WEB APPENDIX 1

Analytical Approaches for Data Generated Under the Causal Graphs Presented in Figure 3

Here we formalize the consideration of analytical approaches for data generated under the assumptions encoded in the directed acyclic graphs (DAGs) presented in **Figure 3** in the text. First, we include an arrow from X' to Y to allow for an overall effect of treatment versus no treatment, a more realistic scenario than the overall sharp null assumed in the first example in the main text and an unnecessary assumption for our purposes. We also add an arrow from X'' to Y; under a treatment specific sharp null ($Y_i^{x=1} = Y_i^{x=2}$ for all subjects i), this arrow would be omitted. These augmented DAGs appear in **Web Figure 1**. We remain interested in estimating the average treatment effect comparing our two treatments: $E[Y^{x=2} - Y^{x=1}]$. Note this effect can be rewritten in terms of X'', i.e., our treatment effect of interest is X'' on Y conditional on $X' = 1$. Under the sharp null, this effect takes the value 0.

Let us first consider data generated under the DAG in **Web Figure 1C**. With an arrow from X' to Y, Z is not an instrument unconditionally: it violates condition 2, as there is a path from Z to X' to Y. When we restrict to $X' = 1$ (i.e., $X = 1$ or $X = 2$), this path gets blocked and all the instrumental conditions are satisfied conditionally: 1) $\Pr[X = 2|Z = 1, X' = 1] \neq \Pr[X = 2|Z = 0, X' = 1]$ because there is an arrow from Z to X''; 2) among those with $X' = 1$, $Y^{z,x''} = Y^{x''}$ for all values x'' and z because there is no path from Z to Y not through X'' when we condition on X'; and 3) $Z \perp\!\!\!\perp Y^{x''} | X' = 1$ for $X'' = 1$ and $X'' = 2$. Because the instrumental conditions are satisfied conditionally, we are able to obtain the effect in the restricted subpopulation using the standard instrumental variable (IV) estimator:

$$E[Y^{x=2} - Y^{x=1} | X \in \{1,2\}] = \frac{E[Y|Z = 1, X \in \{1,2\}] - E[Y|Z = 0, X \in \{1,2\}]}{\Pr[X = 2|Z = 1, X \in \{1,2\}] - \Pr[X = 2|Z = 0, X \in \{1,2\}]}$$
$$\equiv \frac{E[Y|Z = 1, X' = 1] - E[Y|Z = 0, X' = 1]}{\Pr[X = 2|Z = 1, X' = 1] - \Pr[X = 2|Z = 0, X' = 1]}$$

This depends on a further effect homogeneity assumption:

$$E[Y^{x=2} - Y^{x=1} | Z = 0, X = 1, X \in \{1,2\}] = E[Y^{x=2} - Y^{x=1} | Z = 1, X = 1, X \in \{1,2\}]$$

$$E[Y^{x=2} - Y^{x=1} | Z = 0, X = 2, X \in \{1,2\}] = E[Y^{x=2} - Y^{x=1} | Z = 1, X = 2, X \in \{1,2\}]$$

Since $Y^{x=1} \perp\!\!\!\perp X'$ and $Y^{x=2} \perp\!\!\!\perp X'$ (as there are no backdoor paths from X′ to Y), this estimator further

identifies the effect in the full study population. Note we cannot identify the average treatment effect using a non-IV method, as there is unmeasured confounding (i.e., a backdoor path from X'' to U to Y we cannot block). If we were uncomfortable making the effect homogeneity assumption, we may consider bounding the average treatment effect or adapting a monotonicity condition to identify a local average treatment effect.

Next, consider data generated under the DAG in **Web Figure 1B**. Z is not an instrument for X'' (unconditionally or conditional on X'): condition 3 would be violated if we condition on $X' = 1$ due to collider-stratification, while condition 2 would be violated if we did not condition on X' because of the path from Z to X' to Y. We can, however, identify the effect in those who received either treatment 1 or 2 (i.e., conditioning on $X' = 1$) by standardizing over Z:

$$E[Y^{x=2} - Y^{x=1} | X \in \{1,2\}] = \sum_{z \in \{0,1\}} E[Y^{x=2} - Y^{x=1} | X \in \{1,2\}, Z = z] \Pr[Z = z | X \in \{1,2\}]$$
$$= \sum_{z \in \{0,1\}} E[Y | X = 2, Z = z] \Pr[Z = z | X \in \{1,2\}]$$
$$- \sum_{z \in \{0,1\}} E[Y | X = 1, Z = z] \Pr[Z = z | X \in \{1,2\}]$$

We cannot identify the effect in the untreated $E[Y^{x=2} - Y^{x=1} | X = 3]$ because of positivity violations, e.g., $\Pr[X'' = 1 \mid X' = 0] = \Pr[X = 1 \mid X = 0] = 0$. As such, we cannot identify the average treatment effect using either an IV or non-IV approach.

Web Figure 1A merges these themes: we cannot identify the average treatment effect using an IV approach, nor can we identify the effect in those receiving treatments 1 or 2 using a non-IV approach. However, if there was no arrow from Z to X' in any of the causal diagrams presented in **Web Figure 1**, then the IV approach described above would validly identify the average treatment effect in all three scenarios.

These DAGs are somewhat simplified, as we often would have some measured covariates L that may help block some or all of the pathways from U to X' or X'' (**Web Figure 2**). In the context of equivalence randomized trials, Robins (1) considered data generated under the DAG in **Web Figure 1A**,

with a measured L_1 on the pathway from U to X'. Following from his example for identifying the intent-to-treat effect, we could intervene on X' (forcing everybody to take one of the active treatments) using inverse probability weighting with the following stabilized weights:

$$w_1 = \frac{\mathrm{f}[x'|Z = z]}{\mathrm{f}[x'|Z = z, L_1 = l_1]}.$$

The DAG for our pseudopopulation would be **Web Figure 1C**. We could then identify the average treatment effect using an IV analysis in this pseudopopulation for the reasons described above. See Section 4 of the Appendix in Robins (2) for a related discussion.

It is possible that we could measure a set of covariates L_2 that were sufficient to block the path from U to X''. If so, then we could restrict to the treated ($X' = 1$) and instead perform inverse probability weighting to create a pseudopopulation resembling **Web Figure 1B** but without the arrow from Z to X'', using the following weights:

$$w_2 = \frac{\mathrm{f}[x''|X' = 1]}{\mathrm{f}[x''|Z = z, L_2 = l_2, X' = 1]}.$$

In this pseudopopulation, the crude association would identify the effect in the treated under the reasoning described above. Note if we condition on Z in the numerator of the weights, we could instead have created a pseudopopulation that exactly reflected **Web Figure 1B**, but would then need to standardize over Z.

Finally, suppose we have sufficient measured covariates L to block both the U to X′ and U to X'' pathways, i.e., we can do a valid IV or non-IV approach. Does the IV approach offer clear additional value? First, we recognize that IV methods are inefficient relative to most non-IV methods. We also see the targeted estimands differ, and to get a point estimate under an IV method we would need to make an additional assumption not encoded in the DAG. Therefore, the choice should weigh three considerations: the efficiency of the methods, the relevance of the estimand, and the reasonableness of an additional assumption. As the additional assumptions needed for a point estimate from an IV method may not be palatable, and the IV methods are less efficient, doing an IV analysis in addition to or instead of the non-IV analysis would not have clear added utility.

Given all this, the only time an IV analysis would definitively be preferable to a non-IV analysis is if the investigators made the assumptions encoded in **Web Figure 1C** or thought they had measured all L on the pathway from U to X' but not from U to X''.

When an investigator has not measured sufficient L to block either the pathway from U to X' or U to X'', a natural question is how much bias may be incurred in an IV analysis that fails to appropriately account for U in the analysis. Our simulations (described in more detail below) provide a framework toward answering this question. Another approach may be to describe bias formulas in simplified settings. For reasons described above, we cannot repurpose formulas for collider-stratification bias to recover a corrected estimate for the IV numerator had we not restricted analyses to $X' = 1$, as an IV analysis unconditional on X' would violate the exclusion restriction if there was an arrow from X' to Y. We may instead consider repurposing a bias formula for the controlled direct effect to understand the bias in the effect of Z on Y when X' is set to 1 but we fail to adjust for U. In the case where U is binary and under strong homogeneity assumptions, this suggests that the bias in the IV numerator would be a function of 1) the prevalence difference of U across levels of the instrument among those receiving $X = 1$ or $X = 2$, or 2) the difference in the mean of Y across levels of U holding Z and X' constant (and then the bias in the IV estimate would be amplified by the strength of the instrument). However, the homogeneity conditions may often be unlikely to hold, and without a more thorough understanding of how U interacts with Z to influence treatment decisions, it is possible that such a simplified formula would be misleading. For these reasons, the practice of assessing covariate balance by levels of the instrument versus levels of the treatment may not be informative to the relative magnitude or even direction of bias in an IV versus non-IV analysis selecting on treatment. In settings where measured covariates appear well-balanced across levels of the instrument, and the measured covariates are considered good proxies for unmeasured confounders, it could seem reasonable to proceed assuming this selection bias could be minimal — although investigators should proceed cautiously as even small imbalances could be indicative of large biases. Beyond using our simulation framework, investigators may consider using measured covariates in

their dataset as proxies for unmeasured covariates to understand how their estimates change when they omit a measured covariate from the combined IPW-IV approach described above.

Using our statin therapy dataset, we estimated treatment effects of hydrophilic versus lipophilic statin therapy on diabetes risk. In **Web Table 1**, we present the valid estimates under the various sets of assumptions. If we are confident that there is no arrow from U to Z (i.e., the proposed instrument is not confounded), we can confirm the presence of the arrow from Z to X' with the observed data: the probability of receiving statin therapy differs across levels of the instrument (5% versus 6%; $P < 0.001$). As noted in the main text, there are many patient characteristics that are likely to influence the decision to treat with a statin, the decision between types of statins, and diabetes risk, thus **Web Figure 1A** is the most reasonable of these three DAGs to assume.

WEB APPENDIX 2

Description of Simulations

We can generate k samples of N patients from the following general form:

$Z_i \sim \text{bernoulli}(p_1)$

$U_i \sim \text{bernoulli}(p_2)$

$Y_i \sim \text{bernoulli}(a_0 + a_1 U_i)$

$X_i \sim \text{multinom}(b_{10} - b_{11}Z_i - b_{12}Z_iU_i,\ b_{20} + c_1(b_{11}Z_i) + c_1(b_{12}Z_iU_i),\ b_{30} + c_2(b_{11}Z_i) + c_2(b_{12}Z_iU_i))$

where $b_{10} + b_{20} + b_{30} = 1$, $c_1 + c_2 = 1$, and $0 \leq a_0,\ a_0 + a_1,\ b_{10}, b_{20}, b_{30}, c_1, c_2 \leq 1$.

We demonstrated in the main text how bias is a function of the relationship between the confounder and treatment decisions, specifically by varying b_{12}, c_1, and c_2. Bias increases with increases in the risk of the outcome and/or strength of the relationship between the confounder and outcome (a_0, a_1); note if the relationship between U and Y is flipped (e.g., a_1 is negative versus positive), the direction of the bias flips as well. A continuous confounder or outcome could instead be simulated using other distributions (e.g., normal).

To simulate a preference-based instrument, we would need to adapt these simulations to incorporate provider-level measured and unmeasured preference variables. We can generate k samples of N patients seen by M providers. Providers' preferences might be conceptualized as a vector of how strongly the provider prefers treatment 1 versus treatment 2 versus neither for a patient with the reference level of the confounders. The simulation could either allow for providers to have any possible preference with equal probability, or it could restrict this range based on descriptive data of prescribing practices. Measured preference could be as the treatment given to the prior patient seen by that provider who was prescribed either $X = 1$ or $X = 2$.

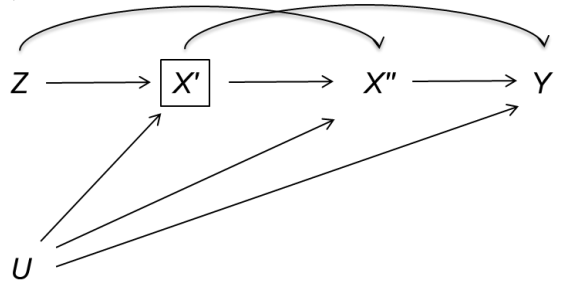
References

1. Robins JM. Correction for non-compliance in equivalence trials. *Stat Med*. 1998;17(3):269–302.

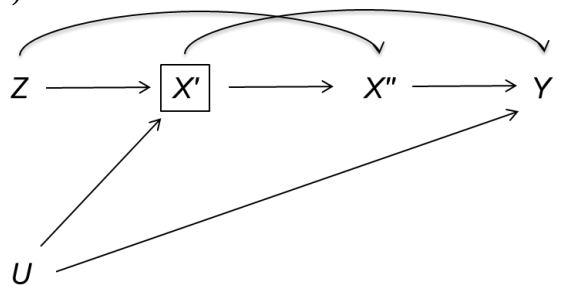
2. Robins JM. Analytic methods for estimating HIV treatment and cofactor effects. In: Ostrow DG, Kessler RC, eds. *Methodological Issues in AIDS Behavioral Research.* New York, NY: Plenum Publishing Company; 1993:213–290.

Web Figure 1. Graphical representation of possible ways an unmeasured confounder may affect treatment decisions in the presence of treatment effects. *Z* indicator of calendar time (1 if post-warning, 0 if pre-warning); *U* indicator of unmeasured confounder; *Y* outcome; *X′* indicator of receiving treatment 1 or 2 versus neither; *X″* indicator of receiving treatment 1 versus treatment 2 or neither.

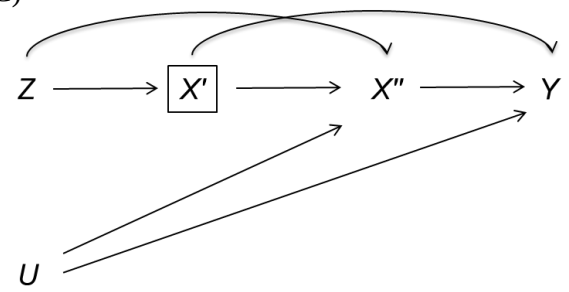
A)



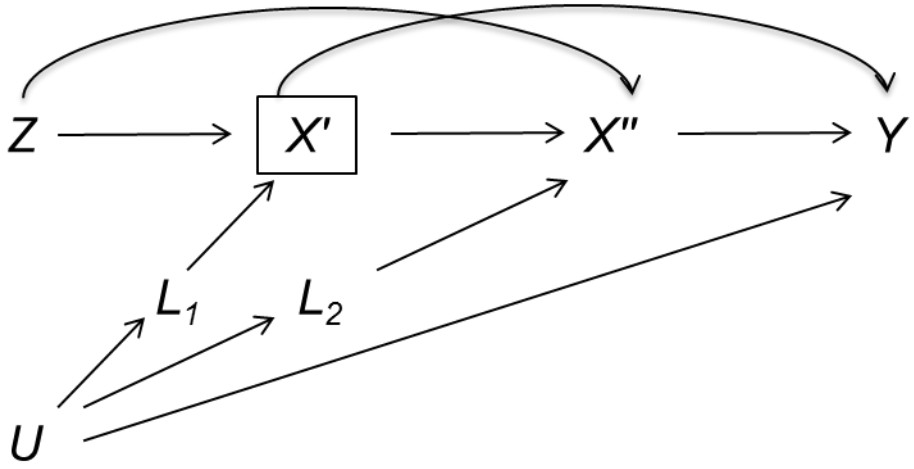
B)



C)



Web Figure 2. Graphical representation of Web Figure 1A with measured covariates.



Web Table 1. Twelve-month risk of diabetes comparing hydrophilic and lipophilic statin therapy using an IV and non-IV approach under the assumptions encoded in the DAGs shown in Web Figure 1

	Risk Differences (95% CI) Derived Using Methods Under the Assumptions Encoded in the Causal Diagram[a]		
	Web Figure 1A	**Web Figure 1B**	**Web Figure 1C**
Non-IV approach, effect in those receiving statin therapy	Invalid	-0.00 (-0.01, 0.01)	Invalid
IV approach with effect homogeneity condition, average treatment effect	Invalid	Invalid	0.02 (-0.00, 0.05)

Abbreviations: CI, confidence interval; DAG, directed acyclic graph; IV, instrumental variable.

[a] Because we measured preference with a proxy, the causal diagrams should technically be augmented to have *Z* be a surrogate (noncausal) instrument.