \mathcal{N}(0, 1)$, $\xi \sim \mathcal{N}(0, 1)$, $\epsilon \sim \mathcal{N}(0, 1)$, $X_2 = X_1 + \xi$ and $S = -X_1 + X_2 + \epsilon$). Second, even when $R_{X_1,X_2|S}^2$ and $\text{MI}(X_1; X_2|S)$ are zero, we can have non-zero interaction information. The zero $R_{X_1,X_2|S}^2$ could be due to perfect balancing between non-zero R_{X_1,X_2} and

R_{S,X_1}, and R_{S,X_2}, which coincides with zero $\mathrm{MI}(X_1; X_2|S)$ due to perfect balancing between non-zero mutual information $\mathrm{MI}(X_1; X_2)$ and interaction information $\mathrm{MI}(S; X_1; X_2)$. For example, if the data generating process is $X_1 \sim \mathcal{N}(0,1)$, $\xi \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,1)$, $X_2 = \gamma X_1 + \xi$ and $S = \sqrt{\gamma}X_1 + \sqrt{\gamma}X_2 + \epsilon$. In this example the marginal dependency between X_1 and X_2 is eliminated when we condition on S.

B.8   Edges with unknown direction

What should researchers do when they do not know whether one variable causes a second variable or if it is the other way around? Can this sort of structural uncertainty be allowed? How might we handle it? First, recall that uncertainty about whether two variables are directly causally related at all should captured by including an edge between these variables, since an included edge can represent a large, a small, or even a null relationship but the absence of an edge can only represent a null relationship. In this way, excluding edges from a graph represents strong causal assumptions. Similarly, even when we know whether two variables are directly causally related, we might not know whether there is some variable that is a common direct cause to both. Accordingly, we should include a bidirected edge between the two variables, unless we can rule out common direct causes. Now, consider instances when we cannot rule out that two variables are directly related (regardless of whether they also share a common direct cause) but we do not know which variable causes which. In terms of our causal graph, this means we do not know which direction the direct edge between the two variables should point. As such we will represent this type of structural uncertainty with an *solid* undirected edge (——; associations created due to sample selection are represented with *dashed* undirected edges, which we call bridges).

A graph with such solid undirected edges should be thought of as a set of DAGs. Every DAG in this set shares the same verticies, directed edges, and bidirected edges, but there is a DAG for each possible combination of edge directions for the solid undirected edges.[39] When we are uncertain about the direction of some set of edges in our causal graph, we want to know if there is a set of covariates that might allow us to identify our causal effect of interest in all the DAGs in the set represented by our causal graph. This means that in each of the DAGs we could extend them to be internal selection graphs and the same set of covariates would satisfy ISAC in each of these internal selection graphs. As might be clear, this will be a more demanding requirement than simply finding a covariate set that satisfies ISAC in a single internal selection graph corresponding to a single DAG. Of course, there are situations in which the uncertainty about edge directions limits our ability to identify causal quantities. See 14(a) for a simple example; in only one of the represented DAGs can we identify $p(Y_d|S = 1)$. There are settings in which ambiguity over whether a path is causal or non-causal; and so it is unclear whether variables should be adjusted for or not. See 14(b); in one of the represented DAGs we should adjust for A but not in the other. There are settings in which it is not clear whether we are able to identify $p(Y_d|S = 1)$ or $p(Y_{d,S=1}|S = 1)$. See 14(c); since it is unclear if S is a mediator or not, we do not know what sort of causal quantity we might be able to identify. Multiple of such problems might arise. See 14(d).

Our first suggestion is simple. When it is feasible, researchers should draw all possible DAGs in the set represented by their causal graph with undertain edge directions. These should then be extended into

---

[39]Some combinations of these edge directions might create cycles; these combinations should not be considered since they are not DAGs.

Figure 14: Examples of Edges with Uncertain Direction
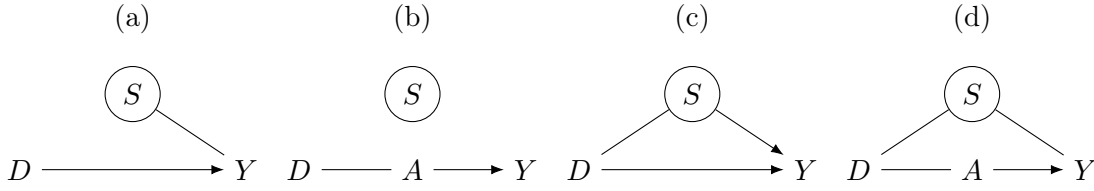
internal selection graphs. They can then evaluate if there is a covariate set that will satisfy ISAC in all of the represented graphs. Drawing all possible graphs will force the researcher to consider the alternatives carefully and explicitly; this will make clear what is and is not identifiable in light of the structural uncertainty. Naturally, this will only be feasible for causal graphs with relatively few edges of unknown direction. As the number of edges of uncertain direction increases, the need for a more abbreviated approach that allows for uncertain edge directions increases.[40] While this is true, it is also likely that, as the number of edges of uncertain direction increases, so does the likelihood that the causal quantitiy of interest will not be possible to identify, since uncertain edge directions are limitations on the causal knowledge with which we imbue the causal model. But we'd still like to have a systematic approach for these settings.

When there are many edges of uncertain direction, we might ask "is there a way to quickly determine that my causal quantity of interest is *not* identifiable?" In this spirit, our second suggestion is that researchers take an adversarial approach to finding a set of covariates that would identify their causal quantity. That is, can we find two internal selection graphs, represented by the causal graph with uncertain edge directions, that have conflicting conclusions about identifiability? Or can we find at least one such internal selection graph in which the causal quantity is not identifiable?[41] To that end, we suggest that researchers look for the following cases in their causal graphs with uncertain edge directions.

1. Sample selection on a possibly causal path:[42] we will not know which of $p(Y_d|S = 1)$ or $p(Y_{d,S=1}|S = 1)$ is identifiable.

2. Possibility that a path could run from Y to D: this means that we don't know if the treatment causes the outcome or the outcome causes the treatment.

3. Causal paths that start with a solid undirected edge out of the treatment: this leads to uncertainty about whether paths are causal (and should not be blocked) or non-causal (and should be blocked).

4. At least one internal selection graph in which identifiability is impossible: see Table 3 for examples. Just one such graph means that we cannot identify the causal quantity of interest.

5. Conflicts between internal selection graphs over which covariate adjustment set would block the non-causal paths: just one such conflict means that we cannot identify the causal quantity of interest. For example, in the graph $D \leftrightarrow A—Y$ with $D \to Y$ (assume S is not causally related to D, A, Y), one

---

[40]There are existing approaches for graphs with uncertain edge directions, but these do not explicitly include sample selection. See Perković et al. (2017).

[41]We urge caution here. Researchers must resist the temptation to assign certain edge directions where evidence and expertise does not allow it. Simply asserting edge directions to have a graph in which the target causal quantity is identifiable does not actually mean this quantity is identifiable in reality. As ever, assumptions must be defended not just asserted.

[42]A possibly causal path is one on which there are directed edges and at least one solid undirected edge, but all directed edges point away from D toward Y. No bidirected edges or bridges are allowed on possibly causal paths.

internal selection graph requires us to adjust for A and the other requires that we do not adjust for A to identify $p(Y_d|S = 1)$.

These will all present problems for identifiability or clear interpretation. If such problems arise, then you know identification or clear interpertability will not be possible. When drawing all represented internal selection graphs is infeasible and we cannot quickly see that identifiablility or clear interpretation is not possible, we might want a formal adjustment criterion for identification of internal causal quantities that allows for uncertain edge directions. Zhang (2008); Maathuis and Colombo (2015); Perković et al. (2017) provide formal criteria for dealing with unknown edge directions but without explicitly considering sample selection.[43] In practice, we believe that the two suggestions above will serve the needs of many researchers.

---

[43]A formal criterion that allows for uncertain edge directions, in addition to explicitly incorporating sample selection, is beyond the scope of this paper, but would be an interesting direction for future work.