(Don't Fear) The Reaper: Bayesian Causal Forests and the Mixed Evidence for Drone Strike Effectiveness

Taylor Cox David Puelz

Working draft – June 24, 2026

Abstract

Do drone strikes reduce terrorist violence? Early studies of the US drone campaign in Pakistan generally answer yes, but these findings depend on research designs with strong assumptions about targeting decisions and baseline violence levels. This paper reassesses the evidence using Bayesian Causal Forests, a flexible estimator designed for observational settings with strong confounding and heterogeneous effects. We first reanalyze two influential studies, Johnston and Sarbahi (2016) and Mir and Moore (2019), and show that their original negative estimates are not recovered under BCF specifications that separate baseline violence from treatment effects. We then extend the analysis to a broader Pakistan panel of group-location-months from 2010 through 2014, using ACLED event data and an explicit selection-on-observables design. Across alternative geographic scopes, outcome definitions, and treatment windows, posterior intervals generally include zero and the estimated treatment effects do not consistently point toward reductions in violence. Unit-level treatment-effect estimates likewise provide little evidence that null average effects conceal a large subgroup of strongly negative effects. The results suggest that the optimistic empirical case for drone-strike effectiveness in Pakistan is more fragile than the early literature implies.

1 Introduction

Recent conflicts have highlighted the expanding role of drones in modern military arsenals: both Russia and Ukraine are rapidly innovating in drone technology to develop long-range munitions capable of striking rear areas and frontline tactical systems that compete along heavily contested frontlines; the United States and its allies in the Gulf are rapidly developing attritable weapons systems to combat Iran's bombardment of Shahed drones; and both China and Taiwan are deploying underwater drones as they prepare for potential conflict in the Pacific theater. However, before drones became central to conventional warfare, they were primarily used by the United States as a counterterrorism tool in the "War on Terror." Following 9/11 and the subsequent US operations in Afghanistan, Iraq, and across the Middle East, studies in political science began exploring whether or not drone strikes were effective at degrading terrorist[1] groups. "Effective" is usually defined as reducing either the incidence or lethality of terrorist attacks against civilians or insurgent attacks against military and civilian targets.[2] Although new studies continue to explore drone strike efficacy from various angles,[3] most early studies conclude that strikes do effectively reduce subsequent terrorist activity, particularly attacks against civilians (Mir and Moore 2019; Johnston and Sarbahi 2016; Jaeger and Siddique 2018).

However, these prior studies may make overly optimistic claims, as their results depend on identification strategies built around assumptions that are hard to defend in this setting. In order to reassesses the evidence for drone strikes, we first attempt to replicate two of the foundational papers in the drone strike literature using a Bayesian Causal Forest, and then extend the analysis by applying the same model to a new dataset to determine whether

[1]Note that, like other literatures on conflict, the literature on drone strikes deals with the distinction between "terrorist," "insurgent" and "militant" in different ways. Here, we use terrorists to describe any group targeted by drone strikes when discussing drone strike efficacy broadly, but purposely transition to using the same term as the authors whose work we replicate in later sections.

[2]Alternatively, studies concerned with decapitation strikes utilize group duration as the outcome variable. For examples of studies examining duration, see Jordan (2014) and Price (2012).

[3]See, for example, Rigterink (2021) on the effect of successful decapitation strikes on terrorist violence; Schwartz, Fuhrmann, and Horowitz (2022) on the cross-country effect of adopting armed drones; and Gartzke and Walsh (2022) on how drone strikes displace militant groups to urban centers.

drone strikes have an effect under varying geographic and temporal conditions. We employ a Bayesian Causal Forest (BCF) because it is better suited to the empirical setting, since it is designed specifically for scenarios with strong confounding, heterogenous effects, and small effect sizes. When attempting to replicate the prior foundational studies, we fail to recover either study's substantive and statistically significant negative effect; the BCF estimates consistently fail to recover precise negative effects, suggesting the original results are fragile to modeling and identification choices. In the extension analysis, the posterior intervals consistently include zero across various operationalizations of terrorist violence, alternative inclusion criteria for "terrorist attack," different geographic levels, and multiple treatment windows. Altogether, we do not find support for prior studies' optimism about drones' ability to degrade terrorist groups and thereby reduce civilian targeting.

2 Drone Strikes and the Existing Evidence

Recent studies have complicated the early consensus that drone strikes were effective at degrading terrorist groups by showing that drone strikes may actually *increase* violence. Yet this newer literature leaves unresolved whether the earlier optimistic findings were the product of specific empirical designs or whether they capture a real effect in the case of Pakistan, where the United States focused its drone campaign after the Taliban fled from Afghanistan.

The first study to assess the affect of drone strikes on terrorist activities was Jaeger and Siddique's (2011, later revised 2018) paper on the relationship between drone strikes and attacks committed by the Taliban and Al Qaeda in Afghanistan and Pakistan. Using a vector autoregressive model, they find strong evidence of deterrence effects in Pakistan: drone strikes, regardless of whether they succeed or fail to hit a leader, appear to reduce Taliban violence in Pakistan, while their effects on Al Qaeda's activities in Afghanistan are less clear. Johnston and Sarbahi's (2016) subsequent analysis also examines the effects of US drone strikes on terrorism in Pakistan, but instead utilizes a two-way fixed-effects (2FE) model with both agency and temporal (week) fixed effects and a spatial lag of drone strikes (2FESL). They find evidence that drone strikes reduce both the incidence and lethality of terrorist attacks, along with the selective targeting of tribal elders, who were frequently targeted for resisting terrorist control in the tribal strongholds of North Waziristan and Khyber Pakhtunkhwa (Fishman 2011). Finally, Mir and Moore, who also focus their analysis on Pakistan but leverage a difference-in-differences model, argue that the US drone program was "associated with monthly reductions of around nine to thirteen insurgent attacks and fifty-one to eighty-six casualties" due to terrorists anticipating drone strikes and consequentially constraining their activites.

However, Rigterink (2021) finds evidence that the number of terrorist attacks may increase after leadership decapitations. When drone strikes successfully eliminate a terrorist leader, groups appear to conduct more attacks in the months that follow compared to when strikes fail and "miss" their intended target. Additionally, drone strikes that kill rank-and-file militants, rather than leaders, may encourage attacks against civilians in urban settings (Gartzke and Walsh, 2022). Mahmood and Jetter (2023) also find evidence that drone strikes increase the incidence of terrorist attacks; using an instrumental variable design, their results suggest that drone strikes encourage terrorism over the upcoming days and weeks.

Wr select Johnston and Sarbahi's (2016) and Mir and Moore's (2019) papers for replication for several reasons. First, Johnston and Sarbahi's paper is one of the most widely-cited papers in drone strike and counterterrorism literature, so it is worthwhile to examine whether their results are robust to an alternative modeling strategy. Second, these two papers represent distinct identification strategies; Johnston and Sarbahi's 2FESL model relies on conditional ignorability while Mir and Moore's difference-in-differences design relies instead on parallel trends. Even if one of these assumptions was violated, we might expect the other to hold and the associated paper's results to be robust to the BCF specification. Third, both of these studies are conducive to replication using BCF, which requires a binary treatment.

We chose not to replicate Jaeger and Siddique (2018) because their results rely partially on tests of joint significance of 21 lags (t = day) of drone strikes, and this more complicated lag structure does not translate well to the binary treatment form required by BCF.[4] We also chose not to replicate any of the three studies occurring after 2020, as each of these

[4]**TC: I probably could replicate the results for the individual lags, I would need to think more about what that would mean.**

studies either extends or contradicts one of the earlier analyses we focus on reanalyzing.

2.1 Johnston and Sarbahi

As mentioned previously, Johnston and Sarbahi (2016) utilize a 2FESL model and their identification strategy is conditional ignorability. While they note that diagnostics suggest that the 2FESL model fit is preferred to the simpler 2FE model, their results remain consistent across both model specifications. As such, we utilize the simpler 2FE model when replicating their analysis:

$$y_{it} = \alpha_i + \beta x_{it} + \eta_t + \epsilon_{it} \tag{1}$$

where y_{it} measures the incidence of terrorism, x_{it} is the number of drone strikes, α_i are unobserved agency fixed effects, and η_t are time (week) fixed effects. Agency and time fixed effects control for all the time-invariant differences between agencies, such as geography, as well as time-specific differences such as seasonal weather patterns and religious events.

After controlling for agency and time fixed effects, Johnston and Sarbahi present their results as "plausibly unbiased estimates of drone strikes' causal effect on terrorism in FATA." (Johnston and Sarbahi, 2016, p. 208) Two "quasi-random" factors support their empirical strategy: day-to-day changes in cloud cover and bureaucratic constraints on drone strike operations. (Johnston and Sarbahi, 2016, p. 209). First, exogenous weather patterns may dictate drone use due to the weapon system's need for good visibility to acquire, surveil, and eliminate a target. Second, a series of logistical constraints, including policymakers' schedules in Washington, the limited availability of armed weapon systems, and competing priorities in other conflict theaters, may hinder operations in Pakistan.

However, it is unlikely that these factors are "quasi-random," in which case the application of drone strikes is also unlikely to be as-if random, even after controlling for agency and time fixed effects. First, Johnston and Sarbahi note that Al Qaeda leadership advised operatives to remain hidden in safe houses on clear days, suggesting that terrorist activities likely increase on cloudy days and decrease on clear days. If drone strikes increase on clear days and decrease on cloudy days, weather is a confounder, not an exogenous factor, and the negative effect of drone strikes on terrorist activities may be the result of terrorist groups understanding US targeting patterns.[5]

Second, there is strong evidence that these bureaucratic constraints did not present real hindrances to the drone campaign in Pakistan. According to classified US military documents, the Joint Special Operations Command had a 60-day window to hit an approved target, meaning policymakers and military personnel had approximately three months to resolve the logistical constraints described by Johnston and Sarbahi. Furthermore, US counterterrorism efforts were centered around Pakistan from 2008 to 2012, so it was unlikely that demands in Somalia, Libya, or Yemen would have pulled resources away from the region (New America 2026).

Additionally, any omitted time-varying agency-specific factor that affects both drone strikes and terrorist violence threatens Johnston and Sarbahi's identification strategy of conditional ignorability (Rosenbaum and Rubin 1983). While the two-way fixed-effects design adjusts for persistent differences across agencies and shocks common to all agencies in a given week, it does not address time-varying agency-specific confounders, such as local military operations, changes in militant capacity, or agency-specific violence trajectories that may affect both drone targeting and subsequent terrorist activity.

2.2 Mir and Moore

Mir and Moore, in constrast, rely on a difference-in-differences design, which does not require conditional ignorability and instead relies on the parallel trends assumption (Angrist and Pischke 2009). In their design, they compare insurgent violence in tehsils in North Waziristan[6] after 2008 to insurgent violence in other FATA agencies in the same period. They specify treatment as North Waziristan post-2008 because the United States was legally only allowed to conduct drone strikes beginning in 2008 in this agency, which comprised the "flight box".

[5] Several studies document the existence of "anticipatory effects," whereby terrorist groups decrease activities in anticipation of being targeted by drone strikes. While potentially resulting in the same outcome, this is a different mechanism than the degradation mechanism that the marginal effect of each additional drone strike modeled by Johnston and Sarbahi suggests **TC: This might be too much context**.

[6] *Tehsils* are local districts within larger administrative *agencies* in FATA.

(Mir and Moore, 2019, p. 846) Their regression conforms to the following structure:

$$Y_{it} = \delta_j(\text{treat}_t \cdot \text{nwa}_i) + \alpha_i + \theta_t + \beta X_{it} + \tau_i \cdot t + \epsilon_{it} \tag{2}$$

where Y_{it} is the outcome variable, treat_t indicates whether time period t occurs in 2008 or later, and nwa_i is a binary variable indicating whether tehsil i is in North Waziristan. In addition, they include tehsil and month fixed effects as well as tehsil-specific linear time trends.

In this setting, the parallel trends assumption requires that, absent the implementation of the drone program, tehsils inside and outside of North Waziristan would have followed the same trends in insurgent violence. After conducting an event study to determine whether the parallel trends assumption holds, Mir and Moore acknowledge that there is strong evidence that the levels of violence in tehsils inside and outside of North Waziristan were already diverging; as such, they caution against a strictly causal interpretation of their results (Mir and Moore, 2019, p. 856). According to their event study, it appears that insurgent violence had begun increasing in other agencies prior to 2008, meaning that their negative results may be due not to a reduction in violence in North Waziristan as a result of the implementation of the drone program, but to rising levels of violence in *other agencies*. However, they argue that "results are not only driven by increasing violence outside the flight box after 2008 but also by the absence within North Waziristan of the large spikes in violence observed during 2006 and 2007 after the commencement of the drone program. A negative causal effect of the program can only be ruled out if one thinks it likely that this observed decline in violence would have occurred anyway—or been more severe—absent the program." In other words, they acknowledge that their findings may over-estimate the negative effect of drone strikes on insurgent violence, but argue that the sharp decline in violence in North Waziristan observed during the treatment period suggests the program has some effect.

TC: There is also something going on with terrorist displacement, ie geographical spillover. They conduct some formal tests, I don't understand them.

3 Estimation Strategy: Bayesian Causal Forests

We use Bayesian Causal Forests (BCF) to replicate Johnston and Sarbahi's (2016) and Mir and Moore's (2019) studies as well as to conduct our own analysis because BCF is especially well-suited to settings with small effect sizes, heterogenous effects, and strong confounding factors (Hahn, Murray, and Carvalho 2020). Attempting to isolate the effect of drone strikes on terrorist violence often encounters these modeling challenges. First, drone strikes are not necessarily expected to have a strong negative effect on weekly or even monthly levels of violence, but rather to have a smaller, marginal effect that compounds over time. Second, drone strikes likely impact various regions and terrorist groups differently; for example, some types of terrorist group structures may be more vulnerable to leadership decapitation than others (Jordan 2009, 2014; Price 2012). Finally, ongoing conflict in the form of militant clashes with the government, other militias, and against tribal leadership make it difficult to separate the effect of strikes from fluctuations in baseline violence.

The problem of small effect sizes is particularly challenging in conflict settings more broadly, because the outcome, violence, usually has a wide range, with some months witnessing extremely high numbers of attacks or civilian casualties, meaning the marginal effect of some intervention has to be determined from a noisy outcome process. BCF addresses this challenge by modeling baseline violence and treatment effect separately, allowing the prognostic function to absorb large baseline differences in violence. This separation avoids falsely attributing variation in the outcome to the treatment when that variation is due to baseline differences in violence.

All of our BCF specifications conform to the following structure:

$$E(Y_i \mid x_i, Z_i = z_i) = \mu(x_i, \hat{\pi}(x_i)) + \tau(x_i) z_i \tag{3}$$

where the functions μ and τ are given independent Bayesian Additive Regression Tree (BART) priors and $\hat{\pi}(x_i)$ is an estimate of the propensity score $\pi(x_i) = \Pr(Z_i = 1 \mid x_i)$ (Hahn, Murray, and Carvalho 2020, p. 969). The treatment-effect forest is typically regularized more strongly than the prognostic forest, reflecting the prior expectation that treatment effects are smaller than baseline outcome variation. The propensity function measures the probability that a unit will receive treatment given the control variables (in this setting, the probability that drone or air strikes will be conducted against the relevant unit of analysis). We estimate the propensity function using logit BART because the treatment is binary. Y_i represents the outcome variable (terrorist violence), z_i is the treatment variable (drone strike exposure), and x_i is the vector of control variables.

Separating $\mu(X)$ from $\tau(X)$ helps mitigate regularization-induced confounding, which can occur when z_i and x_i are highly correlated. In this scenario, a naive BART specification that treats z_i as another predictor may struggle to determine whether a change in the outcome is due to covariate characteristics or treatment; for example, if religious terrorist groups are inherently more violent, but also more likely to be targeted by drone strikes, the model may be unable to determine whether variation in group violence is due to ideology or drone strike intervention. In situations with strong confounding, a regularized model may misallocate variation between treatment and covariates, either attributing baseline differences to treatment or shrinking real treatment effects toward zero. Strong confounding is often present in conflict settings, and almost certainly describes the counterterrorism landscape in Pakistan, where decisions to target specific groups are shaped by the group's prior violence and its size, age, and ideology. Including $\hat{\pi}$ in the baseline risk forest allows μ to split on $\hat{\pi}$ and enables the baseline model to account for the fact that treatment-prone units may have different expected outcomes regardless of treatment.

The treatment-effect estimate follows directly from the model's implied potential outcomes. For an untreated unit, $z_i = 0$, so the estimated untreated potential outcome is

$$\widehat{Y_i(0)} = \widehat{\mu}(x_i, \hat{\pi}(x_i)). \tag{4}$$

For a treated unit, $z_i = 1$, so the estimated treated potential outcome is

$$\widehat{Y_i(1)} = \widehat{\mu}(x_i, \hat{\pi}(x_i)) + \widehat{\tau}(x_i). \tag{5}$$

The estimated treatment effect for unit i is therefore

$$\widehat{Y_i(1)} - \widehat{Y_i(0)} = \widehat{\tau}(x_i). \tag{6}$$

Thus, $\widehat{\tau}(x_i)$ is the unit-level conditional average treatment effect (CATE). The average treatment effect (ATE) is the average of these unit-level effects across observations,

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\tau}(x_i). \tag{7}$$

In computation, we average τ_i across units within each posterior draw and then summarize the posterior distribution of these ATE draws.

Like other estimators, BCF does not by itself identify causal effects; causal interpretation requires that, conditional on the adjustment set, treatment assignment is independent of potential outcomes.

4 Reanalysis of Prior Studies

In the following section, we replicate Johnston and Sarbahi's (2016) and Mir and Moore's (2019) analyses and show that we can recover their initial results when using their main model specifications but not when using BCF. The BCF estimates fail to recover either study's negative effect; across specifications, the BCF posterior intervals include zero, suggesting the original results are fragile to modeling and identification choices.

4.1 Johnston and Sarbahi (2016)

We begin with Johnston and Sarbahi because their study is one of the earliest and most influential findings that drone strikes reduce terrorist violence in Pakistan. The replication targets their main 2FE design, in which terrorist violence is modeled as a function of drone strike exposure, agency fixed effects, and week fixed effects. We examine the same three outcomes emphasized in their analysis: terrorist incident frequency, lethality, and attacks on tribal elders, each measured per 100,000 residents.

As a preliminary check, we recover the original 2FE estimates from Johnston and Sarbahi's replication data; these estimates exactly match the coefficients reported in the original study.

In Johnston and Sarbahi's design, the treatment is a count of drone strikes; however, BCF requires a binary treatment variable. We define treatment as one or more drone strikes in an agency-week:

$$Z_{it} = \mathbf{1}[\text{uav\_a}_{it} \geq 1] \tag{8}$$

Among treated agency-weeks, the majority of units receive only one strike; as Figure 1 illustrates, higher counts are rare.

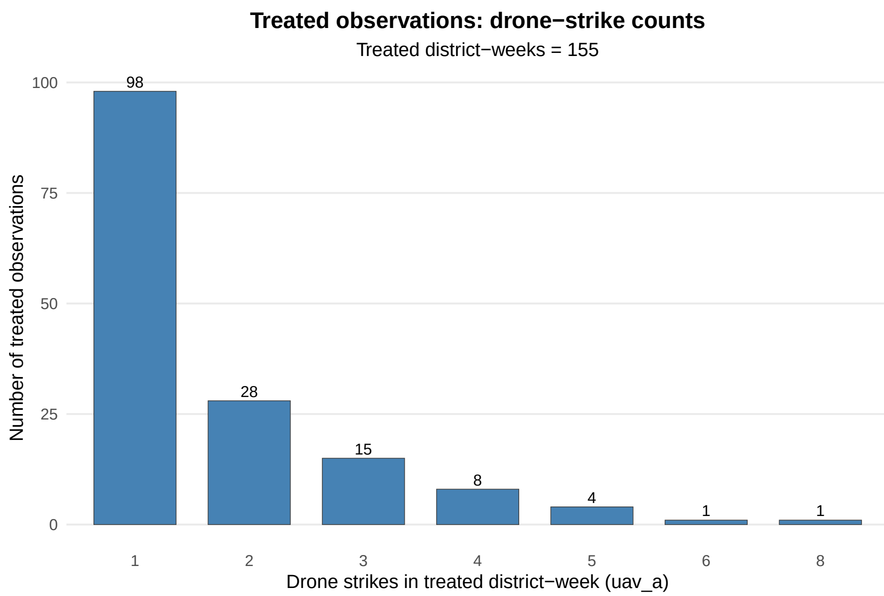

Figure 1: Drone-strike counts among treated agency-weeks in Johnston and Sarbahi's replication data

Our BCF analysis, therefore, estimates the effect of strike exposure rather than the marginal effect of each additional strike. Because multi-strike agency-weeks are uncommon, the binary treatment captures the main treatment contrast in the data. The marginal effect of an additional strike is likely smaller than the effect of any strike, so aggregating in this way is unlikely to drive our null findings.

We estimate BCF under three propensity-score specifications: logit BART, gradient-boosted trees, and elastic-net logistic regression. Varying $\hat{\pi}$ allows us to assess whether the results are sensitive to any one machine-learning specification of the propensity score. The BART-learned propensity remains the main specification, while XGBoost and elastic net provide robustness checks using alternative learners applied to the same fixed-effect design variables.

Table 1: Propensity-score specifications for the Johnston and Sarbahi BCF reanalysis

Specification	Definition	Purpose
Logit BART	Logit BART using agency and week indicators	Main propensity-score specification
XGBoost	Gradient-boosted trees using agency and week indicators	Alternative nonlinear learner for propensity-score robustness
Elastic net	Regularized logistic regression using agency and week indicators	Penalized parametric benchmark for propensity-score robustness

We further define X_i as the same vector of control variables that Johnston and Sarbahi use, including agency and week fixed-effect indicators.

The results of the BCF model do not provide support for Johnston and Sarbahi's initial results. While the posterior means are negative, the 95% posterior credible intervals include zero across outcomes and $\hat{\pi}$ specifications. Note that credible intervals are interpreted slightly differently than confidence intervals: a 95% credible interval means there is a 95% probability that the true estimate lies within the interval. However, like a confidence interval, a credible interval that crosses zero indicates that we cannot be sure whether the treatment has a negative, null, or positive effect; in this case, that means drone strikes may reduce, have no effect on, or increase terrorist violence. The 95 percent credible intervals for the BCF ATE include zero across the propensity-score specifications for incidents, lethality, and attacks on elders. The $\hat{\pi}$ specification has limited effect on this conclusion: changing from logit BART to XGBoost or elastic-net logistic regression changes the posterior mean only modestly and does not restore the original results. Thus, the BCF reanalysis suggests that the original negative estimates are not robust to a flexible model that separates baseline violence from treatment-effect heterogeneity.

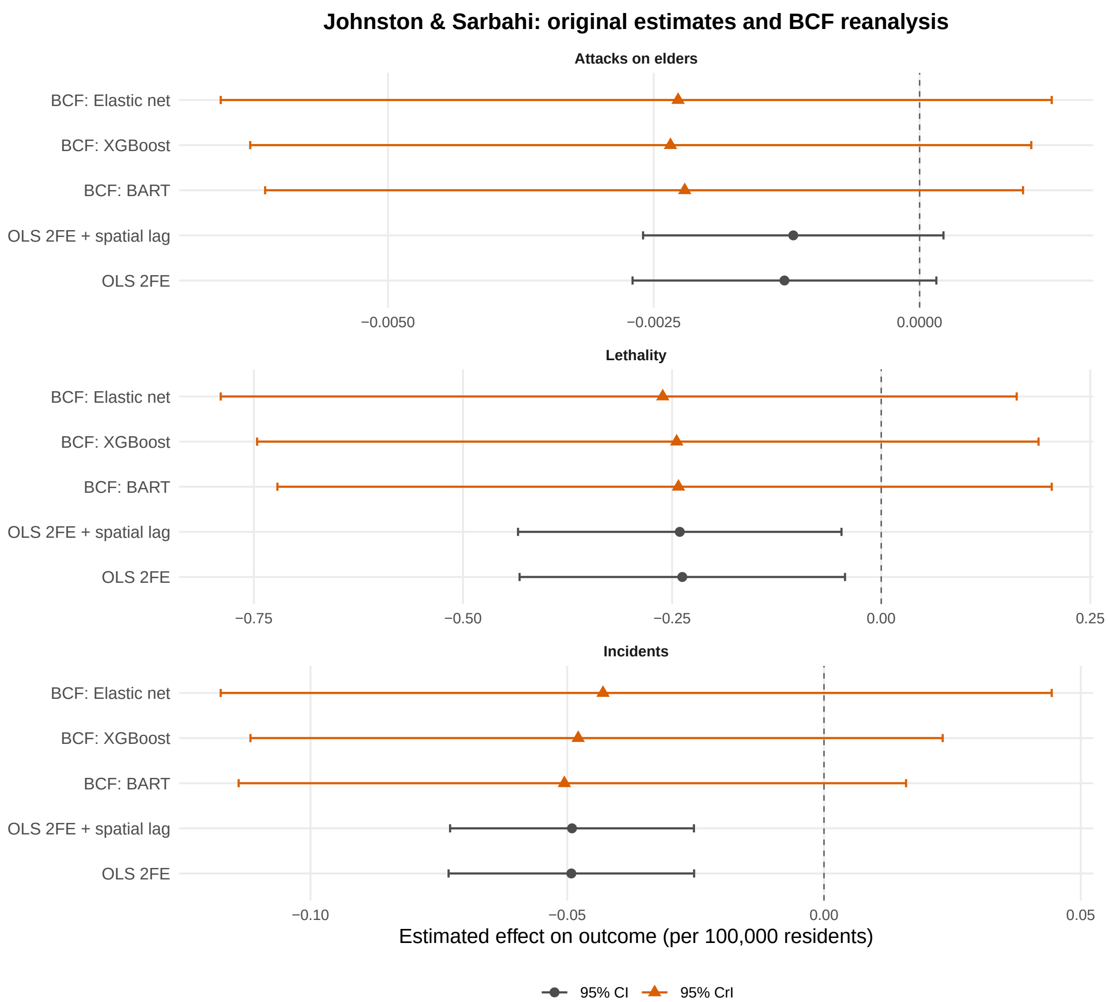

Figure 2: Johnston and Sarbahi original estimates and BCF reanalysis

TC: For incidents and lethality, my estimates are very similar, just more uncertain. Should I discuss that?...

4.2 Mir and Moore (2019)

We next turn to Mir and Moore because their study provides a second influential estimate of drone-strike effectiveness using a different research design. Unlike Johnston and Sarbahi's agency-week fixed-effect model, Mir and Moore estimate a difference-in-differences design, where treatment is defined by the implementation of the US drone program in North Waziristan in 2008. Their estimand is therefore closer to the effect of program exposure in the treated region-period than to the marginal effect of an additional strike. As such, their treatment is already binary; this makes the treatment variable formally compatible with BCF without recoding a count treatment into exposure:

$$Z_{it} = \mathbf{1}[\text{NWA}_i = 1 \text{ and } t \geq 2008] \tag{9}$$

The DiD estimate depends on the assumption that, absent the drone program, tehsils inside and outside North Waziristan would have followed parallel trends in terrorist violence. Mir and Moore acknowledge that this assumption is strained, since violence trends inside and outside the flight box were already diverging before treatment, but they argue their results indicate that the program had some causal effect of reducing terrorist violence.

Again, we start by recovering the original DiD estimates from Mir and Moore's replication data; these estimates exactly match the coefficients reported in the original study.

While the original study design's binary treatment is conducive to a BCF specification, the deterministic nature of the treatment presents a challenge for BCF. Like other estimators, BCF requires overlap between treatment and control units in the covariate space, and treatment mechanically defined by tehsil and year produces very weak overlap. As a result, the model compares violence levels for North Waziristan tehsils after 2008 to untreated observations from other tehsils and periods, and any effect captured may not reflect the drone program, but rather factors unique to North Waziristan in the post-2008 period.

To partially address this problem, we estimate the effect of the drone program on ΔY, the transformed version of each of Mir and Moore's two outcomes, incidents and casualties. For each outcome, ΔY is the *difference* between violence in a post-2008 tehsil-month and that tehsil's pre-program average. We define ΔY as:

$$\Delta Y_{it} = Y_{it} - \bar{Y}_{i,\text{pre}}, \tag{10}$$

where Y_{it} is violence in tehsil i and month t, and $\bar{Y}_{i,\text{pre}}$ is average pre-program violence in tehsil i:

$$\bar{Y}_{i,\text{pre}} = \frac{1}{T_i^{\text{pre}}} \sum_{s \in \mathcal{T}_i^{\text{pre}}} Y_{is}. \tag{11}$$

As a result, for each tehsil in the post-2008 period, the outcome is that tehsil's current level of violence minus the tehsil's pre-program average violence. We construct this pre-program average separately for each of Mir and Moore's two outcomes: incidents and casualties.

The treatment effect for the delta outcome is thus:

$$\begin{aligned}\tau_{it}^{\Delta} &= \mathbb{E}\left[(Y_{it}(1) - \bar{Y}_{i,\text{pre}}) - (Y_{it}(0) - \bar{Y}_{i,\text{pre}}) \mid X_{it}\right] &(12)\\ &= \mathbb{E}\left[Y_{it}(1) - Y_{it}(0) \mid X_{it}\right]. &(13)\end{aligned}$$

Although we define ΔY as the difference between post-2008 violence and pre-program violence, note that τ_{it}^{Δ} can still be interpreted as the effect of the drone program on post-2008 levels of violence, making it directly comparable to Mir and Moore's initial results. Estimating ΔY as opposed to Y partially addresses fixed tehsil-level differences by modeling post-period violence relative to each tehsil's own pre-program baseline.

We then estimate BCF using the difference between post-2008 incidents and casualties and each tehsil's pre-program average. The treatment indicator is Mir and Moore's treatment variable, which is equivalent to indicating North Waziristan membership after restricting the BCF sample to the post-2008 period. The covariate matrix mirrors their preferred specification, including tehsil fixed effects, month-year fixed effects, time trends, tehsil-specific trends, indicators for the Obama review period and the Haqqani network peace deal, and tehsil-specific military operations and peace-deal controls. Although we remove stable baseline differences between tehsils by subtracting each tehsil's pre-program average, the post-period treatment remains highly predictable from the included covariates, likely those associated with tehsil; because of this, there is still less overlap in the Mir and Moore reanalysis setting than in the Johnston and Sarbahi one.

Similarly to the reanalysis of Johnston and Sarbahi's study, we estimate propensity scores using three specifications: logit BART, gradient-boosted trees, and elastic-net logistic regression. For the delta-outcome reanalysis, each learner uses the same time-varying covariates from the Mir and Moore specification. Only the BART and XGBoost propensity scores are used in the BCF reanalysis; the elastic-net propensity score is reported as a support diagnostic, but it recovers a near-deterministic assignment structure in the post-period sample, meaning it is not very useful as a BCF propensity-score regularizer.

Table 2: Propensity-score specifications for the Mir and Moore BCF reanalysis

Specification	Definition	Purpose
Logit BART	Logit BART using time-varying covariates from the Mir and Moore specification	Main propensity-score specification used in the BCF reanalysis
XGBoost	Gradient-boosted trees using the same time-varying covariates	Alternative nonlinear learner used in the BCF reanalysis
Elastic net	Regularized logistic regression using the same time-varying covariates	Support diagnostic; not plotted in the Mir and Moore BCF figure because it produces near-deterministic fitted propensities

The propensity scores also indicate severe separation between treated and control observations. Because treatment assignment is deterministic based on tehsil and year, we estimate the Mir and Moore propensity scores using tehsil-specific time-varying covariates, including linear, tehsil-specific trends in violence, tehsil-specific military operations and peace deals, and the indicators for the Obama review period and the Haqqani network peace deal, which were not applied at the same time in every tehsil. Poor overlap across the learned propensity-score specifications suggests that treatment can still be predicted from these tehsil-specific trends. While this does not by itself invalidate the DiD design, which relies on parallel trends rather than propensity-score overlap, it shows that the treated region-period is highly distinctive and supports caution about any model trying to infer the counterfactual.

The delta-outcome BCF reanalysis does not uniformly recover Mir and Moore's original negative estimates. For attacks, the sign changes across the BART and XGBoost propensity-score specifications, and both credible intervals include zero. For casualties, both posterior means are negative, but the XGBoost interval includes zero, suggesting that the strongest evidence is sensitive to the propensity-score learner. Like with Johnston and Sarbahi's results, the BCF reanalysis shows that the original findings are not stable under a flexible alternative model. Given that Mir and Moore's own diagnostics already call parallel trends assumptions into question, this weakens the evidentiary basis for the argument that drone strikes reduce terrorist violence.

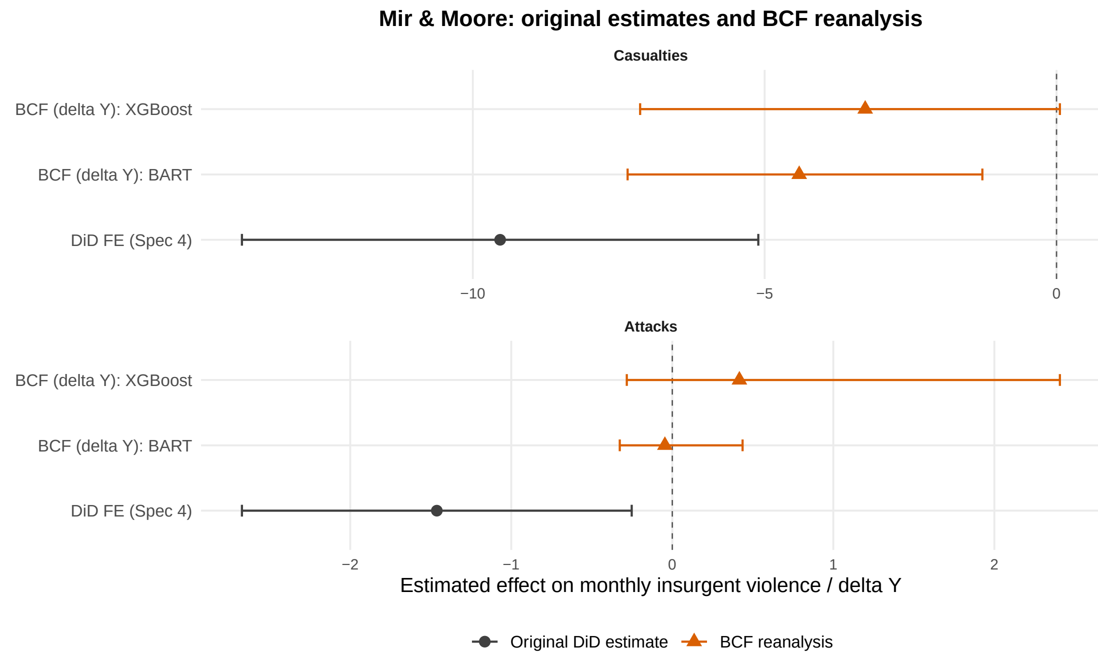

Figure 3: Mir and Moore original estimates and BCF reanalysis

Together, the two reanalyses show that the canonical evidence for drone effectiveness is fragile across two different original research designs: a fixed-effect model and a DiD design. The next section explores whether evidence for drone-strike effectiveness reappears when the analysis is extended beyond the original replication samples to a broader Pakistan panel. This extension is not intended to provide a definitive test of all drone and airstrike effects, but rather to assess whether the negative effects reported in the early literature travel across alternative geographic scopes, outcome definitions, and treatment windows under a common BCF framework with a clearly defined identification strategy.

5 Extension to a Broader Pakistan Panel

5.1 Purpose and Estimand

The replication analyses show that two canonical findings in the literature on drone strike efficacy are fragile under BCF reanalysis. The following extension explores whether evidence for drone-strike effectiveness reappears when the analysis moves beyond the original replication data to evaluate the effect of strikes on a broader geography and over several months.

The unit of analysis is the group-location-month. We use group-location-month because we expect that the probability of being targeted by drone strikes depends not only on location and time, but also on group identity. For each observation i, Y_i denotes subsequent terrorist or insurgent violence in the outcome month,[7] Z_i indicates whether the relevant group-location was exposed to a drone or air strike in treatment window m, and X_i is the vector of pre-treatment covariates used to adjust for differences in baseline risk and treatment propensity. The treatment window m can take four values: one month before the outcome month, one-to-three months before the outcome month, four-to-six months before the outcome month, and seven-to-twelve months before the outcome month. The primary specification uses the one-to-three-month window because it captures short-run effects while allowing more time for militant groups to respond than the one-month specification. We interpret the four-to-six-month and seven-to-twelve-month windows more cautiously, since there were very few treated units in these treatment windows.

Following the BCF specification in Equation 3, the key treatment-effect quantity is $\tau(x_i)$, the estimated effect of strike exposure for observation i given its covariates. The main reported estimand is the average of these unit-level treatment-effect estimates across observations,

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\tau}(x_i). \tag{14}$$

[7] For further information on the criteria used to define "terrorist" and "insurgent" in our extension analysis, see Appendix E.

We also examine the distribution of $\widehat{\tau}(x_i)$ across observations to assess whether the average estimates mask meaningful treatment-effect heterogeneity.

The extension relies on selection on observables: conditional on the adjustment set, strike exposure is assumed independent of potential outcomes. Formally,

$$\{Y_i(1), Y_i(0)\} \perp Z_i \mid X_i. \tag{15}$$

This identifying assumption differs from Mir and Moore's parallel-trends assumption and from Johnston and Sarbahi's fixed-effect design. Rather than relying on a quasi-experiment, the extension makes the adjustment set explicit and asks whether evidence for drone-strike effectiveness reappears under a selection-on-observables framework.

5.2 DAG and Covariate Selection

Figure 4 summarizes the causal framework used to select covariates. The adjustment set conditions on lagged attacks, lagged target type, military operations, population density, location, time, and group capacity. We use group age to proxy group capacity because reliable group size data are not available across the panel. In the DAG, recent attacks and prior target type affect both strike assignment and subsequent violence; military operations affect strike deployment and militant violence; population density shapes targeting constraints, attack opportunities, and expected casualty counts; location absorbs adminstrative unit-specific geographic differences in targeting priorities, violence patterns, and military operations; and time captures period-specific shifts in strike intensity, military operations, political context, and violence.

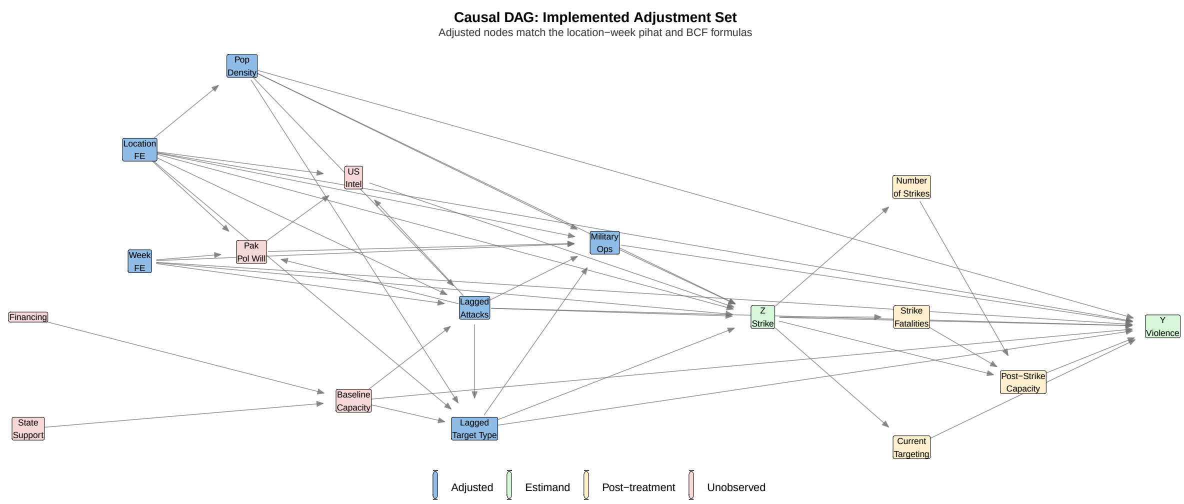

Figure 4: Causal framework for strike exposure and subsequent militant violence

Group type and state capacity are not part of the minimal sufficient adjustment set implied by the DAG. Group type is assumed to shape targeting priorities, but not fatalities directly after conditioning on the adjustment variables. State capacity is treated as affecting strike exposure and violence indirectly through intelligence and military operations. We include both variables in the outcome model to help the prognostic function capture baseline differences in violence across organizations and locations, rather than because they are required for identification. We exclude post-treatment variables such as strike fatalities, the number of strikes in the treatment window, current target type, and post-strike group capacity. Full DAG details, covariate-selection decisions, and excluded variables are reported in Appendix A.

5.3 Data and Panel Construction

We create a dataset of political violence and air/drone strikes in Pakistan from 2010 through 2014 using the Armed Conflict Location and Event Database (ACLED), which records the date, perpetrator, target, location, and fatalities of reported violence and protest events worldwide (Raleigh, Kishi, and Linke 2023).[8] ACLED surveys multiple sources to create a comprehensive database of political violence, including traditional media at the subnational level, reports from international organizations, and verified social media accounts, and it frequently corroborates and updates its event records. Previous studies of civilian victimization frequently use ACLED data, although drone strike research has typically relied on the New America Foundation and the Bureau of Investigative Journalism, both of which restrict their coverage to US drone strikes in limited regions of Pakistan.

To create the dataset for this study, we use ACLED event records on violence against civilians, all insurgent violence, and air/drone strikes.[9] We filter the violence data to construct two outcome scopes: a narrower sample of violence against civilians and a broader sample of all insurgent violence. We also filter the air/drone strike events to identify strikes against the non-state armed groups included in the panel.

The analysis uses four related panels that vary the geographic scope and the definition of violence. Panels A and B use a national geography, covering six provinces or province-equivalent units: Balochistan, FATA, Federal Capital Territory, Khyber Pakhtunkhwa, Punjab, and Sindh. Panels C and D restrict the analysis to FATA and use seven agency-level units: Bajaur, Khyber, Kurram, Mohmand, North Waziristan, Orakzai, and South Waziristan. Panels A and C focus on violence against civilians, while Panels B and D broaden the outcome to all insurgent violence.[10] The main sample applies a five-event threshold for group inclusion. This yields 2,880 observations in Panel A, 5,400 in Panel B, 3,360 in Panel C, and 6,300 in Panel D.

The primary outcome is the number of fatalities in the group-location-month. The extension also examines the number of attacks and fatalities per attack. The adjustment set

[8] ACLED data are available at https://acleddata.com.

[9] ACLED codes the treatment events as "air/drone strikes," which includes both drone strikes and other aerial strikes. We therefore refer to the extension treatment as air/drone strike exposure, while retaining the original authors' drone-strike definitions in the replication analyses.

[10] Violence against civilians refers to ACLED events in which the target actor is coded as civilians. All insurgent violence includes events involving non-state armed actors coded by ACLED as rebel groups, political militias, or identity militias, excluding unidentified armed actors and political parties.

combines covariates constructed from ACLED event histories with external structural and group-level sources. Lagged attacks and military operations are constructed from ACLED; lagged target type is constructed from the Global Terrorism Database (National Consortium for the Study of Terrorism and Responses to Terrorism 2026); population density is taken from WorldPop (WorldPop 2026); state capacity is proxied with VIIRS nighttime lights (NOAA Earth Observation Group 2026); and group age and group type are coded from CISAC's Mapping Militants Project and the South Asia Terrorism Portal (Center for International Security and Cooperation 2026; Institute for Conflict Management 2026). The appendix reports detailed coding rules, group-inclusion decisions, and variable definitions.

5.4 Extension Results

Figure 5 shows that the extension does not recover the negative effects reported in the early literature on drone strike effectiveness. Across the panel definitions, outcome measures, and treatment windows, posterior intervals generally include zero and the point estimates do not consistently move in a negative direction. This pattern holds not only for fatalities, but also for the number of attacks and fatalities per attack. Because strike exposure remains concentrated among a small number of groups and locations, these estimates should be interpreted as support for caution about the strong negative effects presented in earlier literature, rather than as a precise zero effect of drone and airstrikes.

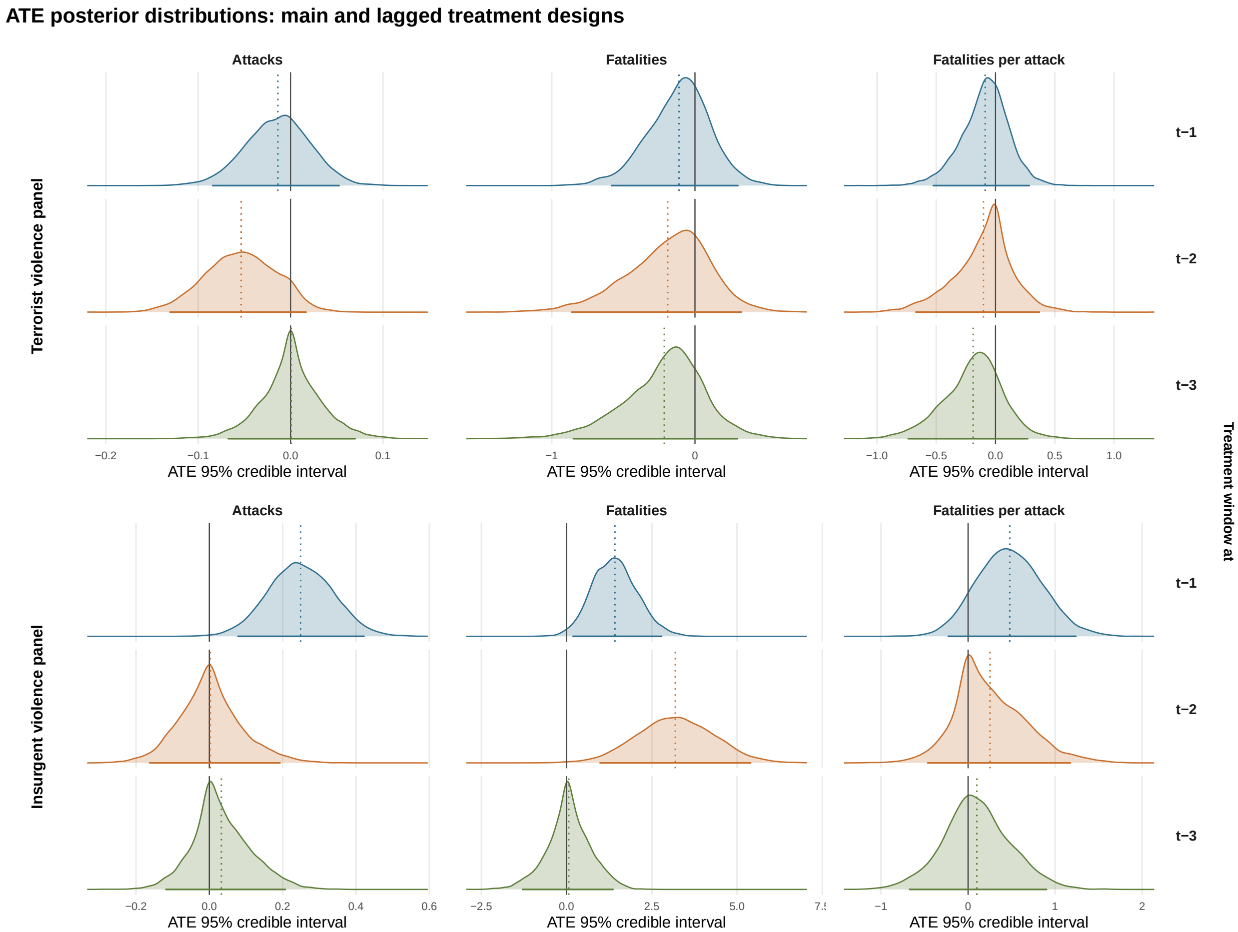

Figure 5: BCF ATE estimates across extension panels, outcomes, and treatment windows

The CATE distributions in Figure 6 provide a similar diagnostic at the unit level. Although BCF allows treatment effects to vary across group-location-months, the posterior mean CATEs are most dense near zero across outcomes, panels, and treatment windows. Thus, the null ATEs do not appear to be masking a large subset of strongly negative treatment effects. The estimates of the effect of drone strikes on the number of attacks show some positive heterogeneity in broader national panels, while the estimates of the treatment effect on fatalities per attack provide the strongest visual suggestion of negative heterogeneous effects in a few specifications. However, neither pattern overturns the main conclusion: the extension does not recover consistent evidence that strike exposure reduces subsequent violence.

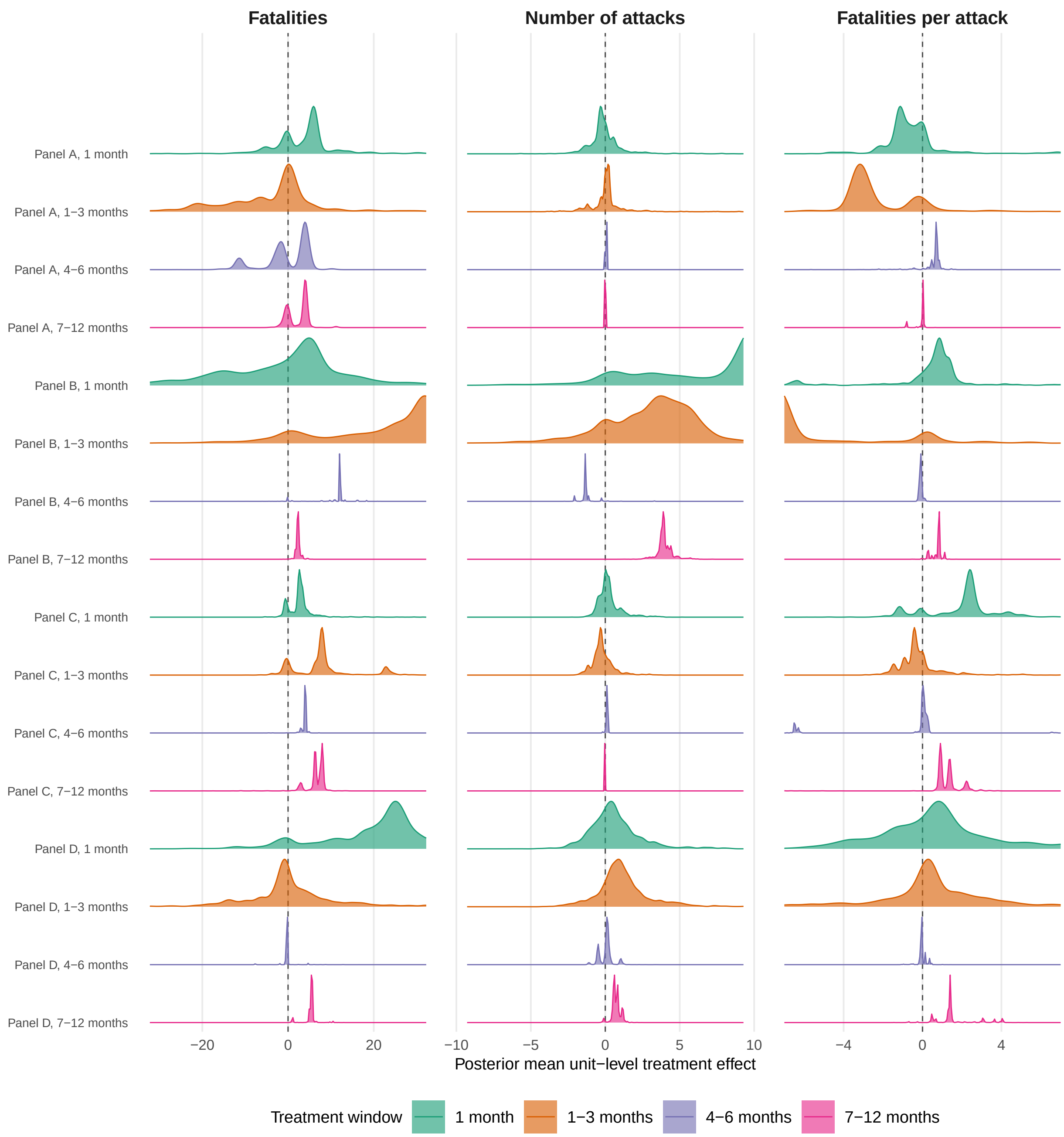

Figure 6: Distribution of posterior mean CATEs across extension outcomes, panels, and treatment windows

6 Discussion and Conclusion

References

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

Center for International Security and Cooperation, Stanford University. 2026. *Mapping Militants Project*. https://cisac.fsi.stanford.edu/mappingmilitants.

Fishman, Brian. 2011. "Militancy and Conflict in North Waziristan." New America Foundation.

Gartzke, Erik, and James Igoe Walsh. 2022. "The Drawbacks of Drones: The Effects of UAVs on Escalation and Instability in Pakistan." *Journal of Peace Research* 59(4): 463–477.

Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho. 2020. "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)." *Bayesian Analysis* 15(3): 965–1056.

Institute for Conflict Management. 2026. *South Asia Terrorism Portal*. https://www.satp.org/.

Jaeger, David A., and Zahra Siddique. 2018. "Are Drone Strikes Effective in Afghanistan and Pakistan? On the Dynamics of Violence between the United States and the Taliban." *CESifo Economic Studies* 64(4): 667–697.

Johnston, Patrick B., and Anoop K. Sarbahi. 2016. "The Impact of US Drone Strikes on Terrorism in Pakistan." *International Studies Quarterly* 60(2): 203–219.

Jordan, Jenna. 2009. "When Heads Roll: Assessing the Effectiveness of Leadership Decapitation." *Security Studies* 18(4): 719–755.

Jordan, Jenna. 2014. "Attacking the Leader, Missing the Mark: Why Terrorist Groups Survive Decapitation Strikes." *International Security* 38(4): 7–38.

Mahmood, Rafat, and Michael Jetter. 2023. "Gone with the Wind: The Consequences of US Drone Strikes in Pakistan." *The Economic Journal* 133(650): 787–811.

Mir, Asfandyar, and Dylan Moore. 2019. "Drones, Surveillance, and Violence: Theory and Evidence from a US Drone Program." *International Studies Quarterly* 63(4): 846–862.

National Consortium for the Study of Terrorism and Responses to Terrorism. 2026. *Global Terrorism Database*. https://www.start.umd.edu/gtd/.

New America. 2026. *America's Counterterrorism Wars: The Drone War in Pakistan*. New America report page.

NOAA Earth Observation Group. 2026. *VIIRS Nighttime Lights*. https://eogdata.mines.edu/products/vnl/.

Price, Bryan C. 2012. "Targeting Top Terrorists: How Leadership Decapitation Contributes to Counterterrorism." *International Security* 36(4): 9–46.

Raleigh, Clionadh, Roudabeh Kishi, and Andrew Linke. 2023. "Political Instability Patterns Are Obscured by Conflict Dataset Scope Conditions, Sources, and Coding Choices." *Humanities and Social Sciences Communications* 10(74).

Rigterink, Anouk S. 2021. "The Wane of Command: Evidence on Drone Strikes and Control within Terrorist Organizations." *American Political Science Review* 115(1): 31–50.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Schwartz, Joshua A., Matthew Fuhrmann, and Michael C. Horowitz. 2022. "Do Armed Drones Counter Terrorism, or Are They Counterproductive? Evidence from Eighteen Countries." *International Studies Quarterly* 66(3).

WorldPop. 2026. *WorldPop Open Population Repository*. https://www.worldpop.org/.

A DAG and Covariate-Selection Details

This appendix provides the full covariate-selection logic behind Figure 4. The DAG was constructed by evaluating whether each candidate variable plausibly precedes and affects strike exposure, subsequent violence, or both. The purpose of the DAG is not to prove that the adjustment set eliminates all confounding, but to make the identifying assumptions explicit.

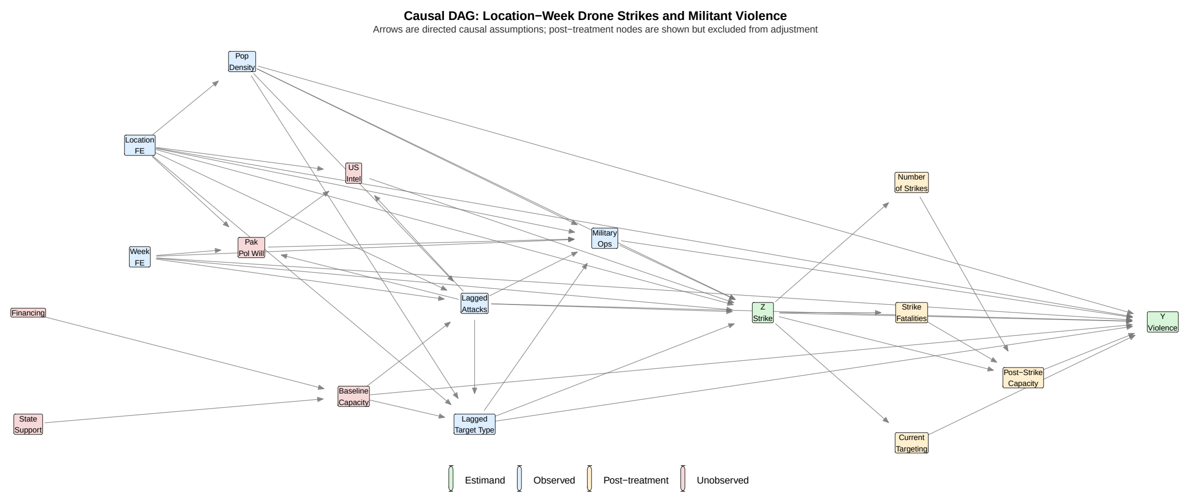

Figure 7: Full causal framework for strike exposure and subsequent militant violence

The minimal sufficient adjustment set implied by the DAG includes lagged attacks, lagged target type, military operations, population density, location, time, and an observed group-capacity proxy. We use group age for this role because a reliable standardized group-size measure is not available across the panel.

Table 3: DAG-based covariate-selection decisions

Variable	Rationale	Role
Lagged attacks	Recent violence increases the likelihood of strike exposure and predicts subsequent violence through persistence in conflict activity.	Adjustment
Lagged target type	Prior target selection captures group tactics and threat profile, which may influence both targeting decisions and later fatalities.	Adjustment
Military operations	Local military activity is correlated with strike deployment and independently affects militant behavior and violence.	Adjustment
Population density	Denser areas affect both targeting constraints and expected casualty counts.	Adjustment
Location	Strike exposure and violence vary sharply across provinces and FATA agencies.	Adjustment
Time	Strike intensity, military operations, and militant violence vary across months and policy periods.	Adjustment
Group age	Older groups may differ in organizational capacity, strategy, and baseline violence; used as the observed group-capacity proxy.	Adjustment
Group type	The DAG treats group ideology as affecting strike and military priorities, but not fatalities directly after conditioning on the core adjustment variables. Included to improve baseline outcome modeling.	Outcome precision

Variable	Rationale	Role
State capacity	The DAG treats state capacity as affecting strikes and violence indirectly through intelligence and military operations. Included to improve baseline outcome modeling.	Outcome precision

Several variables are deliberately excluded because they are downstream of treatment, such as strike fatalities and the number of strikes in the treatment window. Other variables, including US intelligence, Pakistani political will, external financing, state support, and baseline group capacity, are not directly observed. The extension therefore remains a selection-on-observables design: the DAG clarifies the assumptions and observed adjustment set, but it does not eliminate the possibility of residual unobserved confounding.

B Propensity Score Distributions and Support Diagnostics

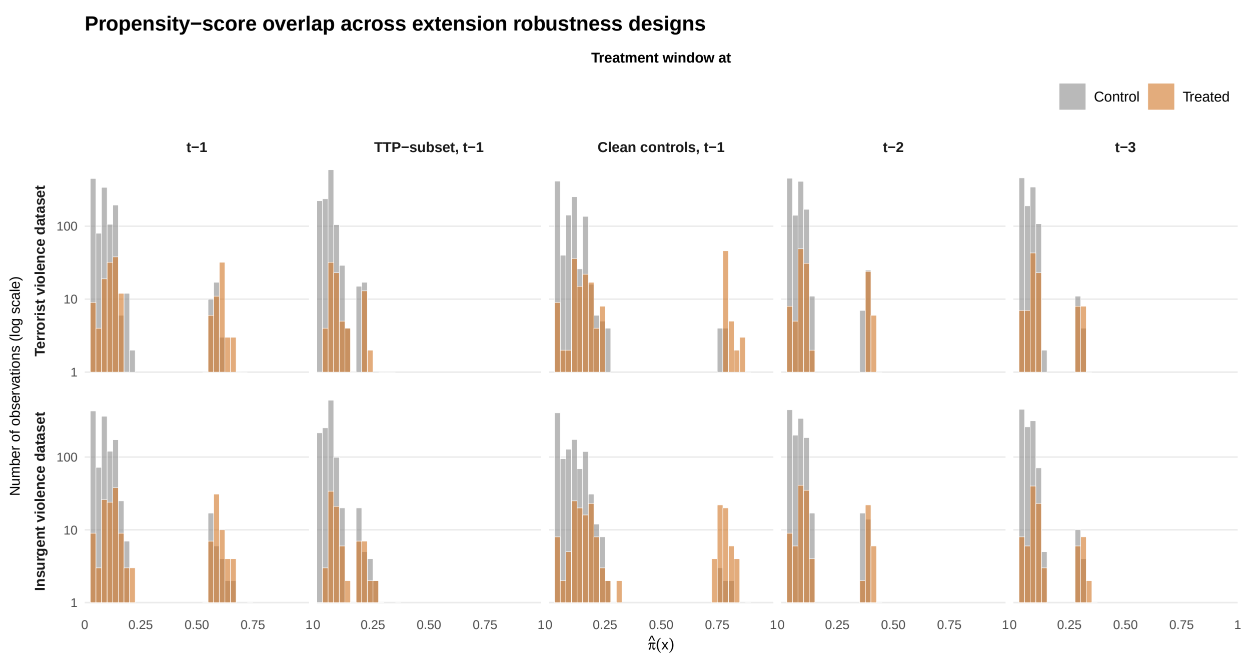

Figure 8: Propensity-score overlap across extension robustness designs

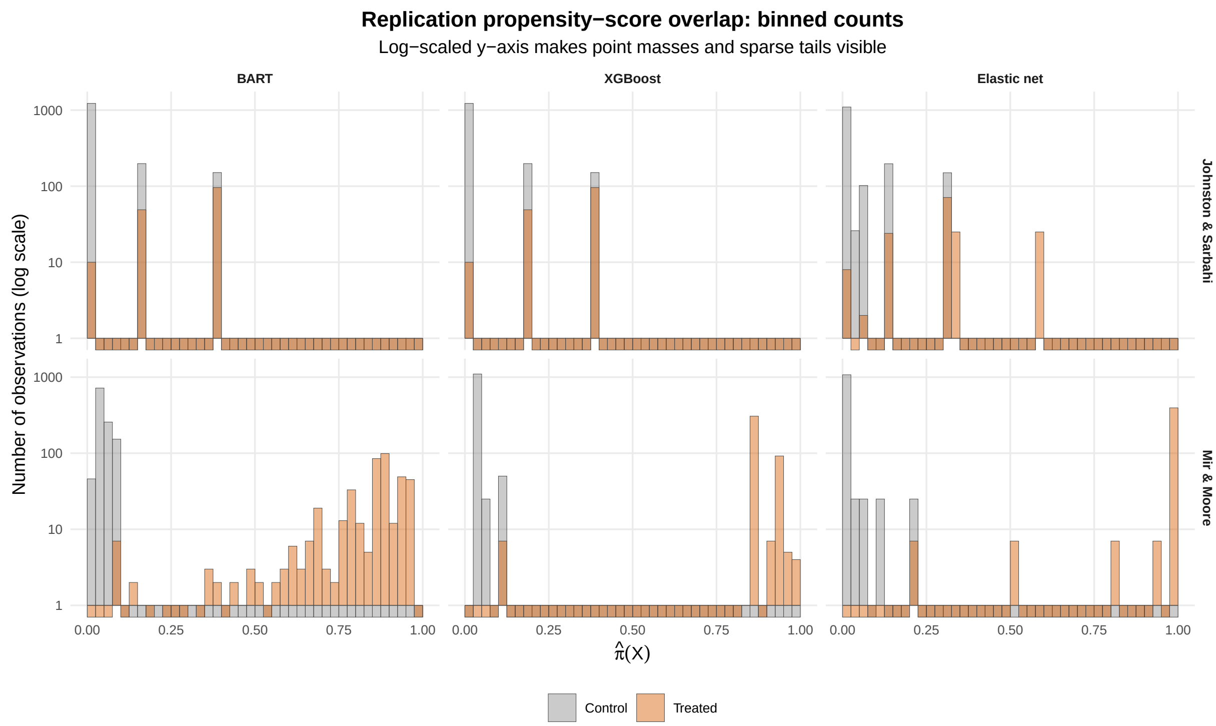

Figure 9: Propensity-score overlap in the replication samples

C Full Replication Tables

Table 4: Mir & Moore original estimates and BCF reanalysis

Outcome	Method	ATE	Interval	p-value	$P(ATE > 0)$
Casualties	DiD FE (Spec 4)	-9.532	[-13.956, -5.109]	< 0.001	
Casualties	BCF (delta Y): BART	-4.408	[-7.348, -1.270]		0.010
Casualties	BCF (delta Y): XGBoost	-3.278	[-7.133, 0.058]		0.108
Attacks	DiD FE (Spec 4)	-1.462	[-2.673, -0.252]	0.019	
Attacks	BCF (delta Y): BART	-0.046	[-0.326, 0.436]		0.240
Attacks	BCF (delta Y): XGBoost	0.417	[-0.282, 2.407]		0.574

Table 5: Johnston & Sarbahi original estimates and BCF reanalysis

Outcome (per 100k)	Method	ATE	Interval	p-value	$P(ATE > 0)$
Incidents	OLS 2FE	-0.049	[-0.073, -0.025]	< 0.001	
Incidents	OLS 2FE + spatial lag	-0.049	[-0.073, -0.025]	< 0.001	
Incidents	BCF: BART	-0.051	[-0.114, 0.016]		0.068
Incidents	BCF: XGBoost	-0.048	[-0.112, 0.023]		0.090
Incidents	BCF: Elastic net	-0.043	[-0.117, 0.044]		0.168
Lethality	OLS 2FE	-0.238	[-0.432, -0.043]	0.017	
Lethality	OLS 2FE + spatial lag	-0.241	[-0.434, -0.047]	0.015	
Lethality	BCF: BART	-0.242	[-0.722, 0.204]		0.159
Lethality	BCF: XGBoost	-0.245	[-0.746, 0.188]		0.150
Lethality	BCF: Elastic net	-0.261	[-0.790, 0.162]		0.145
Attacks on elders	OLS 2FE	-0.001	[-0.003, 0.000]	0.081	
Attacks on elders	OLS 2FE + spatial lag	-0.001	[-0.003, 0.000]	0.099	
Attacks on elders	BCF: BART	-0.002	[-0.006, 0.001]		0.103
Attacks on elders	BCF: XGBoost	-0.002	[-0.006, 0.001]		0.103
Attacks on elders	BCF: Elastic net	-0.002	[-0.007, 0.001]		0.124

D Extension Analysis Robustness Checks

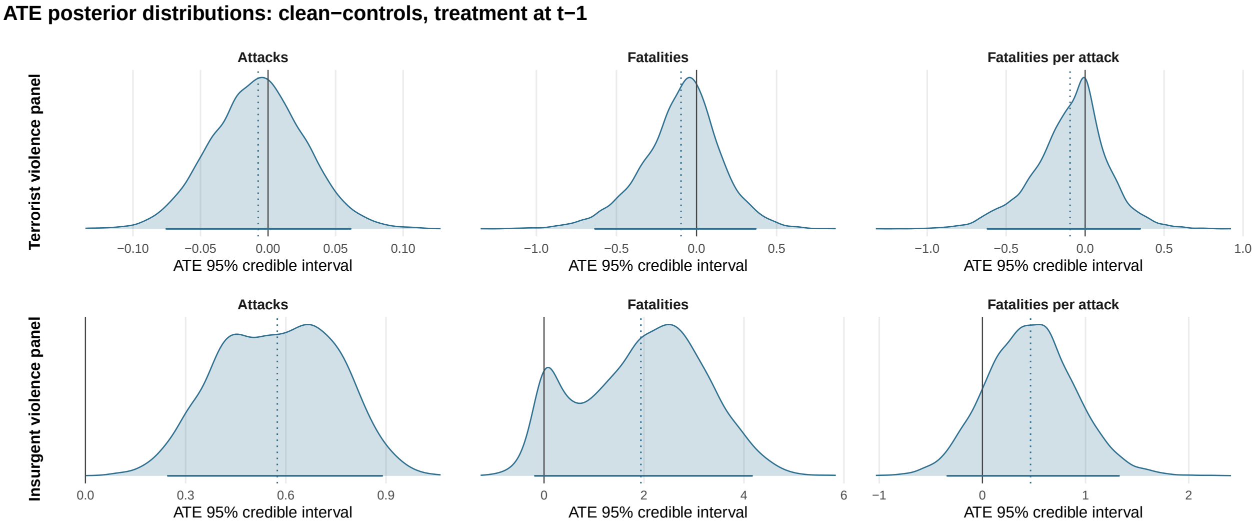

Figure 10: ATE posterior distributions under clean-controls robustness design

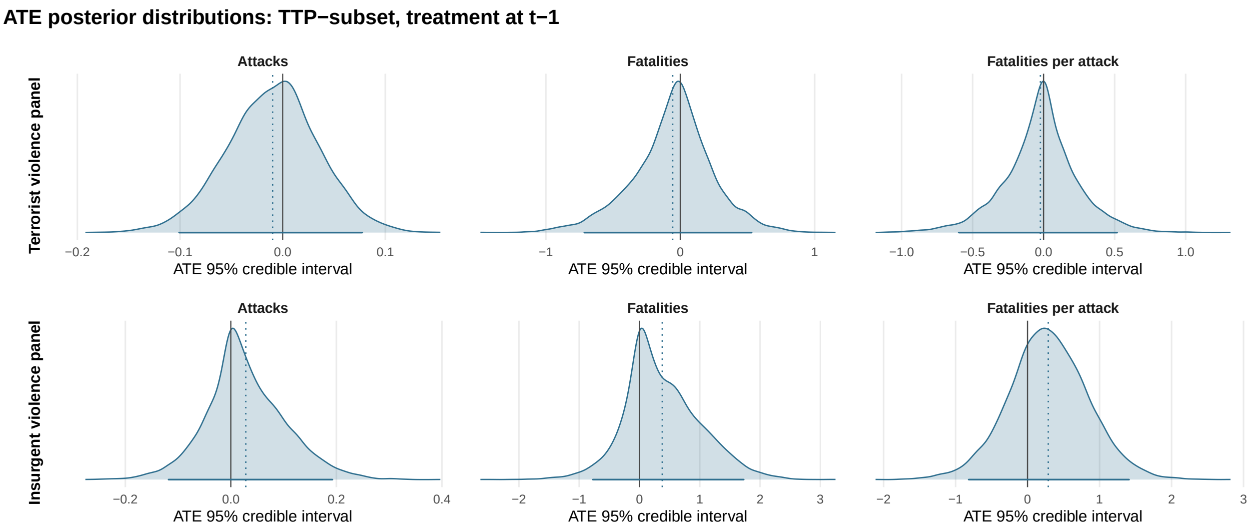

Figure 11: ATE posterior distributions under TTP-subset robustness design

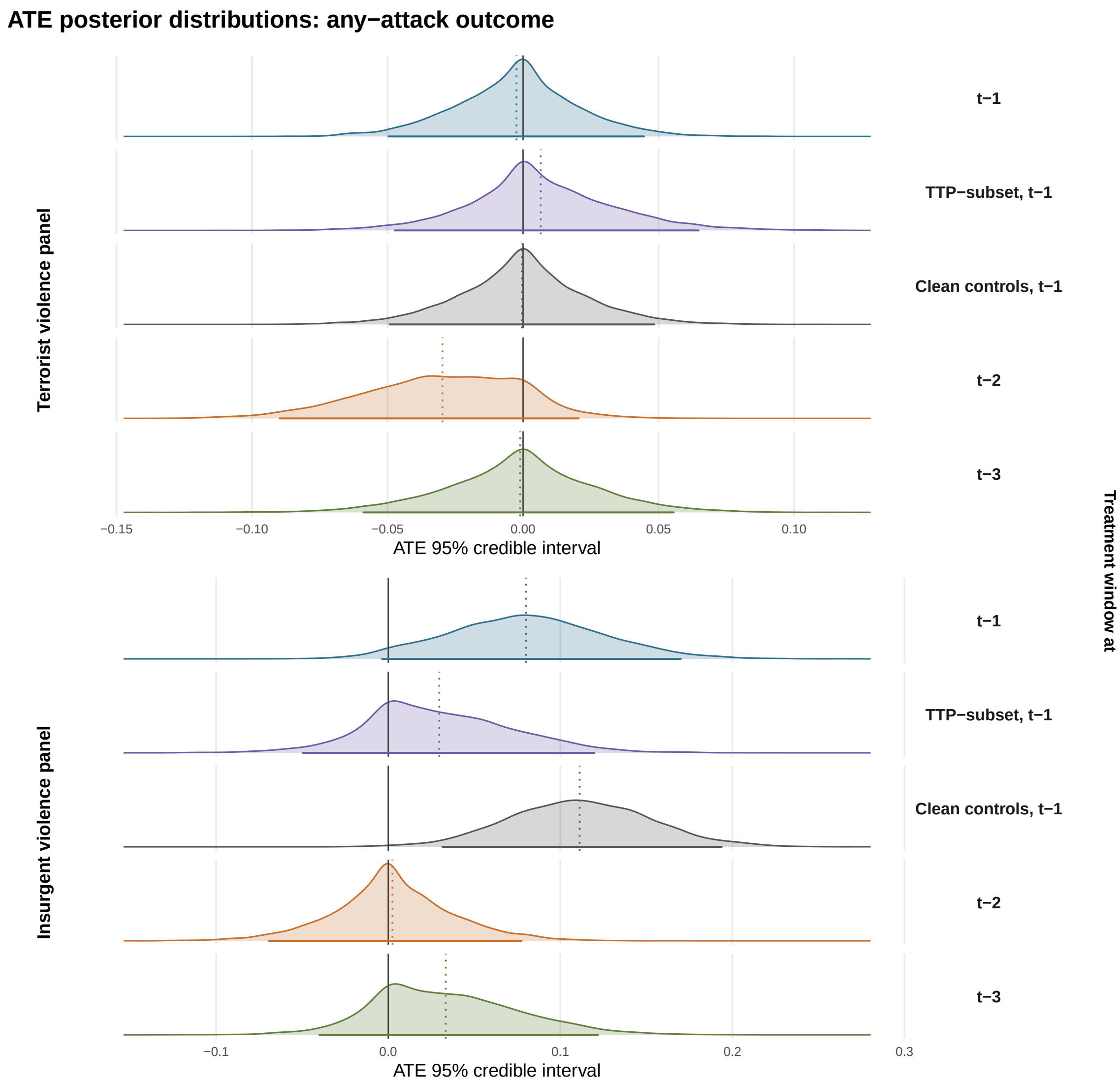

Figure 12: ATE posterior distributions for the any-attack outcome

Table 6: ATE estimates under clean-controls robustness design

Panel	Design	Outcome	Mean	Median	95% CrI	Pr($ATE < 0$)
Terrorist violence	Clean controls, t-1	Attacks	-0.007	-0.007	[-0.075, 0.061]	0.58
Terrorist violence	Clean controls, t-1	Fatalities	-0.097	-0.078	[-0.634, 0.370]	0.65
Terrorist violence	Clean controls, t-1	Fatalities per attack	-0.095	-0.071	[-0.619, 0.349]	0.65
Insurgent violence	Clean controls, t-1	Attacks	0.574	0.579	[0.247, 0.889]	0.00
Insurgent violence	Clean controls, t-1	Fatalities	1.938	2.049	[-0.184, 4.164]	0.07
Insurgent violence	Clean controls, t-1	Fatalities per attack	0.467	0.462	[-0.342, 1.325]	0.14

Table 7: ATE estimates under TTP-subset robustness design

Panel	Design	Outcome	Mean	Median	95% CrI	Pr($ATE < 0$)
Terrorist violence	TTP-subset, t-1	Attacks	-0.010	-0.009	[-0.101, 0.077]	0.58
Terrorist violence	TTP-subset, t-1	Fatalities	-0.057	-0.035	[-0.714, 0.531]	0.57
Terrorist violence	TTP-subset, t-1	Fatalities per attack	-0.023	-0.012	[-0.597, 0.518]	0.53
Insurgent violence	TTP-subset, t-1	Attacks	0.028	0.020	[-0.118, 0.192]	0.36
Insurgent violence	TTP-subset, t-1	Fatalities	0.379	0.288	[-0.773, 1.729]	0.27
Insurgent violence	TTP-subset, t-1	Fatalities per attack	0.287	0.277	[-0.817, 1.410]	0.29

Table 8: ATE estimates for any-attack outcome robustness checks

Panel	Design	Outcome	Mean	Median	95% CrI	Pr($ATE < 0$)
Terrorist violence	Main, t-1	Any attack	-0.002	-0.001	[-0.050, 0.045]	0.54
Terrorist violence	TTP-subset, t-1	Any attack	0.006	0.004	[-0.047, 0.065]	0.41
Terrorist violence	Clean controls, t-1	Any attack	-0.0005	-0.0002	[-0.049, 0.048]	0.51
Terrorist violence	Main, t-2	Any attack	-0.030	-0.028	[-0.090, 0.021]	0.85
Terrorist violence	Main, t-3	Any attack	-0.001	-0.0005	[-0.059, 0.056]	0.51
Insurgent violence	Main, t-1	Any attack	0.080	0.079	[-0.004, 0.170]	0.03
Insurgent violence	TTP-subset, t-1	Any attack	0.030	0.025	[-0.050, 0.120]	0.25
Insurgent violence	Clean controls, t-1	Any attack	0.111	0.111	[0.031, 0.194]	0.00
Insurgent violence	Main, t-2	Any attack	0.002	0.0008	[-0.069, 0.078]	0.48
Insurgent violence	Main, t-3	Any attack	0.033	0.029	[-0.040, 0.122]	0.21

E Data Construction and Variable Definitions

This appendix summarizes the construction of the extension panels and the variables used in the BCF analysis.

Table 9: Extension panel variables and coding rules

Variable	Coding rule	Source
Outcome: fatalities	Sum of fatalities in events attributed to the group-location-month. Panels A and C use violence against civilians; Panels B and D use all insurgent violence.	ACLED
Outcome: attacks	Count of events in the group-location-month.	ACLED
Outcome: fatalities per attack	Mean fatalities per attack in the group-location-month; set to zero when no attacks occur.	ACLED
Strike exposure	Binary indicator for whether the group was exposed to an air/drone strike in the relevant pre-outcome treatment window. Strike exposure is coded at the group-month level and merged into each group-location-month.	ACLED
Lagged attacks	Count of attacks by the same group in the same location in the prior month.	ACLED
Lagged target type	Baseline measure of the group's civilian-targeting profile, used to proxy prior target selection.	Global Terrorism Database
Military operations	Count of state military battle events in the same location-month.	ACLED
Location	Admin1 province or province-equivalent unit for Panels A and B; FATA agency for Panels C and D.	ACLED
Time	Month-year indicators for the panel month.	Constructed
Population density	Location-year population density.	WorldPop

Variable	Coding rule	Source
State capacity	Location-year nighttime lights measure, normalized and used as a proxy for state presence/capacity.	VIIRS nighttime lights
Group age	Years since group founding.	CISAC Mapping Militants Project; South Asia Terrorism Portal
Group type	Group classification used to distinguish religious and separatist organizations.	CISAC Mapping Militants Project; South Asia Terrorism Portal

The main sample includes groups with at least five relevant events in the source event sample. Events by unidentified armed actors are excluded because they cannot be assigned to a group-level panel. Political parties are also excluded from the all-insurgent-violence panels because the analysis focuses on non-state armed organizations rather than electoral parties or party-linked protest activity.