# Do Human Rights Treaties Work? Causal Re-Analysis for the Peace Sciences

While notable research in quantitative peace science scholarship is descriptive (conceptualizing and measuring concepts like war, peace, repression, etc.), most scientific contributions in these literatures are fundamentally interested in examining a causal theory. For example, the majority of entries into the quantitative peace science scholarship attempts to understand the causes, consequences, and solutions to conflict. “Cause”, “consequence”, and “solution” are terms that denote causation. While there a variations within the dominant approach, researchers typically propose a novel independent variable (treatment) and develop a theory explaining why this novel treatment causes some change in outcome (if the word “cause” causes some discomfort, terms like “impacts”, “leads to”, “explains”, and “has an effect on” also imply a causal relationship). However, current conventions in quantitative peace science scholarship yield output that is generally insufficient to examine such theorized causal relationships. Control variables are selected (or omitted) in a fairly arbitrary manner, when causal identification (under a selection on observables approach) requires a systematic evaluation of the necessary set of variables to adjust for (Rohrer 2018, Cinelli et al. 2022, Wysocki et al. 2022, Dworschak 2023). The threat of unobserved confounding is rarely formally examined. Confounding bias generated from past treatment and outcome values when using panel data is rarely sufficiently accounted for. Treatment effects are rarely formally defined, leading to interpretation of results that is either very vague or very misleading. Relatedly, common support (positivity) is rarely examined and made clear to the consumer of research, leading to an incorrect assessment of the effective sample (Aronow and Samii 2016).

One may protest that such concerns are not applicable for quantitative peace science because such research is generally “correlative” and generally cannot be causal due to the reliance on observational data. First, any "correlative" research is itself a type of descriptive research given the correlation simply describes how (and whether) two variables move together and not *why* two variables may move together (Gerring 2012). Second, if a researcher's goal is to examine how two variables correlate, then a simple bivariate correlation between two variables of interest is all that is required and no adjustment for third, fourth, fifth, etc. variables are required (Hernan 2018). However, adjustment for covariates implies that a researcher is seeking to remove variation between treatment/outcome and other variables that may bias a *causal* interpretation of the relationship between treatment and outcome. After all, assuming no issues of reverse causality, isolating all variation between treatment and outcome solely attributable to both treatment and outcome creates the condition for which a statistical correlation can be given a causal interpretation.

To reiterate, most quantitative peace science scholarship is causal scholarship with the caveat that this research is generally not robust enough to deliver strong evidence of causal relationships. Indeed, simply because a final estimate cannot be given a causal interpretation does not render the research as non-causal or descriptive. Causal identification is difficult, and a researcher may not know whether their estimate can be given a causal interpretation until the end of the analysis. Only after a researcher has done their due diligence to identify a causal effect and has failed to do so should they claim that their result is "merely descriptive" (Gerring 2012). However, researchers can also fail to identify causal effects with observational data due to sub-optimal methodological practices, which happen to be often practiced in quantitative peace science scholarship. Much of these sub-optimal practices stem from an incorrect view that causal inference with observational data is unattainable. While causal inference with observational data is certainly *very difficult*, it is far from impossible. To demonstrate the possibility and execution of causal inference with observational data in the peace sciences, this chapter re-analyzes and re-envisions the research design developed by Hill (2010). In particular, this chapter develops a research design in an attempt to answer the question, do human rights treaties work? The remainder of the chapter is organized as follows. First, a brief discussion on causal inference with observational data is provided. Second, I justify the re-analysis of Hill (2010) as a part of a broader need to re-evaluate many foundational questions in the peace sciences using best practices in causal research. Third, an explicitly causal research design is developed to analyze the effect of human rights treaties on human rights. Fourth, the results from this analysis are reported and analyzed, providing mixed results on the effectiveness of human rights treaties and skepticism that the research design of this chapter in particular is sufficient to identify the causal effect of human rights treaties.

**Causal Inference with Observational Data**

Part of the traditional difficulties with causal research has been an epistemological issue, with a clear and standardized understanding of causation lacking. One now-dominant approach that has created a *lingua franca* for speaking in causal terms is the Rubin causal model (Rubin 1974). Under this model, the potential outcomes framework (POF) was developed and such a framework lends the building blocks for subsequent causal research. Under the POF, each unit has (assuming a binary treatment), two potential outcomes. In one potential outcome, the unit is exposed to treatment $Y_{1i}$ and, in another potential outcome, the unit is not exposed to treatment $Y_{0i}$. Under this framework, a causal effect is understood as the difference between the two potential outcomes for a given unit.

This approach is useful not only for defining what a causal effect is, but also for highlighting why causal inference is such a large task. Under the POF, causal inference is seemingly possible. That is, we only ever observe one of two potential outcomes. $\text{Unit}_i$ may be treated or not treated, but it cannot be both. $\text{Unit}_i$ may be exposed to a different treatment status at a later point, but $\text{Unit}_i$ is fundamentally different than $\text{Unit}_{it+1}$ since the latter partially has its outcome influenced by its prior potential outcome. In the absence of two identical $\text{Unit}_i$, causality, as understood by the POF, seemingly cannot be estimated. This conclusion is known as the fundamental problem of causal inference.

Thankfully, there are still clever ways of estimating *average* causal effects. While the estimation of individual-level counterfactuals are not seemingly feasible, one can estimate *population-level* counterfactuals to recover potential outcomes. Conventionally, the randomized controlled trial (RCT) has been viewed as the “gold standard” in this regard. Under a RCT, the researcher has complete control over treatment assignment and distributes treatment randomly. Randomization of treatment is crucial as it (ideally) ensures that, on average, the treated and non-treated groups are near-identical on every observable and non-observable characteristic. Such randomization seeks to ensure that no effect confounding the relationship between treatment and outcome is present. As a result, if both the treated and non-treated units are balanced (each group’s covariate values are statistically the same), any difference in outcome between the two populations can be interpreted as the average treatment effect (ATE).

Conventionally, in the absence of randomization of treatment, causal inference had been relegated as a *de facto* impossibility (Grosz et al. 2020). After all, if treatment is not randomized, any number of factors, observable or unobservable, may lead to treated and control populations that differ fundamentally. Researchers may do their best to adjust for the *known* fundamental differences between treated and control units (through regression adjustment, matching, etc.), but one can never truly *know* if all confounding factors have been accounted for to make a causal inference. Especially in the conflict management and human rights literatures, where treatment is hardly ever randomized due to ethical or practical concerns, this legitimate concern is manifestly present. However, a booming inter-disciplinary movement of novel contributions to the broader causal inference literature is devoted towards the improvement of causal inferences under such scenarios. Such contributions towards more valid causal research designs and causal interpretations of point estimates are demonstrated in the replication sections of this chapter.

## Replication Selection and the Space for Causal Replication

In an attempt to bridge modern causal inference standards and practices into the quantitative peace science literature, this chapter re-examines Hill (2010), an analysis of whether human rights treaties have an effect on the repressive behavior of states. In particular, Hill (2010) examines the International Covenant on Civil and Political Rights (ICCPR), the Convention Against Torture and Other Cruel, Inhumane, or Degrading Treatment or Punishment (CAT), and the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW).

Of course, there are many studies in the quantitative peace science literature to choose from when replicating. However, this chapter replicates Hill (2010) for several reasons. First, program/policy evaluation is inherently a causal aim. Further, analyzing the causal impact of some policy (presumably measured as a binary) synchronizes nicely with many conventional causal research designs. Second, as Hill (2010) notes, selection-bias represents a serious concern in the study of the effectiveness of human rights treaties. Perhaps states already unlikely to commit human rights violations sign onto such treaties. In the opposite direction, perhaps states that anticipate violating human rights use human rights treaties as cover for their behavior. In either case, such aspects mask the true effect of interest the human rights treaties literature is usually interested in; do human rights treaties *cause* a decline in repressive behavior? Lastly, Hill

(2010) was selected because, in many ways, this project already is contextualized within much of the causal inference language. For example, the author demonstrates a clear understanding of counterfactual estimation and matching. However, this chapter seeks to improve on aspects of this study that, with more recent advancements in estimation and best practices, warrant further attention.

Despite the commendable features of Hill (2010), a number of issues are present within this study that warrant further evaluation. First, for causal identification, the researcher must justify that all necessary confounding features have been accounted for in the adjustment strategy (this causal identification assumption is sometimes referred to as the unconfoundedness assumption). While Hill (2010) justifies the inclusion of his control variables, an exhaustive attempt to identify *all* confounding variables is not conducted. As noted earlier, absent adjustment for all confounders, no "standard set" of controls is sufficient for causal inference. Notably, this endeavor can be greatly assisted by the construction of a causal directed graph (such as directed acyclic graphs - DAGs) to identify a set of control variables necessary (and harmful) for causal identification, and the execution of a sensitivity analysis to examine the sensitivity of results to hypothetical unobserved confounding. Further, Hill (2010) did not implement an estimation strategy that accounted for the panel-nature of the data. It is well-established that panel data creates additional complications for causal inference that standard regression adjustment and matching strategies do not account for (Blackwell 2012, Blackwell and Glynn 2018, Imai et al. 2021). As a result, the results of Hill (2010) warrant re-estimation under panel data-specific estimators.

Indeed such modifications suggested above naturally lend support to the enterprise of broader causal replication. The need for causally-informed control variables (incredibly aided by the use of DAGs) in the quantitative peace science literature has recently been directly expressed by Dworschak (2023) who found that 75% of relevant articles published in the Journal of Peace Research and Journal of Conflict Resolution between January 2018 and May 2021 suffered from post-treatment bias, likely leading to misleading results. Further, causal identification is heavily reliant on identification assumptions. Typically, under covariate adjustment strategies such as regression adjustment or matching, the assumption of unconfoundedness is often the most difficult to satisfy. As a result, replication efforts that formally incorporate sensitivity analyses to evaluate unconfoundedness are practically a necessity to substantiate prior findings that lack a sensitivity analysis. Lastly, a booming area of research seeking to estimate causal effects with panel data offers researchers the modern tools to re-evaluate prior findings (Blackwell and Glynn 2018, Abadie 2021, Imai and Kim 2021, Imai et al. 2021, Liu et al. 2022, Rohrer 2023). Given that panel data is often the norm in quantitative peace science studies, replication of studies that used panel-agnostic regression and matching estimators may produce interesting contradictions or further bolster the robustness of certain findings. In sum, studies such as this chapter need not be isolated instances in journals. Rather, there is sufficient reason to believe that long-established findings and hotly-contested estimates *both* could benefit from intentional causal replication.

## Graphical Causal Models

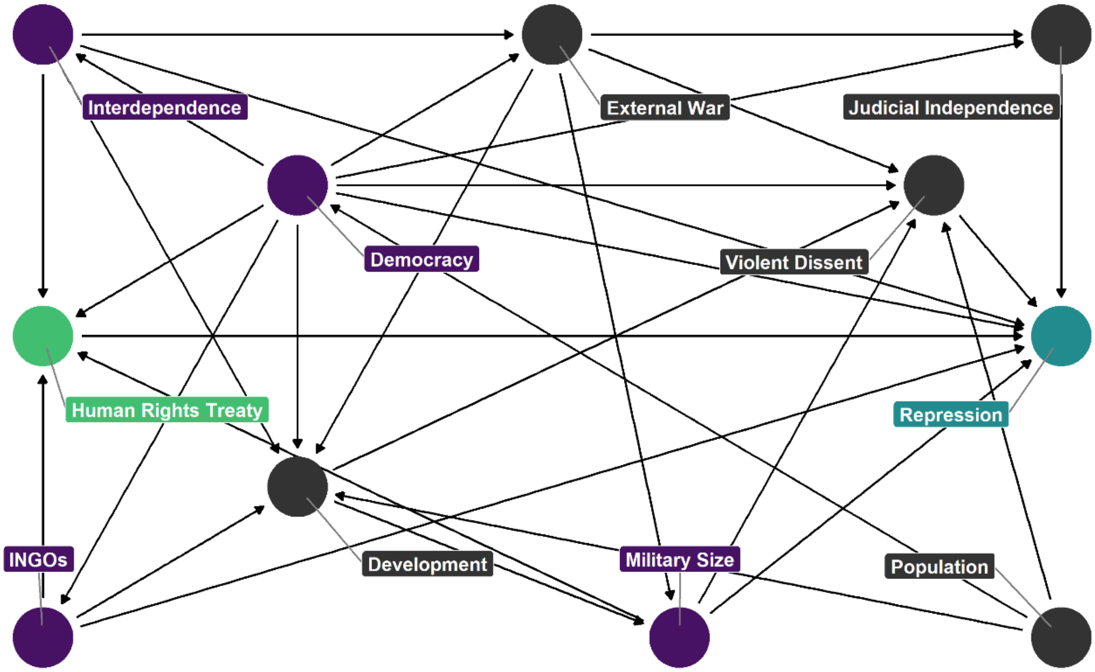
In any project seeking to estimate the effect of some treatment (X) on some outcome (Y), it is generally advisable for a researcher to generate a graphical causal model to reflect their understanding of the data generating process (DGP). Simply put, the DGP is the set of broader processes that lead to the data (concepts of interest) itself. Contextualizing this, the DGP is important to model for any causal project since an understanding of the causes of both treatment and outcome are crucial for causal identification. A popular tool to model the DGP are directed acyclic graphs (DAGs). Under DAGs, nodes are connected by directed arrows which imply a causally ordered relationship, and nodes cannot cause each other (acyclic). The primary benefits of DAGs are twofold. First, careful elaboration of a DAG aids researchers in understanding what identification strategies are available for their research design. In many contexts, researchers will need to operate under the selection on observables identification strategy, where adjustment for certain covariates is required before a causal interpretation is warranted. DAGs are absolutely essential in this process of covariate selection. Second, DAGs make researcher assumptions about the DGP transparent and clear. Such transparency is important for causal identification given that certain assumptions held by a researcher may be harmful for causal identification and DAGs make these incorrect assumptions clear and ready for review.

No DAG is present in Hill (2010). This is unsurprising, given that DAGs still have not received popular attention in peer-reviewed studies across the social sciences. While Hill (2010) justifies the inclusion of his control variables by their relationship to the outcome, the causal inference literature has demonstrated that the status of a control variable as helpful or harmful for causal identification is generally more complex than whether a variable has an effect on just the outcome (Rohrer 2018, Cinelli et al. 2022, Wysocki et al. 2022, Dworschak 2023). For example, certain control variables feature the capability to control for the effect of treatment itself, *introduce* spurious correlations, reduce the precision of estimates, and create difficulties in the construction of counterfactuals. Equally important for causal identification under selection on observables are the control variables *not* included in Hill (2010). Because DAGs are not reliant on any data, a researcher is free to explore the entire DGP, regardless of whether certain data exists or is easily measured for certain concepts. Such a full exploration of the DGP is necessary for causal identification because, if a variable must be accounted for to make an unbiased causal inference and *cannot* be accounted for, then the DAG informs the researcher that an unbiased causal effect cannot be estimated under the selection on observables identification strategy.

While Hill (2010) does not provide a DAG, this chapter does its best in reconstructing the DGP for the effect of human rights treaties on repression. First, Figure 1 demonstrates this chapter's causal model linking human rights treaty ratification to human rights violations.[1] Prior to interpretation of the DAG, it is important to note that a DAG does not require *all* sources of causation to be outlined for treatment and outcome. Instead, this DAG reports covariates specified by Hill (2010) along with other confounding factors detailed in the broader human rights literature. Parents of treatment or outcome (covariates that are causes of only treatment or outcome, but not both) are not included in the DAG as neither are required for causal identification because neither confound the relationship between treatment *and* outcome. Omitting such variables from the analysis does not harm causal identification if these omitted covariates are truly not confounders.

**Figure 1. Causal Model of Human Rights Treaty Ratification and Repression**

[1] Because ICCPR, CAT, and CEDAW are distinct treatments designed to target differing outcomes (physical integrity rights, torture, and women's rights), each distinct research question *should* receive a unique DAG modeling the DGP between the respective human rights treaty and targeted outcome. However, in an attempt to construct a unique DAG for each DGP, the adjustment set (the set of control variables necessary for causal identification) was the exact same across each DAG. Note that this does not imply that the *true* adjustment set for each research question is the same. Instead, veteran researchers in the respective studies of physical integrity rights, torture, and women's rights would likely identify confounders unique to each research question. This approach aligns with current standards in research analyzing multiple hypotheses with a "standard set of controls" implemented in the evaluation of each hypothesis. While this chapter made a greater than normal effort to select control variables that aided causal identification, I acknowledge that using a "standard set" to analyze each distinct hypothesis is a methodological drawback, despite the practice being commonplace in the human rights literature. Critiques are welcome and necessary for future robust causal research in this field.



A number of implications are made from this DAG. First, a minimal sufficient adjustment set (purple nodes) to identify the causal effect of human rights treaties on repression includes democracy, economic interdependence, INGO presence, and military size. Each of these covariates are confounders that potentially transmit spurious associations between treatment and outcome that, left unadjusted, would bias a causal estimate of human rights treaties on repression. Among this adjustment set, there is probably little controversy as the human rights literature offers theoretical justification for each of these covariates jointly causing some change in both a state's propensity to ratify a human rights treaty *and* levels of repression. What is perhaps much more controversial are the covariates *not* included in the adjustment set. Notably, from Hill's (2010) analysis, this chapter's DAG does not suggest that adjusting for development, population, internal conflict, external conflict, or judicial independence is necessary for causal identification. For each of these omitted covariates, a unique causal effect on treatment (ratifying a human rights treaty) is not readily justified, despite obvious theoretical justification for a unique causal effect on outcome (repression).

It is important to note that adjustment for a parent of outcome neither harms nor helps causal identification (Cinelli et al. 2022). Subsequently, many researchers fearful of omitted variable bias may ask, “why not include those omitted covariates just in case?” The primary reason to not simply include any and all variables that are not explicitly harmful for causal identification is the curse of dimensionality. As the number of covariates increases, estimation of counterfactuals to make causal inferences becomes more complex and difficult. Especially under strict matching designs (such as Hill 2010), an increasing number of covariates to match on either decreases the number of matches (therefore, cases) that can be included in the analysis *or* decreases the quality of matches in an effort to conserve sample size. In either case, causal identification is significantly damaged and it cannot be understated how harmful a single haphazard control variable can be. Indeed, the curse of dimensionality likewise applies to non-matching designs. While regression adjustment may not drop cases from the analysis due to a lack of common support between treated and non-treated units along the set of specified covariates, research demonstrates that a lack of common support leads to misleading regression estimates and extreme, model-dependent counterfactual estimation (King and Zeng 2006, Aronow and Samii 2016). Overall, this suggests that, under such designs with so comparatively few cases, researchers should adjust for only the necessary covariates that confound the relationship between treatment and outcome.

## Identification

While any number of variables from a data set can be incorporated into a statistical model of some sort to produce numerical output, the endeavor of causal inference differentiates itself from statistics in that it is tasked with *identifying* a causal effect from descriptive estimates. Causal identification refers to the set of pre-estimation assumptions required for a descriptive estimate to be given a causal interpretation and the post-estimation evaluation of these assumptions. For example, a difference in means under a RCT can be given a causal interpretation because of the randomization identification strategy which asserts that, under complete randomization of treatment, any difference in means between a treated and control group can be attributed solely to the treatment of interest (assuming the assumption of non-interference between treated and control units is satisfied). Clearly, forcing states to adopt human rights treaties at random is not a possibility for this study. Further, natural sources of exogenous randomization of treatment, such as an instrument or an arbitrary discontinuity in treatment assignment are likewise not readily apparent. As a result, the selection on observables identification strategy (identifying and conditioning for all confounders) will suffice as a strategy to estimate the causal effect of human rights treaties.

In practice, a research design reliant on the selection on observables identification strategy is easy in implementation. Researchers simply collect their treatment and outcome variables of interest, create a DAG to identify confounders, collect data on these confounders, and adjust for these confounders (regression adjustment, matching, inverse probability weighting, etc.). However, the satisfaction of the selection on observables identification strategy is much more difficult. In particular, the selection on observables identification strategy can only identify a causal effect if the unconfoundness assumption is satisfied.[2] That is, all confounded relationships between treatment and outcome must be accounted for. If confounding remains

[2] The unconfoundness assumption may also be referred to as the ignorability or conditional exchangeability assumption.

present when proceeding to estimation, a causal effect cannot be identified. Such an assumption is both a difficult one to defend *and* an impossible one to directly test. Again, if a researcher is able to randomize treatment, the assumption of unconfoundness is satisfied because no variable can jointly cause some change in treatment and outcome because the treatment assignment is completely random. However, absent a stronger identification strategy, a theoretically infinite number of confounding factors may exist and it is impossible to *know* if all confounders have been identified. Such limitations justify why selection on observables is perhaps the weakest of identification strategies. Although, many quantitative human rights scholars will undoubtedly find selection on observables as their sole option for identification.

Causal identification under selection on observables necessitates the use of sensitivity analyses. Within the context of selection on observables, sensitivity analysis refers to tests that examine how much an estimate may shift when exposed to hypothetical unobserved confounders of varying magnitudes. Most researchers should assume, regardless of prior theoretical work identifying potential causes of an outcome of interest, that *some* confounders have either not been identified in the literature or have been identified, but are difficult to measure. In either case, a confounder will not be adjusted for in the estimation of a potential causal effect. Sensitivity analyses allow a researcher to toy with varying levels of hypothetical unobserved confounders to examine how large an unobserved confounder would need to be in order to bring an estimate towards zero or flip the direction of the estimate. This step of the causal research design is *necessary* because, although one can identify *many* confounders with a DAG, one cannot ensure that *all* confounders have been accounted for as selection on observables demands. If a researcher finds that their estimate is fairly robust to increasing unobserved confounding sizes, this represents strong evidence of a causal effect. In contrast, if a researcher finds that their estimate is fairly sensitive to varying levels of unobserved confounding, then a causal effect is not identified and subsequent causal interpretations of the estimate should be avoided. As Gerring (2012) notes, results may be interpreted as merely descriptive under such conditions.

Causal identification also requires satisfaction of the stable-unit-treatment-value assumption (SUTVA). SUTVA is comprised of two components. First, a treatment must be *consistent*. That is, the value of treatment itself must be constant across all treated units. Given that the International Covenant on Civil and Political Rights (ICCPR), the Convention Against Torture and Other Cruel, Inhumane, or Degrading Treatment or Punishment (CAT), and the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) are not modified dependent on the state that ratifies them, we can say with confidence that the consistency component of SUTVA is satisfied. If consistency was not satisfied, it would be difficult to identify a causal effect of treatment because treatment itself is not well-defined. For example, consistency would be violated if the ICCPR stipulated different conditions for observance dependent on which states ratified. This would create a natural issue given that treatment itself would no longer be the same for each unit, rendering the attempt to identify a *treatment effect* difficult because treatment itself would be inconsistent across different units. Second, the treatment status of treated units should not impact the outcome of other units. Known as *non-interference*, this component of SUTVA is particularly difficult to satisfy in cross-national research given that many political phenomena tend to “spillover” into neighboring countries. Satisfying non-interference implies that the impact of outcomes from treated units should have no impact on the outcomes of non-treated units. Within the context of conflict and human rights research, non-interference seems *de facto* unlikely given that conflict and repression create conditions that often spread across borders. However, because repression (and not the outcomes of repression), is the outcome of interest in this paper, non-interference appears to be supported. That is, this chapter does not assume that Country A signing onto a human rights treaty will impact the level of repression in Country B. If the outcome of interest was an outcome of repression (such as internal conflict), non-interference may not hold because conflict can be contagious. However, this chapter makes an assumption that repression itself is not contagious. If non-interference did not hold, control units would fail to serve as *true* control units, because their outcome would, in part, be influenced by other unit's treatments. Under such scenarios, researchers should consider alternative, complex designs that allow for some degree of violation of non-interference (such techniques are beyond the scope of this chapter).

Finally, another important aspect of identification is the direct specification of the intended effect to estimate (the *estimand*). The estimand is simply the effect of interest that a researcher seeks to estimate. While a researcher can obtain a *statistical estimate* from a coefficient of a regression model, the task of causal inference is to justify that this statistical estimate reflects the true causal *estimand* that exists in a defined population. Such a deliberate specification of the estimand is crucial for the interpretability of any scientific output, because it is important to know *to whom* an estimated effect applies toward (Lundberg et al. 2021). Whether a descriptive *estimate* is a reflection of the target *estimand* in a population is entirely dependent on the rigor of the causal identification strategy. The decision of which estimand to estimate will be driven by practical considerations and data limitations. For practicality purposes, consider the "default" treatment effect that RCTs generate, the average treatment effect (ATE). The ATE is interpreted as the difference in outcome between two identical populations who differ only because of their treatment status. Such an estimand, while informative, may not be of great interest to many researchers. After all, questions concerning two identical populations who differ only with respect to treatment status is not often the organic question being asked by scholars. Instead, a researcher may specifically be interested in the effect a treatment had *on the units who actually received treatment* (this question would warrant the estimation of the average treatment on the treated - ATT) or, conversely, the effect of withholding treatment for non-treated units (the average treatment on the untreated/control - ATU/ATC).

Second, the decision of which estimand to estimate will likewise be influenced by the data itself. Causal inference requires the satisfaction of the positivity assumption (also known as "common support" or "overlap") where, in all combinations of covariates, the probability of receiving treatment is never exactly zero or one. If a certain covariate or combination of covariates causes a unit or set of units to *never* or *always* experience treatment then a counterfactual cannot be estimated for these units because the probability of the exact same unit existing that differs only in treatment status is zero. If one seeks to estimate the ATE, the positivity assumption must be strictly satisfied, since both treated and non-treated units need to, on aggregate, represent identical populations. However, the positivity assumption can be relaxed with the specification of a different estimand. For example, if a researcher seeks to estimate the ATT, then the positivity assumption only requires that enough similar control units exist to serve as counterfactuals for the treated units. In this scenario, even if there are hundreds or thousands of control units that will *never* receive treatment (such as a country that has never experienced civil or external war receiving a United Nations peacekeeping operation), so long as enough control units exist to serve as counterfactuals for the treated unit, estimation of the ATT can be completed. This chapter explores which estimand this project estimates in the following estimation section.

## Estimation

It is common to see researchers acknowledge the biasing properties of panel data through the implementation of clustered robust standard errors. However, it is fairly uncommon to see studies acknowledge the biasing properties of panel data on their estimates. The causal inference literature has well-documented the complications that panel data creates for causal inference (Blackwell 2012**,** Morgan and Winship 2014**,** Blackwell and Glynn 2018**,** Imai et al. 2021). While many canonical causal inference estimation techniques were designed for the cross-sectional setting, counterfactual estimation becomes increasingly more complex when multiple units receive treatment at different times (sometimes receiving multiple iterations of treatment at different times). What constitutes a valid counterfactual given varying pre-treatment histories that may be biased by prior experience with treatment? When an outcome at time *t* causes a change in a confounder at *t+1*, should this confounder still be adjusted for given that, in a certain temporal iteration, it is now a post-treatment variable (a bad control) when it is a necessary control at time *t*? Left unaddressed, the biasing properties of panel data render causal inference incredibly difficult.

Given these biasing properties, it is still fairly common to see regression adjustment and matching used to estimate unbiased effects from panel data, despite the methods employed being incapable of doing so. Commonly-known attempts to address some of the biasing properties of panel data, such as fixed effects or the incorporation of lagged variables in a model, have proven to be incredibly problematic and reliant on a number of unrealistic assumptions (Bell and Jones 2015, Blackwell and Glynn 2018, Imai and Kim 2019, Plümper and Troeger 2019). Two methods emerging in popularity, panel matching and marginal structural models (MSMs) offer researchers familiar with the matching and regression adjustment framework, a path to estimate causal effects from panel data. This chapter proceeds with the use of panel matching given its ease of implementation and interpretation. Conceptually, panel matching is quite similar to the matching that most human rights researchers may already be familiar with.[3]

In the canonical cross-sectional framework (in which standard matching estimators are designed for), units similar along every observable confounding factor, but differ only with respect to treatment status, are compared to each other, and their difference in outcome is attributed as the causal effect of treatment. However, in the panel data setting, time itself becomes a confounder and a comparison between units at different times with different pre- and post-treatment histories is not sufficient to estimate counterfactuals required for causal inference. In panel matching, the selection of matches is similar to cross-sectional matching in that units who differ "only" with respect to treatment status are matched *if* they have a comparable pre-treatment history. For example, perhaps two units (A - the treated and B - the control) at time *t* share similar characteristics along confounding factors to the extent that A and B would be considered a match under the cross-sectional framework. However, suppose that their pre-treatment history (3 periods prior to *t*) is different such that the pre-treatment history for A is $\{0, 1, 0\}$ and the pre-treatment history for B is $\{0, 0, 0\}$. Under this scenario, the units would not be matched because B does not serve as a valid counterfactual for A. While A has already been effected by past iterations of treatment, B never has, and would serve as a biased counterfactual. Despite this relatively simple framework, a number of subjective decisions must be made on the part of the researcher.

First, a decision must be made on the number of pre-treatment periods to match on. Such a choice reflects a clear bias-variance tradeoff. As the pre-treatment lag criterion increases, the

[3] To compare the differences in results between panel and cross-sectional estimates, this chapter also presents the estimates from standard regression adjustment and coarsened exact matching (CEM) that Hill (2010) applies.

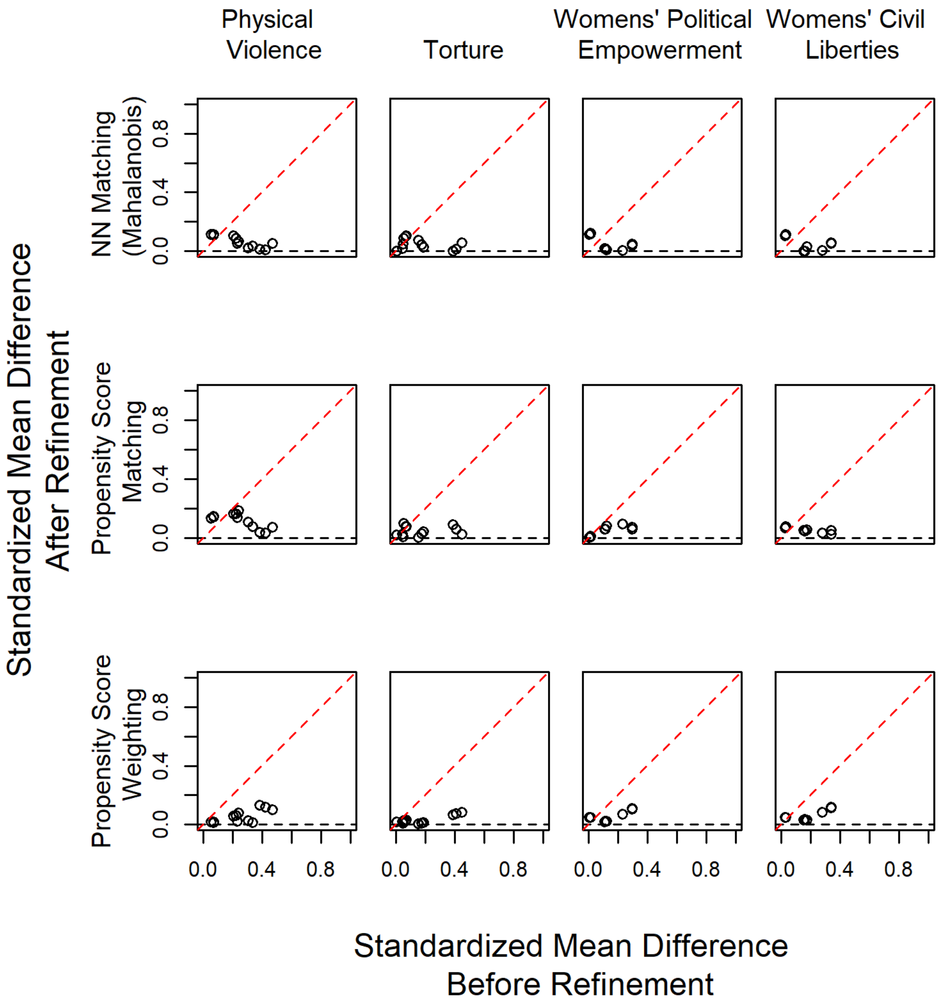
quality of matches increases given that matches have similar pre-treatment histories, therefore reducing bias. However, a consequence of this is a stricter matching criterion for pre-treatment history, which possibly reduces the number of matches available, therefore increasing variance in the estimate. Second, a researcher must decide on their method of "refinement". Refinement refers to the matching or weighting process that attempts to increase covariate balance in comparison to a non-matched/weighted data set. Following Imai et al. (2021), this chapter experiments with nearest-neighbor matching (using the Mahalanobis distance), propensity score matching, and propensity score weighting. To explore the impact of various lag-thresholds, each combination of lags and refinement method for each dependent variable were plotted. These plots, however, are extensive, and can be found in an online appendix.[4]

After examining the impact of moving from one to three pre-treatment periods to match on, the results indicated that the quality of refinement was not damaged with the introduction of an up-to-three pre-treatment history lag criterion. While this may decrease the number of cases, this chapter argues that ensuring higher-quality matches (matches that produce units that are more similar along the set of specified covariates) is more important for causal inference. Increase variability in the estimate changes the estimand from the average treatment effect on the treated (ATT) to a more localized, sample-specific sample average treatment effect on the treated (SATT). While featuring potentially less external validity, this decision increases confidence in causal identification. Concerning refinement, Figure 2 plots the comparative change in covariate balance (standardized mean difference) under the three-year pre-treatment matching criterion for different refinement methods for each of the dependent variables examined in this chapter. Points farther along the right on the x-axis indicate a high degree of difference in covariate value

[4] https://github.com/Brian-Lookabaugh/Human-Rights-Treaties-Project/tree/main/Graphics

between treated and non-treated units pre-refinement. Points higher on the y-axis indicate a higher degree of difference in covariate value between treated and non-treated units post-refinement. If a point is above the dotted red line, this indicates that refinement *harmed* the balance of a specific covariate value. If a point is below the dotted line, this indicates that refinement increased covariate balance.

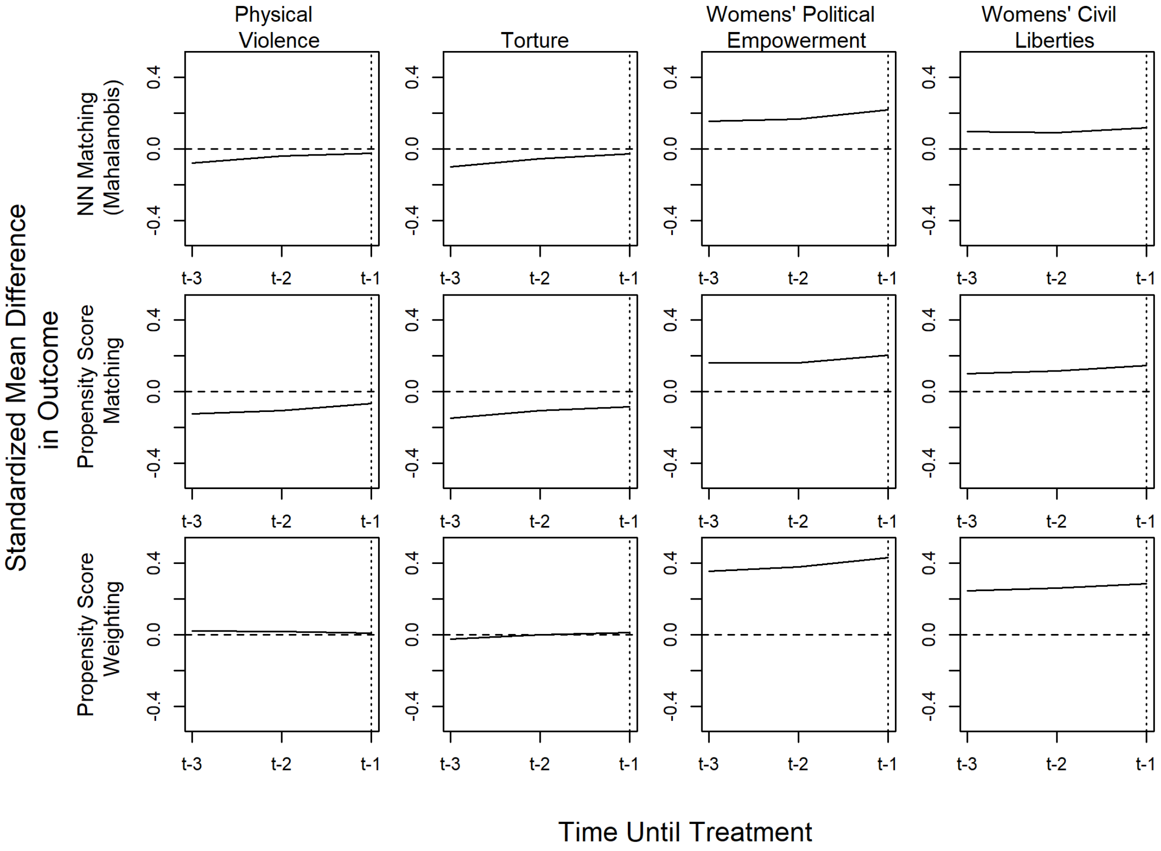
**Figure 2. Covariate Balance Scatter Plots for Different Refinement Methods**



In almost all instances, refinement increases covariate balance. Where refinement does not increase covariate balance, it only marginally increases the difference in values between treated and non-treated units. Overall, this suggests that refinement *of some sort* is an improvement over a non-matched/weighted data set. Further, propensity score weighting (achieving balance by weighting observations by their predicted probability of receiving treatment as specified by a logistic regression where confounders serve as predictors of treatment) preforms the best at reducing differences in covariate values between treated and non-treated units. As a result, propensity score weighting will be used to attempt to estimate a causal effect of human rights treaties. However, because weighting on the propensity score is fundamentally a computationally separate method from matching, this chapter also examines the estimates from propensity score *matching* as well. While propensity score matching and nearest neighbor matching demonstrate near-identical refinement properties, propensity score matching is selected to remain consistent with the decision to likewise utilize propensity score weighting.

A helpful feature of the panel matching design is its ability to examine the long-term effects of a treatment. However, because panel matching relies on the difference-in-differences (DiD) estimator to do so, the satisfaction of the parallel trends assumption is required for any credibility to be given to the estimation of results past time period ***t***. The satisfaction of parallel trends implies that, pre-treatment, the trend in outcome between treated and non-treated units, while potentially different, *must* be increasing/decreasing/remaining constant at the same slope. This assumption is important because, if pre-treatment trends are the same *until* treatment occurs, then any change in trend for the post-treatment outcome between treated and non-treated units can be attributed to treatment. A difference in slope between treated and non-treated units implies that part of the post-treatment "effect" may be driven by unobserved forces that cannot be attributed as a part of the causal effect of treatment. Figure 3 plots the standardized difference in outcome between treated and non-treated units using the three different refinement methods specified earlier. While the baseline difference between treated and non-treated units may vary, confidence in the satisfaction of the parallel trends can only be afforded if the lines remain flat. Unfortunately, in most cases a clear trend is evident, indicating the parallel trends assumption is not supported. As a result, this chapter does not report on the estimated long-term effects of different human rights treaties and focuses solely on the contemporaneous effect, given that this effect does not rely on parallel trends.

**Figure 3. Trend in Difference in Outcome Between Treated and Non-Treated Units**



As mentioned earlier in the chapter, the identification strategy employed for this project relies on selection on observables. Crucial to this design is the assumption that all confounding effects have are being adjusted for. This assumption is both difficult to defend in theory and impossible to directly assess in practice. However, sensitivity analyses allow the researcher to evaluate the impact of varying degrees of magnitude of violation of the unconfoundness assumption. In fact, in any study reliant on selection on observables, a sensitivity analysis is near mandatory for credible causal interpretation. Unfortunately, panel matching does not currently have a corresponding sensitivity analysis. Sensitivity analysis methods by Cinelli et al. (2020) and McGowan (2022) are incredibly valuable tools, but are limited to the regression adjustment framework. Facing the limitations of the panel matching method in this respect, this chapter opts for a "next best" approach. In particular, to approximate the panel matching estimates, a similar (although, not equivalent) estimation is conducted within the regression adjustment framework. As Imai et al. (2021) note, panel matching has similar properties to the two-way fixed effects (TWFE) estimator, which is estimated using regression adjustment. As a result, TWFE estimators are estimated with fixed effects at the country- and year-level on the matched data. Although, it should be stated that this weighted TWFE estimator is still subject to the parametric assumptions of TWFE that are well-documented and critiqued. Hence, why this approach represents a "next-best" approach until a sensitivity analysis is developed specifically for panel matching. Given that the TWFE estimator allows for estimation under the regression adjustment setting, the sensitivity analysis developed by Cinelli et al. (2020) is employed.

### Measurement

Prior to the estimation of results, discussion on how key concepts are measured and where data is collected from is discussed. Information on the ratification status of the ICCPR, CAT, and CEDAW is acquired from the United Nations Treaty Body Database. However, this chapter departs notably from Hill (2010) in the measurement of outcomes (physical integrity rights, torture, and women's rights). While Hill (2010) is very reliant on CIRI Human Rights

Data Project. All of the variables Hill (2010) used were ordinally measured (even though OLS was applied to a nine-point ordinal scale). Given that this chapter uses a fairly novel estimator (panel matching/weighting estimator), for ease of estimation and interpretation, alternatives measured continuously were sought. A discrete outcome would require the estimation of marginal effects to provide meaningfully interpretable results, which are not currently supported with this novel estimator barring complex, hand-coded mathematical transformations. Thankfully, near identical interval-level data on these dependent variables were found using the Varieties of Democracy (V-Dem - v12) data set (Coppedge et al. 2022). The degree to which states violate physical integrity rights is captured by the physical violence index, which measures the extent to which people are free from political killings and torture by the government. V-Dem also offers an interval-level measure for freedom from torture, which specifically focuses on freedom from state torture (Pemstein et al. 2022**)**. Lastly, Hill (2010) uses three separate measures of women's rights (women's political, economic, and social rights). This chapter follows a similar approach and measures women's rights using V-Dem's women's political empowerment index and women's civil liberties index (Sundström et al 2017, Pemstein et al. 2022**)**. While the former is targeted directly towards the political empowerment of women, the latter incorporates economic and social dimensions of women's rights including the right to private property, the freedom of domestic movement, freedom from forced labor, and access to justice.

According to the DAG generated for this chapter (Figure 1), five variables are required (at minimum) for causal identification under selection on observables. First, democracy is measured using the standard Polity V combined score (Marshall and Jaggers 2020). Second, military size is measured as the number of military personnel per capita as provided from the

Correlates of War's National Material Capabilities (v6) data set (Singer et al. 1972, Singer 1987). Third, (because this data is not aggregated dyadically) economic interdependence is measured using a balance between exports and imports (exports minus imports). Data on exports and imports is provided by the Correlates of War's International Trade (v4) data set (Barbieri et al. 2009, 2016). Fourth, the level of INGO activity is measured using V-Dem's civil society organization (CSO) participatory environment variable (Bernhard et al. 2015, Pemstein et al. 2022). This measure is selected over others because it accounts for the monitoring aspects of other non-governmental bodies beyond INGOs that very well may impact the levels of state repression and lobbying for the ratification of human rights treaties.

This chapter also estimates models using Hill's (2010) specified set of non-DAG-informed control variables to contrast with the results obtained by using a DAG. First, economic development is measured using a log-transformation of GDP per capita (Fariss et al. 2021). Second, population is likewise measured using a log-transformation of country-level population estimates (Fariss et al. 2021). Third, violent dissent and external war are both collected from the Uppsala Conflict Data Program's (UCDP)/Peace Research Institute in Olso's (PRIO) Armed Conflict Dataset (v23.1) where violent dissent and external war take on the value of "1" if 25 battle-related deaths are reached (Gleditsch et al. 2002, Davies et al. 2023). Lastly, judicial independence is measured using V-Dem's high court independence measure which evaluates the frequency in which the high court in a judicial system makes decision based on adherence to government desires or a legitimate view of the legal record (Pemstein et al. 2022).

## Results

To illustrate the difference in results between different statistical decisions *and* different causal modeling decisions, each statistical method is estimated with the control variables from

Hill (2010) and the control variables indicated by the DAG generated for this chapter. As indicated previously, panel matching and weighting via the propensity score are used to estimate the potential effect of various human rights treaties on various outcomes. However, to demonstrate differences between popular estimates, regression adjustment and coarsened exact matching (CEM) are also included in the analysis. Regression adjustment represents the *de facto* default method for attempted causal inference in the social sciences and, especially in conflict and conflict-adjacent literatures, coarsened exact matching is a popular matching technique. In fact, CEM represents the matching method used by Hill (2010). However, it should be noted that this chapter does not place much emphasis on the findings of the estimates generated from either regression adjustment or CEM. Regression adjustment does not account for the panel structure of the data, may struggle with counterfactual estimation, and is burdened by difficult functional form assumptions. CEM likewise does not account for the panel structure of the data used in this study, and the harsh exact matching criterion (even if the exact matching is along values within bins rather than exact values) may omit more observations from the analysis than desired. In sum, the estimates presented in Table 1 contrast results along two dimensions. First, the difference in estimates between estimators that account for the panel structure of the data and those that do not. Second, the difference in estimates between the specification of control variables as determined by Hill (2010) or a DAG.

**Table 1. Average Treatment Effects for Countries that Ratified Human Rights Treaties on Human Rights Outcomes**

| | ICCPR - Physical Integrity Respect | CAT - Freedom from Torture | CEDAW - Political Empowerment | CEDAW - Civil Liberties |
|---|---|---|---|---|
| Regression Adjustment (Hill's Covariates) | 0.01<br>[-0.03, 0.05] | -0.003<br>[-0.16, 0.15] | **0.124**<br>**[0.101, 0.147]** | **0.044**<br>**[0.015, 0.072]** |

| | | | | |
|---|---|---|---|---|
| Regression Adjustment (DAG Covariates) | 0.026<br>[-0.023, 0.075] | 0.16<br>[-0.021, 0.342] | **0.122**<br>**[0.1, 0.143]** | **0.031**<br>**[0.007, 0.055]** |
| CEM (Hill's Covariates) | 0.038<br>[-0.003, 0.078] | 0.012<br>[-0.173, 0.197] | **0.106**<br>**[0.082, 0.131]** | 0.025<br>[-0.002, 0.052] |
| CEM (DAG Covariates) | 0.031<br>[-0.019, 0.082] | **0.207**<br>**[0.012, 0.403]** | **0.125**<br>**[0.1, 0.15]** | **0.04**<br>**[0.007, 0.073]** |
| Panel Matching (Hill's Covariates) | -0.002<br>[-0.026, 0.022] | 0.05<br>[-0.047, 0.152] | 0.005<br>[-0.000, 0.01] | -0.002<br>[-0.01, 0.006] |
| Panel Matching (DAG Covariates) | 0.006<br>[-0.018, 0.027] | 0.052<br>[-0.028, 0.144] | 0.002<br>[-0.003, 0.008] | -0.002<br>[-0.01, 0.005] |
| Panel Weighting (Hill's Covariates) | 0.007<br>[-0.012, 0.025] | **0.073**<br>**[0.002, 0.145]** | -0.000<br>[-0.004, 0.003] | -0.002<br>[-0.003, 0.003] |
| Panel Weighting (DAG Covariates) | 0.009<br>[-0.013, 0.027] | **0.079**<br>**[0.003, 0.161]** | 0.000<br>[-0.003, 0.005] | -0.001<br>[-0.005, 0.003] |

Table 1 reports the estimated ATT for countries that ratified respective human rights treaties on related outcomes. Across varying statistical and causal specifications, the ICCPR demonstrates no statistically nor substantively significant effect on the respect of physical integrity rights. In contrast, the evidence for the effectiveness of the CAT is mixed. Using CEM with the DAG-specified covariates and panel weighting for both Hill's (2010) and the DAG-information covariates, a likely non-zero effect is estimated. Indeed, the results under CEM estimation with DAG-informed covariates is noticeably large given that the outcome (freedom from torture) ranges on a 0-1 scale. However, as stated previously, CEM does not account for the panel nature of the data. While technically statistically significant, the panel weighting results indicate a dramatic reduction in the effect size. This is a finding that continues for the impact of CEDAW on both women's political empowerment and civil liberties. While panel-agnostic methods near unanimously estimate statistically significant and substantively significant (only for women's political empowerment), the estimated ATT for these outcomes under panel method is practically zero.

This finding is not incredibly surprising. In an analysis of a prior study, the developers of the panel matching estimator (Imai et al. 2021) found that increased refinement for prior history drove estimated findings closer to zero. However, in Imai et al. (2021), the authors consider the limited spatial-temporal range of treatment as a factor driving results using increased pre-treatment refinement towards zero. In their example, while treatment was found in small clusters across space and time, the near-opposite distribution of treatment is observed in this study, with most states having ratified each of these treaties and the time of ratification varying noticeably as well. As a result, this chapter remains skeptical that the distribution of treatment drives these null results post-refinement. Instead, this chapter views the panel matching technique as valid method to construct counterfactuals that share both similar covariate values *and* pre-treatment histories. When such detail is afforded to the construction of counterfactuals with panel data, the results happen to be pushed towards zero in this instance.

However, like any identification strategy reliant on selection on observables, there is a non-negligible chance that any of the estimates presented so far are completely false. After all, *if* a reader considers that one of Hill's (2010) covariates are a confounder necessary for adjustment and this covariate is not specified in the DAG-informed model, then any results from the DAG-informed models are biased and incorrect. This logic applies for *any* covariate that a researcher can make a theoretically interesting confounding case for. As a result, regardless of the method employed (regression adjustment, non-panel matching, non-panel weighting, panel matching, panel weighting, marginal structural models, etc.), a sensitivity analysis is necessary to examine the threat of unobserved confounding. Because the outcomes in this study are continuous, this chapter employs the sensitivity analysis design {sensemakr} developed by Cinelli and Hazlett (2020). Under this sensitivity analysis design, a researcher specifies a benchmark covariate that varying degrees of unobserved confounding are compared to. This benchmark covariate should be one that is theoretically assumed to, if omitted, heavily bias the results. Following the selection of the benchmark covariate, varying levels of strong unobserved confounders are introduced to the model, and the coefficient on the treatment variable is updated with each iteration of a stronger hypothetical confounder. If increasing levels of hypothetical unobserved confounders do not push a coefficient on treatment at or across zero, this indicates that your findings are fairly robust to unobserved confounding. While this does not *satisfy* the unconfoundness assumption, it does report on the degree to which violations of unconfoundness impact the estimate.

**Table 2. Sensitivity Analysis for Unobserved Confounding**

| | ICCPR | CAT | CEDAW - Political Empowerment | CEDAW - Civil Liberties |
|---|---|---|---|---|
| Original Estimate | 0.008 | 0.066 | 0.026 | 0.007 |
| Unobserved Confounder 1x as Strong as Democracy | -0.01 | -0.07 | 0.024 | 0.004 |
| Unobserved Confounder 2x as Strong as Democracy | -0.028 | -0.207 | 0.021 | 0.002 |
| Unobserved Confounder 3x as Strong as Democracy | -0.046 | -0.345 | 0.019 | -0.001 |

Table 2 reports the results of the sensitivity analysis with the effect of democracy on human rights treaty ratification *and* human rights outcome as the benchmark covariate. Given the strong theoretical links between democracy and normatively positive human rights outcomes and initiatives, the choice of democracy as a benchmark covariate seems appropriate. Additionally, the sensitivity of the DAG-informed models are reported given that sensitivity analyses are designed for models that have explicitly selected covariates for causal identification. Given that this sensitivity analysis is only possible through the estimation of the linear two-way fixed effects (TWFE) estimator, the original TWFE estimate is reported. The TWFE estimator generally does a good job at approximating the estimates from the panel models modeling the effect of human rights treaties on physical integrity rights and torture. However, the TWFE does a comparatively poor job at approximating the panel model estimates where CEDAW is treatment. Given this, less emphasis should be placed on the sensitivity analyses examining the sensitivity of the effect of CEDAW.

The results of the sensitivity analysis suggest that the effect of ICCPR on the protection physical integrity rights using the DAG-specified set of covariates is *highly* sensitive to potential unobserved confounding. A confounder that is as strong as the confounding effect of democracy is enough to push the estimate in a negative (albeit, small) direction. Importantly, this sensitivity analysis does not detail whether an unobserved confounder as strong as democracy *exists*. Although, given that repression and the decision to ratify human rights treaties are complex political phenomena, this chapter maintains that an unobserved confounding effect *at least as strong as democracy* is likely. Further, while this sensitivity analysis is limited towards hypothetical unobserved confounders that push an estimate *downwards*, it is also important to consider that, with the inclusion of an unobserved confounder, an estimate may move *away from* *zero*. However, because scientific research is generally primarily interested in examining whether non-zero relationships exist, this chapter considers the downward-biasing properties of unobserved confounding foremost.

Similarly, the estimated effect of CAT is also highly sensitive. Simply one unobserved confounder as strong as democracy is enough to make the estimate flip. In contrast to the sensitivity of the ICCPR effect, the sensitivity of the effect of CAT dramatically increases with each stronger iteration of an unobserved confounder. While less credibility can be afforded to the TWFE estimates for examining the effect of CEDAW, it is interesting nonetheless that, despite the estimates not being substantively significant, they are generally robust to varying degrees of unobserved confounding. These results may be due to the sensitivity of the parametric assumptions of the TWFE estimator itself, or these results may indicate that adjusting for democracy, NGO presence, economic interdependence, and military size is practically sufficient for causal identification (assuming unobserved confounders 4 times, 5 times, and *k*-times larger than democracy do not exist).

Overall, however, the results of this sensitivity analysis lend some support to the lack of causal interpretability for the results in this chapter. As stated previously, causal inference is *hard*. The task of causal identification only becomes more challenging when one operates under selection on observables. While causal theorizing is *necessary* for causal identification under selection on observables, it alone is not *sufficient*. A single paper's causal model can be wrong and/or incomplete. While it is impossible to directly test the assumption of unconfoundness, these sensitivity analysis represent the next-best approach. Unfortunately, for the results of this chapter, the sensitivity analysis do not suggest that these results *probably* satisfy the ever-demanding assumption of unconfoundness. Undoubtedly, causal identification for this research can be greatly improved by increased attention to the causal modeling of the human rights treaty ratification → repression data generating process (DGP) so that selection on observables designs can become more robust. Alternatively, researchers should continue to examine alternative non-selection on observables strategies and exhaust every source of natural randomization they can find (such as regression discontinuity designs and instrumental variables with exhaustive efforts to demonstrate that selected instruments are *valid* instruments).

**Conclusion**

Do human rights treaties work? Following the implementation of novel panel matching/weighting estimators that match/weight on a set of variables informed by a DAG, the answer to this question remains inconclusive. Point estimates are somewhat sensitive to varying model specifications and sensitivity analyses suggest that unobserved confounding of varying degrees probably exists within this research design, rendering a causal interpretation of results unlikely.

For this research design in particular, a number of improvements can be made in future research to increase the prospects of causal identification. First, and most importantly, the causal model provided in this paper may be *wrong* or *insufficient*. There is no agnostic, mathematical way to confirm whether one causal model is “more correct” than another. However, substantive familiarity in a given field is incredibly valuable in the development of a rigorous causal model that accurately reflects the true DGP. Second, selection on observables is generally the weakest identification strategy and its limitations concerning unobserved confounding are present within this chapter. Future research should explore identification strategies beyond selection on observables that are less reliant on adjusting for *all* confounders to estimate causal effects. However, this remains one of the foundational problems with conflict management and human rights research given that naturally-occurring randomization of treatment is incredibly rare. If stronger identification strategies prove too difficult to find (which this chapter is skeptical of given the overall lack of attention afforded to identification strategies in quantitative peace research), much more careful theorizing is required and much time and effort should be invested in the deliberation and development of explicit causal theories that both explain causal mechanisms for *how* treatment impacts outcome *and* identify as many confounding factors as possible.

One may come to the conclusion and ask, "why add all of these additions to the research design only to end up with non-causal results?" After all, the estimation of non-causal results was, in part, a major critique this paper levies against broader social scientific scholarship. This is a natural critique, but it ignores the primary message of this chapter. The *goal* of almost all scientific inquiry is naturally causal. As a result, our research designs should be structured and designed to estimate causal effects. However, continuing the mantra, causal inference is *hard*. As demonstrated in this paper, not even a carefully crafted research design will always be sufficient for causal identification. The results in this chapter demonstrate this reality. As a result, at best, the results from this paper should be interpreted as merely descriptive. As Gerring (2012) argued, if the goal of a project is not description, and a researcher has carefully crafted a causal research design, *and* the results do not meet the sufficient conditions for causal identification, a researcher may, only then, claim that their results are merely descriptive.

Indeed, researchers ought to be entirely transparent concerning the causal interpretation of their results. After all, in this paper, a DAG was constructed to directly identify confounding effects, a treatment effect was explicitly stated, a method designed explicitly for making causal inferences with panel data was employed, *and* a sensitivity analysis was executed to assess the potential bias of unobserved confounding. Despite all of these steps, causal identification was not achieved. As a result, it is very likely that projects which are not as exhaustive as this chapter also fail to meet the necessary conditions to make causal inferences and interpret their estimates as “effects”, “determinants”, “consequences” or any other term that inherently implies a causal relationship. Where causal inferences are not achieved, researchers ought to be explicit with these limitations while simultaneously crafting their research designs to explicitly estimate causal effects.

## References

Abadie, Alberto. 2021. “Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects.” *Journal of Economic Literature* 59(2): 391–425.

Aronow, Peter M., and Cyrus Samii. 2016. “Does Regression Produce Representative Estimates of Causal Effects?” *American Journal of Political Science* 60(1): 250–67.

Barbieri, Katherine, and Rafael Reuveny. 2005. “Economic Globalization and Civil War.” *The Journal of Politics* 67(4): 1228–47.

Bell, Andrew, and Kelvyn Jones. 2015. “Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data.” *Political Science Research and Methods* 3(1): 133–53.

Bernhard, M. et al. 2015. “The Varieties of Democracy Core Civil Society Index.” *V-Dem Working Paper Series* 13.

Blackwell, Matthew. 2012. “A Framework for Dynamic Causal Inference in Political Science.” *American Journal of Political Science* 57(2): 504–20.

Blackwell, Matthew, and Adam Glynn. 2018. “How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables.” *V-Dem Working Paper Series* 67.

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2022. “A Crash Course in Good and Bad Controls.” *Sociological Methods & Research* 0(0): 1–34.

Cinelli, Carlos, and Chad Hazlett. 2020. “Making Sense of Sensitivity: Extending Omitted Variable Bias.” *Journal of the Royal Statistical Society* 82(1): 39–67.

Coppedge, Michael et al. 2022. “V-Dem Codebook V12.” *Varieties of Democracy (V-Dem) Project*. https://www.v-dem.net/data/the-v-dem-dataset/.

Davies, Shawn, Therése Pettersson, and Magnus Öberg. 2022. “Organized Violence 1989–2021 and Drone Warfare.” *Journal of Peace Research* 59(4): 593–610.

Dworschak, Christoph. 2023. “Bias Mitigation in Empirical Peace and Conflict Studies: A Short Primer on Posttreatment Variables.” *Journal of Peace Research* 0(0): 1–15.

Fariss, Christopher J., Therese Anders, Jonathan N. Markowitz, and Miriam Barnum. 2022. “New Estimates of Over 500 Years of Historic GDP and Population Data.” *Journal of Conflict Resolution* 66(3): 553–91.

Gerring, John. 2012. “Mere Description.” *British Journal of Political Science* 42(4): 721–46.

Gleditsch, Nils Petter et al. 2002. “Armed Conflict 1946-2001: A New Dataset.” *Journal of Peace Research* 39(5): 615–37.

Grosz, Michael P., Julia M. Rohrer, and Felix Thoemmes. 2020. “The Taboo Against Explicit Causal Inference in Nonexperimental Psychology.” *Perspectives on Psychological Science* 15(5): 1243–55.

Hernán, Miguel A. 2018. “The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data.” *American Journal of Public Health* 108(5): 616–19.

Hill, Daniel W. 2010. “Estimating the Effects of Human Rights Treaties on State Behavior.” *The Journal of Politics* 72(4): 1161–74.

Imai, Kosuke, and In Song Kim. 2019. “When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?” *American Journal of Political Science* 63(2): 467–90.

———. 2021. “On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data.” *Political Analysis* 29(3): 405–15.

Imai, Kosuke, In Song Kim, and Erik H. Wang. 2021. “Matching Methods for Causal Inference with Time-Series Cross-Sectional Data.” *American Journal of Political Science* 0(0): 1–19.

King, Gary, and Langche Zeng. 2006. “The Dangers of Extreme Counterfactuals.” *Political Analysis* 14(2): 131–59.

Liu, Licheng, Ye Wang, and Yiqing Xu. 2022. “A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data.” *American Journal of Political Science*: ajps.12723.

Lundberg, Ian, Rebecca Johnson, and Brandon M Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86(3): 532–65.

Marshall, M. G., and K. Jaggers. 2020. “Polity5 Project: Political Regime Characteristics and Transitions, 1800-2018.” http://www.systemicpeace.org/inscrdata.html.

McGowan, Lucy D’Agostino. 2022. “Tipr: An R Package for Sensitivity Analyses for Unmeasured Confounders.” *Journal of Open Source Software* 7(77): 1–6.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Second Edition. Cambridge University Press.

Pemstein, Daniel et al. 2022. “The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data.” *V-Dem Working Paper Series* 21.

Plümper, Thomas, and Vera E. Troeger. 2019. “Not So Harmless After All: The Fixed-Effects Model.” *Political Analysis* 27(1): 21–45.

Rohrer, Julia M. 2018. “Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data.” *Advances in Methods and Practices in Psychological Science* 1(1): 27–42.

Rohrer, Julia M., and Kou Murayama. 2023. “These Are Not the Effects You Are Looking for: Causality and the Within-/Between-Persons Distinction in Longitudinal Data Analysis.” *Advances in Methods and Practices in Psychological Science* 6(1): 1–14.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66: 688–701.

Singer, J. David. 1987. “Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816-1985.” *International Interactions* 14: 115–32.

Singer, J. David, Stuart Bremer, and John Stuckey. 1972. “Capability Distribution, Uncertainty, and Major Power War, 1820-1965.” In *Peace, War, and Numbers*, Beverly Hills: Sage, 19–48. https://correlatesofwar.org/data-sets/national-material-capabilities/.

Sundström, A., P. Paxton, Y.-T Wang, and S. I. Lindberg. 2017. “Women’s Political Empowerment: A New Global Index, 1900–2012.” *World Development* 94: 321–55.

Wysocki, Anna C, Katherine M Lawson, and Mijke Rhemtulla. 2022. “Statistical Control Requires Causal Justification.” *Advances in Methods and Practices in Psychological Science* 5(2): 1–19.