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, the former Secretary General of the United Nation, 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 academics and peace practitioners, which is often called 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

[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).

to provide a theory to answer this question. I argue that the UJ and ICC, by increasing the 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

The conventional wisdom is that the rise of international anti-amnesty regimes deter the use of SV amnesties mainly by creating a 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)). In other words, since SV amnesties can be dishonored by the ICC and other countries practicing Universal Jurisdiction, the instability of SV amnesties would halter the value of amnesty as a peace bargaining tool. Existing studies disregard two important aspects of the SV amnesties in the conflict setting. First, the commitment problem theory rules out the possibility that rebel group can still be free from the commitment problem as long as they stay inside the home country or other neighboring states that are likely to respect the domestic amnesty instead of respecting international norms to hold them accountable. Therefore, rebel groups can still be free from such commitment problem and have incentive to demand SV amnesties. Second, the degree at which the rebel face foreign and international prosecutions vary and hence the effect of the ICC and UJ on their incentive to seek out for SV also amnesties. If a culpable rebel group only faces a threat of domestic prosecutions, the rise of the ICC and UJ would not directly affect its demand to seek out for SV amnesties. In other words, the advent of the ICC and the UJ should change the incentive of demand for SV amnesties only among rebel groups that fear the risk of prosecutions by the ICC and the UJ. Based on the logic, I argue that the advent of the ICC and UJ, by increasing the threat of international prosecutions, increased rebel's incentive to demand SV amnesties which hence fosters the use of SV amnesties.[2] 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

[2] In order for an amnesty deal to be striken, there must be both demand and supply. While I acknowledge the existence of the supply factors (state's capacity/ willingness to grant SV amnesties, this paper theorize mainly focusing on the demand factor.

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' issuance of amnesties for civil wars from 1945 to 2014 (Dancy (2018)). Since my main interest is in examining 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 do 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. The data have observations of 413 dyad conflict of 101 countries.

3.2 Variables and Measures

3.2.1 Response Variable

The dependent variable is coded as 1 if there has been any exchange of SV amnesties in state-rebel group dyad conflict. Among the 413 dyad conflicts in data, 68 dyad wars involved exchanging of SV amnesties. The data show that SV amnesties are usually exchanged once in a state-rebel dyad conflict, if there is any (86.8%). Only nine out of the 68 rebel groups received sv amnesties more than one time, at most five times, probably due to failed attempts to resolve wars even after issuing amnesties.

3.2.2 International Anti-amnesty Regimes (ICC, UJ)

I use the year 1998 to indicate the key independent variable – the emergence of anti-amnesty regimes. In this year, both the ICC and UJ emerged together accidentally, and the 1998-cutoff is widely used in the literature to indicate the transition from the era of impunity to the era of accountability (Dancy (2018), Krcmaric (2018), Daniels (2020)). More specifically, I categorize conflicts by three time coverage: 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* with SV amnesties in *Post98 wars*. 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 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. Hence, amnesty deals in the Ongoing-98 wars are likely to be affected by the ICC and the UJ. In the actual paper, I will report both comparisons, but this pre-analysis mainly discusses the latter comparison using the `Ongoing98` dummy. In the whole data set, pre-98 conflicts comprise about 78% of observations (N =325), post-98 conflicts about 21 % (N =88), and ongoing-98 conflicts about 36% (N = 150).

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. In the data, there are 246 dyads (59.6%) with local rebel groups and 167 dyads (40.4%) with TNRs.

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 was 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 randomized experiments 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 – 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 controlled 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.

However, 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) For the propensity score matching, 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). Using the NSA data, the dummy variable `TNR` is coded 1 if the rebel group operates to at least some extent outside the home country's borders. Among 413 dyads, there are 401 unique rebel groups captured in the dataset, and among them, there are 76 transnational rebel groups (`TNR`) 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. Particularly, I use a Bayesian generalized linear model averaging with the `bayesglm` function (Gelman (2011)) which accounts for the model uncertainty inherent in the variable selection problem by averaging over the best models in the model class according to approximate posterior model probability. While the `glm` model assumes normal distribution of errors, bayesian logistic regression offers a more flexible generalization of ordinary linear regression that does not need the normal distribution of errors. Most importantly, the `glm()` find you the best fitting coefficient while the `bayesglm()` do not give you a single estimated coefficient but instead a complete posterior distribution about how likely different values of coefficient (Chen and Kaplan (2015)).

I specify the propensity score using variables that may affect SV amnesties as suggested in the earlier studies. Borrowing Dancy 2018, I include variables for the number of years at war (`yearsatwar`), territory (`territory`), intensity (`intensity`), ethnic (`ethnic`), number of other groups fighting (`numdyads`), rebel's fighting capacity (`fightcap`), and bloody hands (`blood`) which may affect the number of amnesties (Dancy (2018)). Additionally, I include a variable that indicates 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 generate missing data at random. 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.

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, 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)). Table 1 Table 2 show the structures of matched sets for Pre-98 subgroup and Ongoing-98 subgroup, and they have 101.3 and 50.9 matched pairs (effective sample size) respectively.

	x
10:1	1
8:1	1
6:1	1
5:1	2
4:1	4
3:1	9
2:1	5
1:1	30
1:2	2
1:3	4
1:4	3
1:5	4
1:6	4
1:7	2
1:9	1
1:15	1
1:23	1

Table 1: Structure of Matched Sets for pre98

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 show the test results for the chi-square and the p-value for the pre-98 and ongoing-98 datasets. In the pre-98, the chi-square is 64.83 and p-value is 0.00; in ongoing-98 dataset, the chi-square is 36.10 and p-value is 0.00. 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.

	x
16:1	1
6:1	1
5:1	1
3:1	2
2:1	3
1:1	20
1:2	4
1:3	1
1:4	2
1:5	2
1:6	1
1:8	2

Table 2: Structure of Matched Sets for ongoing-98

Table 3: Balance before Matching for Pre-98

	chisquare	df	p.value
raw	64.83	8.00	0.00

Table 5, Table 6 show the chi-square values and p-values for pre- and ongoing- datasets after propensity score matching. In pre-98 dataset, chi-square value and p-value are 1.72 and 0.99 respectively; in ongoing-98 dataset, the chi-square value and p-value are 6.24 and 0.62 respectively . They suggest that the treatment and control groups are not too different and make 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 most covariates after matching. Hence, I consider the adjustment successful.

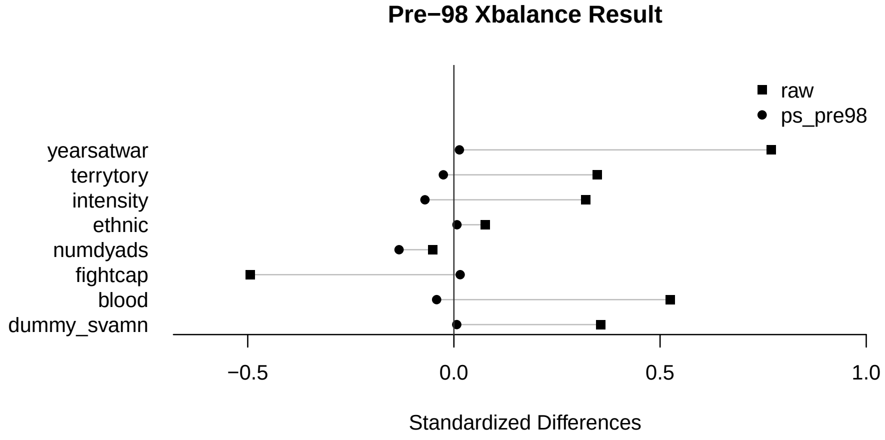


Figure 1: Balance Test for Pre-98

Table 4: Balance before Matching for Ongoing-98

	chisquare	df	p.value
raw	36.10	8.00	0.00

Table 5: Balance of Pre-98

	chisquare	df	p.value
raw	64.83	8.00	0.00
ps_pre98	1.54	8.00	0.99

4 Estimators and Estimand Using Simulated Data

4.1 Creating Simulated Datasets

Before I discuss and draw estimands and estimators, I first create simulated populations based on their original data sets. I have three sets of population: 1) one from the whole dataset encompassing every time periods, 2) one from the Pre-98 dataset, and 3) one from the Ongoing-98 dataset. From each simulated population, I randomly select 500 observations and use them as the simulated datasets for analyses shown afterward. In order to check whether the simulated sampling worked well, I compare the distributions for key variables in the original whole dataset (Figure 3) and its simulated sample dataset (Figure 4). The distributions of the simulated sample data resembles the original data.

4.2 Estimand

In this paper, I use a designed-based inference rather than a model-based inference. The design-based approach involves using information from a random sample to estimate some parameter of the population from which the sample was drawn (Imai (2016)). Compared to the model-based inference, the designed-based approach requires fewer assumptions as

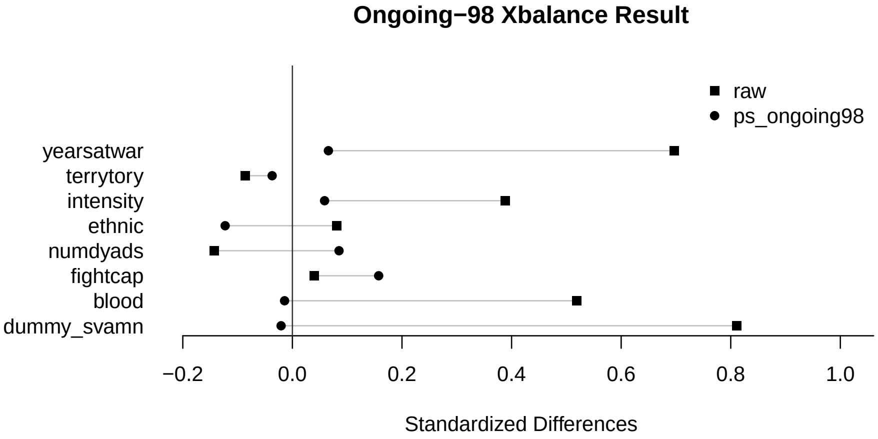


Figure 2: Balance Test for Ongoing-98

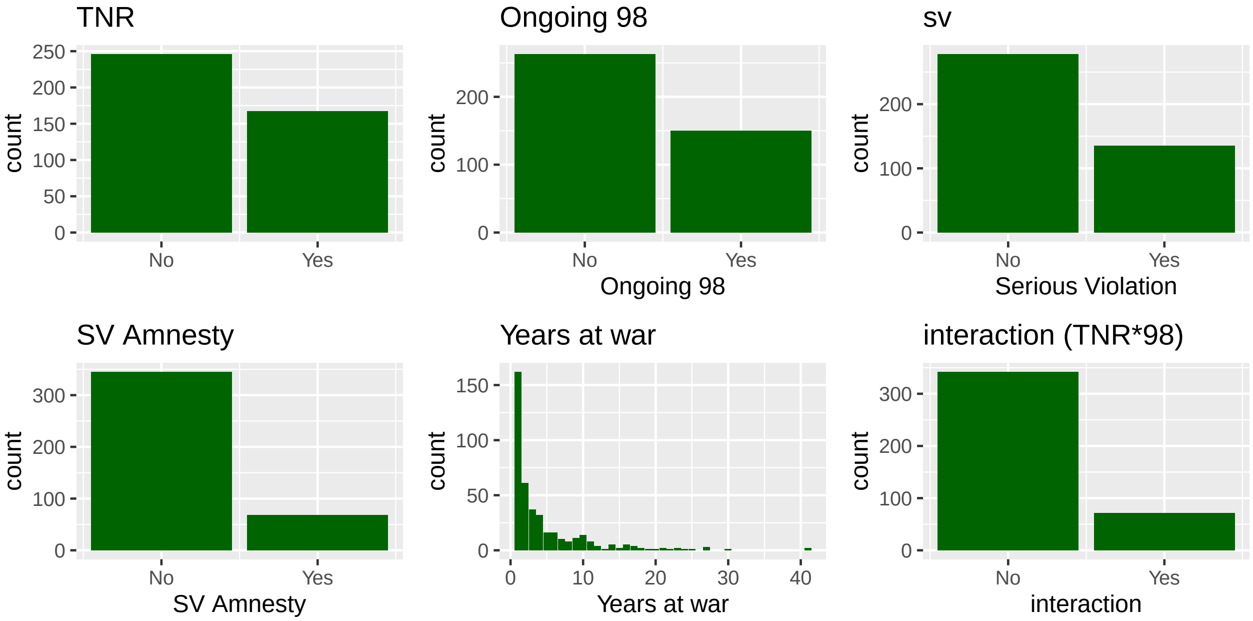


Figure 3: Original Data Population Distribution

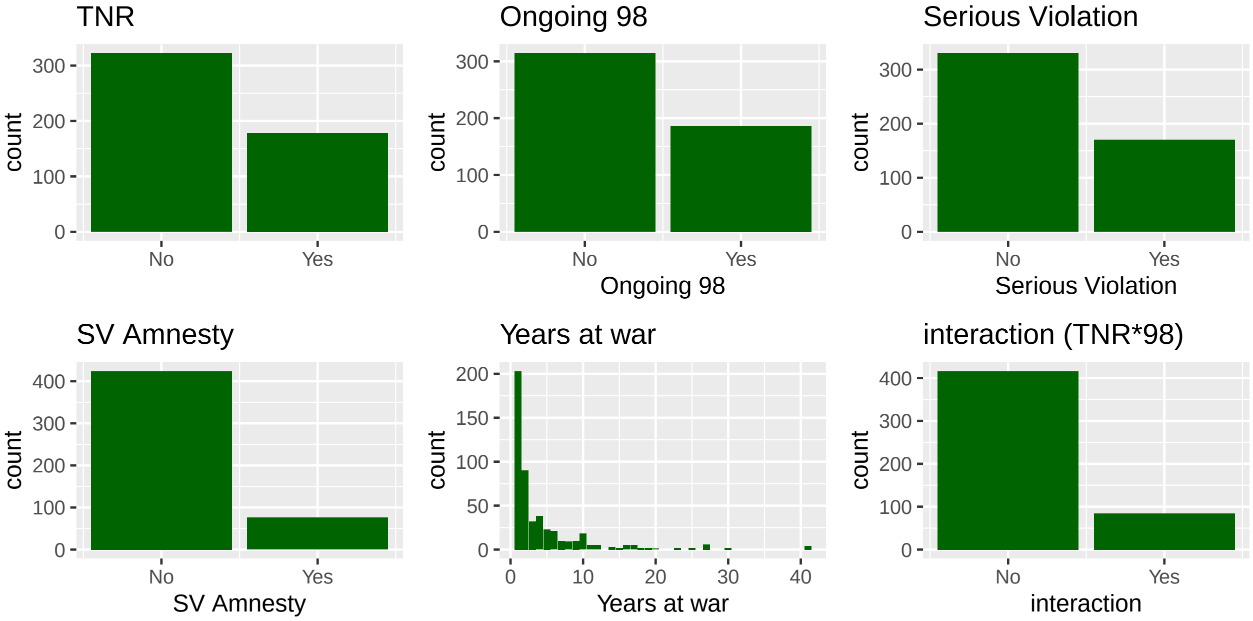


Figure 4: Simulated Sample Distribution

Table 6: Balance of Ongoing-98

	chisquare	df	p.value
raw	36.10	8.00	0.00
ps_ongoing98	1.40	8.00	0.99

it relies on the randomization mechanism to develop estimators. Also, the design-based estimators are unbiased and normally distributed in large samples with simple variance estimator. Using the designed-based inference, the estmand, or the target of estimation, in this study is β_1 – a difference in the probability of SV amnesties (`dummy_svamn`) according to the interaction of the rebel's type (`TNR`) and the advent of the ICC and UJ (`ongoing98`). The basic model that represents the estimand is the following: $y_i = \beta_0 + \beta_(TNR * Ongoing98)$ $+u_i$. I will show the values of the estimand for two estimators in this paper – the logistic regression model and the propensity score matching.

4.2.1 Estimand 1

I first declare the estimand for the logistic regression model using the `DeclareDesign`. I do not disaggregate the data into pre- and ongoing-98 datasets as I can directly obtain the interaction term between the `98-Ongoing` dummy and the `TNR`. The estimand by the interaction term is about 0.35 (Table 7).

Table 7: Estimands1 for interaction

	estimand_label	estimand
interaction	glm	0.3467262

4.3 Estimator 1: Logistic Regression

Logistic regression is a standard probabilistic statistical classification model for dichotomous outcome variables. Different from linear regression, the outcome of logistic regression on one sample is the probability that it is positive or negative, where the probability depends on a linear measure of the sample. However, the linear relationship may not always hold, and hence it is sensitive to the presence of outliers. For this reason, I first diagnose the existence of outliers. Also, the key difference of the logistic regression from linear models is its assumptions. Logistic regression does not require a linear relationship between the dependent and independent variables; the error terms (residuals) do not need to be normally distributed; homoscedasticity is not required; and the dependent variable is not measured on an interval or ratio scale. However, logistic regression requires an appropriate outcome structure (e.g., binary dependent variable for binary logistic regression), independent observations, the absence of multicollinearity (i.e., IVs should not be too highly correlated with each other), linearity of IVs and log odds, and a large sample size (Schreiber-Gregory and Bader (2018)).

4.3.1 Influential Values

From the logistic regression model with interaction term (`TNR` and `ongoing98`), I first check outliers using the Cook's distance. Figure 5 shows five outliers. Since not all outliers are influential observations, I check whether the data contains potential influential observations by inspecting the standardized residual error. As Figure 6 shows, there are six data points with an absolute standardized residuals above 3 –which are highly likely outliers. As a result, I filter the six potential outliers (removed data points: 112, 420, 489, 531, 784, 807).

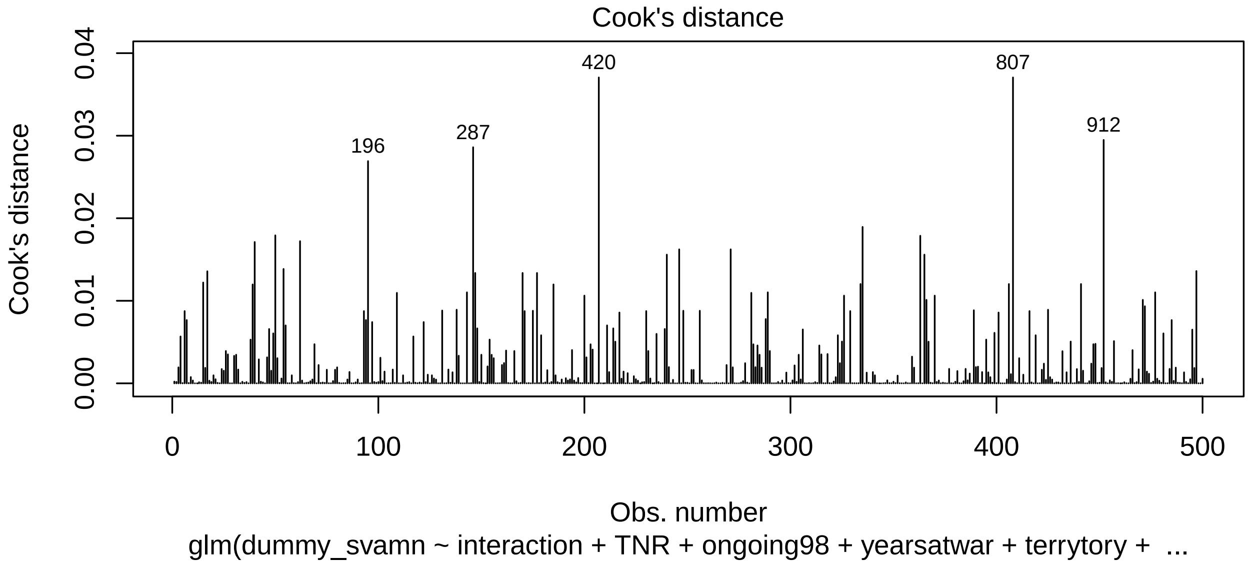


Figure 5: Cook's Distance

4.3.2 Multicollinearity

Multicollinearity corresponds to a situation where the data contain highly correlated predictor variables. It should be fixed by removing the concerned variables. One common way to detect multicollinearity is by looking at the VIF (variable inflation factors). VIF score of an independent variable represents how well the variable is explained by other independent variables. I use `vif()` function from `car package` which computes the VIF. As a rule of thumb, a VIF (variable inflation factors) that exceeds 5 or 10 indicates a problematic amount of collinearity (Kassambara (2018)). Table 8 show the VIF scores of all independent variables in my logistic regression model. As there is no value that exceeds 5, I consider that there is no collinearity problem.

Table 8: Assessing Collenearity Using VIF

	x
interaction	2.716457
TNR	1.987091

	x
ongoing98	1.989306
yearsatwar	1.317877
terrytory	1.237356
intensity	1.388488
ethnic	1.031051
numdyads	1.356099
fightcap	1.182062
blood	1.534997
sv	1.694011

4.3.3 Performance of Estimator 1 (`glm`)

To judge the performance of the estimators, I examine biases and RMSE based on 500 simulations using the `diagnose_design` function in `DeclareDesign`. 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. Bias is the mean of error which is computed through the mean of the difference between the estimate and the estimand (i.e., bias = mean(estimate - estimand)). Hence, an unbiased estimator means that the estimator or test statistic is accurate to approximate the parameter. Table 9 shows the diagnose on estimator 1 for pre-98 dataset. The RMSE is 0.13 which means that the data values quite deviate from the fitted line. The bias is -0.11. An unbiased estimator has bias close to zero, and bias is generally low if the absolute value is below 0.01. Hence, the result shows that there are negative bias on the estimator. Overall, the performance of the logistic regression seems poor.

Table 9: Performance of Estimator 1

	Design	Inquiry	Estimator	Term	N Sims	Bias	RMSE
1	design_glm	interaction	glm	interaction	500	-0.11	0.13
2						(0.00)	(0.00)

4.4 Estimator 2: Propensity Score Full Matching

In order to evaluate the interaction term indirectly, I disaggregate the datasets into pre- and ongoing- datasets for matching estimator with `TNR` as the treatment. I expect its effect to be valid only within the ongoing-98 dataset. As I have done matching in earlier section, I repeat the same process here using the simulated dataset. I did fullmatching using the propensity score and also using a rank-based Mahalanobis distance. Table 10 and Table 11 show the balance after the fullmatching with the ps and the mahalanobis distance respectively, for pre-98. The p-value from matching increased only with the PS matching, and its chi-square decreased significantly. It suggests that the fullmatching with PS performed better than fullmatching with the mahalanobis. The result is similar for ongoing-98 dataset (See Table 12 and 13).

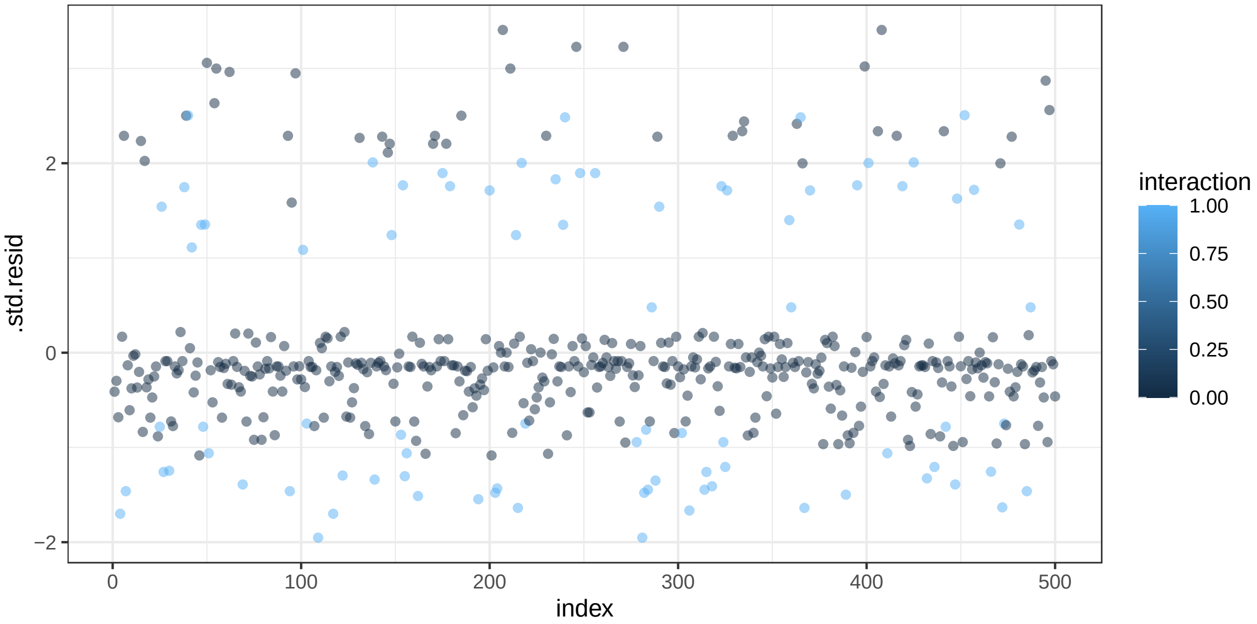


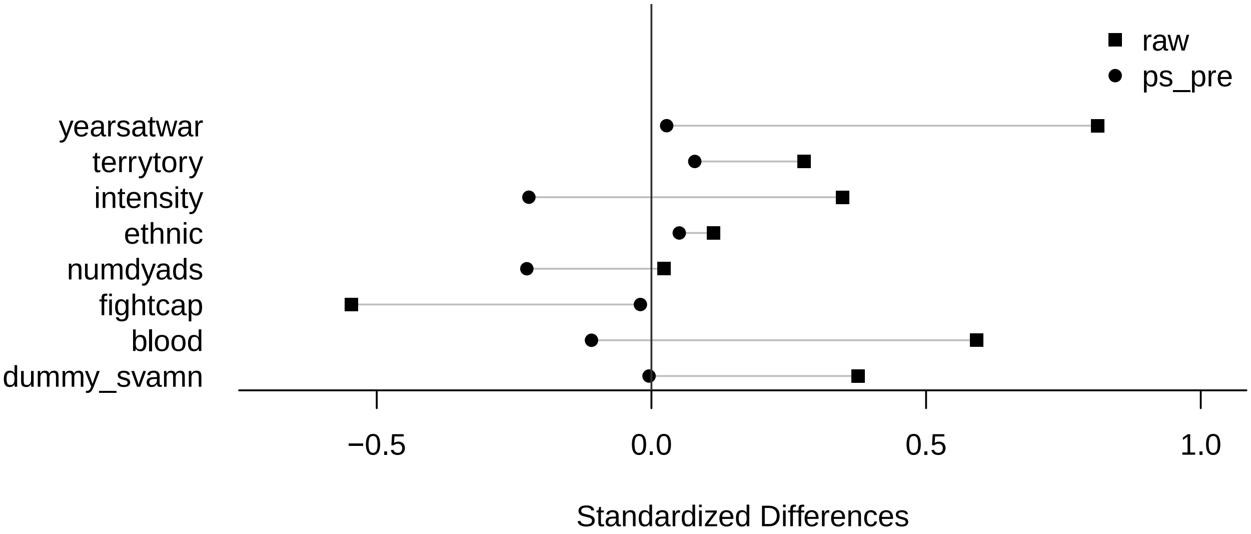
Figure 6: Standardized Residuals

Table 10: Propensity Score Full Matching for Pre98

	chisquare	df	p.value
raw	106.596174	8	0.0000000
ps_pre	8.819881	8	0.3577179

Table 11: Mahalanobis Full Matching for Pre98

	chisquare	df	p.value
raw	106.59617	8	0
mhfull_pre	56.06517	8	0



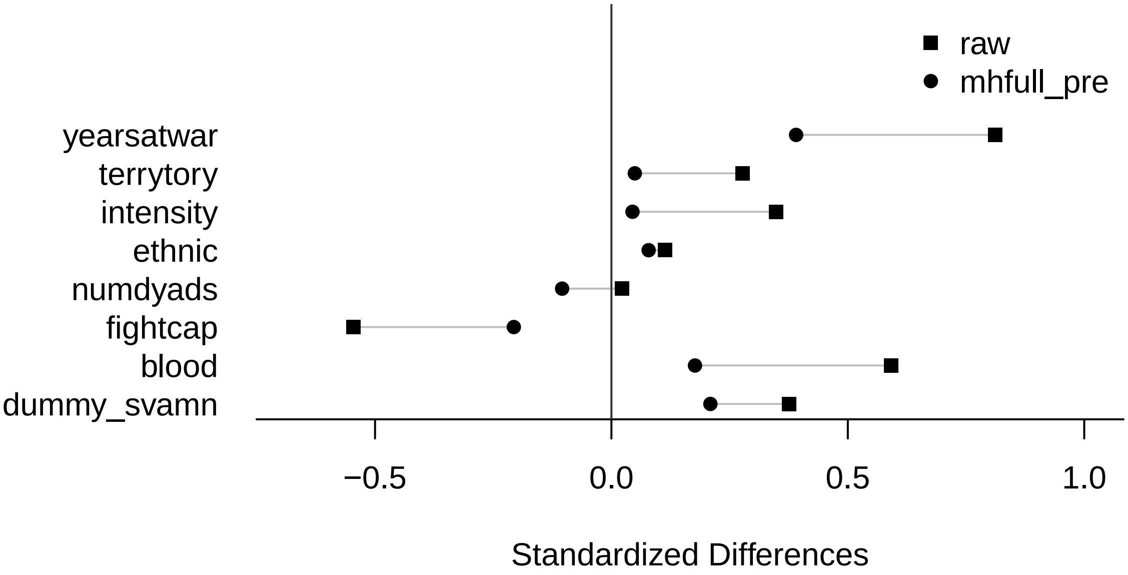
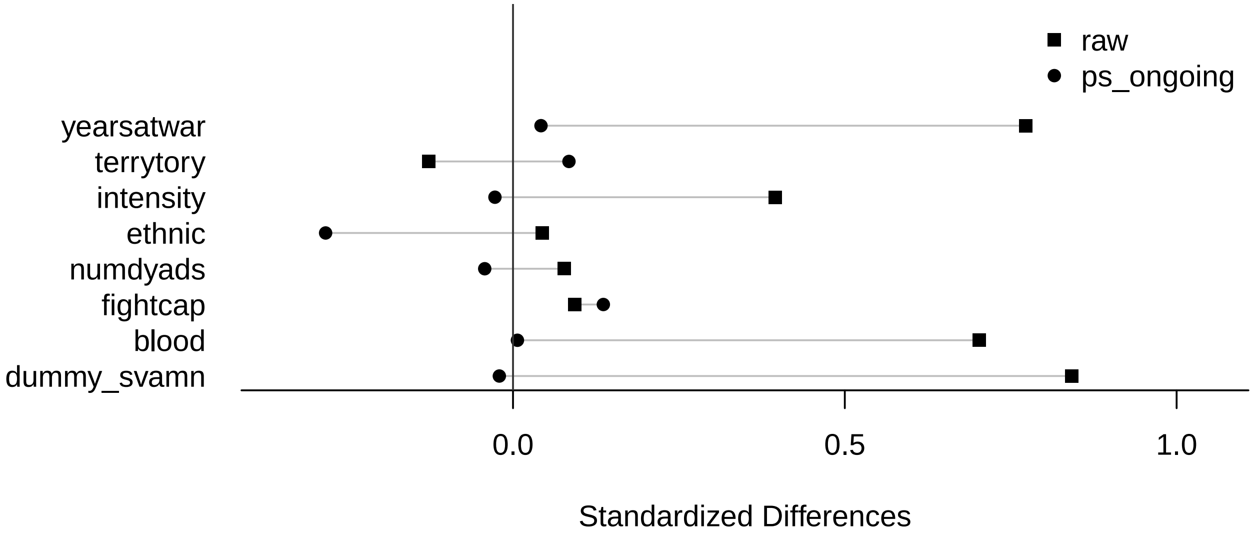


Table 12: Propensity Score Full Matching for Pre98

	chisquare	df	p.value
raw	128.242830	8	0.0000000
ps_ongoing	6.770069	8	0.5616316

Table 13: Mahalanobis Full Matching for Ongoing98

	chisquare	df	p.value
raw	128.2428	8	0
mhfull_ongoing	105.5363	8	0



4.4.1 Estimand 2

In pre-98 dataset, the estimand is about 0.07. In ongoing-98 dataset, the estimand is about 0.35 (Table 14 and 15). The estimand for ongoing98 is almost identical with the estimand 1 (from the interaction term) which makes sense as the interaction term indicates the treatment (`TNR`) in Ongoing98 datasets. The size of the estimand is greater in Ongoing-98 as expected in the theory.

Table 14: Estimands1 for Pre98

	estimand_label	estimand
TNR	TNR	0.0719664

Table 15: Estimands1 for Ongoing98

	estimand_label	estimand
TNR	TNR	0.3504837

4.4.2 Performance of Estimator 2 (Matching)

Again, I examine biases and RMSE based on 500 simulations to evaluate the performance of the matching estimator. Table 19 shows that in Pre-98 dataset, bias is 0.05, and the RMSE is 0.06. Table 20 shows that in ongoing-98 dataset, bias is -0.01, and the RMSE is 0.04. For both pre- and ongoing datasets, the biases and RMSE are smaller than those values using the `glm` etimator. Hence, the estimator 2 performs better than the estimator 1.

Table 16: Performance of Estimator 2 (Pre98)

Design	Inquiry	Estimator	Term	N Sims	Bias	RMSE
designs_match_pre	TNR	matching	TNR	500	0.05 (0.00)	0.06 (0.00)

Table 17: Performance of Estimator 2 (Ongoing)

Design	Inquiry	Estimator	Term	N Sims	Bias	RMSE
designs_match_ongoing	TNR	matching	TNR	500	-0.01 (0.00)	0.04 (0.00)

4.5 Test Statistics and Estimator Diagnosis

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, which I already have conducted.

4.5.1 Performance of the Tests

I judge the performance of tests by looking at the false positive rate and power. The power of a test 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. Table 18, 19 and 20 show the performance of the tests. Both estimators (glm and matching) have high power (all above 0.95). 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. Coverage rates indicate the false-positive rate at alpha = 0.05. 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. The false-positive rate (coverage) of two estimators are all above 0.6.

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.

Table 18: Test Performance on Estimator 1

	Design	Inquiry	Estimator	Power	Coverage	Mean Estimate	Mean Estimand
1	design_glm	interaction	glm	0.94	0.61	0.24	0.35
2				(0.01)	(0.02)	(0.00)	(0.00)

Table 19: Performance of Test (Pre98)

Design	Inquiry	Estimator	Power	Coverage	Mean Estimate	Mean Estimand
designs_match_pre	TNR	matching	0.97	0.61	0.13	0.07
			(0.01)	(0.02)	(0.00)	(0.00)

Table 20: Performance of Test 2 (Ongoing)

Design	Inquiry	Estimator	Power	Coverage	Mean Estimate	Mean Estimand
designs_match_ongoing	TNR	matching	1.00	0.94	0.34	0.35
			(0.00)	(0.01)	(0.00)	(0.00)

5 Mock Result

Table 21, 22, and 23 show the results of the analyses on the estimands using the logistic regression model and the propensity score matching. The result by `glm` on the interaction term (`TNR` and `ongoing98`) shows that the estimate of the interaction term is about 0.123. It indicates that the probability of the exchange of amnesties for serious violations is 0.12 higher to dyads by the transnational rebel groups that ended after the year 1998, compared to other cases. The p-value using t-test is 0.052 which suggests that I can almost reject the null hypothesis at the significant level at 0.05 (Table 21). Hence, if the real outcome were as I have simulated it, then the following table would suggest that there are some evidence to support the proposed theory. On the other hand, the results on the matching estimator (Table 22, and 23) suggest otherwise. The estimates are -0.0014718 and -0.0085794 respectively, which is the opposite direction from the theory. However, the p-values using the t-test are 0.97 and 0.78 which suggest that there is weak statistical evidence to reject the null hypothesis and to believe my theory.

Table 21: Interaction (`TNRx98`) from Estimator 1

	x
Estimate	0.1227070
Std. Error	0.0630019
t value	1.9476715
Pr(>|t|)	0.0520273

Table 22: Treatment (`TNR`) from Estimator 2 (PRE)

	x
Estimate	-0.0014718
Std. Error	0.0353156
t value	-0.0416764
Pr(>|t|)	0.9667777

Table 23: Treatment (`TNR`) from Estimator 2 (Ongoing)

	x
Estimate	-0.0085794
Std. Error	0.0306111
t value	-0.2802708
Pr(>|t|)	0.7794150

5.1 Replication Data

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

6 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))

# Change the name of column
names(df)[names(df) == "max_rebpresosts"] <- "TNR"
names(df)[names(df) == "warend_post98"] <- "ongoing98"

# 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
# Create a column for a binary svamn (indicating whether
# the conflict had an exchange of SV amnesty or not)
df$dummy_svamn <- ifelse(df$sum_hram > 0, 1, 0)  #making SVAmnesty into dummy

# Create a column for interaction of TNR and ongoing98
df$interaction <- (df$TNR) * (df$ongoing98)

# Change NA into 0 for 'SV' columns as NA indicates that
# there is no reported Serious crimes in the dyad by the
# rebel side.
df$sv[df$sv == "NA"] <- "0"
df$sv[is.na(df$sv)] <- 0
names(df)
# Distribution of war-periods (pre98, post98, ongoing98)
# and SV amnesties
table(df$sum_hram)  #Number of wars with SV amnesties: total 76 cases (0:422, 1: 66, 2
table(df$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
```

```r
table(df$cross98war, df$sum_hram)  #27 out of 67 wars involved with svamn.(40.3%)
table(df$ongoing98, df$sum_hram)  #ongoing98 (i.e., cross+post). 44 out of 203 wars inv
names(df)
# Creating reduced working data
wrdf <- df %>%
    dplyr::select(country.x, side_b, dyadid, sum_hram, yearsatwar,
        terrytory, intensity, ethnic, numdyads, fightcap, blood,
        pre98war, post98war, ongoing98, TNR, war_end_yr, war_end_yr,
        interaction, sv, dummy_svamn)
dim(wrdf)  #514, 19

# Removing Missing Data (I discuss this point later)
sum(is.na(wrdf))  #441
wrdat <- na.omit(wrdf)


wrdat$sv <- as.numeric(wrdat$sv)
dim(wrdat)  #dimension: 413, 19
unique(wrdat$country.x)  #101 countries (before removing missing data, there were 105)

table(wrdat$ongoing98)  #263, 150

table(wrdat$sum_hram)  #total dyad: 413, war with sva: 68; 9 of them had more than one

table(wrdat$dummy_svamn)  # 345 wars w/o svamn; 68 wars with.
table(wrdat$interaction)  #71 wars by TNR 'AND' happened ongoing98

# Categorizing conflicts by years of start and end yrs
df_pre98 <- wrdat[which(wrdat$post98war == 0), ]  #325 dyad wars (51 sva)
df_post98 <- wrdat[which(wrdat$post98war == 1), ]  #88 dyad (17 sva)
df_ongoing98 <- wrdat[which(wrdat$ongoing98 == 1), ]  #150 dyad (41 sva)
table(wrdat$TNR)  #246 with local rebel, 167 conflicts with TNRs
unique(wrdat$side_b)  #411
# PREMATCHING Balance Test for Pre-98 Subset
balfmla_pre98 <- reformulate(c(names(df_pre98)[c(5:11, 19)]),
    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(5:11,
    19)]), 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)

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

# 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 101.3

# xtable(ps_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)
```

```r
# There are 50.9 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)

# Below, I also try fullmatching with rank-based
# Mahalanobis distance.

############## Rank-Based Mahalanobis distance #### Make
############## distance matrices
mhdist_ongoing98 <- match_on(TNR ~ pscore_ongoing98, data = df_ongoing98_2,
    method = "rank_mahalanobis")
as.matrix(mhdist_ongoing98)[1:5, 1:5]

# Fullmatchingusing the ps
ps_mhdist_ongoing98 <- fullmatch(mhdist_ongoing98, data = df_ongoing98_2)

ps_mhdist_ongoing98_summary <- summary(ps_mhdist_ongoing98, data = df_ongoing98_2,
    min.controls = 0, max.controls = Inf)
# There are 52.8 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_mhdist_ongoing98 <- xBalance(balfmla_ongoing98, strata = list(raw = NULL,
    ps_mhdist_ongoing98 = ~ps_mhdist_ongoing98), data = df_ongoing98_2,
    report = "all")

plot(xb1_mhdist_ongoing98, main = "Ongoing-98 Xbalance Result")

xtable(xb1_mhdist_ongoing98$overall)
######### PRE-98 ########## Create Simulated Population
fake_population_whole <- declare_model(N = 1000, data = wrdat,
    handler = resample_data)

fake_population_pre98 <- declare_model(N = 1000, data = df_pre98,
    handler = resample_data)
```

```
fake_population_ongoing98 <- declare_model(N = 1000, data = df_ongoing98,
    handler = resample_data)

# sv as numeric
df_pre98$sv <- as.numeric(df_pre98$sv)
df_ongoing98$sv <- as.numeric(df_ongoing98$sv)

# Declare Potential Outcome (using the coeffecients from
# the logistic regression model. )

# summary(glm(dummy_svamn ~ interaction + TNR + ongoing98 +
# yearsatwar +terrytory + intensity+ethnic+ numdyads+
# fightcap+ blood+sv, data=wrdat))
pot.outcome_whole <- declare_potential_outcomes(dummy_svamn ~
    0.23493 * interaction + -0.020289 * TNR + -0.007065 * ongoing98 +
        0.01607 * yearsatwar + -0.079395 * terrytory + -0.015756 *
        intensity + 0.002318 * ethnic + -0.007976 * numdyads +
        0.010228 * fightcap + 0.052514 * blood + 0.085073 * sv +
        0.064828)

# summary(glm(dummy_svamn ~ TNR + yearsatwar +terrytory +
# intensity+ethnic+ numdyads+ fightcap+ blood+sv,
# data=df_pre98))
pot.outcome_pre98 <- declare_potential_outcomes(dummy_svamn ~
    0.008439 * TNR + 0.021124 * yearsatwar + -0.073385 * terrytory +
        -0.054135 * intensity + -0.095101 * ethnic + -0.012384 *
        numdyads + 0.052484 * fightcap + 0.128386 * blood + 0.108249 *
        sv + 0.044678, assignment_variables = "TNR")

# summary(glm(dummy_svamn ~ TNR + yearsatwar +terrytory +
# intensity+ethnic+ numdyads+ fightcap+ blood+sv,
# data=df_ongoing98))
pot.outcome_ongoing98 <- declare_potential_outcomes(dummy_svamn ~
    0.225086 * TNR + 0.016584 * yearsatwar + -0.09681 * terrytory +
        0.055754 * intensity + 0.006644 * ethnic + 0.001769 *
        numdyads + -0.004979 * fightcap + 0.029855 * blood +
        0.040419 * sv + 0.031219, assignment_variables = "TNR")

# Declare assignment
assignment <- declare_assignment(assignment_variable = "TNR")

# Declare how outcomes should be realized
treatment_outcome <- declare_reveal(outcome_variables = "dummy_svamn",
    assignment_variables = "TNR")
```

```r
# Declare design
my_design_whole <- fake_population_whole + pot.outcome_whole

my_design_pre <- fake_population_pre98 + pot.outcome_pre98 +
    assignment + treatment_outcome

my_design_ongoing <- fake_population_ongoing98 + pot.outcome_ongoing98 +
    assignment + treatment_outcome

# New simulated datasets
set.seed(12345)
dat1_whole <- draw_data(my_design_whole)
dat1_pre <- draw_data(my_design_pre)
dat1_ongoing <- draw_data(my_design_ongoing)

## Sampling 500 observations from the population
set.seed(12345)
library(randomizr)
sampling_1 <- declare_sampling(S = draw_rs(N = N, n = 500))

# whole
design_1whole <- fake_population_whole + sampling_1
set.seed(12345)
df1_fake_whole <- draw_data(design_1whole)  #NEW whole******

# pre
set.seed(12345)
design_1pre <- fake_population_pre98 + sampling_1
df1_fake_pre <- draw_data(design_1pre)  #NEW Pre******

# ongoing
design_1ongoing <- fake_population_ongoing98 + sampling_1
set.seed(12345)
df1_fake_ongoing <- draw_data(design_1ongoing)  #NEW Ongoing****
# install.packages('scales') install.packages('ggplot2')
library(scales)
library(ggpubr)

# Descriptive stats for original data
regtnr <- ggplot(wrdat, aes(x = TNR)) + geom_bar(fill = "darkgreen") +
    ggtitle("TNR") + xlab("") + scale_x_continuous(breaks = c(0,
    1), labels = c("No", "Yes"))

regongoing <- ggplot(wrdat, aes(x = ongoing98)) + geom_bar(fill = "darkgreen") +
    ggtitle("Ongoing 98") + xlab("Ongoing 98") + scale_x_continuous(breaks = c(0,
    1), labels = c("No", "Yes"))

regsv <- ggplot(wrdat, aes(x = sv)) + geom_bar(fill = "darkgreen") +
    ggtitle("sv") + xlab("Serious Violation") + scale_x_continuous(breaks = c(0,
    1), labels = c("No", "Yes"))

regdummysvamn <- ggplot(wrdat, aes(x = dummy_svamn)) + geom_bar(fill = "darkgreen") +
    ggtitle("SV Amnesty") + xlab("SV Amnesty") + scale_x_continuous(breaks = c(0,
    1), labels = c("No", "Yes"))

regyears <- ggplot(wrdat, aes(x = yearsatwar)) + geom_bar(fill = "darkgreen") +
    ggtitle("Years at war") + xlab("Years at war")

reginter <- ggplot(wrdat, aes(x = interaction)) + geom_bar(fill = "darkgreen") +
    ggtitle("interaction (TNR*98)") + xlab("interaction") + scale_x_continuous(breaks =
    1), labels = c("No", "Yes"))

ggarrange(regtnr, regongoing, regsv, regdummysvamn, regyears,
    reginter, ncol = 3, nrow = 2)
# Discriptive stats for simulated data
wholetnr <- ggplot(df1_fake_whole, aes(x = TNR)) + geom_bar(fill = "darkgreen") +
    ggtitle("TNR") + xlab("") + scale_x_continuous(breaks = c(0,
    1), labels = c("No", "Yes"))

wholeongoing <- ggplot(df1_fake_whole, aes(x = ongoing98)) +
    geom_bar(fill = "darkgreen") + ggtitle("Ongoing 98") + xlab("Ongoing 98") +
    scale_x_continuous(breaks = c(0, 1), labels = c("No", "Yes"))

wholesv <- ggplot(df1_fake_whole, aes(x = sv)) + geom_bar(fill = "darkgreen") +
    ggtitle("Serious Violation") + xlab("Serious Violation") +
    scale_x_continuous(breaks = c(0, 1), labels = c("No", "Yes"))

wholedummysvamn <- ggplot(df1_fake_whole, aes(x = dummy_svamn)) +
    geom_bar(fill = "darkgreen") + ggtitle("SV Amnesty") + xlab("SV Amnesty") +
    scale_x_continuous(breaks = c(0, 1), labels = c("No", "Yes"))

wholeyears <- ggplot(df1_fake_whole, aes(x = yearsatwar)) + geom_bar(fill = "darkgreen")
    ggtitle("Years at war") + xlab("Years at war")

wholeinter <- ggplot(df1_fake_whole, aes(x = interaction)) +
    geom_bar(fill = "darkgreen") + ggtitle("interaction (TNR*98)") +
    xlab("interaction") + scale_x_continuous(breaks = c(0, 1),
    labels = c("No", "Yes"))
```

```
ggarrange(wholetnr, wholeongoing, wholesv, wholedummysvamn, wholeyears,
    wholeinter, ncol = 3, nrow = 2)  ##fake
# Declare an estimand

### 1. with the interaction term one (glm)
make_estimand1_whole <- function(data) {
    bs <- coef(glm(dummy_svamn ~ interaction, data = df1_fake_whole))
    return(data.frame(estimand_label = c("glm"), estimand = bs[c("interaction")],
        stringsAsFactors = FALSE))
}

estimand1_whole <- declare_inquiry(handler = make_estimand1_whole,
    label = "pop_whole_relationship")

design1_and_estimand_whole <- fake_population_whole + sampling_1 +
    estimand1_whole

kable(estimand1_whole(df1_fake_whole), caption = "Estimands1 for interaction\\label{tab:
glm_fake_whole <- glm(dummy_svamn ~ interaction + TNR + ongoing98 +
    yearsatwar + terrytory + intensity + ethnic + numdyads +
    fightcap + blood + sv, data = df1_fake_whole)

plot(glm_fake_whole, which = 4, id.n = 5)  #196, 287, 420, 807, 912
# not all outliers are influential observations. To check
# whether the data contains potential influential
# observations, the standardized residual error can be
# inspected. Data points with an absolute standardized
# residuals above 3 represent possible outliers and may
# deserve closer attention.  Extract model results:computes
# the standardized residuals (.std.resid) and the Cook's
# distance (.cooksd) using the R function augment() from
# the broom package.

library(broom)
model.data <- augment(glm_fake_whole) %>%
    mutate(index = 1:n())
model.data %>%
    top_n(3, .cooksd)  #420, 807, 912

# plot the standardized residuals:
ggplot(model.data, aes(index, .std.resid)) + geom_point(aes(color = interaction),
    alpha = 0.5) + theme_bw()

# Filter potential influential data points
```

```r
filtered <- model.data %>%
    filter(abs(.std.resid) > 3)  #112, 420, 489, 531, 784, 807 removed

xtable(filtered)
collinearity <- car::vif(glm_fake_whole)
kable(collinearity, caption = "Assessing Collenearity Using VIF\\label{tab:VIF}")
# declare estimator1
glm_estimator1 <- declare_estimator(dummy_svamn ~ interaction +
    TNR + ongoing98 + yearsatwar + terrytory + intensity + ethnic +
    numdyads + fightcap + blood + sv, model = glm, term = c("interaction"),
    inquiry = c("interaction"), label = "glm")

design_glm <- design1_and_estimand_whole + glm_estimator1

set.seed(123345)
sim_full <- simulate_design(design_glm, sims = 500)

diag1_glm <- diagnose_design(sim_full)

estimator1perform <- reshape_diagnosis(diag1_glm, digits = 2,
    select = NULL, exclude = NULL)
xtable(estimator1perform)
xtable(estimator1perform[, c(1, 2, 3, 4, 5, 6, 7)])
# RMSE: BiasL -0.10, 0.13\t, Power: 0.94, Coverage: 0.62,
# Mean Estimate: 0.24, Mean Estimand: 0.35 Pre98

# Create linear predictors for ongoing-98 data
glm_pre98_sampled <- bayesglm(balfmla_pre98, data = df1_fake_pre,
    family = binomial)

df1_fake_pre$pscore_pre98 <- predict(glm_pre98_sampled, type = "link")

# Make distance matrices
psdist_pre98_sampled <- match_on(TNR ~ pscore_pre98, data = df1_fake_pre)
as.matrix(psdist_pre98_sampled)[1:5, 1:5]

caliper(psdist_pre98_sampled, 2)

# Fullmatching using the ps
ps_pre <- fullmatch(psdist_pre98_sampled, data = df1_fake_pre)

xbps_pre <- xBalance(balfmla_pre98, strata = list(raw = NULL,
    ps_pre = ~ps_pre), data = df1_fake_pre, report = c("std.diffs",
    "z.scores", "adj.means", "adj.mean.diffs", "chisquare.test",
    "p.values"))
# The larger chi square value, the greater the probability
# that there really is a significant difference.

# Create a rank-based Mahalanobis distance
mhdist_pre <- match_on(balfmla_pre98, data = df1_fake_pre, method = "rank_mahalanobis")

# fullmatch using a rank-based Mahalanobis distance
mhfull_pre <- fullmatch(mhdist_pre, data = df1_fake_pre)

xb_mh_pre <- xBalance(balfmla_pre98, strata = list(raw = NULL,
    mhfull_pre = ~mhfull_pre), data = df1_fake_pre, report = c("std.diffs",
    "z.scores", "adj.means", "adj.mean.diffs", "chisquare.test",
    "p.values"))
kable(xbps_pre$overall, caption = "Propensity Score Full Matching for Pre98\\label{tab:p
kable(xb_mh_pre$overall, caption = "Mahalanobis Full Matching for Pre98\\label{tab:mahal
df1_fake_pre$ps_pre <- NULL
df1_fake_pre[names(ps_pre), "mhfull_pre"] <- ps_pre
plot(xbps_pre)

df1_fake_pre$mhfull_pre <- NULL
df1_fake_pre[names(mhfull_pre), "mhfull_pre"] <- mhfull_pre
plot(xb_mh_pre)
## Ongoing98

# Create linear predictors for ongoing-98 data
glm_ongoing98_sampled <- bayesglm(balfmla_ongoing98, data = df1_fake_ongoing,
    family = binomial)

df1_fake_ongoing$pscore_ongoing98 <- predict(glm_ongoing98_sampled,
    type = "link")

# Make distance matrices
psdist_ongoing98_sampled <- match_on(TNR ~ pscore_ongoing98,
    data = df1_fake_ongoing)
as.matrix(psdist_ongoing98_sampled)[1:5, 1:5]

caliper(psdist_ongoing98_sampled, 2)

# Fullmatchingusing the ps
ps_ongoing <- fullmatch(psdist_ongoing98_sampled, data = df1_fake_ongoing)

xbps_ongoing <- xBalance(balfmla_ongoing98, strata = list(raw = NULL,
    ps_ongoing = ~ps_ongoing), data = df1_fake_ongoing, report = c("std.diffs",
    "z.scores", "adj.means", "adj.mean.diffs", "chisquare.test",
    "p.values"))
# The larger chi square value, the greater the probability
# that there really is a significant difference.

# Create a rank-based Mahalanobis distance
mhdist_ongoing <- match_on(balfmla_ongoing98, data = df1_fake_ongoing,
    method = "rank_mahalanobis")

# fullmatch using a rank-based Mahalanobis distance
mhfull_ongoing <- fullmatch(mhdist_ongoing, data = df1_fake_ongoing)

xb_mh_ongoing <- xBalance(balfmla_ongoing98, strata = list(raw = NULL,
    mhfull_ongoing = ~mhfull_ongoing), data = df1_fake_ongoing,
    report = c("std.diffs", "z.scores", "adj.means", "adj.mean.diffs",
        "chisquare.test", "p.values"))
kable(xbps_ongoing$overall, caption = "Propensity Score Full Matching for Pre98\\label{t
kable(xb_mh_ongoing$overall, caption = "Mahalanobis Full Matching for Ongoing98\\label{t
df1_fake_ongoing$mhfull_ongoing <- NULL
df1_fake_ongoing[names(mhfull_ongoing), "mhfull_ongoing"] <- mhfull_ongoing
plot(xbps_ongoing)
### 2. for matching estimator Pre
make_estimand1_pre <- function(data) {
    bs <- coef(glm(dummy_svamn ~ TNR + mhfull_pre, data = df1_fake_pre),
        subset = !is.na(mhfull_pre))
    return(data.frame(estimand_label = c("TNR"), estimand = bs[c("TNR")],
        stringsAsFactors = FALSE))
}

estimand1_pre <- declare_inquiry(handler = make_estimand1_pre,
    label = "pop_relationship")

design1_and_estimand_pre <- fake_population_pre98 + sampling_1 +
    estimand1_pre

# View estimand: #0.049011
kable(estimand1_pre(df1_fake_pre), caption = "Estimands1 for Pre98\\label{tab:estmnd2pre

######## Ongoing
make_estimand1_ongoing <- function(data) {
    bs <- coef(glm(dummy_svamn ~ TNR, data = df1_fake_ongoing))
    return(data.frame(estimand_label = c("TNR"), estimand = bs[c("TNR")],
        stringsAsFactors = FALSE))
}
```

```
estimand1_ongoing <- declare_inquiry(handler = make_estimand1_ongoing,
    label = "pop_relationship")

design1_and_estimand_ongoing <- fake_population_ongoing98 + sampling_1 +
    estimand1_ongoing


# View estimand: #0.3504837
kable(estimand1_ongoing(df1_fake_ongoing), caption = "Estimands1 for Ongoing98\\label{ta
# Matching estimator: Common for Pre-98, ongoing98
lm_match_estimator <- declare_estimator(dummy_svamn ~ TNR, inquiry = c("TNR"),
    term = c("TNR"), model = stats::lm, label = "matching")
# lm with matching

######## Pre-98 ###########
designs_match_pre <- design1_and_estimand_pre + lm_match_estimator
set.seed(1232123)
sim_match_pre <- simulate_design(designs_match_pre, sims = 500)

diag2_pre <- diagnose_design(sim_match_pre)
diag2_pre
# 500 simulation, bias: 0.08, RMSE: 0.08, Power: 0.97,
# coverage: 0.35, mean estimate: 0.13, mean estimand: 0.05

######## Ongoing -98 ###########
designs_match_ongoing <- design1_and_estimand_ongoing + lm_match_estimator
set.seed(1232123)
sim_match_ongoing <- simulate_design(designs_match_ongoing, sims = 500)
set.seed(1232123)
diag2_ongoing <- diagnose_design(sim_match_ongoing)
# 500 simulation, bias: -0.01, RMSE: 0.04, Power: 1,
# coverage: 0.94, mean estimate: 0.34, mean estimand: 0.35
es2_pre <- reshape_diagnosis(diag2_pre, digits = 2, select = NULL,
    exclude = NULL)
kable(es2_pre[, c(1:7)], caption = "Performance of Estimator 2 (Pre98) \\label{tab:perf2

es2_ongoing <- reshape_diagnosis(diag2_ongoing, digits = 2, select = NULL,
    exclude = NULL)
kable(es2_ongoing[, c(1:7)], caption = "Performance of Estimator 2 (Ongoing)\\label{tab:
# xtable(estimator1perform[,c(1,2,3,8,9,10,14)])

kable(es2_pre[, c(1:3, 8, 9, 10, 14)], caption = "Performance of Test (Pre98) \\label{ta

kable(es2_ongoing[, c(1:3, 8, 9, 10, 14)], caption = "Performance of Test 2 (Ongoing)\\l
library(stats)

est1 <- glm(dummy_svamn ~ interaction + TNR + ongoing98 + yearsatwar +
    terrytory + intensity + ethnic + numdyads + fightcap + blood +
    sv, data = df1_fake_whole)
matrix_coef_glm <- summary(est1)
est2glm <- matrix_coef_glm$coefficients
coef_glm <- est2glm[2, ]
kable(coef_glm, caption = "Interaction (`TNR`x`98`) from Estimator 1\\label{tab:es1res}"

est2 <- lm(dummy_svamn ~ TNR + ps_pre, data = df1_fake_pre)
matrix_coef_pre <- summary(est2)
est2pre <- matrix_coef_pre$coefficients
coef_pre <- est2pre[2, ]
kable(coef_pre, caption = "Treatment (`TNR`) from Estimator 2 (PRE) \\label{tab:es2respr

est3 <- lm(dummy_svamn ~ TNR + ps_ongoing, data = df1_fake_ongoing)
matrix_coef_ongoing <- summary(est3)
est2ongoing <- matrix_coef_ongoing$coefficients
coef_ongoing <- est2ongoing[2, ]
kable(coef_ongoing, caption = "Treatment (`TNR`) from Estimator 2 (Ongoing) \\label{tab:
```

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.

Chen, Jianshen, and David Kaplan. 2015. "Covariate Balance in Bayesian Propensity Score Approaches for Observational Studies." *Journal of Research on Educational Effectiveness*.

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).

Gelman, A. 2011. *Arm: Data Analysis Using Regression and Multilevel/ Hierarchical Modelsm, Arm: Dataanalysis Using Regression and Multilevel/Hierarchical Models (r Package Version 1.4–13).* http://CRAN.R-project.org/package=arm.

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.*

———. 2016. “Basic Principles of Statistical Inference.” https://imai.fas.harvard.edu/teaching/files/basics Presentation Slides.

Kassambara. 2018. “Logistic Regression Assumptions and Diagnostics in r.” http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/.

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*.

Schreiber-Gregory, Deanna, and Karlen Bader. 2018. "Logistic and Linear Regression Assumptions: Violation Recognition and Control." Henry M Jackson Foundation.

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

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.