DE GRUYTER J. Causal Infer. 2015; 3(1): 131–137

Causal, Casual and Curious

Judea Pearl*

Conditioning on Post-treatment Variables

Abstract: In this issue of the Causal, Casual, and Curious column, I compare several ways of extracting information from post-treatment variables and call attention to some peculiar relationships among them. In particular, I contrast *do*-calculus conditioning with counterfactual conditioning and discuss their interpretations and scopes of applications. These relationships have come up in conversations with readers, students and curious colleagues, so I will present them in a question–answers format.

Keywords: causal effects, back-door condition, *do*-calculus, counterfactuals

DOI 10.1515/jci-2015-0005

Question-1 (Is Rule-2 valid?)

Rule-2 of *do*-calculus does not distinguish post-treatment from pre-treatment variables. Thus, regardless of the nature of Z, it permits us to replace $P(y \mid do(x), z)$ with $P(y \mid x, z)$ whenever Z separates X from Y in a mutilated graph $G_{\underline{X}}$ (i.e. the causal graph, from which arrows emanating from X are removed). How can this rule be correct, when we know that one should be careful about conditioning on a post-treatment variables Z?

Example 1 *Consider the simple causal chain $X \to Y \to Z$. We know that if we condition on Z (as in case control studies) selected units cease to be representative of the population, and we cannot identify the causal effect of X on Y even when X is randomized. Applying Rule-2 however we get $P(y \mid do(x), z) = P(y \mid x, z)$. (Since X and Y are separated in the mutilated graph $X \quad Y \to Z$). This tells us that the causal effect of X on Y IS identifiable conditioned on Z. Something must be wrong here.*

Answer-1

Yes, something is wrong here, but not with Rule-2. It has to do with the interpretation of $P(y \mid do(x), z)$, which will become clear when we prove the validity of Rule-2 in our graph

$$X \to Y \to Z$$

Rule-2 says:

$$P(y \mid do(x), z) = P(y \mid x, z) \qquad \text{If } X \perp\!\!\!\perp Y \mid Z \text{ in } G_{\underline{X}}$$

Indeed, if we go to the definition of $P(y \mid do(x), z)$, we obtain:

$$\begin{aligned} P(y \mid do(x), z) &= P(y, z \mid do(x))/P(z \mid do(x)) \quad \text{by def.} \\ &= P(y, z \mid x)/P(z \mid x) \qquad \text{since } X \text{ is randomized} \\ &= P(y \mid x, z) \end{aligned}$$

which proves Rule-2.

***Corresponding author: Judea Pearl,** Computer Science Department, University of California, Los Angeles, Los Angeles, CA 90095–1596, USA, E-mail: judea@cs.ucla.edu

The same result obtains whenever Z blocks all back-door paths from X to Y, as in the canonical confounding model (Figure 1(a)), as well as in the typical selection-bias model (Figure 1(b)). $P(y \mid do(x), z)$ is identified (by $P(y \mid x, z)$) in both models, despite the fact that in Figure 1(b) Z is a descendant of both treatment and outcome, in double violation of the back-door criterion. $P(y \mid do(x), z)$ is no longer estimable when conditioning on Z opens a back-door path from X to Y as in Figure 2(a), because the condition $(X \perp\!\!\!\perp Y \mid Z)G_{\underline{X}}$ is violated. It is identified in Figure 2(b), where the condition is satisfied.

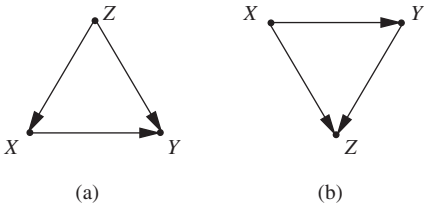

Figure 1: Two models in which $P(y \mid do(x), z) = P(y \mid x, z)$ because Z d-separates X from Y once we remove arrows emanating from X.

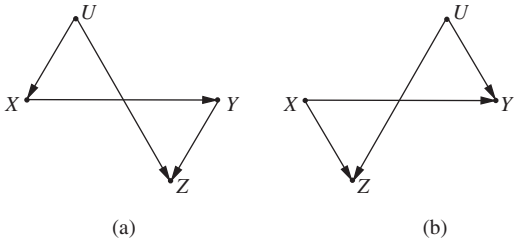

Figure 2: In Model (a), $P(y \mid do(x), z)$ is not identified (when U is unobserved) since Rule-2 in inapplicable. It is identified in Model (b) since X and Y are separated in $G_{\underline{X}}$.

Question-2 (Why back-door prohibition?)

So, when do we need to worry about conditioning on X-affected covariates, virtual colliders, case control studies, etc.? It seems that Rule-2 allows us to circumvent the prohibition that the back-door criterion imposes against conditioning on a treatment-dependent Z.

Answer-2

The two are not contradictory. Rule-2 is always valid, regardless if Z is pre-treatment or post-treatment. At the same time, the prohibition imposed by the back-door cannot be dismissed, it needs to be considered on two occasions. First, whenever we seek a license to use the adjustment formula and write:

$$P(y \mid do(x)) = \sum_z P(y \mid x, z)P(z) \tag{1}$$

Second, whenever we seek to estimate causal effects in a specific group of units characterized by $Z = z$. Contrary to syntactic appearance, the expression $P(y \mid do(x), z)$ in Rule-2, does not represent such effects when Z is post-treatment.

Let us deal with these two cases separately.

2.1 License to adjust

Consider the adjustment formula of eq. (1). This formula is not valid when Z is Y-dependent, as in our causal chain

$$G_1 : X \to Y \to Z.$$

If we apply it blindly, we get the sum in (eq. (1)), instead of the correct answer, which is $P(y \mid do(x)) = P(y \mid x)$.

To see what goes wrong with blind adjustment, let us trace its derivation, for a pre-treatment Z:

$$\begin{aligned} P(y \mid do(x)) &= \sum_z P(y \mid do(x), z) P(z \mid do(x)) \\ &= \sum_z P(y \mid x, z) P(z \mid do(x)) \qquad \text{by Rule-2} \\ &= \sum_z P(y \mid x, z) P(z) \qquad \text{since Z precedes } X \end{aligned}$$

This works fine when we can substitute $P(z \mid do(x))$ with $P(z)$, but not when Z is post-treatment and $P(z \mid do(x))$ depends on x. Thus, Rule-2 in itself is not sufficient for adjustment; blind adjustment will produce erroneous estimands.

If we avoid the substitution $P(z \mid do(x)) = P(z \mid x)$ and proceed cautiously in G_1, we get

$$\begin{aligned} P(y \mid do(x)) &= \sum_z P(y \mid do(x), z) P(z \mid do(x)) \\ &= \sum_z P(y \mid x, z) P(z \mid do(x)) \qquad \text{by Rule-2} \\ &= \sum_z P(y \mid x, z) P(z \mid x) \qquad \text{by Rule-2 on } Z \\ &= P(y \mid x) \end{aligned}$$

which is the correct answer for G_1. But this is obtained though careful derivation, not by blind adjustment.

Blind adjustment is valid, however, when Z is *pure* descendant[1] of X, as in Figure 3. We know that the back-door prohibition against post-treatment covariates is lifted in this case [1, p. 339, 2] and, indeed, if we take Z as a covariate and blindly apply the adjustment formula to G_2, we get the correct result:

$$\begin{aligned} P(y \mid do(x)) &= \sum_z P(y \mid x, z) P(z) \\ &= P(y \mid x) \end{aligned} \tag{2}$$

The latter equality is obtained through the conditional independence $P(y \mid x, z) = P(y \mid x)$ which holds in G_2.

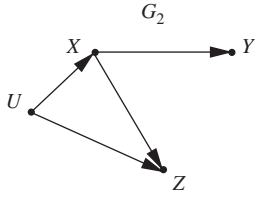

Figure 3: A model in which Z is a pure descendant of X, thus satisfying the (extended) back-door condition and permitting adjustment for Z.

1 By *pure* descendant we exclude variables that are descendants of any intermediate variable between X and Y.

2.2 Identifying unit-specific effects

We are now ready to discuss the second task for which back-door admissibility is needed: estimating unit-specific effects.

In many applications, the query of interest is not to find $Q_{do} = P(y \mid do(x), z)$, but to find $Q_c = P(y_x \mid z)$, where y_x is short for the counterfactual statement $Y_x = y$. Back-door admissibility gives us the license to equate the two queries, and get

$$P(y \mid do(x), z) = P(y_x \mid z) = P(y \mid x, z) \tag{3}$$

By the counterfactual query Q_c we mean: Take all units which are currently at level $Z = z$, and ask what their Y would be had they been exposed to treatment $X = x$. This is different from $Q_{do} = P(y \mid do(x), z))$, which means: Expose the whole population to treatment $X = x$, take all units which attained level $Z = z$ (post exposure) and report their Y' s.

We call Q_c"unit-specific" because, as x varies, Q_c remains focused on the same set of units (i.e. those that are currently at $Z = z$), with (hypothetical) histories that vary with x. Some of these units may not have experienced any of those histories and would have attained different levels of Z if they did. In contrast, Q_{do} focusses on one stratum, $Z = z$, and, as x varies, it allows different units to enter and leave that stratum.

Obviously, when Z is a pre-treatment covariate, we have $Q_{do} = Q_c$, but when Z is post-treatment, the most common question we ask is Q_c: find $P(y_x \mid z)$, not Q_{do}: find $P(y \mid do(x), z))$. The back-door criterion gives us a license to equate both queries with $P(y \mid x, z)$. Here is why: If Z satisfies the back-door condition, the First Law of causal inference[2] dictates the conditional independence $Y_x \perp\!\!\!\perp X \mid Z$, also known as "conditional ignorability" [3], so

$$P(y_x \mid z) = P(y_x \mid z, x) = P(y \mid z, x).$$

This license is similar to Rule-2, but it is applied to a different expression; whereas ignorability allows us to remove a subscript, Rule-2 allows us to remove a *do*-operator.

We can see the difference in graph G_2 of Figure 3. Here Z satisfies the (extended) back-door condition, so we can write

$$P(y_x \mid z) = P(y \mid z, x) = P(y \mid x)$$

Rule-2 in itself does not give us this license because it is applicable to a different query $P(y \mid do(x), z)$ and cannot handle counterfactual expressions.

Question-3 (the key question)

Should we be concerned with the difference between Q_{do} and Q_c? If so, when?

Answer-3

We certainly should, because the two questions have different semantics and deliver different answers, whenever Z does not satisfy the back-door condition. This can be demonstrated in graph G_3.[3]

$$G_3 : X \to Z \to Y$$

2 The First Law of causal inference refers to the structural definition of counterfactuals [1, p. 98, 9], that is, $Y_x(u)$ is defined as the solution for Y in a mutilated model, in which the equation for X is replaced by a constant $X = x$.

3 The failure of "conditional ignorability" in G_3 can also be verified directly from the twin network, as is demonstrated in Ref. [1, p. 214]. See Appendix.

In this graph, Q_{do} gives:

$$P(y \mid do(x), z)) = P(y \mid x, z) \qquad \text{(from Rule-2)}$$
$$= P(y \mid z)$$

While Q_c gives:

$$P(y_x \mid z) = \sum_{x'} P(y_x \mid x', z) P(x' \mid z)$$
$$= P(y \mid x, z) P(x \mid z) + \sum_{x' \neq x} P(y_x \mid x', z) P(x' \mid z)$$
$$= P(y, x \mid z) + \sum_{x' \neq x} P(y_x \mid x', z) P(x' \mid z)$$

which is totally alien to $Q_{do} = P(y \mid z)$.

Intuition supports this inequality. If we let X be education, Z be skill and Y be salary, Q_{do} looks at people assigned to x years of education who subsequently achieved skill level z, and asks how would their salary Y depend on x, assuming that they end up with the same skill $Z = z$. The graph states that skill alone determines salary, not how it was acquired, therefore Q_{do} evaluates to: $P(y \mid do(x), z)) = P(y \mid z)$ namely, education has no effect on salary, once we know z, as shown in the graph.[4] In contrast, Q_c asks for the role that education plays in the salary of one specific group of units, those at skill $Z = z$. In other words, we look at those who are currently at skill $Z = z$ and ask, counterfactually: what their salary would be like had they received x years of schooling. Since some of those at skill $Z = z$ had no schooling, their skill level would be greater than z had they received schooling, and so would their salary. This explains the inequality $Q_{do} \neq Q_c$.

Question-4 (Q_{do} or Q_c)

Which query, Q_{do} or Q_c, is normally asked when Z is affected by X?

Answer-4

Q_{do} is rarely posed as a research question of interest, probably because it lacks immediate causal interpretation. It serves primarily as an auxiliary mathematical object in the service of other research questions. One such research question is the unconditional causal effect of X on Y, denoted $P(y \mid do(x))$, which is fully analyzed using the *do*-calculus [4], namely, using Q_{do}. Another research question benefitting from Q_{do} occurs in transportability problems [5, 6], where the target query is $P^*(y \mid do(x))$ (the causal effect in a new population), and has been fully analyzed in *do*-calculus, again, using Q_{do}. I have not seen Q_{do} presented as a target query on its own right.

Question-5 (selection bias)

What about selection bias problems, where the selection mechanism is often outcome-dependent?

4 In fact, Q_{do} has no immediate causal interpretation; comparing two values of x for the same z amounts to comparing salaries of under-educated highly-talented individuals with those of over-educated un-talented individuals.

Answer-5

If we aim at estimating $P(y \mid do(x))$ from selection biased data under $S = 1$, we are not asking for Q_{do} nor for Q_{do}. Rather, we are asking for $P(y \mid do(x))$ and we are allowed to use all means available, including the rules of *do*-calculus (which invoke $P(y \mid do(x), z)$) as long as we can recover $P(y \mid do(x))$ from selection biased data [7].

To demonstrate, assume that variable Z in Figure 3 stands for "selection" to the data, and our task is to recover the causal effect $P(y \mid do(x))$. Applying Rule-2 (on the null set) we can write

$$\begin{aligned} P(y \mid do(x)) &= P(y \mid x) \\ &= P(y \mid x, Z = 1) \quad \text{using } Y \perp\!\!\!\perp Z \mid X \end{aligned}$$

which established the recovery of the target effect from the biased data $P(y \mid x, Z = 1)$.

As another example, consider the following model (after [8]) $X \rightarrow Y \leftarrow L \rightarrow S$ where L is unobserved and $S = 1$ represents selection. Since S is not separable from Y, $P(y \mid do(x))$ is not recoverable from the data $P(x, y \mid S = 1)$. (For intuition, imagine the confounder L being sex, in a study that excludes girls from participation. Surely, the average treatment effect is not recoverable from male-only data.) Assume moreover that only few cases drop from the study, i.e. $P(S = 0)$ is small and estimable. We can then write

$$P(y \mid do(x)) = P(y \mid do(x), S = 1)P(S = 1 \mid do(x)) + P(y \mid do(x), S = 0)P(S = 0 \mid do(x))$$

and obtain a lower bound

$$P(y \mid do(x)) \geq P(y \mid do(x), S = 1)P(S = 1 \mid do(x))$$

Two points are worth noting (1): the lower bound has the form of $Q_{do} : P(y \mid do(x), z)$ and (2) the lower bound is estimable from the data available, giving $P(y \mid x)P(S = 1 \mid do(x))$.

This bounding method does not work for the graph $X \rightarrow Y \rightarrow S$. Writing:

$$P(y \mid do(x)) > P(y \mid do(x), S = 1)P(S = 1 \mid do(x)),$$

we see that, even if we are given the last term, $P(S = 1 \mid do(x))$, we cannot estimate the first.

It is important to note that, if we set out to estimate this bound, our target of identification would be a Q_{do}-type expression $P(y \mid do(x), S = 1)$ where S is a descendant of X and we could unleash the full power of *do*-calculus, ignoring the fact that we are only in possession of biased data, conditioned on $S = 1$.

Conclusions

Rule-2 of *do*-calculus is valid for both pre-treatment and post-treatment variables. The rule may appear as violating traditional warnings against conditioning on post-treatment variables, but such warnings apply only to stronger claims, not the one made by Rule-2. The stronger claims are (1): the identification of causal effects by adjustment and (2) the identification of unit-specific effects through counterfactual independence (i.e. "ignorability"). The assumptions needed for these two tasks are satisfied by the back-door criterion and that is where the special handling of post-treatment covariates becomes necessary.

Acknowledgment: I thank Elias Bareinboim, Sander Greenland, Karthika Mohan and many bloggers on http://www.mii.ucla.edu/causality/ for being part of these conversations.

Funding: This research was supported in parts by grants from NSF #IIS-1302448 and ONR #N00014-10-1-0933 and #N00014-13-1-0153.

References

1. Pearl J. Causality: models, reasoning, and inference, 2nd ed. New York: Cambridge University Press, 2009.
2. Shpitser I, VanderWeele T, Robins J. 2010. On the validity of covariate adjustment for estimating causal effects. In Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence. Corvallis, OR: AUAI:527–36.
3. Rosenbaum P, Rubin D. The central role of propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
4. Shpitser I, Pearl J. Complete identification methods for the causal hierarchy. J Mach Learn Res 2008;9:1941–79.
5. Bareinboim E, Pearl J. Transportability from multiple environments with limited experiments: Completeness results. In Welling M, Ghahramani Z, Cortes C, Lawrence N, editors. Advances of Neural Information Processing 27 (NIPS Proceedings). 2014, 280–288. http://ftp.cs.ucla.edu/pub/stat_ser/r443.pdf.
6. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. Stat Sci 2014;29:579–95.
7. Bareinboim E, Tian J, Pearl J. Recovering from selection bias in causal and statistical inference. In Brodley CE, Stone P, editors. Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2014. Best Paper Award, http://ftp.cs.ucla.edu/pub/stat_ser/r425.pdf
8. Garcia FM. Definition and diagnosis of problematic attrition in randomized controlled experiments. Working paper, 2013. Available at SSRN: http://ssrn.com/abstract=2267120
9. Balke A, Pearl J. Probabilistic evaluation of counterfactual queries. In Proceedings of the twelfth national conference on artificial intelligence, vol. I. Menlo Park, CA: MIT Press, 1994:230–7.

Appendix (appended to the published version)

To show explicitly that conditional ignorability does not hold in G_3 (see footnote 3) we consider a linear model:

$$G_3 : X \xrightarrow{\alpha} Z \xrightarrow{\beta} Y$$

and show that $E[Y_x|Z = z, X = x']$ depends on x'.

Using the counterfactual formula in *Causality* (p. 389)

$$E[Y_x|e] = E[Y|e] + \tau[x - E(X|e)]$$

we insert $e = \{Z = z, X = x'\}$, and obtain

$$\begin{aligned} E[Y_x|Z = z, X = x'] &= E[Y|z, x'] + \tau(x - E[X|z], x') \\ &= \beta z + \alpha\beta(x - x'). \end{aligned}$$

We see that $E[Y_x|Z = z, X = x']$ depends on x', hence $Y \not\perp\!\!\!\perp X|Z$.