Instrumental Variables with Sample Selection: Opportunities, Threats, and Graphs [DRAFT]

Adam Rohde*, Chad Hazlett†

December 2022

Abstract

In this paper, we discuss interesting cases in which sample selection presents opportunities to use instrumental variables and in which using instrumental variables can be used to overcome sample selection. However, we discuss how these opportunities may arise only in very specific settings, as is the case for credible instrumental variables in general. We also discuss the numerous threats that sample selection can pose to the credibility of instrumental variable approaches. To facilitate this discussion, we revise existing graphical criteria for instrumental variables to highlight the special role that sample selection plays in the instrumental variables setting. We do this by first introducing an extension to typical causal graphs that visualizes how sample selection alters the relationships between variables in the sample. We then provide rules (graphical criteria) that allow researchers to use these extended graphs to evaluate the key assumptions of instrumental variables in their own applications, while responsibly accounting for sample selection. In this way, we generalize recent discussions of sample selection and instrumental variables and connect these to existing graphical criteria. Moreover, we emphasize the importance of including a sample selection node in all causal graphs, warn of the dangers of applying simple heuristics about sample selection, and suggest that practitioners appeal to formal approaches like those we present to understand how sample selection may threaten the validity of instrumental variables in their specific applications.

1 Introduction

When researchers are interested in the causal effect of a treatment (D) on an outcome (Y), for example whether increased trust in government increases support for redistributive policies, they often appeal to a sample of data drawn in some selective way from a larger population. Whether or not an estimated effect obtained from a sample drawn in some non-random way is an unbiased estimate of the causal effect even averaged over the members of the selected group is traditionally referred to as "internal validity" (Campbell, 1957; Campbell and Stanley, 1966; Cook and Campbell, 1979; Shadish et al., 2002). This property is distinct from questions that compare the group in hand to other possible populations ("external validity"). In many studies, a potential threat to internal validity and external validity is the presence of unobserved common causes of both the treatment and the outcome, called confounders. Perhaps trust in government and support for redistributive policies are influenced not only by political party membership and similar measurable factors but also by difficult to measure factors like interest in government, anti-social tendencies, and others. If all such confounding variables have been measured and their effects can be adjusted for, then researchers might be able to attain internal validity using a "no unobserved confounders" type approach. We leave discussions of external validity to other authors.[1] When the ability for researchers to directly adjust for all plausible confounders of this type is limited, then researchers might appeal to an alternative strategies for achieving internal validity. One such strategy is instrumental variables.

The instrumental variables ("IV") identification strategy attempts to leverage the variation in a variable that is associated with the treatment but not directly with the outcome (this variable is called the instrument) to try to understand the causal relationship between the treatment and the outcome. (Imbens and Angrist, 1994; Angrist et al., 1996; Hernán and Robins, 2006; Pearl, 2001, 2009; Baiocchi et al., 2014; Hernán and Robins, 2020) In it's simplest form, this boils down to looking at how the outcome and the instrument are associated and how the treatment and the instrument are associated and then trying to use these components to get at how the treatment and outcome are causally related. To go from associations between the instrument and outcome and the instrument and treatment to a causal relationship between the treatment and outcome requires restrictions on the causal relationships between the three variables. The specific restrictions are discussed in detail

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

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

[1] We agree with Imbens (2014a,b) in putting internal validity before external validity: "Nonetheless, in general the subpopulation of compliers is not chosen for its interest, but because we can hope to learn something about them. It is about the primacy of internal validity over external validity (Shadish, Cook and Campbell, 2002)."

in subsequent sections but at a minimum these include "ignorability," "relevance," and an a third assumption that can take various forms depending on what the researchers believe is plausible (examples are effect homogeniety, monotonicity, and one-sided non-compliance). These restrictions are often demanding and do not hold in many cases. But even when these restrictions do appear to be met, we must also consider how studying a non-random sample of units could threaten the internal validity of an instrumental variables approach.[2][3]

Sample selection can arise at various points in 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. The manner in which the sample was selected can have dramatically different implications for the validity of an instrumental variables design than it does for a simple covariate adjustment approach. Further, we cannot credibly rule out threats to internal validity from sample select for instrumental variables (as is true of other designs) without clearly laying out how sample selection fits within the causal model. Echoing Berk, 1983 and Greenland, 2022, we stress 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.

We are by no means the first authors to point out that sample selection can violate the assumptions of instrumental variables approaches. Canan et al. (2017); Swanson et al. (2015); Swanson (2019); Hughes et al. (2019); Ertefaie et al. (2016); Gkatzionis and Burgess (2018); Hernán and Robins (2020); Elwert and Segarra (2022) all discuss sample selection and instrumental variables. Canan et al. (2017) discusses how one form of sample selection can violate IV assumptions. Hughes et al. (2019) tries to provide a more comprehensive view into how sample selection can violate IV assumptions. They provide several examples, run simulation studies, and provide some guidance and description on the reason violations arise in their examples. They do not provide guidance for an arbitrary causal graph and sample selection mechanism. In their analysis, they do not graphically represent how sample selection alters the relationships between variables and they provide only heuristics not a complete set of rules for dealing with sample selection in the context of instrumental variables. We discuss their examples in detail in a subsequent section. Swanson (2019) discusses broad questions about how sample selection can bias IV studies and provide guidance for applied researchers. However, they do not provide anything systematic for an arbitrary causal model and sample selection mechanism. They do discuss threats posed by sample selection problems that are unique to IV (e.g., selection on the treatment can be a problem for IV whereas it is not typically for a simple covariate adjustment approach). They also mention that not all types of sample selection that might threaten internal validity in the context of other designs threaten internal validity for instrumental variables. Hernán and Robins (2020) in their discussion of instrumental variables mention that sample selection can violate instrumental variables assumptions and also briefly mention some interesting cases that we analyze further in this paper. There are various papers that focus on applications that explore specific examples. Of particular interest is where researchers are interested in comparing two treatment levels and select only units that receive either of these treatment levels, but more than two treatment levels exist. (Swanson et al., 2015) Sheehan et al. (2008) provides good examples of proxy and confounded instruments, as does Hernán and Robins (2020). Swanson and Hernán (2013) suggest a checklist for reporting IV conditions and results. This does not include careful consideration of the sample selection mechanism.

Van Der Zander et al. (2015); Van Der Zander and Liśkiewicz (2016); Kumor et al. (2020) present graphical approaches to finding instruments and their generalizations, but do not focus on how sample selection relates to these. Galles and Pearl (1998); Pearl (2001, 2009); Elwert and Segarra (2022) discuss conditional instruments and graphical criteria for them of the sort we will discuss. While Pearl (2001); Elwert and Segarra (2022) focus primarily on linear models, Galles and Pearl (1998); Pearl (2009) discuss potential outcomes as we do. While our approach could be thought of as asking whether a variable is a conditional instrument (defined by these graphical criteria) where we include S in the conditioning set (the approach used in Elwert and Segarra (2022)), we believe it is instructive and clarifying to treat sample selection as a special variable (since it is the only variable we don't observe but must condition on) and to graphically show how sample selection alters the relationships between variables. Further, it is reasonable for a user to wonder whether sample selection can safely be treated like other variables or not and what this means with respect to internal validity. We also provide graphical criteria that allow us to obtain independence between potential outcomes and the instrument and to obtain relevance of the instrument to the treatment,[4] but we structure these criteria to highlight the role that sample selection is playing, rather than assuming it is clear to the user how sample selection fits into the criteria. Further, Elwert and Segarra (2022) focus on sample selection resulting from conditioning on a descendant of the treatment; we extend our analysis beyond such cases and provide clear graphical guidance on when instrumental variables can be used for any sample selection mechanism. Elwert and Segarra (2022)

[2]The threats that sample selection poses for internal validity in simple covariate adjustment identification strategies are discussed in Rohde and Hazlett (20XX).

[3]It turns out that relevance and ignorability alone are enough to obtain bounds on causal effects. See Pearl (2009) chapter 8 and Balke and Pearl (1994a). We do not discuss this further here, but when such bounds are sufficient no "third" assumption is required.

[4]The graphical criterion that Pearl (2009) refers to originates in Galles and Pearl (1998). This graphical criterion is stated as "every path connecting $[IV]$ to Y must pass through $[D]$, unless it contains arrows pointing head-to-head" (Galles and Pearl, 1998) or "every unblocked path connecting $[IV]$ and Y must contain an arrow pointing into $[D]$" ($(Y \perp\!\!\!\perp IV|X)_{G_{\overline{X}}}$) (Pearl, 2009).

provide exact expressions for sample selection bias in linear models under a few important sample selection mechanisms.

In light of the above discussion, our main contribution is two fold. We restate the graphical criterion from Galles and Pearl (1998); Pearl (2001, 2009) so as to highlight the special role that the sample selection variable plays. As a result of this, we provide a comprehensive guide to how sample selection can threaten the internal validity of instrumental variables approaches. This generalizes the discussions of sample selection as a threat to internal validity found in Canan et al. (2017); Hughes et al. (2019); Swanson (2019); Elwert and Segarra (2022). In this way, we bridge the gap between the graphical criteria for IV that do not discuss sample selection with recent discussions of sample selection and IV. Additionally, we ease analysis by introducing extended causal graphs of the type discussed in Daniel et al. (2012) and Rohde and Hazlett (20XX) that show how sample selection alters the relationships between variables in the sample, we advocate and illustrate the importance of including a sample selection node in all causal graphs and of using formal graphical criteria rather than simple heuristics, and we explore many interesting implications of sample selection for instrumental variables evaluating both threats and opportunities. Our graphical approach emphasizes the potential for many purely statistical relationships to be created by sample selection in an instrumental variables context that are typically not obvious in regular causal graphs. We do not focus on the additional assumptions that lead to identification results for instrumental variables, as these are typically not represented in causal graphs and of which there are many. In Appendix XX, we also illustrate some examples of how identification can proceed in the selected sample, given some additional assumption like monotonicity or one-sided non-compliance. On these points, we believe we provide clarity that does not already exist in the literature.

2 Background, notation, and key quantities

We are interested in the causal effect of a treatment, D, on an outcome, Y, *for units in the selected sample*. For this purpose, 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.[5] We will use a binary variable, S, to denote sample selection.[6]

Our approach is grounded in structural causal models (SCM; Pearl (2009)), potential outcomes (Splawa-Neyman et al. (1990), Rubin, 1974, 1978, 1990), and directed acyclic graphs (DAGs; Pearl (2009)). Potential outcomes are solutions to the equations in SCMs, under intervention. The equations and variables in SCMs correspond to the edges and nodes in DAGs.[7] 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. 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. Researchers might also be interested in the the causal effect for the *sub-population for which the selected sample is a representative sample.* An estimation strategy is said to be "internally valid" if it can unbiasedly or consistently estimate such quantities. In the insturmental variables context, we will typically be targeting a "local" average treatment effect, where the effect is local to the "compliers"[8] in the selected sample or the compliers in the sub-population. Depending on the assumptions that researchers are willing to make, these local average treatment effects might equal more familiar causal effects like the average treatment effect or average treatment effect on the treated in the selected sample or in the sub-population. In what follows, we do not always differentiate units eligible to be in the selected sample from those specifically in the sample in 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.

2.1 Instruments under sample selection

Instrumental variables approaches can be used to identify causal effects of the treatment on the outcome in the presence of relationships between the treatment and outcome other than the causal relationship. This is a powerful capacity, when used

[5] 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)

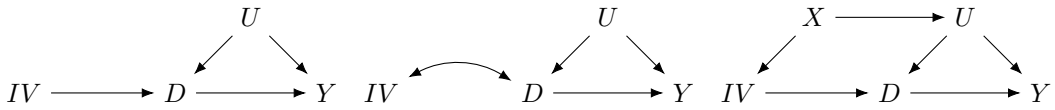
[6] See Rohde and Hazlett (20XX) for a discussion of why we choose to represent sample selection as a separate binary variable.

[7] See Pearl (2009) and Appendix C for formal details.

[8] Compliers are the units that choose to take the treatment only when encouraged to do so by the instrument. See (Imbens and Angrist, 1994; Angrist et al., 1996; Hernán and Robins, 2020) for further discussion. In what follows, we allow for instruments that are not direct causes of the treatment. In this case, so long as ignorability is not violated, there will be an unobserved direct cause of the treatment. This unobserved variable is what defines compliers. See Appendix XXXX for an example.

properly, since ruling out unobserved confounding and other forms of non-causal relationships between the treatment and outcome can be difficult. As we'll see, it turns out that this approach trades one set of challenges for another and is not a silver bullet. The key to an instrumental variables approach is the presence of a variable (that we call an instrument, IV) that is associated with the treatment, D, but that is not otherwise associated with the outcome, Y. We want the instrument to covary with the treatment but only covary with the outcome as a result of the causal association between the treatment and the outcome. Simple instruments are depicted in Figure 1. When such a variable exists (and we possibly make some additional assumptions), we are able to use the association between the instrument and the outcome as well as the association between the instrument and the treatment to identify (or bound) a causal effect of the treatment on the outcome. For such a set up to work we need to be very careful about exactly how the instrument relates to the treatment and to the outcome. We can capture the precise requirements using the conditions of relevance and ignorability.

Figure 1: Simple causal graphs for instruments



We alter the definition of an instrument found in Pearl (2009) to explicitly state that we must restrict ourselves to the selected sample, that is, we must condition on $S = 1$. The following definition is adapted from Pearl (2009), Definition 7.4.1.[9]

Definition 1 (Instruments, Relevance, and Ignorability)**.** A variable IV is an instrument relative to the total effect of D on Y within the stratum $S = 1$ if there exists an X, unaffected by D, such that the following hold.

1. (**relevance**) $D \not\perp\!\!\!\perp IV | X, S = 1$
2. (**ignorability**) $Y_d \perp\!\!\!\perp IV | X, S = 1$

Relevance captures the idea that in order to study the relationship between the treatment and outcome, the instrument must be associated with the treatment. Otherwise, we cannot use the instrument to isolate any of the variation between the treatment and the outcome. Ignorability captures the idea that, while we want the instrument to associate with the treatment, we do not want it to directly cause the outcome or to be related with the outcome other than through its relationship with the treatment. If it were associated with the outcome in one or both of these ways, then we could not disentangle the association between the instrument and the outcome from these relationships and the association that runs from the instrument to the treatment to the outcome. The latter contains the relationship we want to study, namely that between the treatment and the outcome. Our alteration to Pearl's definition make explicit that we want these conditions to hold in the selected sample.

Two points of clarification are useful. First, if conditioning on observed covariates alone can allow for the identification of causal effects of interest, then we may not need an instrumental variables approach at all. For example, see Pearl's back-door criterion (Pearl, 2009) for how this might be done. In this paper, we use covariate adjustment and conditional instruments in order to use an instrumental variables approach. An instrumental variables approach only makes sense when we cannot eliminate all bias between treatment and outcome with covariate adjustment alone, whether this bias is from confounding or sample selection. Obviously, all elements of X must be observed variables for a conditional instrument to be useful. Second, we emphasize once more that, these conditions are not sufficient for identifying a causal effect of interest. They define an instrument and are enough to bound a causal effect, but not point identify them. Assumptions like effect homogeniety, monotonicity, or one-sided non-compliance are also required for point identificaiton. However, these will not be a focus of our discussion, as these are not explicit in causal graphs and our project here is to construct a graphical framework for evaluating relevance, ignorabillity, and the presence of instruments. These additional assumptions are, of course, still vital to the identification of causal effects and researchers need to take care that they are appropriately considering the plausibility of these assumptions. In a the Appendix, we show how such assumptions can be combined with relevance and ignorability to identify causal effects.

We are often not necessarily interested in the causal effect of IV on Y. Rather, we're interested in all the ways that IV associates with Y through D. If IV and D are associated due to confouding or sample selection, this is ok, as long as

[9]This definition is by no means the only way to define an instrument. Pearl (2009) also provides conditions that are purely graphical and do not involve potential outcomes. Angrist et al. (1996); Lousdal (2018) require that the IV - D relationship is causal and unconfounded, which Pearl (2009) points out is unnecessary in general. Hernán and Robins (2006, 2020) split ignorability into two conditions, one of which is the well known "exclusion" restriction: $Y_{d,iv} = Y_{d,iv'} = Y_d$ and $Y_{iv,d} \perp\!\!\!\perp IV$. As shown in the main text, we can combine exclusion and ignorability into $Y_d \perp\!\!\!\perp IV$; see Hernán and Robins (2020). (Proof: $Y_{d,iv'} = Y_d$ and $Y_{iv,d} \perp\!\!\!\perp IV \implies Y_d \perp\!\!\!\perp IV$: $(Y_{iv,d} = Y_d) \perp\!\!\!\perp IV$. $Y_d \perp\!\!\!\perp IV \implies Y_{d,iv'} = Y_d$ and $Y_{iv,d} \perp\!\!\!\perp IV$: Suppose $Y_{iv,d} \neq Y_d$. Then there is a path from IV to Y that does not run through D. This means that IV is a common cause of Y and D and so $Y_d \not\perp\!\!\!\perp IV$, a contradiction. So $Y_{iv,d} = Y_d$, which in turn means $(Y_{iv,d} = Y_d) \perp\!\!\!\perp IV$.) Greenland (2000); Didelez and Sheehan (2007a,b); Sheehan et al. (2008) use somewhat different conditions that invoke an unobserved confounder explicitly. Swanson et al. (2018) discuss bounds that can be found for the ATE under various alternative IV conditions. We adopt what they describe as the "least restrictive" set of IV conditions here.

all the association between IV and Y is thorugh D. So IV and D need only be associated conditional on $X, S = 1$, we do not require $IV \to D$. We might consider a few sub-types of instruments. "Causal" instruments are those for which there is an unconfounded causal path $IV \to D$. "Proxy" instruments are those for which the association between IV and D flows through a path like $IV \leftarrow U^* \to D$, where U^* is a causal instrument and IV is a proxy for U^*. Van Der Zander et al. (2015) define an "ancestral" instrument as one for which conditioning on a variable creates the relevance needed for an instrument. This might look like $IV \to X \leftarrow U^* \to D$, where, again, U^* is a causal instrument and IV is a proxy for U^*, conditional on X. Another interesting case might be $IV \to S \leftarrow U^* \to D$, where S is the selection node; we discuss this further below. One might also consider a "confounded causal" instrument where we have both $IV \to D$ and $IV \leftarrow U^* \to D$, as well as other combinations of these. These sub-types listed are suggestive of the flavors of instruments that might arise, but are not a comprehensive list nor meant to suggest that these types of instruments arise frequently.

Readers may be asking themselves the following questions. How do we know whether relevance and ignorability hold? How are we to determine whether or not relevance and ignorability hold within the selected sample when there is no sample selection node in graphs like those in Figure 1? Answering these questions is difficult, and, in practice, we can never be certain that some set of covariates will provide the relevance and ignorability we need. The onus is on researchers to make plausible arguments for relevance and ignorability. To aid in this, we can build a model of how the treatment and outcome causally relate to each other and relevant covariates. Such a model should capture all the structural information that is available about the causal mechanisms relating important variables, as well as the uncertainty about such relationships. The causal relationships can be non-parametrically encoded in a structural causal model which can be represented graphically as a directed acyclic graph (and extensions thereof). See Pearl (2009) and Appendix C for details and Figure 1 for examples.

DAGs allow us to visualize dependencies and independencies between variables in terms of a path separation criterion, *d-separation*. (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[10] 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 Pearl (2009), chapter 1 for details. Graphical criteria can then be used to determine when relevance and ignorability hold. In the next sections, we will present an extension to the typical causal graphs and an associated graphical criteria built to help researchers determine when relevance and ignorability statements hold in the presence of sample selection.

3 Proposal

In this section, we propose extended causal graphs that explicitly show how sample selection alters the relationships between variables in the sample. We also provide rules (graphical criteria) for using these graphs to determine when relevance and ignorability hold. In doing so, we revise the existing graphical approaches for instruments to highlight the special role of sample selection. At the same time, we formalize and generalize recent discussions of sample selection in the instrumental variables context. In the following section, we review the implications of sample selection for instrumental variables.

3.1 Internal selection graphs

We now detail our simple graphical approach to determining whether relevance and ignorability hold. 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.

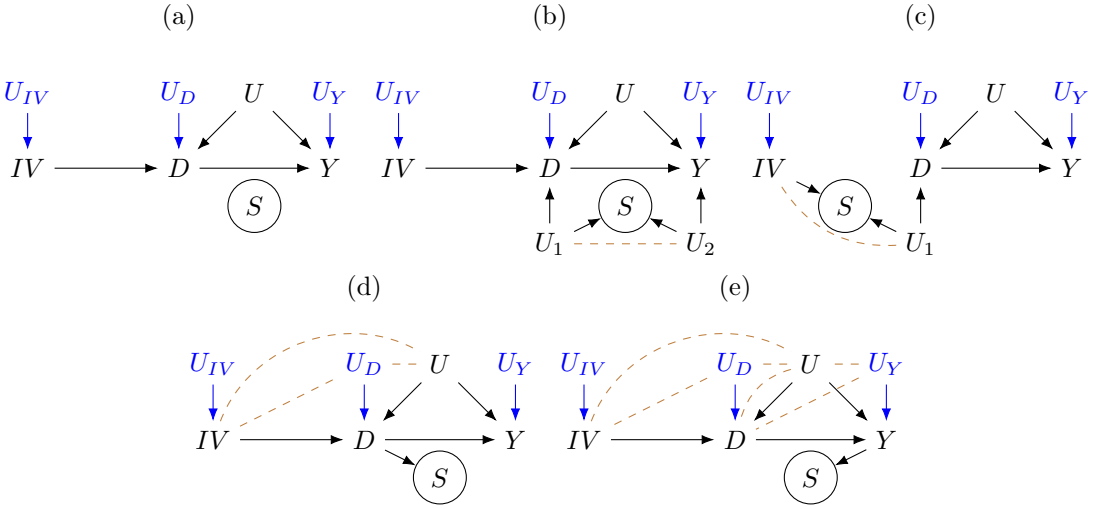
Definition 2 (Internal Selection Graph, G_S^+)**.** Let G be the DAG induced by a SCM.

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. Add to G_S any node which is a parent of the potential instrument or a parent of a descendant of the potential instrument. (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^+. (These graphs are similar to those discussed in Daniel et al. (2012) and Rohde and Hazlett (20XX).)

[10] A collider is a node in the graph into which two arrows point: $A \to S \leftarrow B$. See Pearl (2009) for an introduction to causal graphical models and colliders. Conditioning on a collider or a descendant of a collider can induce an association between the parents of the collider. Shahar and Shahar (2017) discuss the conditions under which such an association is created. Since our approach is non-parametric and graphical, we assume such an association is created when sample selection is a collider or a descendant of a collider.

Figure 2: Examples of Internal Selection Graphs



The key features of internal selection graphs[11][12][13] are the inclusion of an encircled sample selection node, specific background variables, and bridges that capture the statistical associations that result from sample selection. 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 for examples. We will differentiate between a few types of paths. 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 C for details. Generalized paths are any sequence of nodes and edges (directed edges and/or bridges) where each node appears only once (e.g., $D \cdots Z \rightarrow Y$, $D \rightarrow Y$, $D \rightarrow S \leftarrow Z$, $U_D \rightarrow D \rightarrow Y$). Causal paths are any generalized path where all edges between the nodes are directed and point in the same direction (e.g., $D \rightarrow Y$, $U_D \rightarrow D \rightarrow Y$). Generalized non-causal paths are any generalized path that isn't a causal path (e.g., $D \cdots Z \rightarrow Y$). Figure 2(e) provides a clear example of a setting in which internal selection graphs greatly facilitate understanding how sample selection can alter relationships between the variables in the selected sample. The statistical associations created due to sample selection between many variables, as well as some of the variables themselves, would be missing from the corresponding DAG.

3.2 Graphical criteria

So how can we use internal selection graphs to determine whether relevance and ignorability hold? We'll use a set of rules captured in the following graphical criteria.

3.2.1 Relevance

Relevance is the first condition in our definition of instruments and is perhaps the simpler of the two conditions. It captures the idea that, in order to study the relationship between the treatment and outcome using an instrument, the instrument must be associated with the treatment in some way. Otherwise, we cannot use the instrument to understand any of the variation between the treatment and the outcome. The relevance criterion is similar to condition (ii) in the graphical criterion provided in Pearl (2009) and similar to the condition (G1) in the graphical criterion provided in Elwert and Segarra (2022), but altered to indicate the special role that sample selection plays and to work with internal selection graphs.

Definition 3 (Relevance Criterion)**.** A set of nodes X and a possible instrument IV in G_S^+ satisfy the relevance criterion relative to D (treatment), and Y (outcome) if there is at least one (*causal or generalized non-causal*) path between IV and D that does not pass through S and is not blocked by X.

[11] 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 relevance and ignorability, making the inclusion of U_S unnecessary.

[12] 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.

[13] The value of this sort of graph for evaluating sample selection can be seen in **??**, 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. **?** also discusses an approach in which undirected edges are added to the graph.

Theorem 1. *If a set of nodes X and a possible instrument IV in internal selection graph G_S^+ satisfy the relevance criterion relative to D (treatment), and Y (outcome), then $D \not\perp\!\!\!\perp IV|X, S = 1$.*

This result is proved in Appendix C. We need the instrument to be associated with the treatment. Whether this association manifests as a causal relationship (e.g., $IV \rightarrow D$) or a non-causal relationship (e.g., $IV \leftarrow U \rightarrow D$) is not important. For relevance to hold, we just need that the values the instrument takes are related some how to the values that the treatment takes.[14] So we just want there to be some path between these two variables. Moreover, we need this association to persist (or perhaps to arise) when we condition on both X and $S = 1$. Hence, our criterion requires that there is at least one path that does not pass through S. The paths through S become irrelevant in internal selection graphs because if S is not a collider on a path, then the path is blocked and we can ignore it; further, if S is a collider on a path, then we have already drawn a bridge between the nodes that form the collider and so we can bypass the path that actually includes S itself, treating S like any other collider in the graph. If we must condition on X to achieve ignorability, we don't want it to ruin relevance. So we need a path between IV and D that is not blocked by X. Finally, there are situations in which selection and/or conditioning on X can give us relevance. Such instruments are called "ancestral" instruments, as discussed above (Van Der Zander et al., 2015). This might arise due to X being a collider in which case conditioning on X may unblock a path between IV and D; of course such a path would not be blocked by X and would satisfy the relevance criterion. No matter the specific type of path or paths that yield relevance, the key idea of relevance is simple: that IV and D should relate along some unblocked path.

3.2.2 Ignorability

Ignorability is the second condition in our definition of instruments and is somewhat more subtle than relevance. It captures the idea that, while we might not be very particular about how the instrument associates with the treatment, we want to be very careful about how the instrument associates with the outcome. The ignorability criterion is similar to condition (i) in the graphical criterion provided in Pearl (2009) and similar to the conditions (G2) and (G3) in the graphical criterion provided in Elwert and Segarra (2022), but altered to indicate the special role that sample selection plays and to work with internal selection graphs.

Definition 4 (Ignorability Criterion)**.** A set of nodes X and a possible instrument IV in G_S^+ satisfy the ignorability criterion relative to D (treatment), and Y (outcome) if

1. No element of $\{X, S\}$ is a descendant of D and D is not in $\{X, S\}$.
2. X blocks every (*causal and generalized non-causal*) path between IV and Y except
   (a) those that pass through S and
   (b) those ending with a causal path from D to Y (e.g., paths between IV and Y that pass through D but where D or one of its descendants touches a bridge or paths on which D is an ancestor of IV must be blocked by X).

Theorem 2. *If a set of nodes X and a possible instrument IV in internal selection graph G_S^+ satisfy the ignorability criterion relative to D (treatment), and Y (outcome), then $Y_d \perp\!\!\!\perp IV|X, S = 1$.*

This result is proved in Appendix C. We want the instrument to associate with the outcome only along paths that include a causal path between the treatment and the outcome. This is because the causal paths between the treatment and the outcome are those that we ultimately are interested in studying. So other paths that associate the instrument and the outcome are a problem for instrumental variables approaches. If the instrument were associated with the outcome along some other type of path, we would not be able to disentangle the association between the instrument and the outcome from the association that runs from the instrument to the treatment and then to the outcome along causal paths between the treatment and outcome alone. The latter contains the relationship we want to study, namely the causal relationship between the treatment and the outcome. The ignorability criterion formalizes these ideas. Our ignorability criterion leaves the types of paths between IV and Y that we want to leverage unblocked and requires us to block the types of paths that introduce variation between IV and Y that can contaminate our analysis.

4 Discussion

We now have all the necessary machinery in place to start analyzing how sample selection can threaten or provide opportunities for relevance and ignorability and, hence, instruments. Canan et al. (2017); Swanson et al. (2015); Swanson (2019); Hughes et al. (2019); Ertefaie et al. (2016); Gkatzionis and Burgess (2018); Hernán and Robins (2020) all discuss sample selection and instruments. But none of these provide a formal framework for analyzing the implications of sample selection for instrumental

[14] There are also problems associated with weak associations between the instrument and treatment. These are referred to as the weak instruments. In this paper, we focus on whether relevance holds at all and not on whether the instrument is a weak instrument. Fortunately, relevance is a condition that can actually be tested with data. Further discussion of this is also outside the scope of this paper.

variables, where we can have an arbitrary causal graph and selection mechanism. We start in this section by briefly considering some simple examples.

Figure 2 provides a good starting place. When sample selection is not causally related to any of the other variables in the causal model and we have the canonical instrumental variables graph, as in Figure 2(a), we see that we can easily verify relevance and ignorability, where X is the empty set. There is a causal path connecting IV and D, giving relevance. The empty set blocks all paths between IV and Y, except for one that ends in a causal path from $D \rightarrow Y$ and S is not a descendant of D and $D \notin \{X, S\}$, giving ignorability.

Figure 2(b,c) both also meet the relevance and ignorability criteria, as the reader can verify for themselves. Figure 2(b) is interesting in that the instrumental variables approach here can actually be used to overcome both the unobserved confounding between D and Y from U as well as the sample selection bias between D and Y, even when U, U_1, and U_2 are all unobserved. Figure 2(c) is an example of an ancestral instrument that only satisfies the relevance criterion due to the purely statistical association created by sample selection. Perhaps there are opportunities to exploit these types of sample selection mechanisms that have been underappreciated. We will discuss settings like Figure 2(b,c) in more detail in a subsequent section.

Figure 2(d) is interesting in that it shows that selection on the treatment actually leads to a violation of the ignorability criterion. In simple covariate adjustment approaches (i.e., not instrumental variables), selection based on the treatment is, on its own, not biasing. See Rohde and Hazlett (20XX) for details. However, in the instrumental variables case, such a selection mechanism clearly does not satisfy the ignorability criterion. This is perhaps the simplest example for which sample selection does not operate in the same way for covariate adjustment approaches and instrumental variables. Researchers should not assume sample selection can be treated similarly in these two approaches. Figure 2(e) is a simple example of how, as in simple covariate adjustment approaches, selection on the outcome can violate ignorability. Together, these demonstrate that heuristics from other research designs should not be applied to instrumental variables without thoughtful consideration or the use of formal design-specific criteria, like those in this paper and Rohde and Hazlett (20XX). We emphasize to the reader that one cannot credibly ascertain the implications of sample selection on an instrumental variables (or any other design) without laying out how sample selection fits into the causal model and using tools like internal selection graphs and our graphical criteria. Less formal approaches will not confer the same assurance that the researcher has not missed some subtle alteration that sample selection makes to the relationships in the data. In the following discussion, we look at more examples aimed at exploring the various ways that sample selection can alter instrumental variables approaches.

4.1 Interesting cases

We now consider settings in which the way that sample selection interacts with the instrumental variables design is perhaps underappreciated or under-discussed. These cases highlight potential opportunities in which to use instrumental variables and also illustrate how sample selection can threaten internal validity of instrumental variables. We hope they are thought-provoking to the reader.

4.1.1 Instrumental variables can be used to recover from sample selection

There are settings in which sample selection can create generalized non-causal paths between the treatment and the outcome, and hence bias designs other than instrumental variables, but for which an instrumental variables design can overcome the sample selection bias. This setting has been recognized elsewhere in the literature. (Swanson, 2019) See Figure 2(b) for a simple example. The particular form shown in Figure 2(b) is commonly called "M-Bias." More generally, an instrumental variables approach could be used to overcome sample selection bias that takes the form of a generalized non-causal paths running between the treatment and outcome created by sample selection (i.e., containing a bridge) that start with an arrow pointing into the treatment. There are really two equivalent ways to look at such settings. One is that instrumental variables is immune to this type of sample selection bias. The other is that instrumental variables is an approach that could be used when this form of sample selection bias is suspected to threaten the validity of simple covariate adjustment approaches. Both views are interesting. The former is useful to know when a researcher plans to employ instrumental variables before considering sample selection. The latter actually presents opportunities for which instrumental variables approaches might be employed that are not currently common. For example, if a researcher suspects that they have a sample selection M-bias problem, they can use an instrumental variables approach to overcome this bias, when a suitable instrument exists.

4.1.2 Ancestral instruments via sample selection

Another interesting case arises when we consider how sample selection can alter relevance. Above, we mentioned the idea of "ancestral" instruments. These are instruments that meet the relevance criterion only when we condition on some covariate(s). (Van Der Zander et al., 2015) It turns out that sample selection can also create ancestral instruments. This is another setting that has been mentioned in the literature (Hernán and Robins, 2020) but is not widely used or discussed and presents additional opportunities for the use of instrumental variables. A simple version of this appears in Figure 2(c). In this setting,

U^* is an unobserved but causal instrument. Sample selection creates a purely statistical relationship between IV and U^* in the sample at hand. It is then easy to verify that IV satisfies both the relevance and ignorability criteria. So we can view IV as a proxy for the causal instrument U^*. See the Appendix for an example of how, in this setting, we can identify a causal effect using a monotonicity assumption.

While this setting at first sounds quite promising, we urge caution. The nature of this setting makes it likely that there will be violations of ignorability. This setup is only useful if the causal instrument, U^*, is unobserved but is not a common cause to the wrong variables. For instance, ignorability will be violated if U^* is also a parent of Y or U or if there is an unobserved common cause of U^* and Y or various other relationships that might arise in realistic settings. So, while ancestral instruments that arise from sample selection might be intriguing, great care should be taken in evaluating whether ignorability holds for them.

4.1.3 Restricting to units that receive two treatments when more exist

Swanson et al. (2015); Ertefaie et al. (2016) discuss the common practice of employing instrumental variables approaches to study how two particular treatments or treatment levels compare, where the sample is limited to units receiving these two treatments, but where more than two treatments are possible. Such studies suffer from selection on the treatment, which, as we can see in Figure 2(d), can violate ignorability and bias effect estimates. However, the "ensuing selection bias that occurs due to this restriction has gone relatively unnoticed" and is "pervasive." (Swanson et al., 2015) We echo that this and similar practices are a problem as they violate ignorability.

4.1.4 Randomized experiments can have sample selection problems

Random experiments are often held as the gold-standard for causal inference. However, they are not impervious to threats to validity. Often experiments, especially when related to human subjects, have non-compliance with assigned treatments. This means that participants choose not to, say, take a drug when they've been prescribed it. Which participants choose not to comply could have common causes with the outcome, confounding the treatment-outcome relationship. Instrumental variables can be used to address this non-compliance by using assigned treatment as an instrument for whether or not someone is actually treated (e.g., actually takes the drug). A second threat to the validity of randomized experiments is differential attrition from the study. This means that some participants or units drop out of the study and so complete data is not available for all participants. It is possible for the attrition to be post treatment or post outcome. When both non-compliance and attrition both occur, a researcher might be in a scenario where, despite randomly assigning treatment, they are attempting to use instrumental variables to analyze their experiment and also be concerned about sample selection. The simplest forms of this would be captured by Figure 2(d) and (e). Despite the randomization of treatment assignment and no violations of the exclusion restriction, there are violations of ignorability and effect estimates will be biased. Alternatively, if attrition is based on the instrument and there is a common cause of attrition and the outcome, we might also have a violation of ignorability. See Montgomery et al. (2018) for a very useful discussion about post-treatment conditioning and selection in randomized experiments in political science and the bias that can result.

4.1.5 Sample selection can make and break both ignorability and relevance

As we saw with ancestral instruments, sample selection can sometimes help provide relevance. This is also true for ignorability. See Figure 3(a) and (b). However, it can also create ignorability while breaking relevance, as in Figure 3(c). Similarly, sample selection can also create relevance while breaking ignorability. See Figure 3(d) and (e). At first, glance these last two example may seem like they satisfy ignorability. While paths like $IV \cdots D \rightarrow Y$ are allowed since they contain $D \rightarrow Y$, $IV \cdots D$ also creates the path $IV \cdots D \leftarrow U \rightarrow Y$, which breaks ignorability.

4.1.6 Small changes to the causal graph matter

We've stressed the importance of including a sample selection node in every causal graph. We now stress that great care must be taken in constructing causal graphs containing sample selection node. In Figure 4 we see the same graph as in Figure 2(b) but where we flip the direction of just one edge. The conclusions for the two variations on the graph are opposites. In Figure 4(a) we satisfy the ignorability criterion but in Figure 4(b) we do not. 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 threat that sample selection poses to internal validity and what, if anything, might be able to be done. There is also potential for users to intentionally or unintentionally favor one graph over another very similar graph in order to show that ignorability holds. 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.

Figure 3: Sample selection can make and break both ignorability and relevance

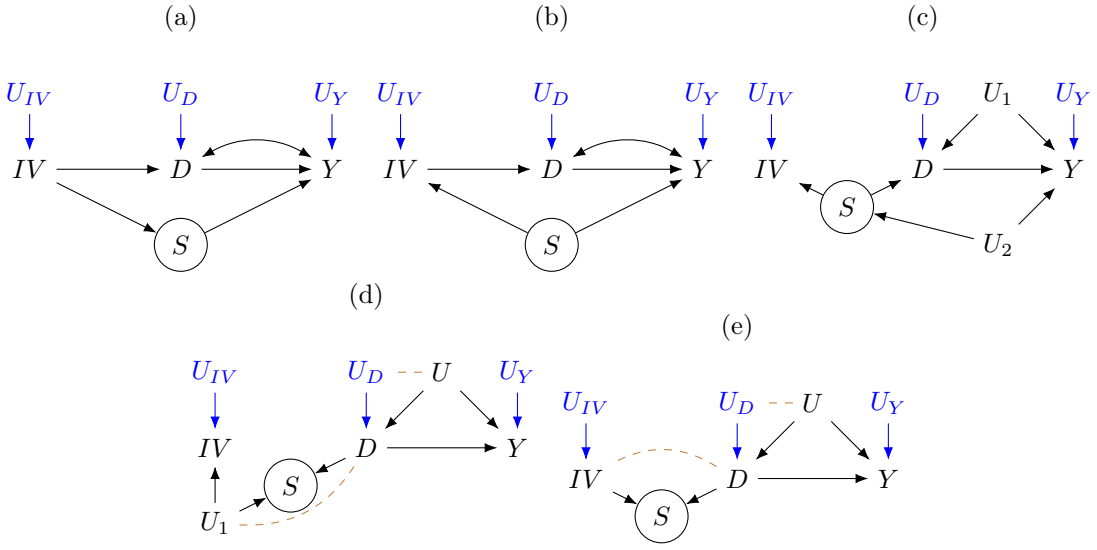
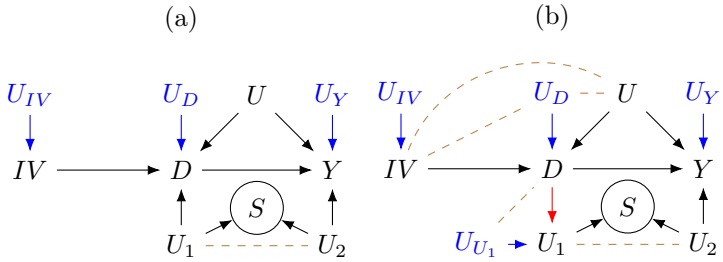


Figure 4: Small changes to the causal graph matter



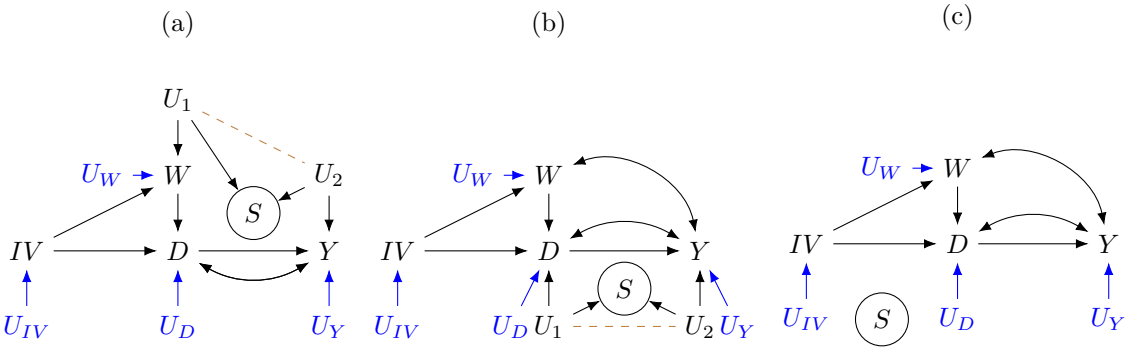
4.1.7 Blocking non-causal paths between treatment and outcome can invalidate instruments

The final interesting case we consider again cautions against applying heuristics from simple covariate adjustment approaches (or other designs) in the instrumental variables setting. In both simple covariate adjustment and instrumental variables, we want to learn about the causal effect of the treatment on the outcome. In a covariate adjustment setting, we attempt to un-confound the true treatment effect by blocking non-causal paths between the treatment and the outcome. We can reduce the bias of our estimates by limiting the number of non-causal paths that run between the treatment and the outcome. Doing so helps to make the treated group comparable to the untreated group (assuming a binary treatment), making a comparison of these groups useful for causal analysis. We employ instrumental variables approaches only when we cannot block *all* the non-causal paths between the treatment and the outcome. And when this is the case, there are settings in which blocking non-causal paths between the treatment and the outcome can actually lead to violations of the ignorability we need for instruments. This can happen even if we would have had a valid instrument before blocking the non-causal paths between the treatment and the outcome. See Figure 5 for three examples. Note that not all of these examples hinge on sample selection creating bias. In these examples, conditioning on W would block a non-causal path between the treatment and the outcome but open others between IV and Y, even though no open non-causal paths existed between IV and Y before conditioning on W.

4.1.8 More examples

Tables 1 - 5 contain numerous example internal selection graphs. Table 1 contains examples in which both relevance and ignorability hold without any covariate adjustment. The examples in this table are similar to things we've already seen. Table 2 also contains examples in which both relevance and ignorability hold without any covariate adjustment. But these examples have a somewhat different flavor. In some, selection is blocking otherwise problematic paths; in others, sample selection does not violate relevance but might weaken the strength of the instrument, which can pose problems for estimation. Table 3

Figure 5: Blocking non-causal paths between treatment and outcome can invalidate instruments



contains examples in which relevance and ignorability hold only after we condition on some covariate. If these covariates are not observed, these settings will have problems, but when they are observed we can use an instrumental variables approach. Table 4 explores various ways that ignorability can be violated by sample selection as well as by things other than sample selection. Sample selection can pose many threats to ignorability, as can common causes or the instrument and outcome or causal paths from the instrument to the outcome. Finally, Table 5 looks at a couple ways that sample selection could violate relevance. When considering sample selection, we should not exclusively worry about threats to ignorability. We additionally provide some more examples in Appendix B, including an internal selection graph for all the DAGs considered in Hughes et al. (2019). These greatly simplify the analysis necessary to determine the threats that sample selection poses.

We hope the settings discussed in this section provide some further insight into how sample selection can effect instrumental variables. We again stress that, without incorporation of the sample selection mechanism into the causal model and use of a formal framework like that presented here, there is no reliable way to determine how sample selection might alter relevance or ignorability for your specific instrumental variables application.

4.2 Lessons

After considering numerous examples, we are in a position to review some of what we've learned. Let us finally review key lessons to round out our discussion.

- Only the incorporation of the sample selection mechanism into the causal model can reliably lead to correct conclusions about how sample selection might effect instrumental variables. Moreover, graphical analysis and a formal framework for the analysis of sample selection can greatly reduce the burden on researcher in analysis of how sample selection alters their instrumental variables approach. 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.
- Informal applications of simple heuristics related to sample selection can be misleading and should be avoided. Further, we should not use heuristics from other research designs, like simple covariate adjustment, in the instrumental variables setting. These might not apply (e.g., like selection on the treatment being non-biasing) and can result in unreliable conclusions.
- Sample selection can influence instrumental variables, even when it is not a collider. Whether selection is a collider, confounder, mediator, or indirectly related to variables of interest, the relevance and ignorability criteria provide clear guidance on this.
- When sample selection manifests as attrition or some other post-treatment type of selection, randomization of treatment or instrument 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 the graphical criteria presented here is the best way to be sure.
- Selection on the outcome is usually a problem for instrumental variables. 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 a problem on its own.
- Indirect association between selection and the outcome or selection and the treatment are typically not problems on their own and also typically not when they appear together for instrumental variables.
- There are various ways that sample selection can threaten relevance and ignorability.
- There are also opportunities presented by sample selection for instrumental varibles as well as by instrumental variables

for sample selection. See the previous section.

- Actual identification of causal effect requires more than just relevance and ignorability. We also need assumptions like homogeneous treatment effects, monotonicity, or one-sided non-compliance. See the Appendix for examples of identification.

In conclusion, instrumental variables approaches leverage specific types of variation between the instrument and treatment and the instrument and outcome to identify causal effects of the treatment on the outcome. We've seen how these associations can be altered in a non-randomly selected sample in a variety of ways. Sample selection can create many wrinkles in an instrumental variables analysis but is not necessarily a death blow. But the already high bar of finding a good instrument is only made higher by responsibly considering how sample selection can influence instruments. We hope that the tools, examples, and lessons in this paper can provide additional caution, clarity, and credibility to researchers that hope to use instrumental variables.

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.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33:2297–2340.

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 *AAAI*.

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. (1966). *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.

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

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).

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale Press, New Haven.

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. and Sheehan, N. (2007a). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330.

Didelez, V. and Sheehan, N. A. (2007b). Mendelian randomisation: Why epidemiology needs a formal language for causality. In Russo, F. and Williamson, J., editors, *Causality and Probability in the Sciences*, pages 5–263.

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.

Ertefaie, A., Small, D., Flory, J., and Hennessy, S. (2016). Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. *The International Journal of Biostatistics*, 12(1):219–232.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3:151–182.

Gkatzionis, A. and Burgess, S. (2018). Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology*, 48(3):691–701.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

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

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

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.

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. (2014a). Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3):323–358.

Imbens, G. (2014b). Rejoinder of "instrumental variables: An econometrician's perspective". *Statistical Science*, 29(3):375–379.

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

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

Kumor, D., Cinelli, C., and Bareinboim, E. (2020). Efficient identification in linear structural causal models with auxiliary cutsets. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5501–5510. PMLR.

Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, 15(1).

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. (2001). Parameter identification: A new perspective (second draft). *Technical Report R-276. UCLA Cognitive Systems Laboratory*.

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

Peyton, K. (2020). Does trust in government increase support for redistribution? evidence from randomized survey experiments. *American Political Science Review*, 114(2):596–602.

Rohde, A. and Hazlett, C. (20XX). Revisiting sample selection as a threat to internal validity: New lessons, examples, and tools. *XXXX*, XX(XX):XXX–XXX.

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.

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.

Sheehan, N., Didelez, V., Burton, P. R., and Tobin, M. D. (2008). Mendelian randomisation and causal inference in observational epidemiology. *PLoS medicine*, 5(8):e177.

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.

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. and Hernán, M. A. (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, 24(3):370–374.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., and Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947. PMID: 31537952.

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.

Van Der Zander, B. and Liśkiewicz, M. (2016). On searching for generalized instrumental variables. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1214–1222, Cadiz, Spain. PMLR.

Van Der Zander, B., Textor, J., and Liskiewicz, M. (2015). Efficiently finding conditional instruments for causal inference. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3243–3249. AAAI Press.

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.

Table 1: **Relevance and Ignorability Criteria Satisfied without Covariate Adjustment - Part 1**

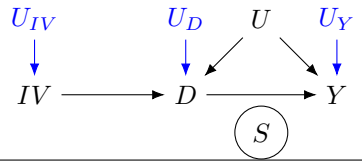
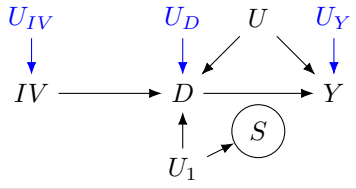
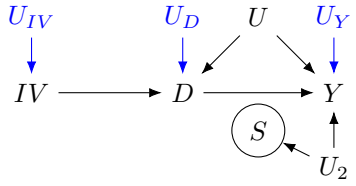
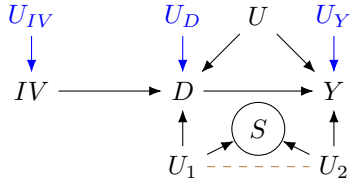
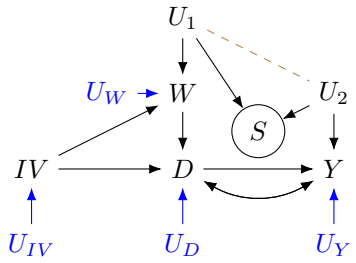
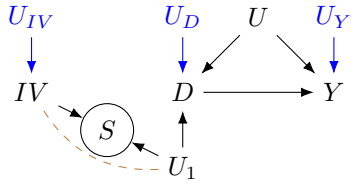
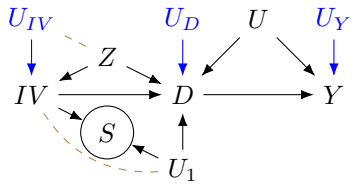
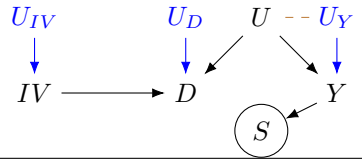
Relations with Selection	Internal Selection Graph	Explanation
Selection unrelated to key variables		Let the covariate conditioning set be $X = \{\emptyset\}$. The relevance criterion is satisfied by the path $IV \to D$. So, by Theorem 1, $IV \not\perp\!\!\!\perp D \mid S = 1$. Conditions 1 and 2 of the ignorability criterion are satisfied and are easy to verify. No generalized non-causal paths between IV and Y that do not pass through S or D exist. No generalized non-causal paths between IV and Y that pass through D but on which D touches a bridge exist. Therefore, condition 3 of ignorability criterion is satisfied and, by Theorem 2, $IV \perp\!\!\!\perp Y_d \mid S = 1$.
Selection indirectly related to treatment or outcome		
		
		
		Same as above. Note that conditioning on W would violate ignorability.
Selection induced ancestral instrument		Same as above.
		
Selection on the outcome; sharp null		

Table 2: **Relevance and Ignorability Criteria Satisfied without Covariate Adjustment - Part 2**

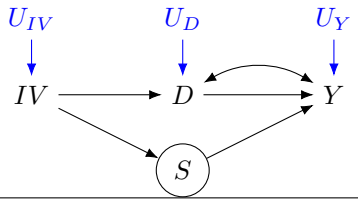
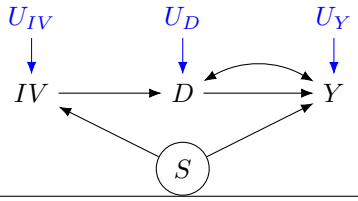
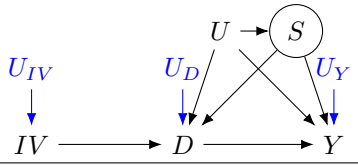
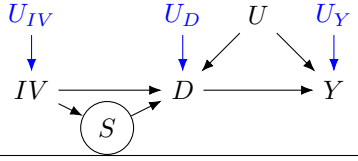
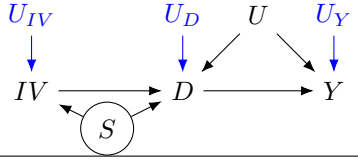
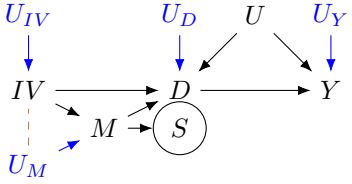
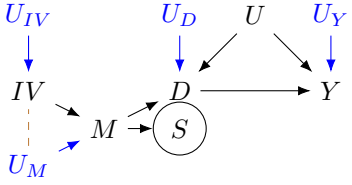
Relations with Selection	Internal Selection Graph	Explanation
Selection blocking paths		Let the covariate conditioning set be $X = \{\emptyset\}$. The relevance criterion is satisfied by the path $IV \to D$. So, by Theorem 1, $IV \not\perp\!\!\!\perp D \mid S = 1$. Conditions 1 and 2 of the ignorability criterion are satisfied and are easy to verify. No generalized non-causal paths between IV and Y that do not pass through S or D exist. No generalized non-causal paths between IV and Y that pass through D but on which D touches a bridge exist. Therefore, condition 3 of ignorability criterion is satisfied and, by Theorem 2, $IV \perp\!\!\!\perp Y_d \mid S = 1$.
		
		
Selection weakening instrument		
		
		
		

Table 3: **Relevance and Ignorability Criteria Satisfied with Covariate Adjustment**

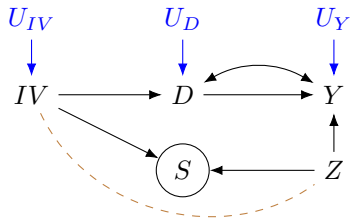
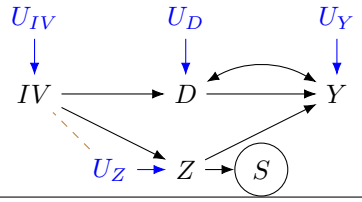
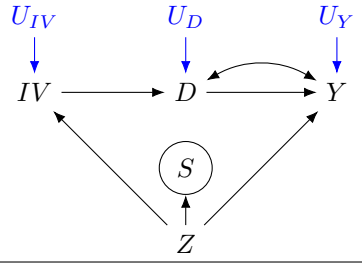
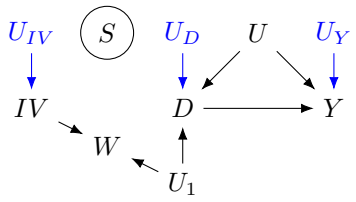
Relations with Selection	Internal Selection Graph	Explanation
Post-instrument selection and selection indirectly related to outcome		Let the covariate conditioning set be $X = \{Z\}$. The relevance criterion is satisfied by the path $IV \to D$. So, by Theorem 1, $IV \not\perp\!\!\!\perp D \mid Z, S = 1$. Conditions 1 and 2 of the ignorability criterion are satisfied and are easy to verify. The generalized non-causal path between IV and Y that passes through Z is blocked by conditioning on Z. Therefore, condition 3 of ignorability criterion is satisfied and, by Theorem 2, $IV \perp\!\!\!\perp Y_d \mid Z, S = 1$.
Selection as descendant of mediator between instrument and outcome		
Selection as child of confounder		
Ancestral instrument		Let the covariate conditioning set be $X = \{W\}$. The relevance criterion is satisfied by the path $IV \to W \leftarrow U_1 \to D$. So, by Theorem 1, $IV \not\perp\!\!\!\perp D \mid W, S = 1$. The ignorability criterion is satisfied and is easy to verify. So, by Theorem 2, $IV \perp\!\!\!\perp Y_d \mid W, S = 1$. Note that in this case, we might consider using W as a proxy instrument, rather than IV as an ancestral instrument, since the association with the treatment is likely to be stronger.

Table 4: **Ignorability Criterion Not Satisfied**

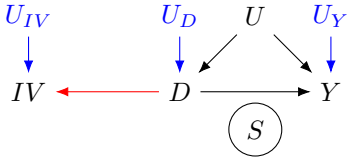
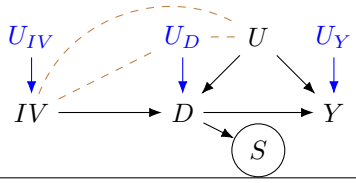
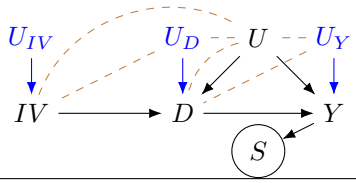
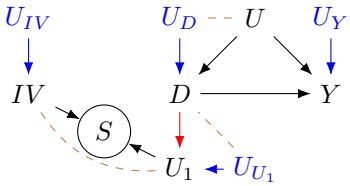
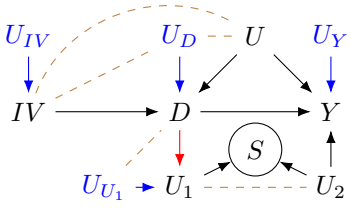
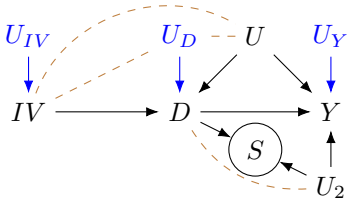
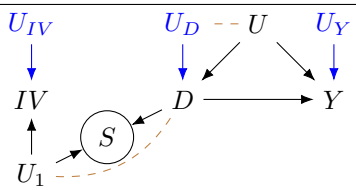
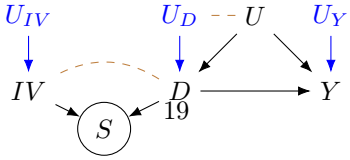
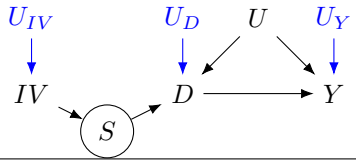
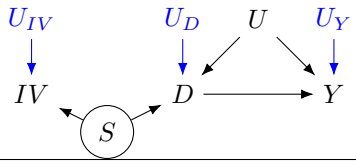
Relations with Selection	Internal Selection Graph	Explanation
Violation of ignorability unrelated to sample selection		We can see that condition 2 of the ignorability criterion is violated. The generalized non-causal path $IV \leftarrow D \leftarrow U \rightarrow Y$ cannot be blocked. So $IV \not\perp\!\!\!\perp Y_d \mid S = 1$. Note that a simple unobserved confounder of IV and Y or a causal path that cannot be blocked from IV to Y would also violate ignorability in a manner unrelated to sample selection.
Sample selection violates ignorability	      	We can see that condition 3 of the ignorability criterion is violated. Generalized non-causal paths between IV and Y that do not run through S or through D exist or generalized non-causal paths between IV and Y that run through D on which D or its descendants touch a bridge exist. So $IV \not\perp\!\!\!\perp Y_d \mid S = 1$.

Table 5: **Relevance Criterion Not Satisfied**

Relations with Selection	Internal Selection Graph	Explanation
Sample selection violates relevance	 	We can easily see that the relevance criterion is violated since sample selection is blocks the path from IV to D.

A Demonstrations of Identification

Here we demonstrate a few ways that we can use relevance, ignorability, and some additional assumptions to identify causal quantities. There are many other additional assumptions that could work. While a causal model may satisfy relevance and ignorability, this does not mean that the additional assumptions of the type discussed in this section will be weak assumptions or believable. Equal care is required in making these assumptions as is required in developing your causal model and assessing relevance and ignorability. Further, not all causal graphs that satisfy relevance and ignorability will be amenable to all additional assumptions that you might want to make; an example is shown in the discussion of monotonicity below, where we need stronger assumptions on the causal graph than are required by just relevance and ignorability alone.

A.1 Homogeneous Treatment Effects

We start simply by considering a homogeneous or constant treatment effects setting. This draws from Angrist and Pischke (2008) and Cunningham (2021). Such a setup assumes that the potential outcomes for each unit in the selected sample can be written as

$$[Y_{d,i} = \alpha + \delta d + \gamma U_i + \epsilon_i | S = 1]$$

where U is some unobserved variable that would appropriately block all non-causal paths between Y and D and give conditional ignorability conditional: $Y_d \perp\!\!\!\perp D | U, S = 1$. Here, δ is the average treatment effect in the sample, as well as the treatment effect for each unit in the sample. We see that this means that the observed Y_i's in the selected sample can be written as

$$[Y_i = \alpha + \delta D_i + \gamma U_i + \epsilon_i | S = 1]$$

We next assume that that we have a variable IV, our instrument, such that $\text{Cov}[IV, D|S = 1] \neq 0$. This is a specific form of relevance: $D \not\perp\!\!\!\perp IV|S = 1$. Simply assuming $D \not\perp\!\!\!\perp IV|S = 1$ does not garauntee $\text{Cov}[IV, D|S = 1] \neq 0$. But assuming $\text{Cov}[IV, D|S = 1] \neq 0 \implies D \not\perp\!\!\!\perp IV|S = 1$. So we're assuming $D \not\perp\!\!\!\perp IV|S = 1$ and a little more. Finally, we assume ignorability: $Y_d \perp\!\!\!\perp IV|S = 1$. $Y_d \perp\!\!\!\perp IV|S = 1 \implies \text{Cov}[IV, U|S = 1] = 0$ and $\text{Cov}[IV, \epsilon|S = 1] = 0$. If $\text{Cov}[IV, U|S = 1] \neq 0$ or $\text{Cov}[IV, \epsilon|S = 1] \neq 0$ then there would be a path from IV to Y (through U or ϵ, respectively) that would violate $Y_d \perp\!\!\!\perp IV|S = 1$. So we have $\text{Cov}[IV, D|S = 1] \neq 0, \text{Cov}[IV, U|S = 1] = 0$ and $\text{Cov}[IV, \epsilon|S = 1] = 0$. We can then identify δ, the average treatment effect in the sample, as follows:

$$\begin{aligned}
\text{Cov}[Y, IV|S = 1] &= \text{Cov}[\alpha + \delta D + \gamma U + \epsilon, IV|S = 1] \text{ by plugging in for } Y \\
&= \delta \underbrace{\text{Cov}[D, IV|S = 1]}_{\neq 0} + \gamma \underbrace{\text{Cov}[U, IV|S = 1]}_{=0} + \underbrace{\text{Cov}[\epsilon, IV|S = 1]}_{=0} \\
&= \delta \text{Cov}[D, IV|S = 1] \\
\implies \delta &= \frac{\text{Cov}[Y, IV|S = 1]}{\text{Cov}[D, IV|S = 1]} = \frac{\text{Cov}[Y, IV|S = 1]/\text{Var}[IV|S = 1]}{\text{Cov}[D, IV|S = 1]/\text{Var}[IV|S = 1]} = \frac{\text{Reg. Coef.}[Y, IV|S = 1]}{\text{Reg. Coef.}[D, IV|S = 1]}
\end{aligned}$$

A.2 One-Sided Non-Compliance

Next we consider heterogeneous treatment effects with a one-sided non-compliance assumption. This will often be applicable in randomized control trials with non-compliance. This draws from Angrist and Pischke (2008). We assume we have a binary treatment, that we have one-sided non-compliance ($P(D = 1|IV = 0, S = 1) = 0$), relevance ($D \not\perp\!\!\!\perp IV|S = 1$ which implies $P(D|IV = 1, S = 1) \neq P(D = 1|IV = 0, S = 1) = 0$), and ignorability ($Y_d \perp\!\!\!\perp IV|S = 1$). We also note that we can write observed outcomes as $Y_i = (1 - D_i)Y_{0,i} + D_i Y_{1,i} = Y_{0,i} + (Y_{1,i} - Y_{0,i})D_i$. Next we show that

$$\begin{aligned}
\mathbb{E}[Y|IV = 1, S = 1] &= \mathbb{E}[Y_{0,i} + (Y_{1,i} - Y_{0,i})D_i|IV = 1, S = 1] \\
&= \mathbb{E}[Y_{0,i}|IV = 1, S = 1] + \mathbb{E}[(Y_{1,i} - Y_{0,i})D_i|IV = 1, S = 1]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[Y|IV = 0, S = 1] &= \mathbb{E}[Y_{0,i} + (Y_{1,i} - Y_{0,i})D_i|IV = 0, S = 1] \\
&= \mathbb{E}[Y_{0,i}|IV = 0, S = 1] + \mathbb{E}[(Y_{1,i} - Y_{0,i})D_i|IV = 0, S = 1] \\
&= \mathbb{E}[Y_{0,i}|IV = 0, S = 1] + \mathbb{E}[(Y_{1,i} - Y_{0,i})D_i|D = 0, IV = 0, S = 1] \\
&\text{since } IV = 0 \implies D = 0 \\
&= \mathbb{E}[Y_{0,i}|IV = 0, S = 1] \\
&= \mathbb{E}[Y_{0,i}|IV = 1, S = 1] \text{ since } Y_d \perp\!\!\!\perp IV|S = 1
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[Y|IV=1,S=1]-\mathbb{E}[Y|IV=0,S=1] &= \mathbb{E}[(Y_{1,i}-Y_{0,i})D_i|IV=1,S=1] \\ &= \mathbb{E}[(Y_{1,i}-Y_{0,i})D_i|D=1,IV=1,S=1]P(D=1|IV=1,S=1) \\ &+ \mathbb{E}[(Y_{1,i}-Y_{0,i})D_i|D=0,IV=1,S=1]P(D=0|IV=1,S=1) \\ &= \mathbb{E}[Y_{1,i}-Y_{0,i}|D=1,IV=1,S=1]P(D=1|IV=1,S=1) \\ &= \mathbb{E}[Y_{1,i}-Y_{0,i}|D=1,S=1]P(D=1|IV=1,S=1) \\ &\text{since } D=1 \implies IV=1\end{aligned}$$

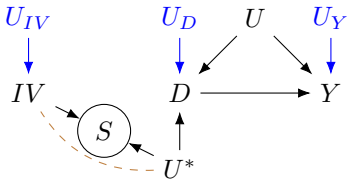
And so we see that we can identify the ATT in the selected sample as

$$\mathbb{E}[Y_{1,i}-Y_{0,i}|D=1,S=1]=\frac{\mathbb{E}[Y|IV=1,S=1]-\mathbb{E}[Y|IV=0,S=1]}{P(D=1|IV=1,S=1)}$$

A.3 Monotonicity

Finally, we consider heterogeneous treatment effects with a monotonicity assumption. For some additional elucidation, we look at an example of an ancestral instrument that achieves relevance only as a result of sample selection. See Figure 6. We can identify causal effects with a monotonicity assumption and an ancestral instrument created by sample selection so long as there is an unobserved *causal* instrument that adheres to the conditions laid out by Angrist et al. (1996). We also draw on Hernán and Robins (2006) here.

Figure 6: Ancestral instrument created by sample selection



We follow the simpler case presented by Hernán and Robins (2006), where the treatment and instruments are all binary. This is for demonstration. Hernán and Robins (2006) argue that a binary causal instrument in this setting won't be believable and discuss a continuous version. But we stick to the simpler case here just to give the flavor of how identification might work in this type of setting.

While relevance and ignorability are required for identification in instrumental variables, they are not sufficient and additional assumptions are required. Additionally, whether or not a specific additional assumption (here the example will be monotonicity) can be used might depend on details of the causal model. That is, different additional identifying assumptions might require additional constraints on the causal model, which may or may not fit with an accurate understanding of the underlying causal mechanisms.

In this example, we assume that IV, D, U^* are all binary. We also make a monotonicity assumption of the form $D_{u^*=0}=1 \implies D_{u^*=1}=1$. We are also assuming the causal model in Figure 6 holds. Here, IV is the observed ancestral instrument and U^* is the unobserved causal instrument. In this graph, we see that

1. Relevance holds: $D \not\perp\!\!\!\perp IV|S=1$ which implies $P(D|IV=1,S=1) \neq P(D=1|IV=0,S=1)$.
2. Ignorability holds: $Y_d \perp\!\!\!\perp IV|S=1$.
3. Three other key conditions hold that will be used in the following identification result: $Y \perp\!\!\!\perp IV|U^*,S=1$, $D \perp\!\!\!\perp IV|U^*,S=1$, and $(Y_d, D_{u^*}) \perp\!\!\!\perp U^*|S=1$.

The last three conditions are extra conditions that are required for identification in what follows, but, as we saw above, are not necessary for all instrumental variables approaches. Note that if there were a path like $IV \to D$, the first two of these would be violated. The last of these says that the $U^* \to D$ relationship is not confounded and that ignorability holds for U^*. So careful thought about the identifying assumptions you are able and willing to make is crucial.

We start by seeing that

$$
\begin{aligned}
&\mathbb{E}[Y|IV=1,S=1]-\mathbb{E}[Y|IV=0,S=1]\\
&=[\mathbb{E}[Y|U^*=1,IV=1,S=1]P(U^*=1|IV=1,S=1)+\mathbb{E}[Y|U^*=0,IV=1,S=1][1-P(U^*=1|IV=1,S=1)]]\\
&-[\mathbb{E}[Y|U^*=1,IV=0,S=1]P(U^*=1|IV=0,S=1)+\mathbb{E}[Y|U^*=0,IV=0,S=1][1-P(U^*=1|IV=0,S=1)]]\\
&=[\mathbb{E}[Y|U^*=1,S=1]P(U^*=1|IV=1,S=1)+\mathbb{E}[Y|U^*=0,S=1][1-P(U^*=1|IV=1,S=1)]]\\
&-[\mathbb{E}[Y|U^*=1,S=1]P(U^*=1|IV=0,S=1)+\mathbb{E}[Y|U^*=0,S=1][1-P(U^*=1|IV=0,S=1)]]\\
&\text{by } Y \perp\!\!\!\perp IV|U^*,S=1\\
&=\mathbb{E}[Y|U^*=1,S=1]P(U^*=1|IV=1,S=1)-\mathbb{E}[Y|U^*=1,S=1]P(U^*=1|IV=0,S=1)\\
&-\mathbb{E}[Y|U^*=0,S=1]P(U^*=1|IV=1,S=1)+\mathbb{E}[Y|U^*=0,S=1]P(U^*=1|IV=0,S=1)\\
&=[\mathbb{E}[Y|U^*=1,S=1]-\mathbb{E}[Y|U^*=0,S=1]]\,[P(U^*=1|IV=1,S=1)-P(U^*=1|IV=0,S=1)]
\end{aligned}
$$

And similarly we see that

$$
\begin{aligned}
&\mathbb{E}[D|IV=1,S=1]-\mathbb{E}[D|IV=0,S=1]\\
&=[\mathbb{E}[D|U^*=1,IV=1,S=1]P(U^*=1|IV=1,S=1)+\mathbb{E}[D|U^*=0,IV=1,S=1][1-P(U^*=1|IV=1,S=1)]]\\
&-[\mathbb{E}[D|U^*=1,IV=0,S=1]P(U^*=1|IV=0,S=1)+\mathbb{E}[D|U^*=0,IV=0,S=1][1-P(U^*=1|IV=0,S=1)]]\\
&=[\mathbb{E}[D|U^*=1,S=1]P(U^*=1|IV=1,S=1)+\mathbb{E}[D|U^*=0,S=1][1-P(U^*=1|IV=1,S=1)]]\\
&-[\mathbb{E}[D|U^*=1,S=1]P(U^*=1|IV=0,S=1)+\mathbb{E}[D|U^*=0,S=1][1-P(U^*=1|IV=0,S=1)]]\\
&\text{by } D \perp\!\!\!\perp IV|U^*,S=1\\
&=\mathbb{E}[D|U^*=1,S=1]P(U^*=1|IV=1,S=1)-\mathbb{E}[D|U^*=1,S=1]P(U^*=1|IV=0,S=1)\\
&-\mathbb{E}[D|U^*=0,S=1]P(U^*=1|IV=1,S=1)+\mathbb{E}[D|U^*=0,S=1]P(U^*=1|IV=0,S=1)\\
&=[\mathbb{E}[D|U^*=1,S=1]-\mathbb{E}[D|U^*=0,S=1]]\,[P(U^*=1|IV=1,S=1)-P(U^*=1|IV=0,S=1)]
\end{aligned}
$$

These two together mean that

$$
\begin{aligned}
&\frac{\mathbb{E}[Y|IV=1,S=1]-\mathbb{E}[Y|IV=0,S=1]}{\mathbb{E}[D|IV=1,S=1]-\mathbb{E}[D|IV=0,S=1]}\\
&=\frac{\mathbb{E}[Y|U^*=1,S=1]-\mathbb{E}[Y|U^*=0,S=1]}{\mathbb{E}[D|U^*=1,S=1]-\mathbb{E}[D|U^*=0,S=1]}\\
&=\frac{\mathbb{E}[Y_0+(Y_1-Y_0)D|U^*=1,S=1]-\mathbb{E}[Y_0+(Y_1-Y_0)D|U^*=0,S=1]}{\mathbb{E}[D|U^*=1,S=1]-\mathbb{E}[D|U^*=0,S=1]}\\
&=\frac{\mathbb{E}[Y_0+(Y_1-Y_0)D_{u^*=1}|U^*=1,S=1]-\mathbb{E}[Y_0+(Y_1-Y_0)D_{u^*=0}|U^*=0,S=1]}{\mathbb{E}[D_{u^*=1}|U^*=1,S=1]-\mathbb{E}[D_{u^*=0}|U^*=0,S=1]}\\
&=\frac{\mathbb{E}[Y_0+(Y_1-Y_0)D_{u^*=1}|S=1]-\mathbb{E}[Y_0+(Y_1-Y_0)D_{u^*=0}|S=1]}{\mathbb{E}[D_{u^*=1}|S=1]-\mathbb{E}[D_{u^*=0}|S=1]}\ \text{by } (Y_d,D_{u^*}) \perp\!\!\!\perp U^*|S=1\\
&=\frac{\mathbb{E}[Y_0+Y_1D_{u^*=1}-Y_0D_{u^*=1}-Y_0-Y_1D_{u^*=0}+Y_0D_{u^*=0}|S=1]}{\mathbb{E}[D_{u^*=1}-D_{u^*=0}|S=1]}\\
&=\frac{\mathbb{E}[(Y_1-Y_0)(D_{u^*=1}-D_{u^*=0})|S=1]}{\mathbb{E}[D_{u^*=1}-D_{u^*=0}|S=1]}\\
&=\frac{1}{\mathbb{E}[D_{u^*=1}-D_{u^*=0}|S=1]}\times\\
&[\mathbb{E}[(Y_1-Y_0)(D_{u^*=1}-D_{u^*=0})|D_{u^*=1}-D_{u^*=0}=1,S=1]P(D_{u^*=1}-D_{u^*=0}=1|S=1)+\\
&\mathbb{E}[(Y_1-Y_0)(D_{u^*=1}-D_{u^*=0})|D_{u^*=1}-D_{u^*=0}=0,S=1]P(D_{u^*=1}-D_{u^*=0}=0|S=1)+\\
&\mathbb{E}[(Y_1-Y_0)(D_{u^*=1}-D_{u^*=0})|D_{u^*=1}-D_{u^*=0}=-1,S=1]P(D_{u^*=1}-D_{u^*=0}=-1|S=1)]\\
&=\frac{\mathbb{E}[Y_1-Y_0|D_{u^*=1}-D_{u^*=0}=1,S=1]P(D_{u^*=1}-D_{u^*=0}=1|S=1)}{P(D_{u^*=1}-D_{u^*=0}=1|S=1)}\ \text{by monotonicity}\\
&=\mathbb{E}[Y_1-Y_0|D_{u^*=1}-D_{u^*=0}=1,S=1]
\end{aligned}
$$

And this is the average treatment effect for "compliers" in the selected sample, where compliers are defined with respect to the causal instrument, not the ancestral instrument, as the units in the selected sample that take the treatment when encouraged by the causal instrument but not otherwise.

B Exercises

B.1 Exercise: Internal Selection Graphs for Hughes et al. (2019) Figure 1

In this section, we show how internal selection graphs and our graphical criteria provide a simple way to visually analyze all the violations to ignorability that Hughes et al. (2019) discuss thorugh simulation and verbal description. We aim to showcase the benefit of using our formal graphical framework, as opposed to case by case analysis and simulation, for understanding how sample selection can influence instruments. Note that in this discussion, we treat C as observed and U as unobserved. Hughes et al. (2019) consider versions of instrumental variables where they do and do not condition on C. Here, we assume that C is used to block generalized non-causal paths when necessary, as in Figure 7**C**. Further, all of these examples meet the relevance criterion. We emphasize that internal selection graphs and our graphical criteria are by no means limited to the examples in Hughes et al. (2019) Figure 1 and provide a comprehensive framework with which researchers can evaluate threats to the internal validity of an instrumental variables approach posed by sample selection for any causal graph and any selection mechanism. No simulation is necessary; nor is any verbal description of the relationships between variables. Also note that while these internal selection graphs can include many bridges, we can determine that ignorability will not hold as soon as we find a single generalized non-causal path that cannot be blocked by covariate adjustment. In these graphs, this can be determined fairly quickly. The internal selection graphs make clear that sample selection on the treatment and/or outcome can create a host of purely statistical relationships in the selected data. These relationships are not immediately clear in more standard causal graphs.

Further, as the internal selection graphs show, many of these examples are redundant in terms of highlighting different implications for instruments from sample selection. Selection only on the treatment or only on the outcome violates ignorability and adding additional edges does not change this. Additionally, none of these examples consider indirect relationships between the variables and selection. Such relationships can also threaten ignorability and are discussed in other sections of this paper. Hughes et al. (2019) also do not consider all possible direct relationships between the variables. For example, selection based on IV and U is not discussed, but this would violate ignorability in a way that no covariate adjustment could resolve. So the analysis by these authors, while shedding some light on the implications of sample selection for instrumental variables, covers a quite limited range of settings that researchers might want to consider. While the authors state that "it was not possible to investigate all possible selection mechanisms even for a single IV analysis example", we argue that our approach can address the needs of researchers, who do not need to consider the implications of sample selection for all possible causal graphs but only those they deem plausible.

In Figures 7, 8, and 9, the first graph in each row is taken directly from Hughes et al. (2019) Figure 1. The second is the corresponding internal selection graph. The examples in Figure 7 all satisfy both relevance and ignorability, where we must condition on C in Figure 7**C** to satisfy ignorability. We can easily see that there are violations of condition 3 of our ignorability criterion in all the examples in Figures 8 and 9, since the generalized non-causal path $IV - U \rightarrow Y$ (along with other paths) cannot be blocked. Our hope is that this section highlights that the graphical framework developed in the present paper goes quite a bit beyond other recent investigations in terms of exploration of how sample selection can influence instrumental variables approaches, makes such analysis simple and graphical, and can illuminate connections not available otherwise.

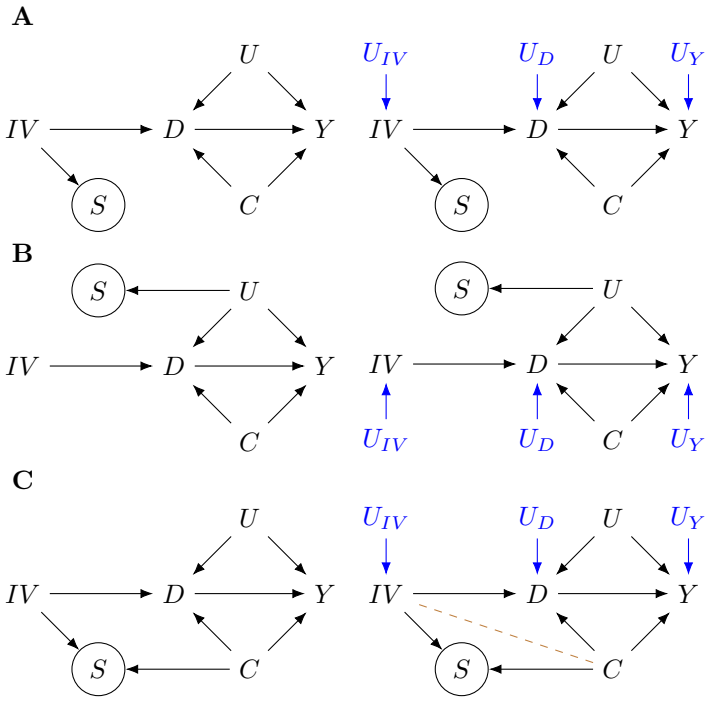
B.2 Exercise: Trust and Support for Redistribution

Peyton (2020) investigates whether political trust increases support for re-distributive policies. The author uses an instrumental variables design using online survey data in which individuals are randomly assigned into instrument groups. The instrument groups are aimed at altering the participant's political trust. Resulting measures of trust in government from the participants are then used as the treatment variable and measures of support for redistribution are used as the outcome. The design can be drawn as in Figure 10. The author conducts placebo experiments to test the exclusion restriction, and randomization ensures no common causes of instrument group and outcome.

The author uses data from three online surveys (two conducted by Amazon Mechanical Turk and one conducted by Qualtrics Panels) in which subjects completed a short survey covering their demographics and partisanship and were then randomly assigned into one of three groups: honest, corrupt, or control. The honest group was provided materials (editorial articles and in one experiment data visualizations) that "emphasized the integrity of government officials and low levels of political corruption." The corrupt group was provided materials that "used contrasting language about the lack of integrity among government officials and the prevalence of political corruption." The control group was provided materials unrelated to politics (related to Anthony Bourdain in two experiments and to recycling in the other). The online appendix for Peyton (2020) includes details on how trust and support for redistribution were measured.

Let us consider the sample selection mechanism. First, while Peyton (2020) states that "MTurk workers tend to skew white, educated, and liberal (see Berinsky, Huber and Lenz, 2012)," this is not necessarily a threat to the internal validity of instrumental variables estimates. For the purposes of this exercise, we assume that these characteristics are not related to attrition from the sample and that the measured versions of these are sufficient to block an violations of ignorability that are

Figure 7: Hughes et al. (2019) Figure 1 A, B, C: Ignorability is Satisfied



unrelated to sample selection. Peyton (2020) shows that the Qualtrics Panels experiment is approximately representative of the US general population.[15]

We turn to components of the sample selection mechanism that might be more threatening to ignorability. In two of three experiments, small numbers of participants "asked to have their data removed from the experiment after learning about deception in the debrief." These amounted to 4.3% (29/672) of participants in experiment and 8.8% (128/1452) of participants in experiment 2. Participants were not allowed to ask to be excluded after the third experiment. These participants were excluded from the study. In both cases, the author finds no statistical evidence that instrument group assignment is associated with these removal requests. It is important to note that the author does not present whether these removal requests were associated with different levels of trust in government or in support for redistribution. These associations (with sample selection and treatment and outcome) are potential threats to internal validity. Though since the instrument effects the treatment and outcome there could be a relationship with the instrument if selection is related strongly to treatment or outcome. However, if the relationship is relatively small or if the first stage is sufficiently weak, we will not see a relationship with the instrument and these individuals.

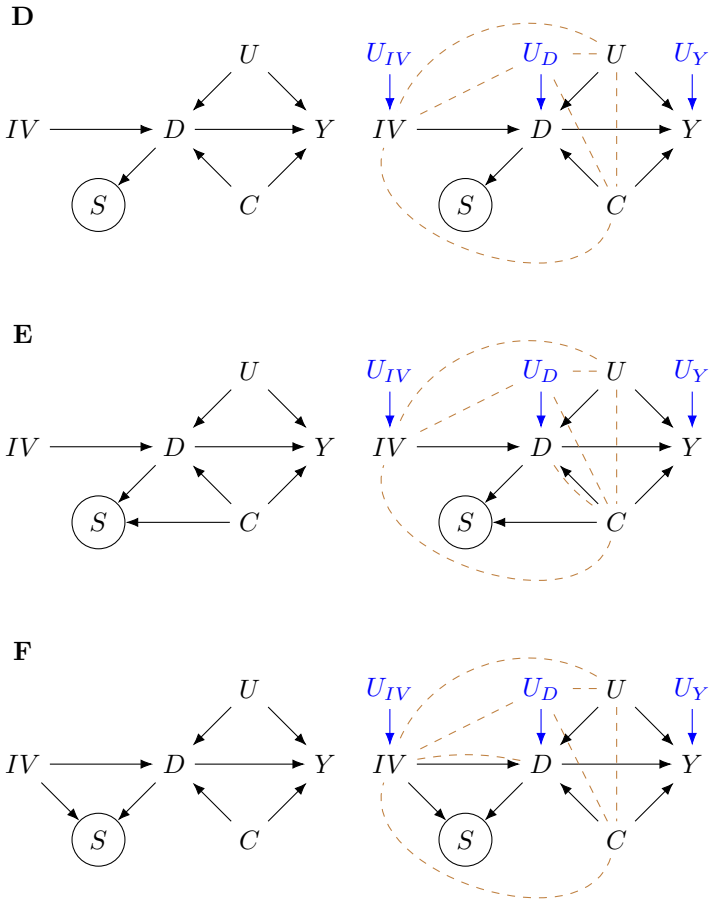
In addition to these notes, we might consider whether pre-existing social, political, and institutional trust are common causes of both political trust as measured in the experiments (after the participants reviewed instrument group material) and of self-selection into the sample. The argument might be that social, political, and institutional trust extend to trust of polling, survey, and academic studies, which would then alter how likely you might be in participating in an experiment like those in Peyton (2020) in the first place and also to asking to be removed from the sample later on. We might also wonder if, as we discuss above, the political trust measured in the experiments was a direct cause of people to ask to be removed from the sample. It seems plausible that pre-existing trust levels might influence the participation decision and certainly would be possible for the "mid experiment" trust to be an influence on asking to be removed from the sample. It also may be that some of the common causes of trust and support for redistribution might also be causes of deciding whether to participate in the study.

Let's look at how these different sample selection mechanisms might threaten ignorability by extending the causal graph to be an internal selection graph and considering our graphical criteria. See Figure 11. In all cases, relevance will not be a problem, and we focus on ignorability. In Figure 11, W represents pre-existing social, political, and institutional trust; U is all unobserved factors; and S is sample selection (participating in the experiments).

In Figure 11(a), we look at the case where pre-existing social, political, and institutional trust are common causes of both political trust as measured in the experiments and of self-selection into the sample and where some of the common causes of

[15] "All quotas were approximately met, so this sample is a reasonable approximation to a nationally representative sample of Americas [sic] on these observables."

Figure 8: Hughes et al. (2019) Figure 1 D, E, F: Selection on the Treatment



trust and support for redistribution also influence deciding whether to participate in the study. In Figure 11(b), we look at the case where political trust measured in the experiments was a cause of people to ask to be removed from the sample. In Figure 11(c), we look at the combination of these.

Based on the evidence cited by the author (i.e., that there was no statistical relationship between the instrument and attrition), it is likely that the purely statistical relationships created as a result of direct selection on the treatment are likely to be small. So Figure 11(a) might present the most realistic causal model. In this case, it turns out that we can indeed satisfy our ignorability criterion, as there are no generalized non-causal paths from IV to Y and all causal paths run through D. So the instrumental variables approach in Peyton (2020) do not seem to face much of a threat to ignorability from sample selection. Moreover, the instrumental variables design actually allows Peyton (2020) to overcome non-causal relationships between the treatment and outcome that result from confounding and sample selection. We emphasize this is perhaps an underappreciated feature of instrumental variables - that it can overcome certain forms of sample selection bias.

This application provides a good example of what might threaten internal validity and ignorability. If the individuals that asked to be excluded from the sample had been strongly associated with the instrument groups or if they are associated with the treatment, despite not being associated with the instrument groups, and this was caused by their political trust measured in the experiments, then ignorability would not hold, as is shown in both Figure 11(b) and (c). In these cases, only conditioning on the unobserved factors, U, would allow us to meet the ignorability criterion. However, the fact that these factors are unobserved was the reason for using an instrumental variables design in the first place. In the next section, we put aside the fact that it is unlikely that there was attrition based directly on the treatment in reality and simulate how things might change if this did occur.

Figure 9: Hughes et al. (2019) Figure 1 G, H, I: Selection on the Outcome

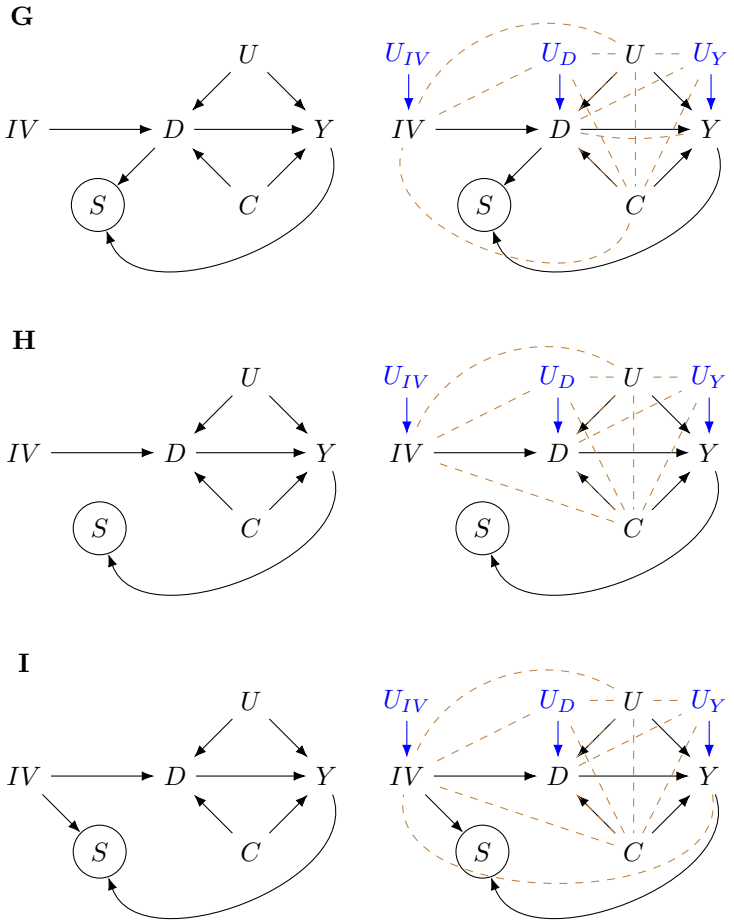
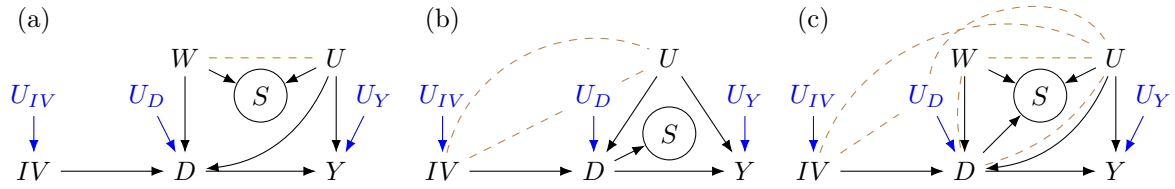


Figure 10: Peyton (2020) DAG

Figure 11: Possible Peyton (2020) Internal Selection Graphs



B.2.1 Selection on Treatment Simulations

We use the results and data from Peyton (2020) as inspiration for simulations in which we impose strong selection on the treatment. This will allow us to see how selection on the treatment can bias effect estimates in a fairly realistic setting. First, we replicate the first stage, reduced form, and 2SLS estimates from the Petyon paper. We directly use the code and data

provided in the supplementary materials for Peyton (2020). This can be seen in the first three rows of Table 6. We see a strong first stage (FS) but no effect for the reduced form (RF) or local average treatment effect (LATE; the instrumental variables estimate).[16]

Our simulations start by fitting a simple linear model for both the treatment and the outcome to the control units (in Peyton (2020) the control units received the placebo treatment with materials not related to political trust). We then use these models to simulate the control potential treatments and outcomes for all units. Next, we simulate the observed treatments as the control potential treatments plus the constant first stage effect we estimated from the experimental data (multiplied by a centered version of the instrument) plus noise. we simulate the observed outcomes as control potential outcomes plus the constant treatment effect we estimated from the experimental data (multiplied by our simulated observed treatments) plus noise. We also simulate an unobserved confounder between the treatment and outcome. This confounder is important, as selection on the treatment will create a purely statistical relationship between this confounder and the instrument. At this point, we have simulated treatments and outcomes in which there are constant first-stage and treatment effects.

The rest of the exercise consists of looking at different types of sample selection to see how the resultant violations (or not) of ignorability bias effect estimates. We first estimate the reduced form and average treatment effect under no selection bias (i.e., using the full sample). These results can be seen in the fourth and fifth rows of Table 6. We again see no reduced form or average treatment effect.

Next we estimate the reduced form and average treatment effect with selection on the treatment. The selection mechanism gives individuals with less than the median value for the treatment (trust) a 20% chance of being selected. The selection mechanism gives individuals with greater than or equal to the median value for the treatment an 80% chance of being selected. Selection here might be thought of as low levels of trust causing individuals to not trust the researchers and hence asking to be excluded from the study more frequently. We recognize that the selection mechanism is extreme and not necessarily realistic. However, our goal is to demonstrate how selection can bias effect estimates, not to be entirely realistic. These results can be seen in the sixth and seventh rows of Table 6. Here, we see a negative reduced form and average treatment effect for which the confidence intervals do not include zero.

Finally, we also estimate the reduced form and average treatment effect from the data selected based on the treatment but where we also adjust for the confounder we built in. Our ignorability criterion shows that if we condition on this, we will achieve ignorability. In practice we would not be able to condition on this sort of variable, as these are generally the reason we need to use instrumental variables in the first place. On that note, we also estimate the average treatment effect using simple covariate adjustment (not instrumental variables). These results can be seen in the last three rows of Table 6. As expected, we again see no effects.

Through this exercise we've seen how sample selection on the treatment can bias effect estimates for instrumental variables by violating ignorability both graphically and through simulation. We've also seen how, if it is possible to gather data on the right covariates, we might be able to statistically adjust for violations of ignorability. In the following section, we'll dig into some more interesting cases.

Table 6: Peyton (2020) Selection on Treatment Simulation

Selection	Estimand	Estimate	SE	CI Low	CI High	DF
True Value in Simulation (i.e., Peyton Results)	FS	0.611	0.041	0.532	0.691	3716
	RF	0.006	0.032	-0.057	0.070	3718
	LATE	0.011	0.053	-0.093	0.116	3716
No Selection Simulation	RF	-0.015	0.063	-0.138	0.109	3718
	ATE	-0.025	0.107	-0.234	0.185	3718
Selection on Treatment Simulation	RF	-0.240	0.081	-0.399	-0.081	1884
	ATE	-0.477	0.191	-0.851	-0.102	1884
Selection on Treatment Simulation, Adjust for U	RF	0.008	0.011	-0.014	0.030	1883
	ATE	0.013	0.019	-0.024	0.051	1883
	ATE – Covariate Adjustment Only	0.026	0.014	-0.002	0.054	1883

[16] See [CITES] for discussion of first stage, reduced form, and local average treatment effects.

C Technical Appendix

Here we provide technical details and prove the results found in the main text. First, we introduce a series of definitions. These are followed by a series of lemmas. Finally we state our main results in a set of theorems that follow directly from the lemmas.

C.1 Definitions

Definition C.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.*

Definition C.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 C.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 C.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 C.11 below) for $p(v)$.*

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

Definition C.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.*[17]

Definition C.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 C.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. 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.

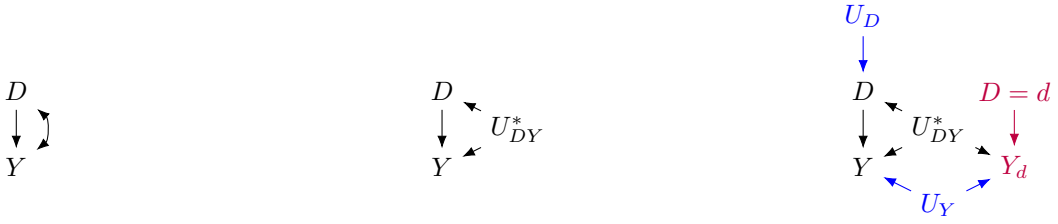
[17] We do not consider a node to be a descendant of itself.

Definition C.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 12 for a simple example. Correa et al. (2018) make a similar alteration to the causal graphs they consider.

Figure 12: Twin network with no bidirected edges

(a) DAG with bidirected edge (b) DAG without bidirected edge (c) Twin Network without bidirected edge



Definition C.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: $A \to C \leftarrow B$.*

Definition 2 (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. Add to G_S any node which is a parent of the potential instrument or a parent of a descendant of the potential instrument. (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*. (These graphs are similar to those discussed in Daniel et al. (2012) and Rohde and Hazlett (20XX).)

Definition C.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. 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 C.7 and C.8 that shows 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.

Twin network graphs can become pretty complicated, even when the original causal graph only contains three nodes. This is what makes graphical criteria like the one presented in this paper attractive for simplifying the analysis that leads to ignorability statements. 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.

Definition C.15 (Paths and Generalized Non-Causal Paths)**.** *We revise Definition C.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 C.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.*

Definition C.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 3 (Relevance Criterion)**.** A set of nodes X and a possible instrument IV in G_S^+ satisfy the relevance criterion relative to D (treatment), and Y (outcome) if there is at least one (*causal or generalized non-causal*) path between IV and D that does not pass through S and is not blocked by X.

Definition 4 (Ignorability Criterion)**.** A set of nodes X and a possible instrument IV in G_S^+ satisfy the ignorability criterion relative to D (treatment), and Y (outcome) if

1. No element of $\{X, S\}$ is a descendant of D and D is not in $\{X, S\}$.
2. X blocks every (*causal and generalized non-causal*) path between IV and Y except
   (a) those that pass through S and
   (b) those ending with a causal path from D to Y (e.g., paths between IV and Y that pass through D but where D or one of its descendants touches a bridge or paths on which D is an ancestor of IV must be blocked by X).

C.2 Lemmas

Lemma C.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 C.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.*

Lemma C.3. *If X and a possible instrument IV in G_S^+ satisfy the relevance criterion relative to D (treatment) and Y (outcome), then X does not d-separate D and IV in G_S^+.*

Proof. D and IV are d-separated in G_S^+ if and only if there are no unblocked paths connecting them. Consider an internal selection graph G_S^+ in which there are no paths between IV and D other than those that run through S or that are blocked by X. In such a graph, any path between IV and D that runs through S on which S is not a collider is blocked as a result of our having to condition on S and any path running through S on which S is a collider (or a descendant of a collider) corresponds to a path that is identical with the exception that the edges forming the collider are replaced with a bridge connecting the immediate parents of the collider. Any path like this that could connect D to IV must then be blocked by X by our construction. And by construction, all other paths that might connect D to IV are also blocked by X. In such a graph we can clearly see that IV and D are d-separated and we violate the relevance criterion. If we take the same graph and add one or more paths between IV and D that do not run through S and are not blocked by X, then IV and D are not d-separated and we also satisfy the relevance criterion. □

Lemma C.4. *If X does not d-separate D and IV in G_S^+, then $\{X, S\}$ does not d-separate D and IV in G_S.*

Proof. If X does not d-separate D and IV in G_S^+, then there is a path that connects D to IV on which S does not appear that is not blocked by X. This is because any path that passes through S is either blocked, when S is not a collider, or, when S is a collider or a descendant of a collider, corresponds to a path that is identical except that that the edges forming the collider are replaced with a bridge connecting the immediate parents of the collider. Such paths cannot be blocked by X since X does not d-separate D and IV in G_S^+. Given a path that connects D to IV on which S does not appear and that is not blocked by X, we can find the corresponding path in G_S (which may include S as a collider or a descendant of a collider). This path will then also not be blocked when we condition on $\{X, S\}$ since S can only be either a collider or a descendant of a collider on this path in G_S and we know that X does not block it. Therefore, there is a path between D and IV in G_S that is not blocked by $\{X, S\}$ and so $\{X, S\}$ does not d-separate D and IV in G_S. □

Lemma C.5. *If $\{X, S\}$ does not d-separate D and IV in G_S, then $D \not\perp\!\!\!\perp IV | X, S = 1$ for every model inducing G_S.*

Proof. This follows from Lemma C.1 and Definitions C.11 and C.12. □

Lemma C.6. *If X and a possible instrument IV in G_S^+ satisfy the ignorability criterion in G_S^+ relative to D (treatment) and Y (outcome), then X d-separates IV and Y_d in N_S^+.*

Proof. We 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 X, an unblocked path from IV to Y_d in N_S^+ implies that the ignorability criterion is violated in G_S^+.

Let π be an unblocked path from IV to Y_d in N_S^+. We assume without loss of generality that π intersects IV only at the endpoint. Elements of X are only in the pre-intervention world, since we can only condition on observed variables never counterfactual variables. So no descendant of D in the post-intervention world is conditioned on. Any unblocked path from IV to Y_d that "lands" in the post-intervention world must descend, along arrows pointing to Y_d, to Y_d. Therefore, an unblocked path from IV to Y_d in N_S^+ has three parts: π_1 (an unblocked path in G_S^+), π_3 (a causal path in $(G_S)_{\overline{D}}$ on which every node is a descendant of D), and π_2 (a single edge connecting π_1 and π_3 in N_S^+). N_S^+ contains copies of nodes in G_S^+: one copy corresponding to the variable in the pre-intervention world (G_S^+) and one copy corresponding to the variable in the post-intervention world ($(G_S)_{\overline{D}}$). So π may contain two such copies that refer to the same node in G_S^+.

Sample selection means we condition on $S = 1$. This is a pre-intervention world or observed 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 S, though π may contain bridges in π_1. Post-intervention versions of S cannot appear on π_3, since for this to occur S must be a descendant of D, otherwise it would not be in the post-intervention world, but this is a violation of the criterion.

Let π' be a route in G_S^+ created by the following two steps:

1. replace all occurrences of post-intervention variables in π by the nodes in G_S^+ these variables correspond to, coupled with the appropriate occurrence numbers
2. replace all occurrences of a single variable twice in a row in the sequence from above by a single occurrence of that node (with the occurrence numbers and all subsequent occurrence numbers appropriately decremented)

Let π^* be the direct route of π' in G_S^+. The portions of π' corresponding to π_1 and π_3 in π must be active. The remaining portion of π' is node triple, where the middle node had been pointed to by π_2 in π. The second edge in this triple must be pointing away from the middle node, since all edges in π_3 point toward Y_d. If this node triple exists in π (it may not if there are no edges in π_3), it must also be active in π'. For example, say that G_S^+ contains $IV \to D \to Y$ and $IV \to D \to X \to Y$ and sample selection does not connect to any other node. Then suppose that π is taken to be $IV \to D \to X \leftarrow U_X \to X_d \to Y_d$. Here π_1 is $IV \to D \to X \leftarrow U_X$, π_2 is the edge between U_X and X_d, and π_3 is $X_d \to Y_d$. So π' is $IV \to D \to X \leftarrow U_X \to X \to Y$. The node triple in π' that does not correspond to π_1 or π_3 is $U_X \to X \to Y$. This is blocked since we condition on X. However, conditioning on X is a violation of the criterion since X is a descendant of D 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 the criterion 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 the criterion. If the node triple is not active, it is like the previous

example or was obtained from four consecutive nodes in π, with the two middle nodes being node copies connected by π_2. Since this involves a post-intervention copy (and hence a node on a causal path from $D = d$ to Y_d), this means that these middle nodes cannot intersect or be ancestral to X or they would violate the criterion. (e.g., say we have a bidirected arrow in the previous example rather than U_X and so π is $IV \to D \to X \leftrightarrow X_d \to Y_d$ and π' is $IV \to D \to X \to Y$, which is also the node triple in π' that does not correspond to π_1 or π_3 and is blocked since we condition on X; but again this is a violation of the criterion.) Either the node triple is active or it isn't. But, if it isn't, then it could only have resulted from a violation of the criterion. So π' is an active route. By Lemma C.2, π^* is an active path in G_S^+.

If π^* does not end with a causal path from D to Y, then we immediately violate the criterion since such paths must be blocked by X. If π^* does end with a causal path from D to Y, then we must consider how such a π^* could have arisen from π. Since no edges can point into the post-intervention copy of D, the copy of D that we see on π^* must have resulted from the pre-intervention copy of D being on π_1. If π^* ends with a causal path from D to Y, then we assume without loss of generality that it is a proper causal path from D to Y. Since π^* ends with a causal path from D to Y, the first edge in π between D and Y must be a directed edge pointing away from an element in D. If π_2 was bidirected in π, then the only way π could be unblocked in N_S^+ would be for the second node in π between D and Y to be an ancestor of X (or a member of X itself). By construction of π^*, the second node in π is also the second node in π^*. This is a violation of the criterion. But we've also replaced all the bidirected edges in the twin network, so this case should not appear. If π_2 was directed in π, then the nodes it connects in G_S^+ are a parent-child pair, with the parent copy (node P) in the G_S^+ part of N_S^+ and the child copy (node C) in the $(G_S)_{\overline{D}}$ part of N_S^+. So P cannot be a descendant of D, otherwise it would be in the $(G_S)_{\overline{D}}$ part of N_S^+. If π_1 and π_3 do not share nodes (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^+. Otherwise, P would have to be a descendant of D, a contradiction. If π_1 and π_3 share nodes, then the only way to reach P from D is via a collider unblocked by X. This would mean that the second node in π between D and Y (and the second node in π^* between D and Y) is an ancestor of X, which violates the criterion. □

Lemma C.7. *If X d-separates IV and Y_d in N_S^+, then $\{X, S\}$ d-separates IV and Y_d in N_S.*

Proof. We very closely follow the proof of Lemma 3 in the Web Appendix for Daniel et al. (2012), with changes for sample selection. Suppose that this statement is false. Then there must be a twin network, N_S, and a X for which there is a path in N_S from IV to Y_d that is not blocked by $\{X, S\}$ but all paths from IV to Y_d in N_S^+ are blocked by X. Let p be such a path in N_S. p 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 p is not blocked after conditioning on $\{X, S\}$ in N_S but is blocked after conditioning on X in N_S^+, then either

- A variable in p is a member of X (and hence blocks p in N_S^+) but is not a member of $\{X, S\}$ (and hence does not block p in N_S). But this is a contradiction, since $X \subset \{X, S\}$.
- p contains a collider, C, such that either C is in $\{X, S\}$ or has descendants in $\{X, S\}$ (so that p is not blocked in N_S), but C is not in X and does not have descendants in X (so that p is blocked in N_S^+). This means that p must contain a collider C with either $C = S$ or C is an ancestor of S. In either case, p corresponds to a generalized non-causal path, p', in N_S^+, which is identical to p except that C is not on p' and the parents of C on p are connected with a bridge on p'. If p is not blocked in N_S then p' must not be blocked in N_S^+, since none of the variables (except for C) on p is in $\{X, S\}$. Thus, none of the variables on p' is in X. This is a contradiction.

□

Lemma C.8. *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 C.7. □

Lemma C.9. *If $\{X, S\}$ d-separates IV and Y_d in N_S, then $Y_d \perp\!\!\!\perp IV | X, S = 1$ for every model inducing G_s.*

Proof. This follows from Lemma C.1 and Definitions C.11 and C.12. □

C.3 Theorems

Theorem 1. *If a set of nodes X and a possible instrument IV in internal selection graph G_S^+ satisfy the relevance criterion relative to D (treatment), and Y (outcome), then $D \not\perp\!\!\!\perp IV | X, S = 1$.*

Proof. Lemmas C.3, C.4, and C.5 prove the result. □

Theorem 2. *If a set of nodes X and a possible instrument IV in internal selection graph G_S^+ satisfy the ignorability criterion relative to D (treatment), and Y (outcome), then $Y_d \perp\!\!\!\perp IV | X, S = 1$.*

Proof. Lemmas C.6, C.7, and C.9 prove the result. □

C.4 Violations of exclusion restriction and the definition of instruments

In this section we shed some light on the usefulness of the specific definition of instruments that we use. Suppose that we have the causal graph in Figure 13(a). If we consider the definition of an instrument in which the exclusion restriction is seperated from ignorability (i.e., $Y_{d,iv} = Y_{d,iv'} = Y_d$ and $Y_{iv,d} \perp\!\!\!\perp IV$), then we have a violation of exclusion but not ignorability in this graph. (Hernán and Robins, 2006, 2020) If we consider the definition of an instrument in which these are combined in to a single ignorability condition ($Y_d \perp\!\!\!\perp IV$ or $Y_d \perp\!\!\!\perp IV|X$), then we have a violation of ignorability.

We might consider conditioning on M to fix the problems. When we condition on M, we see that exclusion is not achieved ($Y_{d,iv} \neq Y_d$) and further we have $Y_{iv,d} \not\perp\!\!\!\perp IV|M$. However, when we condition on M, we get $Y_d \perp\!\!\!\perp IV|M$. These can be seen in Figure 13(b,c). We're not actually interested in $Y_{d,iv}$ in its own right. There is nothing that requires that conditional ignorability statements for Y_d follow those for $Y_{d,iv}$. This is just such an example where they don't follow. Indeed, conditioning on M can actually fix problems for ignorability with Y_d but creates problems for ignorability with $Y_{d,iv}$.

So we see that while it might be intuitive to consider exclusion separately from ignorability, writing the instrument definition in this way actually imposes some unnecessary limitations on the type of conditional instruments that might work. Below, our graphical criterion allows the user to think intuitively in terms of separately ruling out causal paths and non-causal paths between IV and Y, but does not impose unnecessary restrictions as a result of writing the instrument definition in a certain way. This is a key distinction since much of the literature on instruments discusses these conditions seperately.

As we show above, we can say the following for the unconditional statements: $Y_d = Y_{iv,d}, Y_{d,iv} \perp\!\!\!\perp IV \iff Y_d \perp\!\!\!\perp IV$, but this does not hold for the conditional versions of these. In particular, $Y_d \perp\!\!\!\perp IV|M \not\Longrightarrow Y_d = Y_{iv,d}, Y_{d,iv} \perp\!\!\!\perp IV|M$. Though $Y_d = Y_{iv,d}, Y_{d,iv} \perp\!\!\!\perp IV|M \implies Y_d \perp\!\!\!\perp IV|M$.

Figure 13: A Violation of the Exclusion Restriction