Bias mitigation in applied peace and conflict studies: A short primer on posttreatment variables

Christoph Dworschak

Norwegian University of Science and Technology &
University of Essex

christoph.dworschak@ntnu.no

13th August 2021

Work in progress. Please do not share without the author's permission.

Posttreatment variables are covariates that are preceded by the main explanatory variable (treatment). Their inclusion in a statistical model does not "control" for their influence on the relationship between treatment and outcome, and does not substitute for a mediation analysis. Likewise, the coefficient estimate of an appropriate "control variable" cannot be interpreted as a causal effect estimate. While these facts are well-established in various fields across the social sciences, their recognition in the field of peace and conflict studies is more limited. Originally collected data on recent publications from leading peace and conflict journals show that a review of the fallacies of posttreatment variables can help to significantly improve future research. Using simulated examples and graphical approaches, I offer an intuitive explanation of the logic of posttreatment variables and clarify common misconceptions. The article concludes with a discussion of the implications for applied researchers.

I am grateful to Moritz Marbach, Charles Butcher, Clara Neupert-Wentz, Patrick Bayer, Christoph V. Steinert, and Tobias Böhmelt for many useful comments, suggestions, and discussions. Special thanks goes to my team of research assistants, namely Christoffer Andersen, Lasse Holtar, Conor Kelly, Andreas Lillebråten, and Niclas Weischner, for their help in coding the article sample. All mistakes are my own.

Introduction

Which variables should researchers *not* condition on (not "control for") in empirical research on peace and conflict? Research design and variable selection are areas in which there are no easy answers available. While the computation of a regression is usually just one click away, which covariates to include in that regression no computer can tell (King, Keohane and Verba 1994).[1] Therefore, questions of designing research and selecting variables have been studied abundantly. This paper reviews key lessons that are particularly relevant to applied research, and attempts to raise renewed awareness to the challenge of variable selection. The target audience of this article are applied empirical peace and conflict researchers. While retaining as much methodological rigor as possible, more emphasis is given to straight-forward and accessible explanations than to statistical depth.

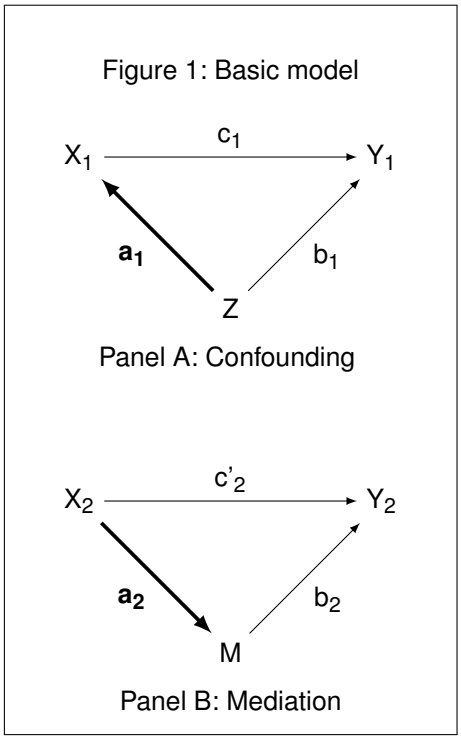
I argue that empirical research in the field of peace and conflict studies does not pay enough attention to the question of causal sequence. The causal ordering of variables matters for the estimation of causal effects. One well-known example of the importance of sequence is the topic of "reversed causality": when estimating the effect of an explanatory variable of interest (the "treatment"; X) on a variable to be explained ("outcome"; Y), the estimate may be distorted by a reverse effect, i.e., not only does X influence Y, but also Y influences X. An example is the effect of democratization on economic prosperity, and vice versa. The importance of this issue is well-established and commonly considered in the curriculum, reviewer comments, conference discussions, and editor reports.

This manuscript addresses an issue of similar importance that is also related to causal effect direction, but which receives much less appreciation: treatment effect estimates are sensitive to the causal sequence of the covariates included for conditioning (also called "confounders" or "control variables"; Z). If the treatment variable causally precedes a covariate, this covariate is called a "posttreatment variable" and its inclusion in the analysis biases the treatment's total effect estimate. This concern over causal direction between the treatment and covariates should receive just as much attention as the question of causal direction between the treatment and the outcome.

1. This and other generalizing statements in this manuscript assume a quantitative research design seeking to approximate causal claims in the context of a standard regression framework. While the lessons drawn here equally apply to other empirical approaches, including qualitative comparison, I choose to narrow the manuscript's language for the sake of simplicity. Also, variable selection in the context of machine learning tasks is an entirely different endeavor and is not considered here.

In this article I discuss why the causal sequence of covariates matters for unbiased estimation. Using graphical approaches and example regressions, I review core concepts and provide an accessible explanation of why peace and conflict research should care about posttreatment variables. While examining the logic of posttreatment variables in applied research, I reflect on avenues for how to avoid common sources of bias related to variable selection, including omitted variable bias, selection bias, and over-control bias. Based on a random sample of recent publications from the Journal of Conflict Resolution and the Journal of Peace Research, and by replicating one study, I show how considering the topic of posttreatment variables provides an opportunity to further improve research in the field of peace and conflict. The article concludes with a summary of implications.

Confounder or mediator?



Observational peace and conflict studies that are interested in the estimation of a directional effect of a treatment variable on an outcome variable include additional covariates in their analysis. These covariates are included for the purpose of mitigating confounding. Not conditioning on a confounder means risking omitted variable bias (OVB).

However, not all variables are confounders and qualify to be held constant, and their causal ordering can give important clues on their adequacy (Pearl, Glymour and Jewell 2016; Gelman and Hill 2007). As a rule of thumb, it is useful to condition on covariates that affect *both* the treatment and the outcome (Imai 2017, 57-58)[2] – see Panel A in Figure 1 for a schematic. A researcher interested in estimating the total effect of X on Y, here denoted as c_1, would condition on Z in the analysis (that is, include Z as a "control variable"). Some covariates, however, do not (only) influence

2. As is common for a rule of thumb, this ignores certain caveats for the sake of simplicity. In terms of the potential outcomes framework, we wish to include covariates so that the potential outcomes are conditionally independent of treatment assignment. In terms of causal graph theory, we wish to condition on a set of covariates as to satisfy the backdoor criterion (cf. Pearl, Glymour and Jewell 2016).

X, but are directly or indirectly influenced by X – see Panel B, noting the change in direction of path a. Such variables are causally preceded by the treatment and are therefore called "posttreatment variables" (M). If they also influence Y (here b_2), then they may constitute a channel through which the effect of X on Y is relayed. In a setup like the one in Panel B, instead of confounding the relationship between X and Y, they "mediate" the relationship. In other words, these variables are the mechanisms that link the explanatory variable of interest to the outcome variable.[3]

Most studies are interested in estimating the overall (total) effect of X on Y, which is denoted in Panel A as c_1. In Panel B the total effect c_2 is not shown, because a part of the total effect goes *through* M. Borrowing from mediation analysis language, the effect c'_2 is called the "direct effect" of X_2 on Y_2, independent of M (Hayes 2018, 107-108).[4] The effect a_2b_2 is referred to as the "indirect effect" of X_2 on Y_2. Taken together, the direct and indirect effects can be combined as $c_2 = c'_2 + a_2b_2$, which is the total effect of X_2 on Y_2 in Panel B (Hayes 2018; Pearl 2014; Imai, Keele and Tingley 2010). This decomposition of the total effect is illustrated via simulation in the Online Appendix, assuming linearity and using simple Ordinary Least Squares regression.

In summary, determining whether a variable precedes or succeeds the treatment is necessary to understand whether the variable acts as a pretreatment confounder or a posttreatment mediator. Therefore, when discussing the "control variables" in a research design, it is vitally important to exercise transparency over the assumed direction between the treatment and each covariate (paths a in Figure 1). The following section details why this distinction is important for unbiased estimation.

Bias by (not) conditioning and what to do about it

Conditioning on a posttreatment variable M can be problematic.[5] First, and intuitively, conditioning on a posttreatment variable means to partial out a part of the treatment effect itself. The mediating

3. "Posttreatment variable" is an umbrella term encompassing all variables causally preceded by the treatment. Not all of them influence the outcome (missing b_2) or are directly affected by the treatment (missing a_2). I specifically focus on mediating posttreatment variables as shown in Figure 1's Panel B, because applied research on peace and conflict seldom conditions on covariates that do not influence the outcome. In fact, while covariates' relationship with the treatment variable is sometimes ignored altogether, special emphasis is usually put on covariates' effect on the outcome. In addition, estimates are potentially distorted even when b_2 is missing due to endogenous selection bias (Elwert and Winship 2014).

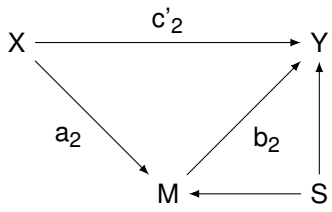
4. Also referred to as the "natural direct effect" (NDE; Imai et al. 2011) as opposed to the "controlled direct effect" (CDE; Acharya, Blackwell and Sen 2016).

5. See e.g. Cinelli, Forney and Pearl (2020), Montgomery, Nyhan and Torres (2018), Acharya, Blackwell and Sen (2016), Pearl, Glymour and Jewell (2016), Elwert and Winship (2014), Mayer et al. (2014), Westreich and Greenland (2013), Angrist and Pischke (2009), Imai, King and Stuart (2008), Gelman and Hill (2007), Rubin (2004), Pearl (2000), King, Keohane and Verba (1994) and Rosenbaum (1984). Including a covariate in a regression model is a special case of conditioning, which also subsumes subsetting and, more generally, missingness. For simplicity, the language in this paper focuses on covariate inclusion in regression models.

variable acts as a mechanism that relays a part of the effect of X on Y, which is why the arrows in Panel B indicate that part of the effect of X on Y "flows through" M. Conditioning on it means to exclude a portion of (i.e., biasing) the total treatment effect (Cinelli, Forney and Pearl 2020; Pearl, Glymour and Jewell 2016; Gelman and Hill 2007). This kind of bias has the intuitive name of "over-control bias" or "posttreatment bias" (Elwert and Winship 2014). The bias can go in either direction, either inflating or attenuating the coefficient estimate. Under ideal circumstances, the effect researchers are left with is c'_2 instead of c_2, as illustrated above and in the Appendix Table A.1. However, for reasons discussed in the following paragraphs, even this is rarely the case.

What about instances in which researchers *wish* to isolate a certain mechanism by partialing out another mechanism, or discern the magnitude of one path independent of another? Acharya, Blackwell and Sen (2016) find that this is the case for 23% of publications that condition on posttreatment variables, out of a sample of publications from three top political science journals[6] between 2010 and 2015. However, the "isolated" direct effect c'_2 is only representative of observations in which X has no influence on M. Such observations may not exist, and the results may therefore not be reflective of reality.[7]

Figure 2: Basic extension



Assuming that c'_2 was indeed the quantity of interest and there were any observations of which this was representative, causality in applied research is rarely as simple as in Figure 1. Even slight modifications to this basic setup can add significant bias to the estimate of c'_2 if this is not modeled using proper mediation analysis techniques (Hayes 2018; Pearl 2014; Imai, Keele and Tingley 2010). For example, consider the simple extension visualized in Figure 2. Here, the variable S is added, influencing both M and Y. M is now a "collider variable" on the path between X and S: the paths from the two variables meet in M, meaning that M blocks an effect transmission between X and S (Pearl, Glymour and Jewell 2016). Therefore, in a simple bivariate regression of Y on X, S does not confound the total effect c_2. However, a researcher looking to isolate c'_2 by condi-

6. *American Political Science Review*, the *American Journal of Political Science*, and *World Politics*.

7. This is the case when conditioning on M in a "standard" regression framework, yielding the natural direct effect as coefficient estimate for c'_2 (Acharya, Blackwell and Sen 2016). To relax this assumption, the controlled direct effect may be estimated using sequential *g*-estimation. Also, see Acharya, Blackwell and Sen (2016) for an illustration of potential uses of the natural direct effect.

tioning on M, without taking S into account, opens a non-causal path from X to S. The estimate of c'_2 will be biased (Acharya, Blackwell and Sen 2016; Elwert and Winship 2014; Imai, Keele and Yamamoto 2010; Rosenbaum 1984). This is a form of selection bias also known as "collider-stratification bias". Depending on functional form and model setup, more bias accumulates quickly (Pearl 2014; Glynn 2012; Imai, Keele and Tingley 2010).

In sum, whether the aim is to estimate the total treatment effect or the direct treatment effect, simply conditioning on a posttreatment variable without considering the underlying assumptions is probably adding bias rather than reducing it. When the quantity of interest is the total treatment effect, as is the case in most empirical peace and conflict research, a variable that is purely posttreatment should just be disregarded altogether. Unfortunately, many confounders in applied research are not purely posttreatment and exhibit directional ambiguity, which is more complicated to address and will be discussed in the next section. When the interest lies with the direct treatment effect, this needs to be appropriately modeled: while, in a most basic setup and under strong assumptions, the coefficient estimates may still be recovered by a simple regression model, in a more realistic research scenario it is almost certainly necessary to implement a mediation analysis.[8]

The total treatment effect in an imperfect world

The implications of the previous section are clear: most applied research seeking to estimate a total treatment effect should "control for" variables that resemble Z in Figure 1's Panel A (pretreatment), and not include variables that resemble M in Panel B (posttreatment). So far, so easy.[9] Unfortunately, the real world of potential covariates does not divide into pretreatment and posttreatment.

What to do when the causal direction between treatment and covariate is ambiguous and may go both ways? A version of this problem is exemplified by Angrist and Pischke (2009) under the name "proxy control": in order to account for an unobserved confounder, the researcher has the option to use a proxy variable. That proxy is, however, only observed after the treatment. A

8. For more information on estimators and assumptions in mediation analysis, see e.g., Hayes (2018), Pearl (2014, 2012), Shpitser and VanderWeele (2011), Imai, Jo and Stuart (2011), Imai, Keele and Tingley (2010) and VanderWeele (2009). For setups in which fully modeling the relationship between mediator and outcome is not attainable, the direct treatment effect can be estimated using either a simple instrumental variable approach as shown in Aklin and Bayer (2017), or via the Average Controlled Direct Effect (ACDE) approach presented in Acharya, Blackwell and Sen (2016).

9. Needless to say, upon closer inspection it is not. This does not consider the many instances in which conditioning on pretreatment variables is not advisable (including M-bias), and fringe cases in which it is beneficial to condition on a posttreatment variable. While elaborating on such and other caveats goes beyond the scope of his manuscript, I encourage interested readers to further engage with, e.g., Cinelli, Forney and Pearl (2020) and Elwert and Winship (2014), for an overview of various variable constellations.

dilemma occurs in which including the covariate in the analysis alleviates omitted variable bias because it proxies for a pretreatment confounder, but at the same time induces over-control bias because it was measured post treatment.[10] Not including it means to avoid posttreatment bias, but to risk omitted variable bias. The researcher is caught between a rock and a hard place.

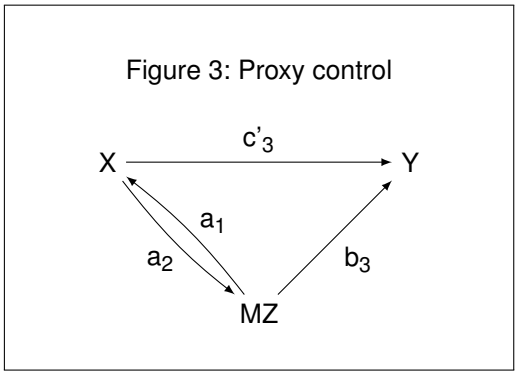


Figure 3 visualizes an adaptation of the problem, showing the variable MZ in an interdependent relationship with X.[11] Recalling the effect decomposition outlined earlier, it is apparent that a_1 makes an inclusion of MZ in our model desirable lest the researcher risks omitted variable bias, while a_2 cautions against including MZ to avoid over-control bias ($c'_3 \neq c_3$).[12] Estimating one model with and one without MZ does not, in the absence of very strong assumptions, bound the true parameter (Groenwold, Palmer and Tilling 2021; King 2010). When discussing solutions to this issue in his 2010 presentation, Gary King ended on the quip "Is there hope? There's always hope; just no answers!" (King 2010). Based on a review of more recent academic work, I identify a few scenarios in which the total effect can still be recovered or approximated in applied research on peace and conflict.

The total effect decomposition (TED) approach developed by Aklin and Bayer (2017) offers an intuitive solution to recover an unbiased total effect estimate for the case of a binary treatment (dummy) variable ($X = \{0, 1\}$).[13] The offending variable MZ is completely left out in this approach, which uses only pretreatment covariates (Z) that are unrelated to MZ to predict counterfactuals of the outcome. Assuming linearity, the total treatment effect can be estimated in a three-step procedure:

1. The outcome Y is regressed on all covariates that are pretreatment (all "proper" confounders; Z) and that are unrelated to X, but only for the subset of untreated observations (X = 0). Also

10. This is in addition to endogenous selection bias by opening a non-causal path through the unobserved confounder, and residual omitted variable bias due to being a mere proxy measure of the original confounder. See Elwert and Winship (2014) for a detailed review of this and other scenarios.

11. I say "adaptation" because this is, of course, not a dyadic acyclic graph (DAG) and is completely unsolvable. However, displaying the problem this way allows me to convey its intuition while skipping the introduction of unobserved confounders.

12. Slight changes to the underlying setup can result in very different dynamics (Groenwold, Palmer and Tilling 2021; Cinelli, Forney and Pearl 2020; Elwert and Winship 2014). Therefore, it is not advisable to "assume away" either kind of bias based on a theoretical consideration of relative effect sizes.

13. Accordingly, in Aklin and Bayer (2017) X is denoted as D. The approach is designed to not only recover the total effect, but also the direct and indirect effects. With a focus on the total effect, I limit the explanation below to the first three steps of the approach.

written as regressing:
$Y_{X=0} = \beta_0 + \beta_1 Z_{X=0} + \epsilon$

2. Using the estimates from the previous step, predicted values are calculated for all treated observations (X = 1). Also written as predicting: $\hat{Y}_{X=1} = \hat{\beta}_0 + \hat{\beta}_1 Z_{X=1}$
3. The total effect estimate c is the mean difference between the predicted values of the previous step ($\hat{Y}_{X=1}$) and the observed outcome values among the treated $Y_{X=1}$. Also written as calculating: $c = E[\hat{Y}_{X=1} - Y_{X=1}]$

Depending on the nature of the explanatory variable of interest, it may be necessary to dichotomize the treatment. Uncertainty can be estimated via bootstrapping.

Workshop note: in a similar manner, I will proceed to discuss solutions in the context of longitudinal or panel data, as well as the importance of sensitivity analysis.

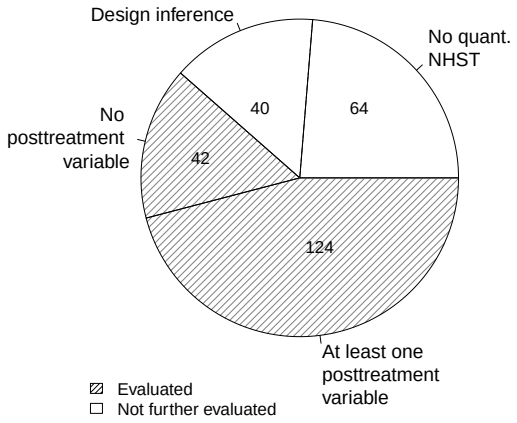
Does it matter?

Posttreatment bias is a known issue in the social sciences and scholarship on it dates back to, at least, Rosenbaum (1984). To understand the extent to which research on peace and conflict takes this into account, I collected data on recent publications in the *Journal of Conflict Resolution* and the *Journal of Peace Research*. Of a random sample of articles that employ a "standard" regression framework (hypothesis test with identification based on observables)[14] and were published between 2018 and 2021, 75% condition on covariates that may be influenced by the treatment (124 out of 166).[15] The proportions are visualized in Figure 4 on the next page. In other words, almost three quarter of relevant articles may report substantially biased results. The most common offenders are so-called "standard control variables", like GDP per capita, which are often included in regression models without considering their necessity and appropriateness as covariates. As discussed above, using an inappropriate "control variable" in research on peace and conflict is just as problematic as failing to account for relevant confounders, and means risking substantively wrong conclusions and misleading policy recommendations.

14. Design-based inference and qualitative studies can suffer from posttreatment bias just as quantitative inferential approaches that rely on identification based on observables (Montgomery, Nyhan and Torres 2018). However, to streamline coding and in line with the emphasis of this manuscript, data gathering focused solely on the latter type of articles.

15. Based on the latest update in an ongoing coding process. Coding covariates' potential for posttreatment bias requires careful theoretical and empirical consideration of each variable with respect to the relevant treatment variable. Therefore, while data are reported at the article level, they were coded on the level of individual hypotheses to ensure transparency. When the coding process is finished, each article will have been independently coded by two coders. In case of deviations, the more conservative coding is chosen (minimizing false positives, increasing false negatives). Details of the coding process, coder training, and coded variables are documented in the codebook.

Figure 4: Article sample



Numbers refer to the absolute number of articles corresponding to each slice. Articles without quantitative null hypothesis significance testing (NHST), and those that conducted inference using experimental or quasi-experimental setups (design inference), were excluded from further coding.

However, as mentioned earlier, it can be painstaking to decide on the inclusion of covariates. Sometimes, choosing to condition on a "bad control" (Angrist and Pischke 2009) is a conscious decision in favor of mitigating omitted variable bias and accepting the potential of posttreatment bias. Such a decision requires careful consideration of the assumed data generating process and empirical model. To learn the degree to which the pattern of inclusion of posttreatment variables is the product of a conscious decision-making process, information on manuscripts' operationalization and results discussion were also coded. Only 12 of all 71 relevant articles show any awareness of this issue by either explicitly mentioning it, or by taking (even implicit) steps to mitigate posttreatment bias.[16] This lack of transparency in the face of possibly large bias that may, in some cases, completely distort the results is concerning and suggests unawareness among peace and conflict scholars. This is especially the case when looking beyond individual authorship and considering the production of research as a whole: the lack of transparency over this potential source of bias raises the question whether it was not noticed among colleagues at draft stage, or by reviewers or editors at publication stage.

In sum, the topic of "control variables" and of their causal sequence requires renewed attention. As the sample of coded articles shows, this is not an issue pertaining to any individual paper, but is something that the field of peace and conflict research faces collectively. For example, there is no reason why conference discussions and peer review should solely focus on the risk of omitted variable bias (the notorious "Have you controlled for...?") while disregarding over-control bias (asking "How do you justify controlling for...?").[17] For any research project, even if

16. As with all other coded information, awareness for the potential of posttreament bias was coded as favorable as possible. For example, lagging the covariates to temporally precede the treatment is, even in absence of an accompanying justification, treated as showing awareness.

17. Also see Clarke (2005, 2009) for the importance of well-justified model specification.

accepting the potential for posttreatment bias was a tenable option amid worse alternatives, such meaningful decisions warrant a transparent discussion.

Conclusion and implications for applied research

In summary, the direction of the effect between the treatment variable and individual covariates matters. Even in the simplest setup, conditioning on a posttreatment variable biases the total treatment effect. If the aim is to isolate a direct treatment effect, due care has to be given to the modeling strategy to avoid bias. A substantial share of recent publications in the field of empirical peace and conflict research does not consider the question of whether individual covariates precede the explanatory variable of interest. Just as reversed causality between the treatment and the outcome is an important point of consideration for authors and reviewers, reversed causality between the treatment and a covariate should be as well.

Workshop note: what follows here is a "checklist" for applied research to facilitate model specification.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. “Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects”. *American Political Science Review* 110 (3): 512–529.

Aklin, Michaël, and Patrick Bayer. 2017. “How Can We Estimate the Effectiveness of Institutions? Solving the Post-Treatment versus Omitted Variable Bias Dilemma”. [Working paper. Accessed March 2, 2021. Semantic Scholar ID 218841486.]

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion.* Princeton: Princeton University Press.

Cinelli, Carlos, Andrew Forney and Judea Pearl. 2020. “A Crash Course in Good and Bad Controls”. [Working paper. Accessed March 3, 2021. SSRN doi 10.2139/ssrn.3689437.]

Clarke, Kevin A. 2005. “The Phantom Menace: Omitted Variable Bias in Econometric Research”. *Conflict Management and Peace Science* 22 (4): 341–352.

———. 2009. “Return of the Phantom Menace: Omitted Variable Bias in Political Research”. *Conflict Management and Peace Science* 26 (1): 46–66.

Elwert, Felix, and Christopher Winship. 2014. “Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable”. *Annual Review of Sociology* 40 (1): 31–53.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research).* Cambridge: Cambridge University Press.

Glynn, Adam N. 2012. “The Product and Difference Fallacies for Indirect Effects”. *American Journal of Political Science* 56 (1): 257–269.

Groenwold, Rolf H. H., Tom M. Palmer and Kate Tilling. 2021. “To Adjust or Not to Adjust? When a "Confounder" Is Only Measured After Exposure”. *Epidemiology* 32 (2): 194–201. eprint: 33470711.

Hayes, Andrew F. 2018. *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach (Methodology in the Social Sciences).* New York: The Guilford Press.

Imai, Kosuke. 2017. *Quantitative Social Science: An Introduction.* Princeton: Princeton University Press.

Imai, Kosuke, Booil Jo and Elizabeth A. Stuart. 2011. “Commentary: Using Potential Outcomes to Understand Causal Mediation Analysis”. *Multivariate Behavioral Research* 46 (5): 861–873.

Imai, Kosuke, Luke Keele and Dustin Tingley. 2010. “A general approach to causal mediation analysis”. *Psychological Methods* 15 (4): 309–334. eprint: 20954780.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. “Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies”. *American Political Science Review* 105 (4): 765–789.

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects”. *Statist. Sci.* 25 (1): 51–71.

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. “Misunderstandings between experimentalists and observationalists about causal inference”. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502.

King, Gary. 2010. “A Hard Unsolved Problem? Post-Treatment Bias in Big Social Science Questions”. [Presented at the "Hard Problem in Social Science" Symposium, Harvard University, 4/10/2010.]

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry.* Princeton: Princeton University Press.

Mayer, Axel, Felix Thoemmes, Norman Rose, Rolf Steyer and Stephen G. West. 2014. “Theory and Analysis of Total, Direct, and Indirect Causal Effects”. *Multivariate Behavioral Research* 49 (5): 425–442. eprint: 26732357.

Montgomery, Jacob M., Brendan Nyhan and Michelle Torres. 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, Judea. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge: Cambridge University Press.

———. 2012. “The causal mediation formula – A guide to the assessment of pathways and mechanisms”. *Prevention Science* 13 (4): 426–436. eprint: 22419385.

Pearl, Judea. 2014. “Interpretation and identification of causal mediation”. *Psychological Methods* 19 (4): 459–481. eprint: 24885338.

Pearl, Judea, Madelyn Glymour and Nicholas P. Jewell. 2016. *Causal Inference in Statistics: A Primer.* Chichester: Wiley.

Rosenbaum, Paul R. 1984. “The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment”. *Journal of the Royal Statistical Society* 147 (5): 656–666.

Rubin, Donald B. 2004. “Direct and Indirect Causal Effects via Potential Outcomes∗”. *Scand. J. Stat.* 31 (2): 161–170.

Shpitser, Ilya, and Tyler J. VanderWeele. 2011. “A complete graphical criterion for the adjustment formula in mediation analysis”. *International Journal of Biostatistics* 7 (1): 16. eprint: 21556286.

VanderWeele, Tyler J. 2009. “Marginal structural models for the estimation of direct and indirect effects”. *Epidemiology* 20 (1): 18–26. eprint: 19234398.

Westreich, Daniel, and Sander Greenland. 2013. “The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients”. *Am. J. Epidemiol.* 177 (4): 292–298.

A. Online Appendix

A.1. Simulated regression example

This data generating process is an example based on Panel B in Figure 1 in the manuscript. The explanatory variable of interest X_2 is independent, meaning that it is not influenced by any other variable in this setup.[1a] The posttreatment variable M is influenced by X_2. For this example, I decided that $a_2 = 0.5$, so that:

$$\begin{aligned} M &= a_2 X_2 \\ &= 0.5 X_2 \end{aligned} \tag{1}$$

The outcome variable Y_2 is influenced by X_2. This influence is both direct (path c'_2) and indirect (through M, paths a_2 and b_2). For this example, I set $c'_2 = 2$ and $b_2 = 0.5$, so that:

$$\begin{aligned} Y_2 &= c'_2 X_2 + b_2 M \\ &= 2X_2 + 0.5M \end{aligned} \tag{2}$$

Examining Table A.1, the effect decomposition described above holds true for this simulated example. The total effect of interest, c_2, is shaded gray and decomposed as:

$$\begin{aligned} c_2 &= c'_2 + a_2 b_2 \\ &= 1.988 + (0.490 * 0.515) \\ &= 2.240 \end{aligned} \tag{3}$$

Table A.1: Simulated regression to illustrate Panel B in Figure 1

	Dependent variable:		
	M	Y	Y
	(1)	(2)	(3)
X	0.490 (a_2)	1.988 (c'_2)	2.240 (c_2)
	(0.010)	(0.011)	(0.011)
M		0.515 (b_2)	
		(0.010)	
Constant	–0.010	0.005	0.0001
	(0.020)	(0.020)	(0.023)
Observations	10,000	10,000	10,000

Standard errors are in parentheses.

1a. More generally speaking, it is independent of the error term. All variables are drawn from a random normal distribution $\mathcal{N}(0, 4)$, which is not included in the DGP notation to reduce jargon and increase accessibility. As the regression is only to showcase unbiased effect decomposition, I avoid a Monte Carlo simulation for the same reason.

Codebook

Bias mitigation in applied peace and conflict studies: A short primer on posttreatment variables

Christoph Dworschak
Norwegian University of Science and Technology
christoph.dworschak@ntnu.no

August 13, 2021

This file describes the data and coding process accompanying the manuscript “Bias mitigation in applied peace and conflict studies: A short primer on posttreatment variables,” including the steps taken in order to minimize the degree of coding subjectivity. The data contains information on articles that were published in the Journal of Peace Research (JPR) and the Journal of Conflict Resolution (JCR) between 2018 and 2021. The unit of observation is at the article-hypothesis level (one entry per hypothesis).

The coding process

Coding was performed with the help of research assistants (RAs) at the University of the German Armed Forces Munich and the Norwegian University of Science and Technology. At the start of the coding process, the journal data frames (one for JPR and one for JCR) contained web-scraped information on each manuscript, including the article’s issue number, manuscript title, and the link to the manuscript. Author names are omitted due to irrelevance and as an attempt to minimize familiarity, gender, or minority biases of the human coders. The order of articles was randomly shuffled to minimize time contiguities.

The coders (RAs and me) proceeded through the data frames row by row, opening the article web link and skimming the manuscript. Author names were to be scrolled past without heeding them. Coders then proceeded to sequentially code the variables in the data frame, as listed below. Manuscripts without a positivist, quantitative research design that aims at causal inference (usually engaging in null-hypothesis significance testing; NHST) were excluded from further coding. For all other articles, the coders proceeded to record features of the research design and results discussion as described in detail below. All variables’ default value is `NA`, which was kept if no value could be attributed based on the coding rules. Coders expanded the data by additional rows based on the number of hypotheses in each article. Therefore, the unit of observation of the final dataset is at the article-hypothesis level. At the far end of this file, a "coder section" is attached as it was used by the RAs to help with coding (supplementing this codebook).

Inter-coder reliability and coding uncertainty

Prior coder training consisted of:

1. Background readings. Excerpts in Kellstedt and Whitten (2018), Imai (2020), Cinelli, Forney, and Pearl (2020), etc.

2. Training session, led by me. Recap of causal inference and research design, as well as revisiting topics such as counterfactuals and treatment assignment mechanism.
3. Joint discussion and Q&A session on causal inference.
4. Going through a few example papers together, discussing the coding process and criteria.

The coding process is set up as a checks-and-balances system. Four RAs worked in two "opposing" teams. To facilitate exchange and feedback while coding, they would meet up with their team partner. Articles were assigned in a complimentary fashion, so that no article would be coded by both at any point. Each team worked independently. Therefore in the end, each manuscript was independently coded by two coders with minimal spill-over. I reviewed all manuscripts that showed deviations in the RAs' codings, as well as a random subsample of all others. I regularly met up with both teams to discuss questions and ambiguities throughout the process.

In some instances, identifying a potential for posttreatment bias is straight-forward. In other cases, however, it may require a certain degree of judgment by the coder. This is especially the case when the manuscript's discussion of the covariates and treatment assignment mechanism is limited, making it difficult to infer the causal sequence. To minimize coding subjectivity, coders were asked to follow up on any uncertainties by, e.g., drawing on the manuscript's appendix, doing online searches, engaging in peer discussions, and through our joint coding meetings.[1] Still, in some instances fundamental uncertainty may remain. For the coding process in this project, such uncertainty is not an issue as long as there is at least one "unambiguous" case of potential posttreatment bias related to the same hypothesis, as this gets aggregated to the article-hypothesis level by the coder. However, in a truly ambiguous scenario,[2] coders were able to quantify their certainty through an additional certainty indicator (and leave a qualitative note explaining the uncertainty). Generally, ambiguity was treated conservatively to minimize false positives.

Something is wrong and needs fixing

Should it occur to a reader that a colleague's or their own manuscript has been misclassified in these data, they should please reach out to me. Transparency and coding reliability is my foremost concern. Despite extensive efforts to reduce errors and especially false positives, there is no safeguard against the occasional inaccuracy. When reaching out, information on the article and hypothesis that require review should be included, as well as a short explanation for why the current coding is perceived as erroneous. I will then review the entry and change the coding accordingly, or follow up with further questions.

[1] Individual studies were not replicated as this would go beyond the scope of this project.
[2] For example, when there is a substantial difference in the treatment's and posttreatment covariate's volatilities, so that it may be argued that the covariate is simply too "slow-moving" to be credibly affected by the treatment.

Variables

Variable	Type	Description
quant.testing	binary, 0/1	Code (1) if article contains a positivist & quantitative research design aiming at directional causal inference, code (0) if otherwise. If (0), further coding is discontinued.
other.inference	binary, 0/1	Code (1) if no covariates are included for conditioning in a regression model (inference by design or other statistical means), code (0) if otherwise. If (1), further coding is discontinued.
n.hypotheses	integer	Number of hypotheses tested in the article's main text body.
hypothesis	character	Quote the article's alternative hypothesis. If an article tests more than one hypothesis, new rows are added equal to the number of hypotheses(-1) so that each hypothesis makes up one row (unit of observation: article-hypothesis).
outcome.var	character	Quote the hypothesis' outcome variable.
treatment.var	character	Quote the hypothesis' treatment variable.
interaction.var	character	If present, quote the interaction variable. If not, leave as `NA`.
posttreat	binary, 0/1	Code (1) if any of the covariates in any of the hypothesis' tests may induce posttreatment bias, code (0) if otherwise (i.e., there are only pretreatment covariates).
posttreat.var	character	If present, quote the article's posttreatment covariate. If more than one posttreatment covariate is identified, use a semicolon delimiter. If none are identified, leave as `NA`.[3]
posttreat.acknwl	binary, 0/1	Code (1) if the authors acknowledged potential posttreatment bias, code (0) if otherwise.
posttreat.as.med	binary, 0/1	Code (1) if the authors interpret any of the covariates in `posttreat.var` as "mediator", "mediating", "concomitant", "surrogate variable", "intermediate", "indirect effect", or anything similar, code (0) if not.
posttreat.certainty	ordinal, 1-5	Indicator of `posttreat` coding certainty. Code (1) if not sure at all about the coding. Code (5) if entirely certain about the coding. Code (2), (3), or (4) for a degree of certainty in-between.
int.cov	binary, 0/1	Code (1) if the parameter estimates of any of the covariates included for conditioning are interpreted by the author(s), code (0) if otherwise.
int.cov.var	character	If present, quote the article's covariates which are actively interpreted. If more than one interpreted covariate is identified, they are separated by a semicolon. If all covariates receive interpretation, write 'all'. If none are interpreted, leave as `NA`.
coder.notes	character	Additional information provided by the coder.[4]
coder.abbr	character	Coder abbreviation.
PI.notes	character	Additional information or comments by the PI.

[3] Make sure not to use a semicolon for any other purpose.
[4] List of common comments below.

Standardized coder notes

- No hypothesis could be identified despite an NHST being present, so the hypothesis is paraphrased based on context.
- At least one posttreatment covariate is another hypothesis' treatment, but was entered in the same model.
- The operationalization of at least one model covariate was not mentioned in the paper, inducing ambiguity in its coding wrt the potential for PTB.

Coder section

In addition to the above codebook, RAs received this "coding roadmap" to streamline processing of manuscripts, as well as the Q&As below to answer common questions.[5]

Coding roadmap

1. Understand what the article is about. Read the title and abstract, and skim the introduction and conclusion.
2. Does the article engage in quantitative hypothesis testing? If necessary, scroll through the article to find out. If it does not (e.g., a review article, or a theoretical, qualitative, or exclusively predictive article), code it accordingly and continue to the next article.
3. Does the article exploit a research design that does not rely on covariate conditioning in a "traditional" regression setup (e.g., an experiment or quasi-experiment)? If so, code it as such and continue to the next article.
4. Awesome; if you made it until here, then you have an article in front of you that we are interested in coding! Let's dive into it: what are the hypotheses being tested? If they are not visibly stated, you can usually find them at the end of the theory section. Expand the dataset accordingly, with one hypothesis per row. Determine each hypothesis' outcome variable, treatment variable, and potentially interaction variable, and follow the steps below for each of the hypotheses.
5. Understand the research design and model. Read as much of the empirics section as necessary/online search to understand how the authors attempt to estimate their treatment effect. What is the unit of analysis? Is it panel data or cross-sectional?
6. Which are the covariates that the authors condition on? Start with the analysis table to get an overview, and then dive into the text to double-check on how the variables were coded. You usually find this information in a "control variables paragraph".
    a) How are the covariates related to the treatment variable? Code whether the treatment variable may causally precede any of the other covariates included in the model, or in other words: is there a chance that the treatment variable may influence at least one of the covariates?
    b) Check whether the authors show any awareness of posttreatment/overcontrol bias or covariate sequence (usually discussed in the "control variable paragraph(s)" or in the robustness section). If the covariates are lagged to temporally precede the treatment, we also treat this as awareness.
    c) Check whether the authors call the posttreatment covariates "mediators", "mediating variables", "concomitant variables", "surrogate variables", "intermediate variables", or as having an "indirect effect", or anything similar.
    d) How certain are you that there is at least one covariate that may be (partly) influenced by the treatment, in any of the analysis/analyses used to test this hypothesis?
7. Check whether the authors interpret the coefficients of the covariates that they condition on (usually in the results discussion).

[5] The wording in the coder section is kept as straight-forward as possible, sometimes at the expense of methodological rigor, to facilitate the coding process and increase coding consistency.

8. Finally, remark any caveats or ambiguities, or whatever else should be added.
9. At the end, add your coder abbreviation. You are done with this article!

Questions and answers

- What about covariates that are lagged?
  - For the purposes of this project, we interpret temporal sequence as causal sequence (never do this in your own projects, but for the coding process we have to simplify things at the cost of false negatives). Therefore, do not flag covariates that are lagged in a way that they temporally precede the treatment (which may not be a fix for PTB, but we'll be conservative and treat it as such); for example if the treatment is measured at t-1 and the time-variant covariates are measured at t-2, these covariates do not contribute to your coding of `posttreat`. If there are no other covariates except the lagged ones, you can code `posttreat` as 0 and code `posttreat.acknwl` as 1. Note that it does not suffice if all variables (treatment and covariates) are uniformly measured at t-1; there needs to be a time difference between treatment and covariates for this rule to apply.
- There are multiple treatment variables (connected to separate hypotheses) tested with the same regression model. Do I code them as posttreatment?
  - When a regression model tests multiple treatment effects simultaneously, make sure to go hypothesis by hypothesis (as always) and simply consider the other treatment variables (those not related to the hypothesis you are currently coding) as covariates. If they may plausibly be influenced by the treatment variable of the hypothesis you are currently coding, code them as posttreatment covariates. If not, then don't.
- There are multiple treatment variables (connected to separate hypotheses) tested with the same regression model. Do I code them as posttreatment?
  - When a regression model tests multiple treatment effects simultaneously, make sure to go hypothesis by hypothesis (as always) and simply consider the other treatment variables (those not related to the hypothesis you are currently coding) as covariates. If they may plausibly be influenced by the treatment variable of the hypothesis you are currently coding, code them as posttreatment covariates and make a note of this in the coder notes column. If not, then don't.
- The authors are talking about testing a relationship of one variable on another, and the entire context makes it clear that they are assessing a hypothesis, but there is no hypothesis clearly stated in the paper.
  - If you are sure that there is a hypothesis test happening to which our criteria apply, but there is no hypothesis visibly stated by the authors, quote a sentence from the paper which postulates the directional relationship as clearly as possible. If there is not such sentence

that you can use: paraphrase the hypothesis based on the paper's context in the form "X leads to an increase/decrease in Y", or as similar as possible to this, and make a note of this in the coder notes column.

- Should we code a quadratic treatment variable as an interaction term?
  - We will not code a quadratic term as an interaction variable.
- Does it suffice to code other.inference="1" based solely on the condition that at least one of the following is true: (a) the study includes no regression models (and carries out some other quantitative analysis) or (b) any regression models included in the study are bivariate (i.e., for any given regression model there is only the treatment variable)?
  - Yes. Condition (a) means we will miss some potentially relevant cases; other quant approaches (as well as qualitative ones) are just as prone to posttreatment bias as a standard regression, but dismissing them ensures that coding remains as straight-forward as possible and that we err on the conservative side. Condition (b) is (usually) only possible when the authors were able to use some form of treatment randomization, so we disregard those cases.
- The regression table in the main paper just shows the row "controls" without specifying which covariates are included. Should we refer to the paper's appendix if necessary to find out which covariates the authors conditioned on?
  - Yes, if there is not enough evidence in the main paper to know which covariates the authors conditioned on in their main analysis, but it is clear that they *did* condition on something, we have to do some investigative work.
- How do we deal with fixed effects?
  - We treat "fixed effects" like we treat any other covariate.
- What if there are multiple treatment variables used jointly to measure a single explanatory concept (one hypothesis)?
  - If they are not interpreted separately (like "treatment variable A shows this effect, meanwhile treatment variable B shows that effect"), but jointly (e.g., joint predicted probabilities), then we also do not code those variables separately. This is a bit arbitrary, but we aim to be as conservative as possible here.
- The authors acknowledge the potential for posttreatment bias for one included covariate but not for another. Do we still code acknowledgement as "1"?
  - Yes, this is sufficient to show the authors' awareness of the issue. After all, this remains a theoretically informed judgment call for each variable. Therefore, them not discussing it with regards to another variable that we would classify as potentially inducing posttreatment bias may simply be due to them having arrived at a different conclusion than we did. In this case, we code both mentioned and unmentioned variables as posttreatment variables if they conform with our criteria.

- Somewhat related to the previous point: what if the authors decidedly conclude that a covariate is *not* posttreatment and include it in their model, but we would normally judge it to be – do we code that variable as being potentially posttreatment?
  - Please email me if you come across such a case, as it may merit a review of our criteria. In either case, we will not code this variable as potentially posttreatment if the authors decided it is not, because the authors are experts in their field and can make this theory-informed judgment call better than we can for their paper. However, if the authors discuss it but express uncertainty in their judgment (expressing resolve equal to or less than "it is not very likely that this variable induces posttreatment bias"), and the variable would merit coding based on our criteria, we do code it. In either case, remember to code the authors' acknowledgement of the issue.
- The authors are doing crazy methods stuff that I do not feel comfortable coding.
  - No worries, just highlight the entry in your data and I will review it. Should the issue persist and it is just impossible for me to make a call, I will code it as not experiencing the potential for posttreatment bias (again, we wish to err on the conservative side).
- The first half of the empirics is qualitative, the second half is quantitative with a standard regression. Do I proceed to code it?
  - Yes, if the main manuscript includes a regression analysis to test an hypothesis we proceed to evaluate it against our criteria.
- When there is no potential for posttreatment bias to be coded (posttreat = 0) and the authors show awareness of the topic, do I still code the acknowledgment variable as 1?
  - Yes, if the authors show awareness of the topic, either by explicitly discussing it or by, for example, lagging their covariates so that they temporally precede the treatment, we code posttreatment acknowledgement as "1".
- What if there is no way the treatment could possibly affect the covariate, but the covariate is realized after the treatment (posterior to the treatment)?
  - If you can think of no mechanism through which the treatment may influence the covariate, we do not code it as posttreatment variable.
- Imagine there is no possible *direct* effect of the treatment on a posttreatment covariate Z1. However, there is another posttreatment covariate Z2 that clearly is affected by the treatment and that itself may plausibly affect Z1.
  - In this case, coding of most variables in made simple by Z2 being a clear candidate for possibly inducing posttreatment bias. The only question that remains is whether Z1 should also be listed in posttreat.var next to Z2. Since you can think of mechanisms through which the treatment may influence the covariate Z1 (even if it is through Z2), we code this as a potential posttreatment variable. This requires no coder note.

- What if the authors refer to their covariates specifically with the term "pretreatment"? As in "Our pretreatment controls are..."?
  - If you encounter this word, please pay special attention to the authors' discussion and whether they show any other signs of "awareness" for the topic of temporal or causal sequence among covariates. However, if they do not: just the use of the word "pretreatment" alone, without further contextualization or justification, should not affect your coding. The use of this jargon word by itself is not enough to suggest the authors' awareness, and is not enough to exclude a potential posttreatment variable from being coded as such. If however there is also some (however small) mention of why they are pretreatment (or anything else suggesting that the authors use the term "pretreatment" as actually referring to temporal or causal sequence): do code acknowledgement as "1" and do not code the as "pretreatment" discussed covariate as being potentially posttreatment (even if you disagree with their judgment).
- The authors acknowledge that "the introduction of control variables can introduce bias" – does this suffice to code their awareness as "1"?
  - While we try to be as favorable in our coding as possible, this alone would not suffice to code posttreat.acknwl as "1". There are other sources of bias, and we only code this if this is then followed up by an explicit discussion of temporal/causal covariate sequence or by any implicit steps to mitigate PTB.