PS531 Pre-Analysis Plan

Negotiating Justice: Conflict Amnesties in the Era of Accountability

Myung Jung Kim

1 Introduction

1.1 Research Question

"The failure to prosecute … perpetrators such as Pol Pot, Idi Amin, and Saddam Hussein convinced the Serbs and Hutus that they could commit genocide with impunity" (Akhavan (2009), 629). To fight against such vicious cycle of injustice, the international community has been striving to end impunity for grave human rights violations. The effort culminated around 1998 with the rise of the International Criminal Court (ICC) and Universal Jurisdiction (UJ)[1] which enabled the overriding of domestic amnesties for serious crimes against international law including genocide, war crimes, and crimes against humanity. This meant that even if a perpetrator has been amnestied by his home country, he can now still be prosecuted before international and foreign courts. Ban Ki-moon even claimed that such change brought forth the transition from the "era of impunity" to "era of accountability" (Ki-moon (n.d.)). Indeed, the rise of the ICC and UJ (hereafter, the anti-amnesty international regimes) stirred up a fierce discussion among academia and peace practitioners, which is often called as the 'peace versus justice debate' which was based on the conventional belief that the advent of the ICC and UJ would complicate states' use of amnesty as a peacemaking tool in conflicts (Goldsmith and Krasner (2003), Snyder and Vinjamuri (2003), Ginsburg (2009), Prorok (2017), Reiter (2010), Kim and Sikkink (2010), Simmons and Danner (2010)).

Contrary to the traditional belief of legal and political science scholars, however, recent studies find that states not only persistently grant SV amnesties, but even increase its usage after the rise of the ICC and UJ (Mallinder (2012), 95). This raises a puzzle: why do we witness a persistent use of SV amnesties despite the advent of ICC and UJ? What explains the mismatch between the conventional wisdom and the recent findings? This paper aims to provide a theory to answer this question. I argue that the UJ and ICC, by increasing the

[1] The term "Universal Jurisdiction (UJ)" refers to the idea that a national court may prosecute individuals for serious crimes against international law –such as crimes against humanity, war crimes, genocide, and torture –based on the principle that such crimes harm the international community or international order itself, which individual States may act to protect (International Justice Resource Center). To date, 163 out of the 193 UN member states that incorporate Universal Jurisdiction under national law, and they can potentially overrule amnesties for serious violations to act like an international court to prosecute international crimes (Amnesty International 2012, 2).

risk of foreign and international prosecutions, increases the demand of SV amnesties from the perpetrators of international crimes and hence the use of it.

2 Theory

In order for an amnesty deal to be stricken, it requires the willingness to grant the amnesty in the perspective of the granter's side and the demand to receive such amnesty in the recipient's side. Hence, many scholars, although largely untested, have suggested that the persistent use of SV amnesties can be traced by the amnesty granter's side (the supply factors) and the recipient's side (the demand factors). The supply factors that may prolong state's use of SV amnesties include states' "perceived utility [of amnesties] for ending violence" (Jeffery (2014)), and "the desire to protect certain perpetrators from prosecution" (Sikkink (2012), 8). The demand factor include the increasing demand of amnesties from perpetrators who now should feel "the credible threat [of international prosecution and] ... at least the perception of the value [of amnesty]"(Sly (n.d.), Mallinder (2012)). While acknowledging the existence of the supply side, this paper focuses on the demand side – how the rebel group's risk of the ICC and UJ prosecutions affect the use of SV amnesties.

The conventional wisdom is that the new rise of international anti-amnesty regimes deter the use of SV amnesties mainly by creating the commitment problem between the amnesty granters (i.e., states) and potential recipients (i.e., culpable rebels) (Goldsmith and Krasner (2003), Snyder and Vinjamuri (2003), Ginsburg (2009), Prorok (2017)). However, a culpable rebel can still be free from the commitment problem by expecting an amnesty to let him stay safe at least in the home country or other neighboring states that is likely to respect the domestic amnesty more than international norms to punish him at its own yard (using UJ) or by handing him over to the ICC. In other words, if a culpable rebel group face a high risk of ICC/UJ prosecution, it can seek out for amnesty and stay safe as long as he is in the amnesty-granting country. If this theory holds, rebel groups that face higher risk of foreign and international prosecutions should demand more SV amnesties and hence have higher possibility of receiving SV amnesties than groups that face lower risk of ICC/UJ prosecutions. Based on the theory, I come up with the following hypothesis.

Hypothesis: With the advent of the anti-impunity regimes, rebel groups that face greater risk of foreign and international prosecutions receive more SV amnesties than rebel groups that face lower risk of foreign and international prosecutions.

3 Research Design

The main comparison of this study is SV amnesties before and after the rise of the anti-amnesty regimes. More specifically, in empirical terms, this study hypothesizes that there is an interaction effect between the rise of the anti-amnesty regimes and a rebel group's risk of foreign and international prosecutions on the likelihood of the rebel groups' receiving of SV amnesties. To test the hypothesis, I make an as-if randomized comparison using propensity score matching with observational data. I run the propensity score estimation separately for two time-periods – before and after the rise of the anti-amnesty regimes. Research design and identification strategies are discussed in great detail below.

3.1 Data

I use Dancy's Conflict Amnesty Data which provide information on states' issue of amnesties for civil wars from 1945 to 2014 (Dancy (2018)). Since my main interest is to examine SV amnesties which usually occur once or twice in a state-rebel dyad conflict, I collapse the original data's *yearly* observations of dyad (a state-a rebel) civil conflicts into *event* observations to prevent overfitting (i.e., years of a state-rebel dyad conflict is one observation). Additionally, while the original data identify whether the amnesties cover serious crimes or not, they not identify whether the amnestied rebel groups indeed committed serious crimes. It means that some rebel groups may have received amnesties that cover a wider coverage of crimes (i.e., serious crimes) than the actual crimes that they have committed. To complement this issue, I identify rebel groups' reported involvement in serious crimes including civilian killing, child soldier, and sex crimes using the UCDP One-sided violence data set(Eck (2007)), the Haer and Böhmelt (2017) data set (Haer and Böhmelt (2017)) and the SVAC data set (Cohen and Nordås (2014)) respectively. The unit of analysis is a state-rebel dyad. My data have observations of 514 dyad conflicts of 105 countries.

3.2 Variables and Measures

3.2.1 Response Variable

The dependent variable is coded 1 if there has been any exchange of SV amnesties in state-rebel group dyad conflict. Data show that SV amnesties are usually exchanged once in a state-rebel dyad conflict, if there is any (86.8%). Yet, in some wars, SV amnesties were granted multiple times (at most five times), probably due to failed attempts to resolve wars even by issuing amnesties. Among 514 dyad conflicts in data, 76 cases involved with exchanging of SV amnesties.

3.2.2 International Anti-amnesty Regimes (ICC, UJ)

I use year 1998 to indicate the key independent variable – the emergence of anti-amnesty regimes. In this year, both ICC and UJ emerged together accidentally, and the 1998-cutoff is widely used in literature to indicate the transition from the era of impunity to the era of accountability (Dancy (2018), Krcmaric (2018), Daniels (2020)). Using the indicator, I categorize conflicts into three types: Pre-98 wars, Post-98 wars, and Ongoing-98 wars. They represent wars that ended before 1998, wars that started after 1998, and wars that were ongoing in 1998 (i.e., that started before 1998 and ended after 1998 (e.g., 1980-2010)) respectively. Using them, I make two comparisons: First is to compare SV amnesties in

Pre98 wars withSV amnesties in *Post98.* This comparison would be the sharpest since Pre-98 and Post-98 amnesties are clearly without and with the potential effect of the ICC and UJ respectively. Second, I can compare SV amnesties in *Pre-98 wars* with SV amnesties in *Ongoing*-98 *wars.* This comparison is also theoretically suitable because states generally grant amnesties at the end-stage of a conflict. In the actual paper, I will report both comparisons, but this pre-analysis mainly discusses the latter comparison using `Ongoing98` dummy. In the whole data set, pre-98 conflicts comprise about 59% of observations (N =295), post-98 conflicts about 27 % (N =136), and cross-98 conflicts about 13% (N = 67) (i.e., ongoing-98 conflicts (N = 203; 40%).

3.2.3 Rebel's Risk of Prosecutions

To test for the conditional impact of anti-amnesty regimes, I interact the impact of anti-amnesty regimes with a measure of rebel's risk of foreign and international prosecutions. In order to measure the level of risk, I use the binary indicator of rebel's type – whether a rebel group is a transnational rebel groups (TNRs) that operate across state borders with foreign sanctuaries or local rebel groups. This is based on my theoretical claim that TNRs face greater risk of foreign and international prosecutions than local rebel groups that operate only within its national territory. State boundaries are *de facto* lines of defense against foreign aggression (Salehyan 2007, 220), and international and foreign courts require state cooperation to apprehend suspects. For this reason, amnestied perpetrators are most likely to stay safe from arrest by foreign and international actors as long as they stay in the amnesty-granting home country. This makes local rebel groups face a lower risk of foreign or international prosecutions than TNRs. Local rebel groups have little worry whether amnesties would be overridden by the ICC or UJ. Yet, TNRs with foreign-based assets and facilities are more likely to linger outside the home country and hence confront a higher risk of arrests of external actors. Indeed, many high-ranking rebels indicted by the foreign and international courts were arrested in foreign territories, including Straton Musoni (head of the FDLR (Rwanda) arrested in Germany), Mohammed Jabbateh (a high-ranking officer of ULIMO (Liberia) arrested in the U.S.), and Charles Blé Goudé (former leader of Congrès Panafricain des Jeunes et des Patriotes (Ivory) arrested in Ghana) to name a few.

3.3 Identification Strategy

To draw a causal inference (i.e., to understand an effect of any treatment), a researcher should be able to answer what would have happened to a group that is not treated (i.e., the counterfactual). In other words, one needs a precise comparison group – which are equivalent except for the fact that one of them received the treatment. Such setting is possible in an experimental setting in where a researcher has a control over data generation. However, this condition is difficult to be met in an observational study in which "[a] investigator cannot control the assignment of treatments to subjects" (Rosenbaum (2010), vii). Since the treated subjects and non-treated subjects are not randomly selected, the studies suffer from biases (i.e., differences between treated and control groups) before treatment. In other words, it is difficult to say whether the differences in outcome between the treated and control groups are due to “chance” or “the real treatment effect.” Hence, while observational studies can draw information on key variables and their associations with its low complexity, low cost, and low ethical constraints, they are far limited in drawing a causal inference compared to a randomized experimental design.

3.3.1 Propensity Score Matching

One approach to account for this limitation is to conduct a propensity score matching which enables an as-if randomized comparison by drawing a more sensible comparison group. Propensity score matching pairs subjects based on their propensity score – the conditional probability of treatment given the observed covariates (Rosenbaum (2010), 72). By this, it effectively reduces observed biases and makes it possible to draw and compare the treated and control subjects. As the single variable summarizes relevant information in all observed control variables, one only needs to match on this scalar variable. For this reason, there is no limit on the number of covariates for adjustment, and it makes matching simpler and free from the curse of dimensionality. Most importantly, researchers can assess whether the adjustment is done enough by looking at the balance of observed covariates between control and treated units. Researcher can change model specification until a good balance is achieved. Such advantages are something unthinkable in usual regression analysis. Yet, a propensity matching strategy still has its limits. In most cases, the true propensity score is unknown, and hence it has to be estimated by modeling the receipt of treatment given observed covariates (Imai (2005)). It means that bias can still arise from the process of researcher’s choice of covariates in specifying the propensity score and unobserved covariates (Rosenbaum (2010), 73). Also, it discards unmatched units (Rubin (2002)). Lastly, it is difficult to see the effect of matching variables on the outcome variable (Thavaneswaran (2008)). Despite the limitation, this study attempts to overcome potential bias from observable covariates and reduce doubt of the result by transparently explaining the model specification and choices.

3.3.1.1 The Treatment (TNRs) I use the binary indicator of rebel group’s type being transnational (`TNR`) as a treatment. The control group is the observations of local rebel groups (TNR = 0). The hypotheses predict that the treatment effect (`TNR`) on SV amnesties is only valid after the rise of the anti-amnesty regimes (post-1998). To examine the treatment effect heterogeneity, I test treatment effects for pre-1998 conflict observations (hereafter, pre98 subgroup) and post-1998 conflict observations (hereafter, post98 or ongoing98 subgroup depending on the cutoff point). I use the NSA data to distinguish whether the rebel group is a transnational. The NSA data’s variable `Rebpresosts` indicates whether the rebel group operates to at least some extent outside the home country’s borders. While the variable in the original dataset is trichotomous (‘no,’ “some,” and “extensive”), I make them dichotomous. Among 414 dyads, there are 187 unique rebel groups captured in the dataset, and among them, there are 76 transnational rebel groups (TNRs) and 98 local rebel groups (no info about 13 groups). There are 201 amnesties granted to local-rebel group and 139 amnesties to TNRs.

3.3.1.2 PS Score Model Specification To estimate the propensity score, I use logistic regression where I include available covariates that would statistically balance the covariates between the treated and control groups. Brookhart et al. (2006) shows that the best PS model is the model that include all predictors of outcome regardless of whether they are associated with exposure. Accordingly, I specify the propensity score using variables that may affect SV amnesties as suggested in the earlier studies. Dancy 2018 suggest that judicial independence (`judicialinde`), democratic transitions (`demtrans`), number of years at war (`yearsatwar`), territory (`territory`), intensity (`intensity`), ethnic (`ethnic`), number of other groups fighting (`numdyads`), rebel strength (`rebcap`), fighting capacity (`fightcap`), and bloody hands (`blood`) can affect the number of amnesties (Dancy (2018)). Additionally, I include a variable that indicates an involvement of a third-party mediation (`mediation`) and rebel groups' actual involvement of serious violations (`sv`).

3.3.1.3 Missing Data Theories behind propensity score analysis assume that the covariates are fully observed (Paul R. Rosenbaum and Rubin (1983)). However, in practice, missingness in the covariates is sometimes inevitably. The two common solutions to deal with the missingness are 1) imputation such as filling the mean values or zero to missing observations. and 2) omitting the observations. In this study, missing data are mainly caused by merging of multiple data sets which cover different time periods. Hence, imputing the missing values as 0 or mean value would be inappropriate. As long as missingness does not depend both on the outcome variable and treatment variable, this bias is generally small. Since there is no theoretical base to believe that the missingness in this study is related to any of these, I ignore the missing data. After removing the missing data, there are 240 observations for `pre98` wars and 100 observations for `ongoing98` wars.

3.3.1.4 Matching Method There are multiple ways of matching treated and untreated units such as nearest neighbor matching, Mahalanobis metric matching, and caliper matching. Among various options, I use the full matching to form weights and to analyze the outcome (Stuart EA and KM (2008)). The matched sets are created in a way that minimizes the global PS difference, defined as the sum of the distances between the PS of all pairs of treated and comparison individuals within each matched set, across all matched sets (Stuart EA and KM (2008)). Full matching makes use of all units in the data by forming a series of matched sets in which each set has either one treated unit and one or more control units or one control units and one or more treated units (B. B. Hansen (2004)). The exposed units that have many comparison units with similar propensity scores will be grouped with many comparison units, whereas exposed units with few similar comparison units will be grouped with relatively fewer comparison units (Kerry M. Green and Stuart (2014)). Full matching uses original scores just to create the subclasses, not to form the weights directly (Hansen Ben B. and Klopfer (2006)), and hence it is less sensitive to the form of the propensity score model and known to form the subclasses in an optimal way (B. B. Hansen (2004)). Lastly and most importantly, while other distance matching methods cannot estimate the average treatment effect (ATE) but only the average treatment effect of the treated (ATT), the full matching can be used to estimate the ATE (Peter C Austin and Stuart (2015))–which this paper aims to estimate. Table 1 Table 2 show the structures of matched sets for Pre-98 subgroup and Ongoing-98 subgroup, and they have 94.1 and 33.9 matched pairs (effective sample size) respectively.

	x
10:1	1
9:1	1
7:1	1
5:1	1
4:1	1
3:1	3
2:1	8
1:1	32
1:2	9
1:3	5
1:4	5
1:5	1
1:7	1
1:9	2
1:14	2

Table 1: Structure of Matched Sets for pre98

	x
11:1	1
6:1	2
5:1	1
2:1	1
1:1	7
1:2	4
1:3	2
1:4	1
1:6	1
1:8	1
1:9	1

Table 2: Structure of Matched Sets for ongoing-98

3.3.1.5 Balance of Covariates

If the propensity score is estimated properly, the distribution of covariates should be similar between treated and matched control units (Ben B. Hansen and Bowers (2008), Imai (2005)). I will judge the success of the adjustment by looking at the balance of covariate distributions in the treatment and control groups after matching. I first conduct a balance test before matching to calculate standardized differences across covariates without the stratification. Table 3 and 4 test results are shows that the chi-square value and the p-value are 59.5 and 4.58e-09 for pre-98 data subset and 24.9 and

0.00554 for ongoing-98 subset. They suggest that there are considerable differences between the treatment and control groups for both pre- and ongoing- datasets. Such difference makes it difficult to induce a good comparison, and hence shows why propensity score matching can be useful in this study.

Table 3: Balance before Matching for Pre-98

	chisquare	df	p.value
raw	59.46	10.00	0.00

Table 4: Balance before Matching for Ongoing-98

	chisquare	df	p.value
raw	24.90	10.00	0.01

After propensity score matching, the chi-square and p-value are 1.04 and 1.00e+00 for pre-98 dataset and 1.66 and 0.99836 for ongoing-98 dataset (Table 5, Table 6). They suggest that the treatment and control groups are not too different and hence a good comparison group. The balance of each covariate distributions before and after matching are nicely visualized in Figure 1 and 2 which illustrate the `xBalance` results for Pre-98 and Ongoing-98 war observations (Ben B. Hansen and Bowers (2008)). For both Pre-98 and Ongoing-98 datasets, the standardized differences of control and treatment group became closer to 0 for all covariates after matching. Hence, I consider the adjustment successful. Table 7 and Table 8 show the pre- and post-matching balances in more detail for individual covariates for the pre- and ongoing-98 datasets respectively.

Table 5: Balance of Pre-98

	chisquare	df	p.value
raw	59.46	10.00	0.00
ps_pre98	1.04	10.00	1.00

3.4 Statistical Tests

A test statistic summarizes the relationship between treatment and observed outcomes using a simple number (i.e., a point estimate). However, relying on a single test statistic and a p-value from it can be misleading because the observed test statistic can be a extreme one from the perspective of the distribution of test statistics. This can cause an incorrect rejection of null hypothesis which is called the false positive error. Hence, the better way of estimating the false positive errors would be by repeating the study, calculating the test statistics, and then assessing the distribution of the test statistics that could have occurred if the null hypothesis were true. This process can be done by simulation. I plan to use the mean difference test statistic and conduct a simulation to get a confidence interval.

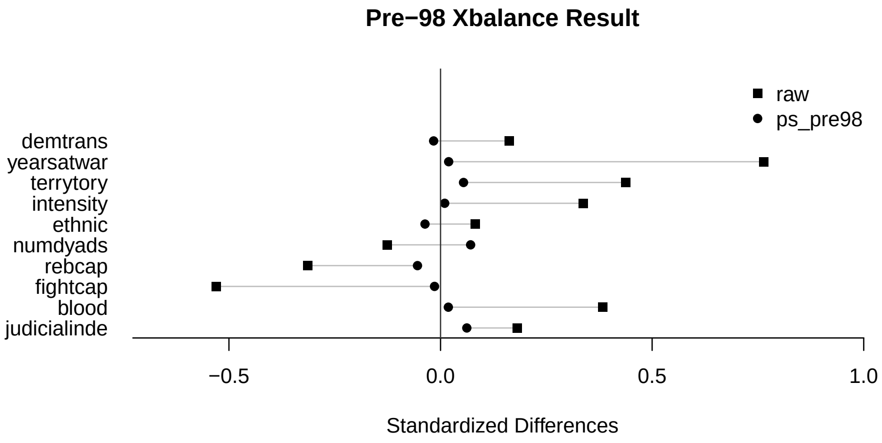

Figure 1: Balance Test for Pre-98

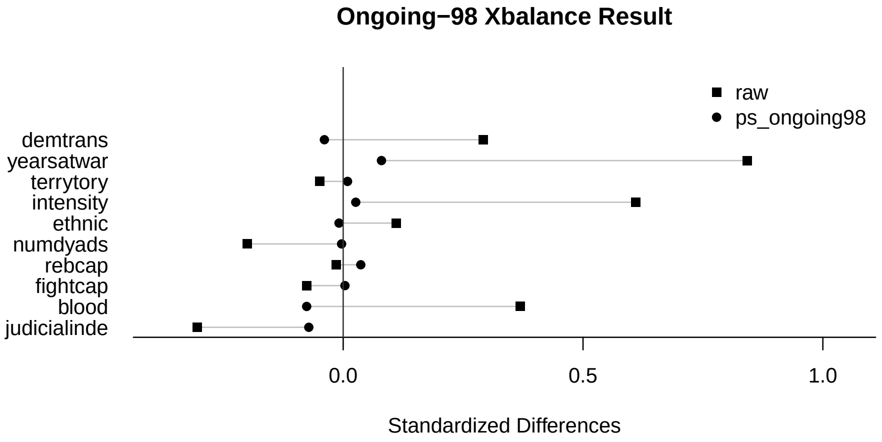

Figure 2: Balance Test for Ongoing-98

Table 6: Balance of Ongoing-98

	chisquare	df	p.value
raw	24.90	10.00	0.01
ps_ongoing98	1.66	10.00	1.00

3.4.1 Performance of the Tests

I will judge the performance of tests by looking at the false positive rate and power. The power of a test is denoted by "(1-β)," and it is the probability of a true positive or the probability of avoiding a false negative. It ranges from 0 to 1, and as the power increases, the probability of making type II error (false negative) decreases. A false positive rate is the probability of a type I error. The false-positive rate of the test that makes up the confidence interval is the same as the coverage probability of a confidence interval. I will not judge the test performance by Family-wise error rate (FWER). Family-wise error rate (FWER) is the probability of making one or more false discoveries, or Type I errors (i.e., incorrectly rejecting the null hypothesis when the null hypothesis is true). This is usually inflated when performing multiple hypotheses tests. In this case, p-value has to be adjusted using Bonferroni correction or adjusting false discovery rate. However, this study does not involve any multiple testing. Also, I collapse all the yearly observations into a state-rebel dyad, so there is little concern with overfitting issue.

3.4.1.1 False Positive Rate and Power Table 5 shows powers and false positive rates using two different estimators for Pre- and Ongoing-98 data sets. Here, I show the `lm_robust` and `lm` estimator not propensity score matching OLS, since I can easily compare the model using DeclareDesign. Both are based on 0.05 thresholds for p-value (i.e., power = mean(p.value <0.05); the same for false positive rates) and are based on 500 times of simulations.

The powers of the estimators are both very low–all four power values are less than 0.1. It indicates that these models are likely to cause a high false negative. The false positive rates are also low–all less than 0.1. It indicates that the two models are less likely to cause type I error. The low power could be caused due to frequent zero outcome variables or due to the change of outcome variable as a binary outcome.

3.4.2 Estimator and Estimand

This paper aims to estimate (estimand) the average treatment effect (ATE) which can be characterized as $E((Y_i(1) - Y_i(0))$. In plain words, the ATE is the likelihood of receiving of SV amnesty of the treated units (TNR) minus the likelihood of receiving of SV amnesty of the control units (local rebel groups). I calculate the ATE for both Pre-98 subset and Ongoing-98 subset. As discussed earlier, a full matching is a good option for study with ATE as a target of estimand, and it once again explains why I aim to use a full matching.

If the outcome of a randomized trial is dichotomous, the treatment effect can be expressed as several effect estimates, including the risk difference, the risk ratio (RR), the odds ratio(OR),

Table 7: Differences in Observed Characteristics between Treatment and Control Group Before and After Matching Adjustment for Pre-98 Wars

<Before>	TNR=0	TNR=1	adj.diff	adj.diff.null.sd	std.diff	z	
judicialinde	0.24	0.28	0.04	0.03	0.18	1.49	
demtrans	0.10	0.15	0.05	0.04	0.16	1.33	
yearsatwar	2.77	6.73	3.95	0.67	0.76	5.88	***
terrytory	0.21	0.41	0.20	0.06	0.44	3.52	***
intensity	0.50	0.66	0.16	0.06	0.34	2.73	**
ethnic	0.06	0.08	0.02	0.03	0.08	0.67	
numdyads	1.72	1.58	-0.14	0.13	-0.13	-1.03	
rebcap	-0.94	-1.23	-0.29	0.12	-0.31	-2.55	*
fightcap	0.61	0.30	-0.31	0.07	-0.53	-4.21	***
blood	0.12	0.26	0.15	0.05	0.38	3.10	**
mediation	0.08	0.15	0.06	0.04	0.20	1.67	.
<After>	TNR=0	TNR=1	adj.diff	adj.diff.null.sd	std.diff	z	
judicialinde	0.26	0.25	-0.01	0.03	-0.04	-0.28	
demtrans	0.10	0.09	-0.01	0.05	-0.03	-0.21	
yearsatwar	4.10	4.22	0.12	0.35	0.02	0.35	
terrytory	0.32	0.30	-0.02	0.07	-0.06	-0.38	
intensity	0.58	0.59	0.00	0.07	0.01	0.06	
ethnic	0.08	0.08	-0.01	0.04	-0.03	-0.16	
numdyads	1.61	1.71	0.10	0.15	0.09	0.65	
rebcap	-1.15	-1.15	-0.01	0.12	-0.01	-0.06	
fightcap	0.43	0.41	-0.01	0.06	-0.02	-0.18	
blood	0.17	0.19	0.03	0.05	0.07	0.55	
mediation	0.12	0.14	0.02	0.05	0.08	0.52	

and the hazard ratio which is used for time-to-event-data ((**Knol_etal2011?**)). This study will use the risk difference which is one of the common different measures of effect for study with binary outcomes (Peter C Austin and Stuart (2015)). The estimator of the risk difference is $p_1 - p_0$ when p_1 and p_0 denote the probability of the outcome in treated and control subjects, respectively.

The two primary methods that have been shown to perform well in matched samples are using cluster-robust standard errors and the bootstrap.

Standard error estimation involves a cluster-robust standard error as recommended by Austin and Stuart 2017.

I can use a weighted generalized linear model regressing the outcome on the treatment with a link function appropriate to the effect measure of interest. I will use `sandwich` package which allows **.

I will conduct a series of Monte Carlo simulations to examine the performance of full matching on the propensity score for estimating the effect of treatment on binary outcomes (Peter C Austin and Stuart (2015)). The methods' performances were assessed using the two criteria:

Table 8: Differences in Observed Characteristics between Treatment and Control Group Before and After Matching Adjustment for Ongoing-98 Wars

<Before>	TNR=0	TNR=1	adj.diff	adj.diff.null.sd	std.diff	z	
judicialinde	0.32	0.25	-0.08	0.05	-0.30	-1.51	
demtrans	0.08	0.17	0.09	0.07	0.29	1.45	
yearsatwar	2.00	5.79	3.79	0.97	0.84	3.89	***
terrytory	0.32	0.30	-0.02	0.09	-0.05	-0.25	
intensity	0.47	0.74	0.28	0.09	0.61	2.92	**
ethnic	0.06	0.09	0.03	0.05	0.11	0.55	
numdyads	1.90	1.70	-0.20	0.20	-0.20	-1.00	
rebcap	-1.25	-1.26	-0.01	0.13	-0.01	-0.07	
fightcap	0.36	0.32	-0.04	0.10	-0.08	-0.38	
blood	0.43	0.62	0.18	0.10	0.37	1.82	
mediation	0.36	0.47	0.11	0.10	0.22	1.11	
<After>	TNR=0	TNR=1	adj.diff	adj.diff.null.sd	std.diff	z	
judicialinde	0.28	0.29	0.01	0.06	0.02	0.09	
demtrans	0.06	0.06	-0.00	0.08	-0.01	-0.05	
yearsatwar	2.65	2.88	0.23	0.56	0.05	0.41	
terrytory	0.20	0.29	0.08	0.11	0.18	0.75	
intensity	0.64	0.61	-0.03	0.09	-0.07	-0.33	
ethnic	0.05	0.06	0.02	0.07	0.06	0.25	
numdyads	1.96	1.77	-0.19	0.23	-0.18	-0.82	
rebcap	-1.33	-1.40	-0.07	0.17	-0.10	-0.42	
fightcap	0.36	0.29	-0.07	0.13	-0.14	-0.56	
blood	0.51	0.45	-0.06	0.13	-0.13	-0.49	
mediation	0.44	0.36	-0.08	0.13	-0.16	-0.61	

(1) bias in estimating the true treatment effect; and (2) the mean squared error (MSE) of the estimated treatment effect.

After creating a population, I randomly slected 200 observation to generate my simulated data.

```
## [1] 0.26
```

Table 9: Power and False Positive Rates

	estimator_label	power	se(power)	false_positive_rate
1	lm_robust	0.08	0.01	0.08
2	lm	0.04	0.01	0.04
2	lm_robust	0.00	0.00	0.00
2	lm	0.06	0.01	0.06

3.5 Estimators and Estimand

Based on the information presented in Table 5 and 6, and considering the benefit of keeping the treatment (`democracy index`) as continuous variable, I will use `lm_robust` as the statistical estimators. As seen in Table 1, this observational study suffers bias from background confounders. In order to tackle this issue, adjustment for covariate is very critical. While propensity score matching would be a great way to overcome the problem, the nature of outcome variable in this study makes it difficult to use the matching strategy. `lm_robust` has efficiency in large samples and low bias in small samples as well as similarities to design-based randomization estimators (Samii and Aronow 2012).

3.6 Performance of Estimators

The performance of estimators can be evaluated by looking at characteristics like the bias, consistency, coverage, and mean squared error (MSE). The RMSE (Root Mean Squared Error) is the standard deviation of the residuals that measures how well the data values fit the line of best fit. Unbiased estimator means that the estimator or test statistic is accurate to approximate the parameter. Lastly, coverage rates refer to the coverage probability of the confidence intervals. The covarge probability shows how often we obtain a confidence interval that contains the true population parameter if we were to repeat the entire sampling and analysis process. I will judge the performance of estimators using bootstrapping (i.e., resampling with replacement).

3.6.1 Biases and MSE

I use the `diagnose_design` function in `DeclareDesign` to generate a table showing the characteristics discussed in 11. Table 6 shows the diagnose for estimators using two different estimators and each with 500 simulations. Biases are generally low–0.01 or below. Since bias is close to zero, it indicates that the estimators for the treatment coefficients are unbiased.Also, the RMSEs are below 0.2 under both estimators. It means that the data values do not deviate much from the fitted line.

Table 10: Diagnosing Estimators

	estimator_label	bias	rmse	power	coverage
1	Pre-LM	-0.07	0.07	0.00	1.00
2	Pre-GLM	-0.28	0.28	0.00	1.00
3	Post-LM	0.16	0.16	0.00	1.00
4	Post-GLM	0.99	0.99	0.00	1.00

Table 11: Diagnosing Estimators

	estimator_label	se(bias)	rmse	power	se(power)
1	lm_robust	0.01	0.11	0.08	0.01
2	lm	0.01	0.18	0.02	0.01
1	lm_robust	0.00	0.06	0.00	0.00
2	lm	0.01	0.19	0.04	0.01

4 Mock Result

Using fake data, I show a mock figure which shows the real data values in black, and the fitted values of the lm_robust model in red. I use propensity score matching and fixed effects to see the number of SV amnesties by democracy score. This figure shows that the actual data points are below the predicted line is above – which means that democratic countries grant more sv amnesties. This is the opposite direction of what the theory is predicting (the figure here does not show the interaction effect).Hence, if the actual study exhibit similar figure, it would mean that the theory has weak statistical evidence.

4.1 Replication Data

All data and codes (in .Rmd) can be found in the following github repository: https://github.com/mjkim12/Preanalysis

5 The Appendix

```
library(formatR)
library(knitr)
library(readr)
library(tidyverse)
library(car)
library(optmatch)
library(Matching)
library(RItools)
library(pscl)
library(DeclareDesign)
library(mosaic)
library(estimatr)
library(tidyverse)
library(xtable)
library(fabricatr)
library(randomizr)
library(WeightIt)
library(cobalt)
library(arm)
library(stats)
# Load Data from Github
urlfile = "https://raw.githubusercontent.com/mjkim12/Preanalysis/main/amnesty_mjk_220109

df <- read_csv(url(urlfile))

# Original Data Composition (before removing missing data)
dim(df)  # 514 observations, 57 variables
unique(df$country.x)  #105 countries
table(df$sum_hram)  #Number of wars with SV amnesties: total 76 cases

# Distribution of war-periods (pre98, post98, ongoing98)
# and SV amnesties
table(df$pre98war, df$sum_hram)  #32 out of 295 dyad-conflicts involved with sv amnesti
table(df$post98war, df$sum_hram)  #17 out of 136 dyad-conflicts involved with sv amnest
table(df$cross98war, df$sum_hram)  #27 out of 67 wars involved with svamn.(40.3%)
table(df$warend_post98, df$sum_hram)  #war_end_98 refers to ongoing98 (i.e., cross+post

# Make some variables as binary
table(df$sum_hram)  #Number of wars with SV amnesties: total 76 cases (0:422, 1: 66, 2
df$dummy_svamn <- ifelse(df$sum_hram > 0, 1, 0)  #making SVAmnesty into dummy

wrdf <- df %>%
    dplyr::select(sum_hram, demtrans, yearsatwar, terrytory,
        intensity, ethnic, numdyads, rebcap, fightcap, blood,
        judicialinde, post98war, warend_post98, max_rebpresosts,
        mediation, mean_v2xpolyarchy, war_end_yr, )

# I excluded sv because it reduces sample size into 340 ->
# 100.

# Changed the variable name (max_rebpresosts is the
# indicator for TNR)
names(wrdf)[names(wrdf) == "max_rebpresosts"] <- "TNR"

# Removing Missing Data
sum(is.na(wrdf))  #601
wrdat <- na.omit(wrdf)
dim(wrdat)  #dimension: 340, 17

table(wrdat$warend_post98)  #240, 100
table(wrdat$sum_hram)
wrdat$dummy_svamn <- ifelse(wrdat$sum_hram > 0, 1, 0)
table(wrdat$dummy_svamn)  # 287 wars w/o svamn; 53 wars with.
table(wrdat$TNR)  #201 conficts vs. local, 139 conflicts vs. TNRs

# Categorizing conflicts by years of start and end yrs
df_pre98 <- wrdat[which(wrdat$post98war == 0), ]  #276 dyad wars
df_post98 <- wrdat[which(wrdat$post98war == 1), ]  #64 dyad
df_ongoing98 <- wrdat[which(wrdat$warend_post98 == 1), ]  #100 dyad
# PREMATCHING Balance Test for Pre-98 Subset
balfmla_pre98 <- reformulate(c(names(df_pre98)[c(2:11)]), response = "TNR")

xb0_pre98 <- xBalance(balfmla_pre98, strata = list(raw = NULL),
    data = df_pre98, report = c("std.diffs", "z.scores", "adj.means",
        "adj.mean.diffs", "chisquare.test", "p.values"))
# xtable(xb0_pre98$overall) PREMATCHING Balance Test for
# Ongoing098 Subset
balfmla_ongoing98 <- reformulate(c(names(df_ongoing98)[c(2:11)]),
    response = "TNR")

xb0_ongoing98 <- xBalance(balfmla_ongoing98, strata = list(raw = NULL),
    data = df_ongoing98, report = c("std.diffs", "z.scores",
        "adj.means", "adj.mean.diffs", "chisquare.test", "p.values"))
# xtable(xb0_ongoing98$overall)
df_pre98_2 <- df_pre98

# Create linear predictors for pre-98 data
glm_pre98 <- bayesglm(balfmla_pre98, data = df_pre98_2, family = binomial)

## Here, I use the function `bayesglm` -- a Bayesian
## Generalized Linear Model averaging -- which accounts for
## the uncertainties of the model parameter.

df_pre98_2$pscore_pre98 <- predict(glm_pre98, type = "link")

# pscore_pre98 <- predict(glm_ps_pre98, type = 'response')

# Make distance matrices
psdist_pre98 <- match_on(TNR ~ pscore_pre98, data = df_pre98_2)
as.matrix(psdist_pre98)[1:5, 1:5]

# Fullmatching using the propensity score
ps_pre98 <- fullmatch(psdist_pre98, data = df_pre98_2)
ps_pre98_summary <- summary(ps_pre98, data = df_pre98_2, min.controls = 0,
    max.controls = Inf)
ps_pre98_summary  #Effective sample size 94.1

# xtable(fm1_pre98_summary$matched.set.structures, caption
# = 'Structure of Matched Sets for pre98') xBalance to
# assess the balance properties of the match pre-98
xb1_ps_pre98 <- xBalance(balfmla_pre98, strata = list(raw = NULL,
    ps_pre98 = ~ps_pre98), data = df_pre98_2, report = "all")
plot(xb1_ps_pre98, main = "Pre-98 Xbalance Result")
xtable(xb1_ps_pre98$overall)
df_ongoing98_2 <- df_ongoing98

# Create linear predictors for ongoing-98 data
glm_ongoing98 <- bayesglm(balfmla_ongoing98, data = df_ongoing98_2,
    family = binomial)

df_ongoing98_2$pscore_ongoing98 <- predict(glm_ongoing98, type = "link")

# Make distance matrices
psdist_ongoing98 <- match_on(TNR ~ pscore_ongoing98, data = df_ongoing98_2)
as.matrix(psdist_ongoing98)[1:5, 1:5]

# Fullmatchingusing the ps
ps_ongoing98 <- fullmatch(psdist_ongoing98, data = df_ongoing98_2)
ps_ongoing98_summary <- summary(ps_ongoing98, data = df_ongoing98_2,
    min.controls = 0, max.controls = Inf)
# There are 33.9 effective sample size

xtable(ps_ongoing98_summary$matched.set.structures, caption = "Structure of Matched Sets
##### xBalance to assess the balance properties of the
##### match
xb1_ps_ongoing98 <- xBalance(balfmla_ongoing98, strata = list(raw = NULL,
    ps_ongoing98 = ~ps_ongoing98), data = df_ongoing98_2, report = "all")

plot(xb1_ps_ongoing98, main = "Ongoing-98 Xbalance Result")
xtable(xb1_ps_ongoing98$overall)

# I tried a rank-based Mahalanobis distance matching, but
# full matching with propensity score produced a better
# balance

############## Rank-Based Mahalanobis distance ####
mhdist <- match_on(balfmla_pre98, data = df_pre98_2, method = "rank_mahalanobis")
fm22 <- fullmatch(mhdist, data = df_pre98_2)
```

```r
fm22sum <- summary(fm22, data = df_pre98_2, min.controls = 0,
    max.controls = Inf)

# Add matched set indicators back to data
df_pre98_2$fm22 <- NULL
df_pre98_2[names(fm22), "fm22"] <- fm22

## Balance test to see if I 'adjusted enough' xb22 <-
## xBalance(TNR ~ judicialinde + demtrans +yearsatwar +
## terrytory+ intensity + ethnic + numdyads +rebcap +
## fightcap + blood+ mediation, strata = list(raw = NULL,
## fm22 = ~fm22), data = df_pre98_2, report =
## c('std.diffs', 'z.scores', 'adj.means','adj.mean.diffs',
## 'chisquare.test', 'p.values'))


xb22 <- xBalance(TNR ~ judicialinde + demtrans + yearsatwar +
    terrytory + intensity + ethnic + numdyads + rebcap + fightcap +
    blood + mediation, strata = list(raw = NULL, fm22 = ~fm22),
    data = df_pre98_2, report = c("std.diffs", "z.scores", "adj.means",
        "adj.mean.diffs", "chisquare.test", "p.values"))

# plot(xb22)## Balance is worse than fullmatching

df_fake = ""
set.seed(1234)
fake_population_pre98 <- declare_population(df_pre98)
set.seed(1234)
df_fake <- fake_population_pre98()

set.seed(12345)
## Sampling 500 observations from the population

sampling_1 <- declare_sampling(S = draw_rs(N = N, n = 200))
design_1 <- fake_population_pre98 + sampling_1
set.seed(12345)
df1 <- draw_data(design_1)

######################### DeclareDesign PRE-98 and
######################### Ongoing-98 1-1. Define population
pop_pre98 <- declare_population(df_pre98_2)
pop_ongoing98 <- declare_population(df_ongoing98)

# 1-2. Declare Potential Outcomes
```

```
pot.outcome_pre98 <- declare_potential_outcomes(Y = rnorm(N))  # In pre98 wars, TNR tre
pot.outcome_ongoing98 <- declare_potential_outcomes(Y ~ 0.26 +
    0.02 * Z)  #In Ongoing98 wars, I expect treatment effect.

mean(df_ongoing98$dummy_svamn)  #0.26 is the mean possibility for a dyad to have sv amn

# 2. Inquiry: what is research question I want to answer?
estimand <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))

# 3. Treatment Assignment 1: random assignment to half of
# the units.
assignment_pre98 <- declare_assignment(Z = complete_ra(N = N,
    m = 138))
assignment_ongoing98 <- declare_assignment(Z = complete_ra(N = N,
    m = 50))

# 3-2: Treatment assignment by matching: treatment is also
# correlated with the covariates
assignment <- declare_assignment(handler = function(data) {
    prob <- with(data, pnorm(TNR + mediation + judicialinde +
        demtrans + yearsatwar + terrytory + intensity + ethnic +
        numdyads + rebcap + fightcap + blood))
    data$Z <- rbinom(nrow(data), 1, prob)
    return(data)
})

revewal_Y <- declare_reveal(Y, Z)


# 4. Answer Strategy
estimator_1 <- declare_estimator(Y ~ Z, model = lm_robust)
estimator_2 <- declare_estimator(Y ~ Z, model = difference_in_means)

estimator_pre1 <- declare_estimator(Y ~ TNR, model = lm_robust,
    term = "TNR", inquiry = "ATE", label = "lm_robust")

estimator_pre_match <- declare_estimator(Y ~ TNR, model = lm_robust,
    term = "TNR", inquiry = "ATE", label = "lm_robust")

estimator_pre2 <- declare_estimator(Y ~ max_rebpresosts + mediation +
    judicialinde + demtrans + yearsatwar + terrytory + intensity +
    ethnic + numdyads + rebcap + fightcap + blood, model = lm,
    term = "TNR", inquiry = "ATE", label = "lm")
```

```
# 5. Design Declare
design_pre98 <- pop_pre98 + pot.outcome_pre98 + estimand + assignment_pre98



reveal_pre <- declare_reveal(Y, Z)

samp_pre <- declare_sampling(S = draw_rs(N = N, n = 100), filter = S ==
    1)


design_pre1 <- pop_pre98 + pot.outcome_pre98 + estimand + assignment_pre98 +
    reveal_pre + samp_pre + estimator_pre1

design_pre2 <- pop_pre98 + pot.outcome_pre98 + estimand + assignment_pre98 +
    reveal_pre + samp_pre + estimator_pre2

# diagnose1 <- diagnose_design(design_pre1 ,design_pre2)

# xtable(head(diagnose1$diagnosands_df[,c(3,5,6, 7,8,
# 9,10)]))

######################### DeclareDesign POST 98
######################### ####################

# Declare Population
pop_ongoing98_2 <- declare_population(df_ongoing98)  #fm1_ongoing98

pot.outcome_post <- declare_potential_outcomes(Y ~ -0.01 * Z +
    dummy_svamn)

estimand_post <- declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0))

assignment_post <- declare_assignment(Z = conduct_ra(N = N, m = 50))

reveal_post <- declare_reveal(Y, Z)

samp_post <- declare_sampling(S = draw_rs(N = N, n = 100), filter = S ==
    1)

estimator_post1 <- declare_estimator(Y ~ max_rebpresosts, model = lm_robust,
    term = "TNR", inquiry = "ATE", label = "lm_robust")
```

```
estimator_post2 <- declare_estimator(Y ~ max_rebpresosts + mediation +
    judicialinde + demtrans + yearsatwar + terrytory + intensity +
    ethnic + numdyads + rebcap + fightcap + blood, model = lm,
    term = "TNR", inquiry = "ATE", label = "lm")


design_post1 <- pop_ongoing98_2 + pot.outcome_post + estimand_post +
    assignment_post + reveal_post + samp_post + estimator_post1


design_post2 <- pop_ongoing98_2 + pot.outcome_post + estimand_post +
    assignment_post + reveal_post + samp_post + estimator_post2


# diagnose1_post <- diagnose_design(design_post1
# ,design_post2)
# xtable(head(diagnose1_post$diagnosands_df[,c(3,5,6, 7,8,
# 9,10)])) design_pre <-diagnose_design(design_pre1,
# design_pre2, diagnosands = declare_diagnosands(power =
# mean(p.value <= 0.05), false_positive_rate = mean(p.value
# <= 0.05)))

# design_post <-diagnose_design(design_post1, design_post2,
# diagnosands = declare_diagnosands(power = mean(p.value <=
# 0.05), false_positive_rate = mean(p.value <= 0.05)))

# xtable(head(design_pre$diagnosands_df[,c(3,5,6,7)]))
# xtable(head(design_post$diagnosands_df[,c(3,5,6,7)]))
library(pscl)
# ALL Propensity Score#
wrdat_zero <- wrdat
# glm_ps <- glm(max_rebpresosts ~ mediation+ judicialinde +
# demtrans +yearsatwar + terrytory + intensity + ethnic +
# numdyads +rebcap + fightcap + blood + sv, data =
# wrdat_zero, family = binomial())

# pscore <- predict(glm_ps, type = 'response') psdist
# <-match_on(max_rebpresosts~ pscore,data=wrdat_zero)
# fm_all <- fullmatch(psdist, data = wrdat_zero) fm_summary
# <- summary(fm_all, data = wrdat_zero, min.controls = 0,
# max.controls = Inf)

# ps into the dataset wrdat_zero$fm_all <- NULL
# wrdat_zero[names(fm_all),'fm_all'] <- fm_all
# head(wrdat_zero) diagnose1 <- diagnose_design(design_pre1
# ,design_pre2) diagnose1_post <-
# diagnose_design(design_post1 ,design_post2)

# xtable(head(diagnose1$diagnosands_df[,c(3,6,7,9,10)]))
# xtable(head(diagnose1_post$diagnosands_df[,c(3,6,7,9,10)]))
# fakedata <- draw_data(design_pre1) fakedata$sum_hram

# ALL Propensity Score# glm_ps <- glm(DEM ~ judicialinde
# +yearsatwar + numdyads + ethnic + sv, data = wrdat,
# family = binomial())

# pscore <- predict(glm_ps, type = 'response') psdist
# <-match_on(DEM~ pscore,data=wrdat) fm_all <-
# fullmatch(psdist, data = wrdat) fm_summary <-
# summary(fm_all, data = wrdat, min.controls = 0,
# max.controls = Inf)

### lm_robust model1 <-lm_robust(sum_hram ~
### mean_v2xpolyarchy, fixed_effects = ~fm_all, data =
### wrdat)

# plot(wrdat$mean_v2xpolyarchy[wrdat$warend_post98==0],
# wrdat$sum_hram[wrdat$warend_post98==0], xlab='Level of
# Democracy', ylab = 'Number of SV Amnesties', ylim =
# c(0,1.2), main = 'M4') abline(plot(model1$fitted.values,
# col = 2))

## What is the biggest difference within set.
## diffswithinsets<-df_pre98_2 %>% group_by(fm22) %>%
## summarize(meandiff = mean(sum_hram[DEM==1]) -
## mean(sum_hram[DEM==0]))
## summary(diffswithinsets$meandiff)

# DIFF PRE MATCHING with(df_pre98_2, mean(sum_hram[DEM==1])
# - mean(sum_hram[DEM==0])) What are the distances like?
# tmp2 = df_pre98_2$sum_hram names(tmp2) <-
# rownames(df_pre98_2) absdist2 <- match_on(tmp2, z =
# df_pre98_2$DEM) qtl2 <-
# quantile(as.vector(absdist2),seq(0,1,.1))

## declare population
pop_pre98 <- declare_population(df_pre98_2)
pop_ongoing98 <- declare_population(df_ongoing98)
```

```
## declare estimator estimator22 <-
## declare_estimator(sum_hram~DEM+judicialinde + yearsatwar
## + numdyads + ethnic + sv, model = lm, term = 'DEM',
## estimand = 'DEM', label = 'OLS')

## declare estimand make_estimands22 <- function(data){ bs
## <- coef(lm(sum_hram~DEM, data=df_pre98_2))
## return(data.frame(estimand_label= 'DEM',
## estimand=bs['DEM'], stringsAsFactors = FALSE))}
## estimand22 <- declare_inquiry(handler=make_estimands22,
## label='Pop_Relationships')

# design1_plus_estimands22 <- pop_pre98_2 + estimand22
# kable(estimand1(df_pre98_2), caption = 'Estimands 22')

# design_full22 <- design1_plus_estimands22 + estimator22

# run_design(design_full22) What is the biggest difference
# within set. diffswithinsets<-df_pre98_2 %>%
# group_by(fm22) %>% summarize(meandiff =
# mean(sum_hram[DEM==1]) - mean(sum_hram[DEM==0]))
# summary(diffswithinsets$meandiff)

## DIFF PRE MATCHING with(df_pre98_2,
## mean(sum_hram[DEM==1]) - mean(sum_hram[DEM==0])) What
## are the distances like?  tmp2 = df_pre98_2$sum_hram
## names(tmp2) <- rownames(df_pre98_2) absdist2 <-
## match_on(tmp2, z = df_pre98_2$DEM) qtl2 <-
## quantile(as.vector(absdist2),seq(0,1,.1))


## declare estimator estimator22 <-
## declare_estimator(sum_hram~DEM+judicialinde + yearsatwar
## + numdyads + ethnic + sv, model = lm, term = 'DEM',
## estimand = 'DEM', label = 'OLS')

### declare estimand make_estimands22 <- function(data){ bs
### <- coef(lm(sum_hram~DEM, data=df_pre98_2))
### return(data.frame(estimand_label= 'DEM',
### estimand=bs['DEM'], stringsAsFactors = FALSE))}
### estimand22 <- declare_inquiry(handler=make_estimands22,
### label='Pop_Relationships') pop_pre98_2 <-
### declare_population(df_pre98_2)
```

```
# design1_plus_estimands22 <- pop_pre98_2 + estimand22
# kable(estimand1(df_pre98_2), caption = 'Estimands 22')

# design_full22 <- design1_plus_estimands22 + estimator22

# run_design(design_full22)
```

References

Akhavan, Payam. 2009. "Are International Criminal Tribunals a Disincentive to Peace?: Reconciling Judicial Romanticism with Political Realism." *Human Rights Quarterly* 31 (3): 624–54. https://doi.org/10.1353/hrq.0.0096.

———. 2009. "Are International Criminal Tribunals a Disincentive to Peace?: Reconciling Judicial Romanticism with Political Realism." *Human Rights Quarterly* 31 (3): 624–54. https://doi.org/10.1353/hrq.0.0096.

Ben B. Hansen, and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statist. Sci* 23 (2): 219–36.

Cohen, Dara Kay, and Ragnhild Nordås. 2014. "Sexual Violence in Armed Conflict Dataset." http://www.sexualviolencedata.org.

Dancy, Geoff. 2018. "Deals with the Devil? Conflict Amnesties, Civil War, and Sustainable Peace." *International Organization* 72 (2): 387–421. https://doi.org/10.1017/S0020818318000012.

———. 2018. "Deals with the Devil? Conflict Amnesties, Civil War, and Sustainable Peace." *International Organization* 72 (2): 387–421. https://doi.org/10.1017/S0020818318000012.

Daniels, Lesley-Ann. 2020. "How and When Amnesty During Conflict Affects Conflict Termination." *Journal of Conflict Resolution* 64 (9): 1612–37. https://doi.org/10.1177/0022002720909884.

———. 2020. "How and When Amnesty During Conflict Affects Conflict Termination." *Journal of Conflict Resolution* 64 (9): 1612–37. https://doi.org/10.1177/0022002720909884.

Eck, Kristine & Lisa Hultman. 2007. "Violence Against Civilians in War." *Journal of Peace Research* 44 (2).

Ginsburg, Tom. 2009. "The Clash of Commitments at the International Criminal Court." *Chicago Journal of International Law.*

———. 2009. "The Clash of Commitments at the International Criminal Court." *Chicago Journal of International Law.*

Goldsmith, Jack, and Stephen D. Krasner. 2003. "The Limits of Idealism." *Daedalus.*

———. 2003. “The Limits of Idealism.” *Daedalus.*

Haer, Roos, and Tobias Böhmelt. 2017. “How Child Soldiering Prolongs Civil War.” *Cooperation and Conflict* 52 (3): 332–59.

Hansen, Ben B. 2004. “Full Matching in an Observational Study of Coaching for the SAT.” *Journal of the American Statistical Association* 99.

Hansen, Ben B., and Stephanie O. Klopfer. 2006. “Optimal Full Matching and Related Designs via Network Flows.” *Journal of Computational and Graphical Statistics.*

Imai, Kosuke. 2005. “Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments.” *American Political Science Review.*

Jeffery, Renée. 2014. *Amnesties, AAmnesties, Accountability, and Human Rights.* Universty of Pennsylvania Press.

Kerry M. Green, and Elizabeth A. Stuart. 2014. “Examining Moderation Analyses in Propensity Score Methods: Application to Depression and Substance Use.” *Journal of Consulting and Clinical Psychology.*

Kim, Hunjoon, and Kathryn Sikkink. 2010. “Explaining the Deterrence Effect of Human Rights Prosecutions for Transitional Countries.” *International Studies Quarterly* 54 (4): 939–63.

Ki-moon, Ban. n.d.

Krcmaric, Daniel. 2018. “Should I Stay or Should I Go? Leaders, Exile, and the Dilemmas of International Justice.” *American Journal of Political Science* 62 (2): 486–98. https://doi.org/https://doi.org/10.1111/ajps.12352.

———. 2018. “Should I Stay or Should I Go? Leaders, Exile, and the Dilemmas of International Justice.” *American Journal of Political Science* 62 (2): 486–98. https://doi.org/https://doi.org/10.1111/ajps.12352.

Mallinder, Louise. 2012. “Amnesties’ Challenge to the Global Accountability Norm?” In *Amnesty in the Age of Human Rights Accountability*, edited by Francesca Lessa and Leigh A. Payne, 69–96. Cambridge: Cambridge University Press.

———. 2012. “Amnesties’ Challenge to the Global Accountability Norm?” In *Amnesty in the Age of Human Rights Accountability*, edited by Francesca Lessa and Leigh A. Payne, 69–96. Cambridge: Cambridge University Press.

Paul R. Rosenbaum, and Donald B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* Volume 70 Issue 1: Pages 41–55.

Peter C Austin, and Elizabeth A Stuart. 2015. “Estimating the Effect of Treatment on Binary Outcomes Using Full Matching on the Propensity Score.” *Statistical Methods in Medical Research.*

Prorok, Alyssa K. 2017. “The (in)compatibility of Peace and Justice? The International Criminal Court and Civil Conflict Termination.” *International Organization.*

———. 2017. “The (in)compatibility of Peace and Justice? The International Criminal Court and Civil Conflict Termination.” *International Organization*.

Reiter, Trica D Olsen; Leigh A. Payne; Andrew G. 2010. *Transitional Justice in Balance - Comparing Processes, Weighting Efficacy*. nited States Institute of Peace.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. Springer.

Rubin, Donald B. 2002. “Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation.” *Health Services & Outcomes Research Methodology*.

Sikkink, Kathryn. 2012. “Amnesty in the Age of Human Rights Accountability: Comparative and International Perspectives.” In, edited by Franceesca Lessa and Leigh A. Payne. Cambridge University Press.

Simmons, Beth Ann, and Allison Danner. 2010. “Credible Commitments and the International Criminal Court.” *International Organization* 64 (2): 225–56.

Sly, Ron. n.d.

Snyder, Jack, and Leslie Vinjamuri. 2003. “Trials and Errors: Principle and Pragmatism in Strategies of International Justice.” *International Security* 28 (3): 5–44.

Stuart EA, and Green KM. 2008. “Using Full Matching to Estimate Causal Effects in Nonexperimental Studies: Examining the Relationship Between Adolescent Marijuana Use and Adult Outcomes.” *Developmental Psychology*.

Thavaneswaran, Arane. 2008. “Propensity Score Matching in Observational Studies.” University of Manitoba, Manitoba Centre for Health Policy.