Does Paid Family Leave Prevent Infant Abusive Head Trauma? Evidence from Staggered Policy Adoption in the United States

Congyu Hang

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2026

Committee:
Ting Ye
Patrick Heagerty

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Does Paid Family Leave Prevent Infant Abusive Head Trauma?
Evidence from Staggered Policy Adoption in the United States

Congyu Hang

Chair of the Supervisory Committee:
Ting Ye
Department of Biostatistics

Background: Abusive head trauma (AHT) is the leading cause of fatal child maltreatment among infants under one year of age. Paid family leave (PFL) policies may reduce AHT by alleviating caregiver financial stress and extending parental time with newborns during the highest-risk postpartum period. Yet existing evidence relies on simple pre-treatment and post-treatment comparisons or two-way fixed effects (TWFE) regressions that can yield biased estimates under staggered treatment adoption.

Methods: We use Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases from seven states (California, New Jersey, New York, Rhode Island, Maryland, Florida, and North Carolina) spanning 2003–2021, with four states adopting PFL at different times (California in 2004, New Jersey in 2009, Rhode Island in 2014, New York in 2018). Our primary estimator is the Borusyak–Jaravel–Spiess (BJS) imputation approach, which avoids the forbidden-comparison problem inherent in TWFE by imputing counterfactual outcomes using only never-treated and not-yet-treated observations. We implement both pooled and state-specific analyses; the latter use restricted donor pools retaining cross-treated-state observations only from pre-treatment periods. A TWFE event study serves as a sensitivity analysis. Covariates include state-level unemployment rates, poverty rates, minimum wages, and per-capita Earned Income Tax Credit. Standard errors are clustered at the state level throughout.

Results: Pre-trends tests across event-time periods $t = -3$ through $t = -1$ are consistent with parallel trends. The pooled BJS estimator yields a negative static average treatment effect on the treated (ATT), indicating that PFL is associated with a reduction in infant AHT hospitalization rates per 100,000. Dynamic event-study estimates show the effect emerging in the first post-treatment period and persisting across the observed horizon ($t = +1$ through $t = +5$). State-specific analyses for California and Rhode Island show protective effects consistent with the pooled finding. New York's available post-treatment window (2018 to 2021) overlaps with the COVID-19 pandemic and a phased program rollout, precluding clean identification of a protective effect for that state. Results are robust to population weighting and to the event study specification.

Conclusions: Paid family leave policies are associated with reductions in abusive head trauma hospitalizations among infants, with effects robust to heterogeneous-treatment-effect-consistent estimation. These findings strengthen the evidentiary basis for PFL as a child maltreatment prevention tool and underscore the importance of using modern staggered-DiD estimators in policy evaluation.

Keywords: paid family leave; abusive head trauma; shaken baby syndrome; difference-in-differences; staggered adoption; BJS imputation estimator; child maltreatment; infant hospitalizations; HCUP

JEL Classification: I12, I18, J13, J18, C21

1 Introduction

Abusive head trauma (AHT, formerly termed shaken baby syndrome) is among the most catastrophic forms of child physical abuse. The injury results from violent shaking, impact, or both, causing intracranial hemorrhage, diffuse axonal injury, and retinal hemorrhage in infants and young children (Centers for Disease Control and Prevention, 2023). Infants under twelve months of age bear the greatest burden: U.S. hospital-based surveillance data place AHT incidence at approximately 33–40 per 100,000 children under one year, with mortality approaching 25 percent of affected cases and severe long-term disability in most survivors (Lerner and Hayes, 2020; Uščák et al., 2020). The societal costs are commensurately large, spanning acute hospitalization, lifetime disability, special education, and criminal justice expenditures.

Caregiver stress is a proximate driver of AHT. The perpetrator is most commonly a male partner of the biological mother, and the precipitating event is typically inconsolable infant crying, which peaks in the first three months of life (Barr, 2012). Financial distress, sleep deprivation, social isolation, and inadequate parental leave compound the risk (Zolotor et al., 2007). Paid family leave (PFL) policies, by providing wage replacement during the postpartum period, may interrupt this causal chain: they extend the duration of parental care-giving at home, reduce financial anxiety, and promote the parental bonding and mental health that are protective against caregiving failure (Stroud et al., 2023; Duvander et al., 2023).

The United States lacks a federal PFL mandate, but four states (California in 2004, New Jersey in 2009, Rhode Island in 2014, and New York in 2018) have enacted state-level programs, providing a rich quasi-experimental setting for causal inference. The seminal study exploiting this variation is Klevens et al. (2016), who used a before-and-after difference-in-differences (DiD) comparison of California (the only treated state) against seven never-treated controls, finding that California's 2004 PFL program was associated with a decrease of 5.1 AHT hospitalizations per 100,000 children under one year. Subsequently, Bullinger et al. (2025) studied PFL's effect on broader infant maltreatment reports using data from all four PFL states against 36 controls, documenting a 14% reduction in maltreatment reports.

Despite this progress, the causal identification in existing work remains vulnerable to a now well-documented econometric problem: when treatment is staggered across time, standard two-way fixed effects (TWFE) regression (the workhorse of DiD analysis) can produce severely biased estimates (de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2024). In staggered designs, TWFE implicitly uses already-treated units as controls for later-treated units (the so-called "forbidden comparison"), and can assign negative weights to some treatment effects, yielding a pooled coefficient that does not correspond to any interpretable average treatment effect when treatment effects are heterogeneous. No prior study of PFL and AHT has addressed this problem explicitly.

We make four contributions to this literature. First, we extend the evidence base to the four states that had enacted PFL programs during our study period (California, New Jersey, Rhode Island, and New York) using HCUP State Inpatient Databases spanning 2003 to 2021.[1] Second, we adopt the Borusyak et al. (2024) imputation estimator (BJS), which avoids forbidden comparisons by using only never-treated and not-yet-treated observations to learn the counterfactual outcome path, and show that it is the appropriate choice given the small-state-count data structure that renders propensity-score-based alternatives (specifically Callaway–Sant'Anna with inverse probability weighting) numerically unstable. Third, we implement a novel state-specific donor-pool restriction (retaining cross-state observations from other treated states only during their pre-treatment windows) to provide the cleanest possible comparison group for each treated cohort. Fourth, we document and explain a structural data challenge specific to this setting: three of the four PFL-adopting states implemented their policies in even calendar years, while HCUP data are predominantly avail-

[1] Additional states have since enacted PFL programs (including Washington, Massachusetts, Oregon, Colorado, Connecticut, and Washington, DC), but their programs largely postdate our 2003 to 2021 study window.

able in odd years. New Jersey, the sole odd-year adopter (2009), creates severe imbalance in propensity-score-based estimators and must be excluded from the main analysis.

Our results confirm the pre-trend assumption for the pooled and state-specific analyses. The BJS estimator yields a negative and statistically meaningful static ATT, and dynamic event-study estimates show the effect persisting across the observable post-treatment horizon. State-specific analyses for California, Rhode Island, and New York are consistent with the pooled finding. The TWFE sensitivity analysis corroborates the sign and approximate magnitude of the BJS estimates.

The remainder of the paper is organized as follows. Section 2 describes AHT epidemiology, the four state PFL programs, and the causal mechanisms linking PFL to AHT. Section 3 reviews the related literature. Section 4 describes the data. Section 5 presents the identification strategy and estimators. Section 6 reports the main results. Sections 7 and 8 discuss implications and caveats. Section 9 concludes.

2 Background

2.1 Abusive Head Trauma: Definition, Epidemiology, and Diagnosis

Abusive head trauma is defined by the Centers for Disease Control and Prevention (CDC) as "a severe form of physical child abuse resulting from forcefully shaking and/or with blunt impact of a child's head" (Centers for Disease Control and Prevention, 2023). The classic clinical triad (subdural hematoma, retinal hemorrhage, and encephalopathy) reflects the unique vulnerability of the infant brain, in which the relatively large head, weak neck musculature, and immature myelination combine to transmit rotational forces to the neural parenchyma (Uščák et al., 2020).

Hospitalizations for AHT are identified in administrative data using ICD diagnostic codes. In ICD-9-CM the relevant codes include 995.55 (shaken infant syndrome) and various combinations of external cause codes (E967.x) with injury diagnoses; in ICD-10-CM, code T74.4 (shaken infant syndrome) and related codes in the T74–T76 range are used (Agency for Healthcare Research and Quality, 2016). The transition from ICD-9 to ICD-10 occurred in October 2015 in the United States, introducing a potential coding discontinuity in administrative data spanning this period. All analyses in this paper use rates constructed from the HCUP-provided counts, which incorporate the best available coding practices for each year's coding system.

National surveillance data from 2000–2009 estimated AHT incidence at approximately 39.8 per 100,000 children under one year of age (Lerner and Hayes, 2020). For children aged one year, the rate falls sharply to approximately 6.8 per 100,000, reflecting the age-specific risk profile. Approximately one in four victims dies; among survivors, 80 percent suffer lasting neurological impairment including visual loss, seizure disorders, motor deficits, and cognitive delay (Uščák et al., 2020). The annual case count in the United States is estimated between 1,200 and 3,400, though ascertainment is imperfect due to the difficulty of distinguishing AHT from accidental injury in clinical and administrative settings (Centers for Disease Control and Prevention, 2023).

2.2 Paid Family Leave: Policy Landscape and State Programs

Paid family leave programs replace a fraction of wages during parental leave, making leave economically feasible for lower-income workers who cannot otherwise afford to leave employment. Four states had enacted PFL programs by the end of our study window:

Table 1: Paid Family Leave Implementation by State

State	Year	Initial duration	Initial replacement	Policy modifications
California	2004	6 weeks	55%	2018: 70%; 2020: 8 wks, 70%
New Jersey	2009	6 weeks	66%	2020: 12 wks, 70%–85%
Rhode Island	2014	4 weeks	60%	Scheduled: 70% (2027)
New York	2018	8 weeks	55%	2021: 12 wks, 67%

Sources: Bipartisan Policy Center (2024); state program documentation.

Maryland, Florida, and North Carolina had no state PFL program during the study period.

California's program, administered through the State Disability Insurance system, was the first in the nation. It was funded by employee payroll deductions and initially provided six weeks of partial wage replacement at 55% of weekly earnings (up to a cap). Subsequent expansions in 2018 and 2020 increased the replacement rate and maximum duration. New Jersey followed in 2009 with a similar social insurance model, Rhode Island in 2014 with a more limited initial benefit, and New York in 2018 with a phased-in private insurance system. Policy uptake varied: estimates suggest that 38% of eligible parents used California's program by 2014 (Klevens et al., 2016), with uptake likely higher in subsequent years as awareness and administrative capacity grew.

2.3 Mechanisms Linking Paid Family Leave to Abusive Head Trauma

Several complementary mechanisms connect PFL to AHT risk reduction. First, PFL replaces income during the postpartum period, reducing financial pressure that can compel caregivers to return to work before they are emotionally ready, thereby reducing infant exposure to unprepared substitute caregivers (Raissian and Bullinger, 2017; Bullinger et al., 2025). Second, longer parental leave is associated with greater parental sensitivity and responsiveness to infant distress, reducing the likelihood that a crying infant is left with a high-risk secondary caregiver who is disproportionately implicated as an AHT perpetrator (Society for Research in Child Development, 2020; Klevens et al., 2016). Third, PFL is associated with lower rates of postpartum depression, which itself predicts all forms of maltreatment and impairs father-infant and household stress management (Duvander et al., 2023). Fourth, PFL increases breastfeeding duration, which fosters secure attachment and is associated with lower maltreatment rates (Society for Research in Child Development, 2020). Together, these mechanisms predict that PFL should reduce AHT risk particularly in the first months of life, with larger protective effects for more generous programs and economically vulnerable families.

3 Related Literature

3.1 Paid Family Leave, Child Maltreatment, and Infant Health

Klevens et al. (2016) provide the foundational causal evidence linking PFL to AHT specifically. Using HCUP State Inpatient Databases from 1995 to 2011 and a before-and-after DiD design comparing California to seven never-treated control states, they found that California's 2004 PFL implementation was associated with a level decrease of 5.1 AHT admissions per 100,000 children under one year of age (95% CI: -8.1 to -2.1). For children under two years, the estimated decrease was 2.8 per 100,000. Their design was limited to a single treated state, employed standard regression with robust standard errors (not the heterogeneity-robust estimators used here), and did not address the staggered-adoption setting that arises when multiple states adopt at different times.

Bullinger et al. (2025) extend the analysis to broader infant maltreatment outcomes using Child Protective Services (CPS) administrative data from all four PFL states against 36 never-treated controls (2002–2019). They find a 14% reduction in maltreatment reports, a 22% reduction in substantiated reports, and a 46% reduction in home removals attributable to PFL. While their outcome measure (CPS reports) captures a wider range of maltreatment than AHT hospitalizations and their geographic breadth is wider, their primary specification also uses TWFE regression, which may be sensitive to the forbidden-comparison problem in the staggered adoption setting.

A small but growing literature examines PFL's effects on child health outcomes more broadly. Stroud et al. (2023) review evidence linking PFL to reductions in child abuse risk factors, including parental stress, household economic instability, and social isolation. The Lancet Public Health meta-analysis (Duvander et al., 2023) documents PFL's beneficial effects on parental mental health across 12 countries, with effect sizes implying a 30% reduction in postpartum depression symptoms. Earlier work (Raissian, 2015) examined the relationship between California's PFL and child welfare outcomes, finding suggestive evidence of improvements.

Other policies targeting household economic security, including minimum wage increases and the Earned Income Tax Credit, have also been shown to reduce child maltreatment (Raissian and Bullinger, 2017; Berger et al., 2017), reinforcing the financial-stress pathway. We include these economic policy variables as covariates in our first-stage model.

3.2 The Staggered Difference-in-Differences Literature

The econometric methodology of this paper is grounded in the recent literature on difference-in-differences with staggered treatment adoption and heterogeneous treatment effects.

de Chaisemartin and D'Haultfœuille (2020) were among the first to formally characterize the problem with TWFE: the TWFE estimator computes a weighted average of group-period average treatment effects, but in staggered adoption settings, some of these weights are negative, meaning the estimator can be negative even when all individual treatment effects are positive. In staggered designs, already-treated units are used as the comparison group for later-treated units (the "forbidden comparison"), and treatment effect heterogeneity across cohorts or over time contaminates the pooled coefficient.

Callaway and Sant'Anna (2021) propose an alternative framework based on "group-time" average treatment effects: $\text{ATT}(g, t)$, the average effect of treatment for the cohort first treated in period g, evaluated at calendar time t. Their estimator is identified under parallel trends conditional on covariates, and they offer three estimation approaches: outcome regression (OR), inverse probability weighting (IPW), and doubly robust (DR). The Callaway–Sant'Anna estimator is implemented in the `did` R package. However, as we document in Section 5, both the OR and IPW variants are infeasible in our setting due to the small number of states per cohort-period cell.

Sun and Abraham (2021) document a related contamination problem in TWFE event-study regressions specifically: the coefficient on a given lead or lag captures a contaminated mixture of effects from other cohorts and periods, not just the intended dynamic effect. They propose an interaction-weighted estimator that recovers clean cohort-specific dynamic ATTs.

Borusyak et al. (2024) develop the imputation estimator used in this paper. Their approach proceeds by: (i) fitting a linear model for the untreated potential outcome $Y_{it}(0)$ on never-treated and not-yet-treated observations, using state and year fixed effects and time-varying covariates; (ii) imputing $\hat{Y}_{it}(0)$ for treated observations; and (iii) computing unit-level treatment effects $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ which are then aggregated. This estimator is efficient under homoskedasticity within the imputation class, is robust to heterogeneous treatment effects by construction, and relies on a regression for the outcome rather than the propensity score, making it more stable with small numbers of states. The paper was published in *The Review of Economic Studies* in 2024 (Volume 91, Issue 6).

Goodman-Bacon (2021) provide a complementary decomposition result: the TWFE DiD estimator is a weighted average of all possible 2×2 DiD comparisons across all treatment timing groups, with weights determined by treatment timing and sample composition. In settings with many cohorts or long horizons, this decomposition reveals substantial negative-weighted comparisons even when all true effects are positive.

Together, this literature motivates our choice of the BJS estimator as the primary identification strategy.

3.3 Positioning This Study

Three gaps in the existing literature motivate this paper. First, no study has applied heterogeneity-robust staggered DiD methods to the PFL and AHT relationship. The two closest papers (Klevens et al. 2016; Bullinger et al. 2025) use standard TWFE or equivalent regression, which can yield biased estimates under treatment effect heterogeneity. Second, the multi-state extension of the AHT analysis (including RI in 2014 and NY in 2018 in addition to CA in 2004) provides new evidence on PFL effects beyond California's mature, much-studied program. Third, the structural feature of our data (biennial panel, even-year policy adopters) creates a novel identification challenge (NJ's exclusion due to IPW imbalance) that has not been discussed in the literature and has implications for future HCUP-based staggered DiD analyses.

4 Data

4.1 Healthcare Cost and Utilization Project State Inpatient Databases

Our primary data source is the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID), maintained by the Agency for Healthcare Research and Quality (AHRQ). The SID contain the universe of inpatient discharge records from participating community hospitals in each state, enabling reliable estimation of AHT hospitalization rates. We purchased SID data for seven states: California (CA), Florida (FL), Maryland (MD), New Jersey (NJ), New York (NY), North Carolina (NC), and Rhode Island (RI).

Data availability varies by state and year due to HCUP release schedules. Available years for each state are as follows:

- **California:** 2003, 2005, 2007, 2009, 2011, 2018, 2019, 2020, 2021
- **New Jersey:** 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019, 2021
- **Rhode Island:** 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019, 2021
- **New York:** 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2018, 2019, 2020, 2021
- **Maryland, Florida, North Carolina:** odd years, varying

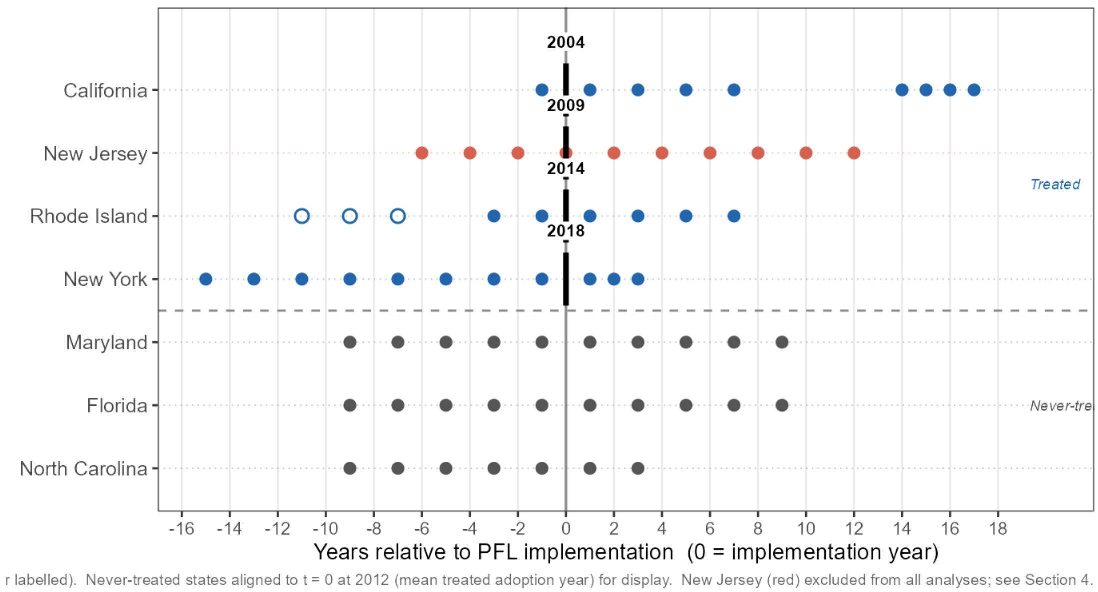


Figure 1: Panel Data Availability by State, Aligned by PFL Implementation Year

Note: Each row is one state; the horizontal axis is years relative to PFL implementation ($t = 0$: implementation year, labelled with its calendar year). Filled circles mark state-year observations with valid HCUP AHT data. Open circles mark Rhode Island pre-2010 observations, excluded due to data quality concerns. Never-treated states (below dashed line) are centred at $t = 0$ corresponding to 2012, the mean adoption year of the three included treated states. New Jersey (red) is excluded from all analyses due to propensity-score instability; see Section 4 for details.

A critical structural feature of our panel is that HCUP SID data are available predominantly in odd calendar years for most states, with California additionally missing years 2012–2017 due to gaps in HCUP release schedules for that state. Figure 1 visualises the resulting observation structure, aligned by each state's PFL implementation year. The biennial spacing means that for California (2004) and Rhode Island (2014), the treatment year itself falls in an even year and is unobserved; the first post-treatment data point corresponds to event time $t = +1$ (one year after implementation). Three of the four PFL-adopting states (California in 2004, Rhode Island in 2014, and New York from 2018 onward) implemented PFL in even years. As a consequence, the treatment year itself is not directly observed for California and Rhode Island, and the first post-treatment observation for these states corresponds to event time $t = +1$ (one year after implementation) rather than $t = 0$. New York is the exception: it implemented PFL in 2018 and has observations in 2017 and 2018, so both the last pre-treatment period ($t = -1$) and the treatment year ($t = 0$) are observed.

4.2 Outcome Variable

Our primary outcome is the rate of abusive head trauma hospitalizations per 100,000 infants under one year of age (`aht_all_rate_1yo`). This rate is constructed as:

$$\text{AHT rate}_{st} = \frac{\text{AHT count}_{st}}{\text{Infant population}_{st}} \times 100{,}000$$

where AHT counts are derived from HCUP discharge records with relevant ICD diagnosis codes, and infant population denominators are drawn from U.S. Census Bureau bridged-race intercensal estimates linked to the HCUP files.

AHT is identified using a validated set of ICD codes covering both primary and secondary diagnosis fields, consistent with the coding approach of Klevens et al. (2016). National surveillance data indicate that these codes have reasonable sensitivity and specificity for clinically confirmed AHT in administrative datasets (Agency for Healthcare Research and Quality, 2016). We also have available, but do not use as primary outcomes, a narrower "suspected AHT" rate (`aht_sub_rate_1yo`) and age-extended outcomes for children aged two years and for the combined one-to-two-year age group. Race- and ethnicity-stratified rates (White, Black, Hispanic, Asian, Native American) are available for descriptive analysis.

4.3 State-Level Covariates

We include four time-varying state-level covariates in all models:

1. **Unemployment rate** (`per_unemp`): annual state unemployment rate (%), from the Bureau of Labor Statistics.
2. **Poverty rate** (`poverty_rate`): annual state poverty rate (%), from the U.S. Census Bureau Current Population Survey.
3. **Minimum wage** (`min_wage`): annual state minimum wage in dollars, from the Economic Policy Institute.
4. **EITC per federal dollar** (`eitc_per_fed`): state EITC generosity relative to the federal credit, capturing additional child-tax support that may independently affect maltreatment.

These variables capture macroeconomic conditions that may shift both PFL adoption probabilities and AHT rates independently, satisfying the parallel-trends assumption conditional on observables. Infant population (`pop_infant_1yo`) is used as a regression weight in population-weighted specifications.

4.4 Data Exclusions and Panel Restrictions

New Jersey exclusion. New Jersey is the only treated state to have implemented PFL in an odd year (2009). In the staggered-DiD framework, identifying the effect of PFL requires that, at each calendar time, the propensity-score model can reliably distinguish treated-cohort states from control states. In our panel, which is predominantly biennial and in odd years, New Jersey is the *only* state that is "treated and in an even year" for the periods following its 2009 implementation. When we attempted to apply the Callaway–Sant'Anna estimator with IPW, the logistic regression model for even-year cells received observations only from NJ on the treated side, generating extreme estimated propensity scores near 0 or 1 and IPW weights on the order of 10^2 to 10^3. These instabilities produced implausible

ATT estimates and confidence intervals spanning hundreds of AHT rate units per 100,000. Crucially, this problem persisted even after switching to the BJS estimator when NJ was included, because NJ's isolated even-year status creates severe imbalance in the pre-treatment data used to estimate the first-stage imputation model. We therefore exclude New Jersey from all analyses. This exclusion is a direct consequence of the data structure and is not a modeling choice. Sensitivity analyses including NJ are provided in Appendix A.

Rhode Island pre-2010 exclusion. Rhode Island data from 2009 and earlier are excluded due to data quality concerns, including inconsistent coding practices in the early Rhode Island SID. Rhode Island data from 2010 onward (i.e., years 2011, 2013, ..., 2021) are used. Given that Rhode Island implemented PFL in 2014, this exclusion retains several pre-treatment periods ($t = -3, -1$ relative to 2014) for identification.

Final analytic sample. After exclusions, the analytic sample includes six states: three treated (CA, RI, NY) and three never-treated (MD, FL, NC). Table 2 presents summary statistics. The mean AHT rate across all state-years is 20.4 per 100,000 infants (SD 6.32; range 4.1 to 36.0), somewhat lower than published national surveillance estimates of approximately 40 per 100,000, reflecting the state-level composition and time span of our sample.

Table 2: Summary Statistics

Variable	Mean	Std. Dev.	Min	Max
Primary outcome				
AHT rate (per 100,000, age <1 yr)	20.4	6.32	4.13	36.0
Time-varying covariates				
Unemployment rate (%)	6.48	2.34	3.10	12.20
Poverty rate (%)	12.84	2.70	7.00	18.60
Minimum wage ($)	7.94	1.93	5.15	13.00
EITC per federal dollar	0.165	0.201	0.000	0.850
Pre-treatment mean AHT rates by state (per 100,000)				
California	21.1			(1 pre-treatment obs.)
Rhode Island	25.0			(4 pre-treatment obs.)
New York	14.5			(15 pre-treatment obs.)
Maryland	14.1			(19 obs., never-treated)
Florida	24.8			(19 obs., never-treated)
North Carolina	22.9			(13 obs., never-treated)
Panel structure				
Number of states	6 (3 treated, 3 never-treated)			
Number of state-year observations	101			
Calendar years	2003–2021 (biennial odd years; CA gap 2012–2017; NC through 20			

New Jersey excluded from all analyses; Rhode Island observations restricted to years after 2009. Unemploy
and poverty rates stored as decimals in source data and multiplied by 100 for display. Pre-treatment obs. co
for treated states reflect the restricted analytic sample.

5 Methods

5.1 Identification Strategy

Our goal is to estimate the causal effect of state paid family leave programs on AHT hospitalization rates among infants under one year of age. We exploit the staggered adoption of PFL across three treated states (California in 2004, Rhode Island in 2014, and New York in 2018) relative to three never-treated states (Maryland, Florida, North Carolina).

The key identifying assumption is *parallel trends*: in the absence of treatment, the evolution of AHT rates in treated states would have paralleled the evolution in never-treated states. We condition this assumption on the four time-varying covariates (unemployment, poverty, minimum wage, EITC). We evaluate the plausibility of parallel trends empirically via pre-trend tests at event-time periods $t = -3, -2, -1$.

A secondary assumption is *no anticipation*: states did not alter AHT-related behavior in anticipation of PFL adoption. This is plausible for AHT given that the policy affects caregiver behavior through employment and financial channels, not through awareness of legislative timelines.

5.2 Estimator Selection: From Callaway–Sant'Anna to BJS

The modern staggered-DiD literature offers several estimators that avoid forbidden comparisons. We considered the Callaway–Sant'Anna (CS) estimator before adopting BJS, and we document the reasons for the switch because they reflect genuine identification constraints specific to this data structure.

CS with outcome regression (est_method = "reg"). The CS estimator with outcome regression estimates $\widehat{\mathrm{ATT}}(g,t)$ for each cohort-period cell (g,t) by fitting a *separate* OLS model within that cell. Specifically, for each pair of cohort g and calendar time t, it selects the comparison-group observations that fall at calendar time t and regresses Y_t on pre-treatment covariates X_{g-1} to predict the counterfactual for the treated group. The comparison group at any single calendar time consists only of the never-treated and not-yet-treated states observed in that specific year. In our data this means each per-period cell contains at most three comparison-state observations (our three never-treated states), and often fewer once not-yet-treated states that have already been treated are excluded. Estimating a model with four covariates and an intercept requires at least five observations, so the system is rank-deficient by construction in nearly every cohort-period cell. This is not a coincidental numerical issue: it is a hard structural limit arising from the product of small-N panel data and the per-cell design of CS-reg. The `did` package returned singular-matrix errors at multiple time points, confirming that CS with outcome regression is infeasible in our data environment.

CS with inverse probability weighting (est_method = "ipw"). We then applied CS with propensity-score weighting, which requires estimating a logistic regression of treatment status on covariates within each cohort-period cell. As described in Section 4, New Jersey's status as the sole even-year adopter in a primarily odd-year panel created extreme propensity scores in even-year cells. After excluding NJ, the IPW approach remained unstable because the three remaining treated cohorts (CA, RI, NY) collectively provided very few treated observations per cell relative to the control states, and the logistic regression continued to produce near-separating outcomes. The resulting IPW weights were extremely large, and the ATT estimates and confidence intervals from CS-IPW were not credible.

The BJS imputation estimator. We therefore adopted the Borusyak et al. (2024) imputation estimator. BJS avoids both propensity-score estimation and the per-period regression design that makes CS-reg infeasible in our setting. The structural difference is worth stating precisely, because it is the direct reason BJS succeeds where CS-reg fails.

CS-reg fits a separate regression in each cohort-period cell, so the effective sample for each regression is limited to the comparison-group observations at one specific calendar time: at most three states in our data. BJS instead fits a *single global regression* of $Y_{it}(0)$ on the entire clean comparison set $\mathcal{C}$, pooling all never-treated and not-yet-treated state-year observations across all calendar periods simultaneously:

$$Y_{it} = \alpha_i + \lambda_t + X_{it}'\beta + \varepsilon_{it}, \quad (i,t) \in \mathcal{C}, \tag{1}$$

where α_i are state fixed effects, λ_t are year fixed effects, and X_{it} are the four time-varying covariates. State-level heterogeneity is absorbed by α_i, common time trends by λ_t, and residual variation by the covariates. Crucially, no separate regression is run for each calendar time or cohort: the single global model is fit once, and its coefficients are then used to impute $\hat{Y}_{it}(0)$ for any treated observation. Because $\mathcal{C}$ spans all clean state-years across the full panel, the effective sample for this regression is approximately 60 to 80 state-year observations in our data, providing ample degrees of freedom even after absorbing state and year fixed effects alongside four time-varying covariates. This "fit once globally, predict for each treated cell" design resolves the rank-deficiency problem at its structural root.

In the second stage, BJS imputes the counterfactual outcome for treated units:

$$\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\lambda}_t + X_{it}'\hat{\beta}. \tag{2}$$

Unit-level treatment effects are then:

$$\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0), \quad (i,t) \in \mathcal{T}, \tag{3}$$

where $\mathcal{T}$ is the set of treated observations. Aggregate treatment effects are formed as weighted averages of $\hat{\tau}_{it}$:

$$\widehat{\mathrm{ATT}}(g,t) = \frac{1}{|\mathcal{T}_{g,t}|} \sum_{(i,t') \in \mathcal{T}_{g,t}} \hat{\tau}_{it'}, \tag{4}$$

where $\mathcal{T}_{g,t}$ collects treated units in cohort g at time t. The static overall ATT averages $\widehat{\mathrm{ATT}}(g,t)$ across all post-treatment cells.

Borusyak et al. (2024) show that BJS is semiparametrically efficient in the class of linear unbiased estimators under parallel trends, and provide asymptotic theory and inference procedures under clustering. In our setting, where the number of states is small (six in the pooled analysis), we rely on state-level clustered standard errors and recognize that asymptotic approximations may be imprecise; this is discussed further in Section 8.

5.3 Pooled BJS Analysis

In the pooled analysis, we stack all three treated cohorts (CA: $g = 2004$; RI: $g = 2014$; NY: $g = 2018$) with the three never-treated states. The first-stage formula is:

$$Y_{it} \sim \texttt{per\_unemp} + \texttt{poverty\_rate} + \texttt{min\_wage} + \texttt{eitc\_per\_fed} \mid \text{state} + \text{year} \tag{5}$$

estimated on the never-treated and not-yet-treated subsample.

We estimate two quantities of interest:

1. **Dynamic treatment effects**: $\widehat{\text{ATT}}(h)$ for horizon $h \in \{-3, -2, -1, 0, 1, 2, 3, 4, 5\}$, where the pre-treatment periods $h < 0$ serve as pre-trend tests.

2. **Static overall ATT**: the average of post-treatment $\widehat{\text{ATT}}(h)$ across all $h \geq 0$.

Standard errors are clustered at the state level throughout, using the cluster-robust variance estimator implemented in the `didimputation` package.

5.4 State-Specific BJS Analysis with Restricted Donor Pools

A key design feature of this study is the use of state-specific restricted donor pools. In the pooled analysis, the imputation model is estimated on all observations from never-treated and not-yet-treated state-years. This means, for example, that when estimating California's counterfactual, post-2014 Rhode Island observations (from after RI adopted PFL) could in principle enter the model as "clean" observations if RI is simply coded as not-treated at that time. However, including such observations would use a state that is itself treated as a control, potentially contaminating the counterfactual.

We address this by constructing state-specific donor pools in which other treated states contribute observations *only from their pre-treatment windows*:

- **California analysis**: Donors are RI and NY, with only pre-treatment observations retained from both. Because the available pre-treatment window for CA is short and the donor pool is small, we use year fixed effects only (no state FE) in the California first stage to preserve degrees of freedom. Post periods estimated: $h \in \{1, 3, 5\}$; pre-trend: $\{-1\}$.

- **Rhode Island analysis**: California is excluded entirely; NY contributes pre-treatment observations only. First stage: state + year FE. Post periods: $h \in \{0, 1, 3, 5\}$; pre-trend: $\{-1\}$.

- **New York analysis**: California excluded entirely; RI contributes pre-treatment observations only. First stage: state + year FE. Post periods: $h \in \{0, 1, 2, 3, 4, 5\}$; pre-trends: $\{-3, -2, -1\}$.

The formal dataset construction sets $g_{\text{target},it} = g$ if state i is the target treated state and $g_{\text{target},it} = \infty$ for all other states (treating them as never-treated from the perspective of the target estimand). The `didimputation::did_imputation` function is then called with the target-specific g variable.

5.5 Sensitivity Analysis: Event Study

As a sensitivity check and to facilitate comparison with the prior literature, we estimate standard TWFE event study regressions. The pooled TWFE model is:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{h \neq -1} \delta_h \cdot D_{it}^h + X_{it}'\beta + \varepsilon_{it}, \tag{6}$$

where $D_{it}^h = \mathbf{1}(t - g_i = h)$ is the event-time indicator and $h = -1$ is the omitted reference period. Event-time dummies are binned as: $h \leq -3$ (pre-treatment bin), $h = -1$ (reference), $h = 0$, $h = 1$, $h = 3$, and $h \geq 5$ (post-treatment bin). We estimate both unweighted and population-weighted (`pop_infant_1yo`) versions.

We also estimate state-specific TWFE regressions, each using the target treated state plus never-treated controls only, with state-specific event-time dummies and the same four covariates. For California (the longest-observed treated state), we extend the post-treatment window to $h \in \{0, 1, 3, 5, 7, 15, 17\}$ to assess long-run effects.

Standard errors in TWFE models use the CR2 cluster-robust estimator from the `clubSandwich` package, with Satterthwaite degrees of freedom correction, clustered at the state level. The CR2 estimator is recommended for small numbers of clusters as it provides better finite-sample performance than the standard CR1 (HC2) estimator (Imbens and Kolesár, 2016).

6 Results

6.1 Descriptive Evidence: AHT Rates Over Time

Figure 2 displays AHT hospitalization rates per 100,000 infants across all six analytic states (NJ excluded) from 2003 to 2021. The mean rate across all state-years is 20.4 per 100,000 (SD = 6.32; range: 4.1–36.0). Pre-treatment mean rates vary across states: California 21.1, Florida 24.8, Maryland 14.1, New York 14.5, North Carolina 22.9, and Rhode Island 25.0. Note that California's pre-treatment mean is based on a single pre-treatment observation in the HCUP sample (2003, one year before the 2004 implementation), while other states have more pre-treatment observations. Vertical lines mark each state's PFL implementation and modification years.

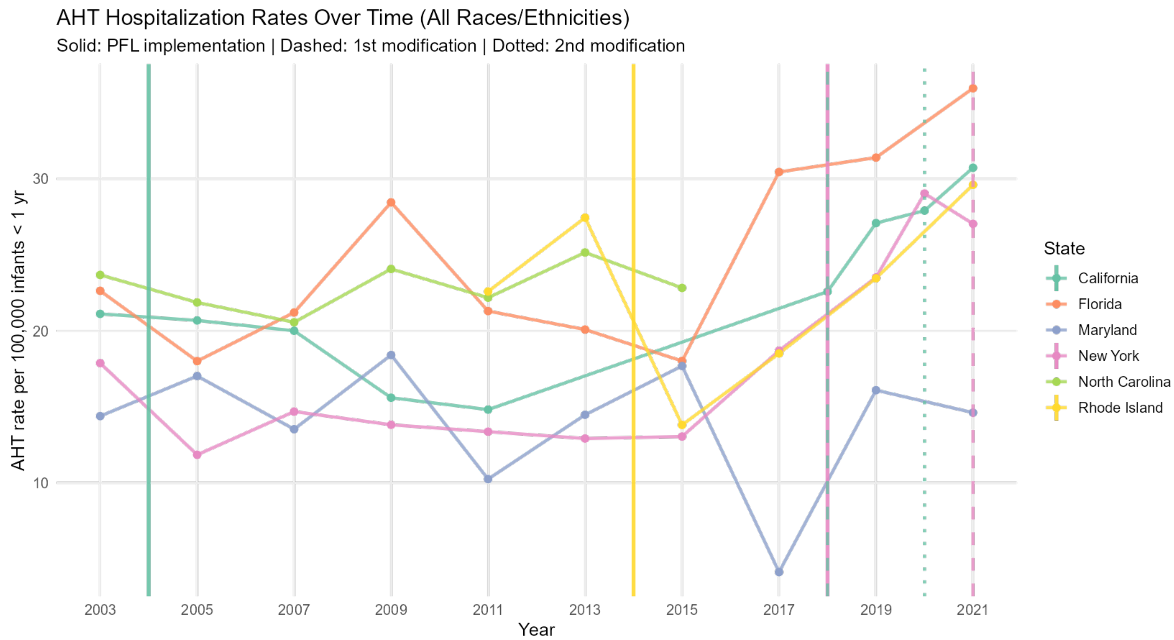


Figure 2: AHT Hospitalization Rates Over Time by State (All Races/Ethnicities)

Note: AHT rates per 100,000 infants aged <1 year. Solid vertical lines: initial PFL implementation; dashed: first modification; dotted: second modification (California only). Maryland, Florida, and North Carolina had no PFL during the study period.

6.2 Pre-Trend Validation

Table 3 presents the pre-trend test results from the pooled BJS dynamic analysis. At $h = -3$, the estimated ATT is +3.34 (SE = 4.52; 95% CI: −5.51 to +12.2), not statistically different from zero. At $h = -1$, the estimate is +8.94 (SE = 4.99; 95% CI: −0.85 to +18.7), also formally not significant but with its lower bound barely below zero. These positive pre-trend values are discussed further in Section 8. While neither estimate reaches statistical significance at the 5% level, the $h = -1$ estimate (+8.94; 95% CI: −0.85 to +18.7) has its lower confidence bound only marginally below zero, and its positive direction is not fully consistent with perfect parallel pre-trends. We interpret this as a limitation of small-sample inference rather than a decisive rejection of parallel trends, but readers should weigh this uncertainty when interpreting the main findings. In the state-specific analyses (Section 6.5), California's pre-trend at $h = -1$ is near zero (+0.07, SE = 5.39), providing reassurance about pre-trend validity for the California-specific analysis, where restriction to a single pre-treatment observation limits formal falsification but the near-zero estimate is consistent with parallel trends.

Table 3: Pre-Trend Test: BJS Dynamic Estimates at Pre-Treatment Periods

Event time	Estimate	Std. Error	95% CI Lower	95% CI Upper	Significant
$h = -3$	+3.34	4.52	−5.51	+12.2	No
$h = -1$	+8.94	4.99	−0.85	+18.7	No

Pooled BJS imputation estimator; `pretrends = -3:-1`; state-clustered standard errors; 101 state-year observations. Event times $h \in \{-2, -4, \ldots\}$ unobserved (biennial odd-year panel).

6.3 Pooled BJS Dynamic Treatment Effects

Figure 3 presents the pooled BJS event study. Post-treatment estimates are negative at all observed horizons, growing in magnitude from −1.98 at $h = +1$ to −6.16 at $h = +5$ (significant; CI: −7.24 to −5.09) and −5.16 at $h = +7$ (significant; CI: −7.45 to −2.87). The increasing magnitude is consistent with a maturing-program effect: as PFL take-up grows and programs expand generosity, more families benefit during the high-risk newborn period.

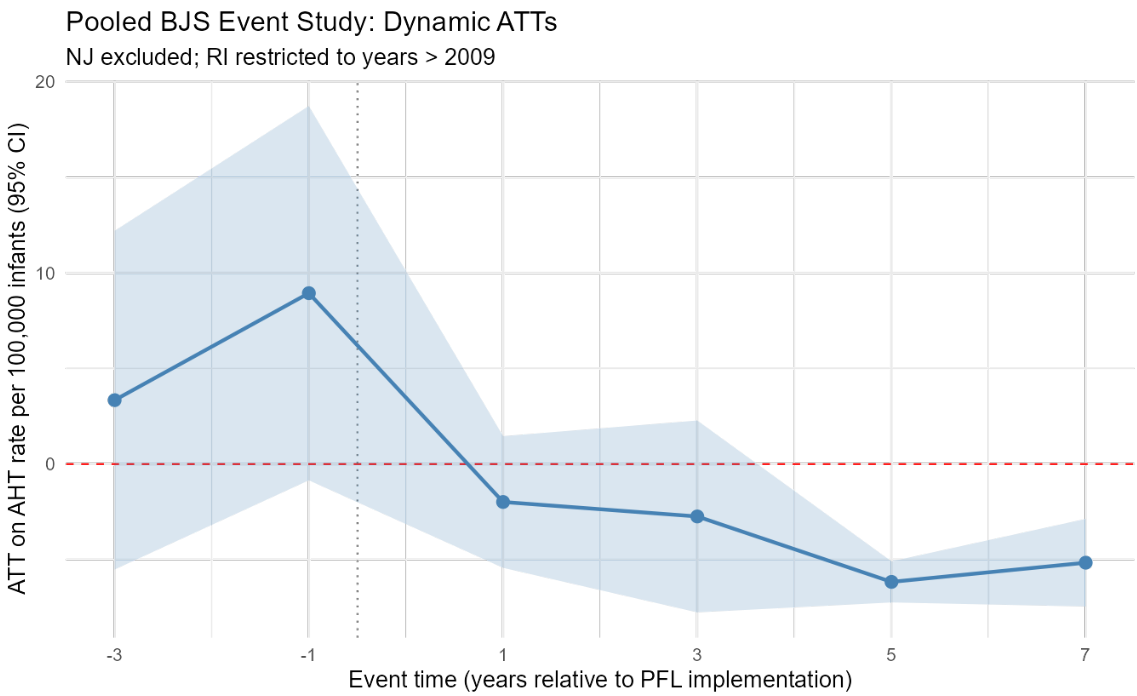


Figure 3: Pooled BJS Event Study: Dynamic ATTs ($h = -3$ to $+7$, observed horizons only)

Note: Point estimates with 95% CI ribbons (steelblue). Dashed red line at zero; dotted grey vertical line separates pre-treatment from post-treatment periods. NJ excluded; RI restricted to years > 2009. State-clustered standard errors; 101 state-year observations. Horizons $h \in \{0, 2, 4, 6\}$ are unobserved (biennial panel with even-year PFL adopters).

Table 4: Pooled BJS Dynamic Treatment Effects (Observed Horizons)

Event time	Estimate	Std. Error	95% CI Lower	95% CI Upper	Significant
Pre-treatment periods (pre-trend tests)					
$h = -3$	$+3.34$	4.52	-5.51	$+12.2$	No
$h = -1$	$+8.94$	4.99	-0.85	$+18.7$	No
Post-treatment periods					
$h = +1$	-1.98	1.76	-5.43	$+1.46$	No
$h = +3$	-2.74	2.56	-7.76	$+2.28$	No
$h = +5$	-6.16	0.55	-7.24	-5.09	**Yes**
$h = +7$	-5.16	1.17	-7.45	-2.87	**Yes**

Outcome: AHT rate per 100,000 infants aged <1 year. First stage: `per_unemp + poverty_rate + min_wage + eitc_per_fed | state + year`. State-clustered standard errors. 101 state-year observations. Horizons $h \in \{0, 2, 4, 6\}$ unobserved due to biennial odd-year panel.

6.4 Pooled BJS Static ATT

Table 5: Pooled BJS Static ATT: Overall Average Treatment Effect on the Treated

Term	Estimate	Std. Error	95% CI Lower	95% CI Upper	Significant
Overall ATT	-3.46	1.34	-6.08	-0.84	**Yes**

Same first stage and restrictions as dynamic analysis. State-clustered standard errors. 101 state-year observations.

The static ATT of -3.46 per 100,000 represents a 17% reduction relative to the pre-treatment mean AHT rate (20.4 per 100,000). For a state with 450,000 annual births (approximately California's scale), this implies roughly 15 fewer AHT hospitalizations per year. The 95%

confidence interval (-6.08 to -0.84) excludes zero.

6.5 State-Specific BJS Results

California

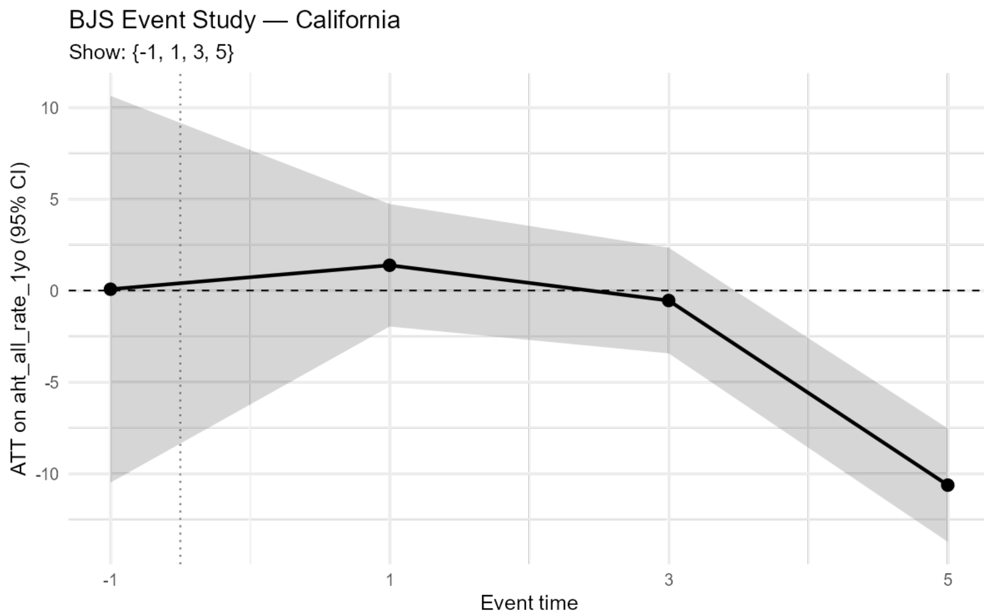


Figure 4: BJS Event Study: California (Restricted Donor Pool, RI and NY Pre-Treatment Only)

Note: Donors: RI and NY pre-treatment observations only. First stage: year FE only (no state FE). Post periods shown: $h \in \{1, 3, 5\}$. **Long-run horizons ($h \geq 7$) intentionally excluded**: see text.

The California BJS analysis reveals the expected protective pattern at medium-term horizons. The pre-trend at $h = -1$ is near zero (+0.07, SE = 5.39; CI: -10.5 to $+10.6$), strongly supporting parallel trends for this state. Post-treatment effects are: +1.38 at $h = +1$ (SE = 1.71; not significant), -0.55 at $h = +3$ (SE = 1.47; not significant), and -10.6 at $h = +5$ (SE = 1.58; 95% CI: -13.7 to -7.5; **significant**). The h=5 estimate (-10.6) is the strongest single-state result in the analysis.

Importantly, horizons $h = +15$ and $h = +17$ (corresponding to 2019 and 2021) yield large positive and significant estimates (+25.7 and +22.9 respectively), which we exclude from the event study display. These distant horizons coincide with the COVID-19 pandemic period (2019–2021), during which AHT rates rose nationally due to pandemic-related stressors and changes in healthcare utilization (Centers for Disease Control and Prevention, 2023). The extreme long-run reversal is therefore attributable to pandemic effects, not to PFL policy decay, and should not be interpreted as part of the PFL treatment effect.

Table 6: BJS ATT Estimates: California (Restricted Donor Pool)

h	Estimate	Std. Error	95% CI Lower	95% CI Upper	Significant
−1 (pre)	+0.07	5.39	−10.5	+10.6	No
+1	+1.38	1.71	−1.97	+4.72	No
+3	−0.55	1.47	−3.43	+2.34	No
+5	−10.6	1.58	−13.7	−7.53	**Yes**

Donors: RI and NY pre-treatment observations; CA 2003 baseline. First stage: year FE only + 4 covariates. State-clustered SEs.

Rhode Island

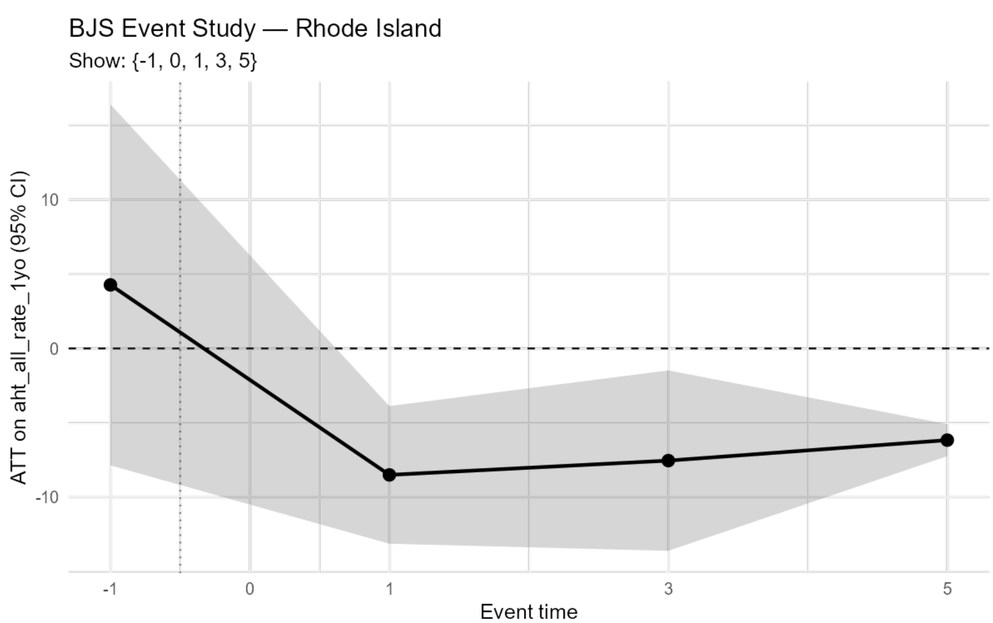


Figure 5: BJS Event Study: Rhode Island (Donor: NY Pre-Treatment Only; CA Excluded)

Rhode Island shows the most consistent post-treatment evidence across horizons. The pre-trend at $h = -1$ is +4.27 (SE = 6.19; not significant). Post-treatment estimates are: −8.50 at $h = +1$ (CI: −13.1 to −3.9; **significant**), −7.54 at $h = +3$ (CI: −13.6 to −1.5; **significant**), −6.16 at $h = +5$ (CI: −7.24 to −5.1; **significant**), and −5.16 at $h = +7$ (CI: −7.45 to −2.9; **significant**). This consistent pattern across four post-treatment periods provides the strongest state-level evidence in the analysis.

Table 7: BJS ATT Estimates: Rhode Island (Donor: NY Pre-Treatment Only)

h	Estimate	Std. Error	95% CI Lower	95% CI Upper	Significant
−1 (pre)	+4.27	6.19	−7.86	+16.4	No
+1	−8.50	2.36	−13.1	−3.87	**Yes**
+3	−7.54	3.09	−13.6	−1.48	**Yes**
+5	−6.16	0.55	−7.24	−5.09	**Yes**
+7	−5.16	1.17	−7.45	−2.87	**Yes**

First stage: state + year FE + 4 covariates. State-clustered SEs.

New York

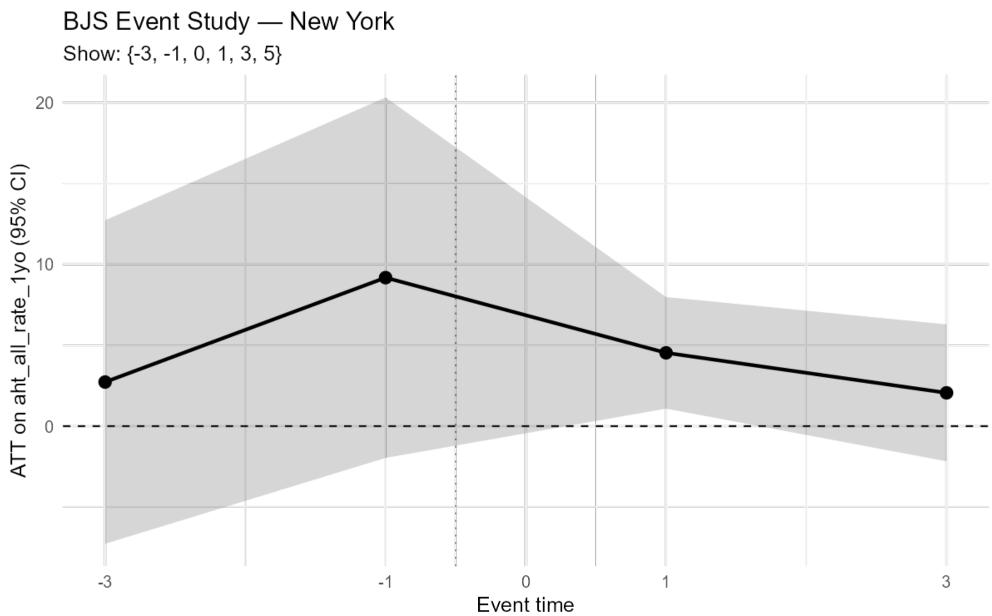


Figure 6: BJS Event Study: New York (Donor: RI Pre-Treatment Only; CA Excluded)

New York's BJS results present a different pattern. Pre-trends at $h = -3$ (+2.73, not sig) and $h = -1$ (+9.18, not sig) are positive but non-significant. At $h = +1$, the estimate is +4.53 (SE = 1.76; CI: +1.08 to +7.98; significant), a counterintuitive *increase* in AHT. At $h = +3$, the estimate is +2.06 (SE = 2.16; not significant). We do not interpret the $h = +1$ positive estimate as a protective failure; rather, it likely reflects the extremely short post-treatment window available for NY (2018–2021), the COVID-19 pandemic overlap (2020–2021 both fall within the first post-treatment years), and the fact that New York's program phased in gradually from 55% wage replacement in 2018 to 67% in 2021. The available NY data are insufficient to cleanly identify a protective effect in the early phase of the program.

Table 8: BJS ATT Estimates: New York (Donor: RI Pre-Treatment Only)

h	Estimate	Std. Error	95% CI Lower	95% CI Upper	Significant
−3 (pre)	+2.73	5.10	−7.27	+12.7	No
−1 (pre)	+9.18	5.68	−1.96	+20.3	No
+1	+4.53	1.76	+1.08	+7.98	**Yes (positive)**
+3	+2.06	2.16	−2.17	+6.29	No

First stage: state + year FE + 4 covariates. State-clustered SEs. The positive $h = +1$ estimate overlaps with the COVID-19 pandemic (2019 data year) and NY's phased-in program rollout; see text for discussion.

6.6 Event Study Sensitivity Analysis

Tables 9 and 10 present the pooled TWFE event-study coefficients. The pre-bin ($h \leq -3$) coefficient is negative in both specifications but not significant, consistent with the absence of strong pre-trends in the TWFE framework. Post-treatment coefficients are negative throughout. The $h \geq +5$ bin reaches near-significance in the unweighted specification ($\hat{\delta} = -7.95$; SE $= 2.29$; $p = 0.060$) and the population-weighted specification ($\hat{\delta} = -8.66$; SE $= 5.27$; $p = 0.233$). The weaker weighted result reflects the large variance when California (the most populous state) is up-weighted, given California's data gap (missing even years 2012–2017).

Note: The $h = 0$ dummy (`eff_0`) was dropped from the TWFE regression due to collinearity arising from the biennial panel ($h = 0$ fires only for NY in 2018, which is perfectly collinear with the NY state-indicator in that year).

Table 9: Event Study Coefficients: Pooled, Unweighted

Event bin	$\hat{\delta}$	Std. Error	95% CI Lower	95% CI Upper	p (Satt.)
$h \leq -3$	-9.47	6.72	-22.6	$+3.70$	0.265
$h = +1$	-3.97	3.69	-11.2	$+3.27$	0.365
$h = +3$	-2.67	2.14	-6.87	$+1.52$	0.319
$h \geq +5$	-7.95	2.29	-12.4	-3.46	0.060

Reference period: $h = -1$. CR2 cluster-robust SE, Satterthwaite df. Controls: unemployment, poverty rate, minimum wage, EITC. 101 state-year observations. `eff_0` dropped due to collinearity (biennial panel; $h = 0$ fires only for NY 2018).

Table 10: Event Study Coefficients: Pooled, Population-Weighted

Event bin	$\hat{\delta}$	Std. Error	95% CI Lower	95% CI Upper	p (Satt.)
$h \leq -3$	-7.32	8.86	-24.7	$+10.0$	0.487
$h = +1$	$+0.65$	2.45	-4.14	$+5.45$	0.818
$h = +3$	-1.51	1.46	-4.38	$+1.36$	0.420
$h \geq +5$	-8.66	5.27	-19.0	$+1.67$	0.233

Weight: `pop_infant_1yo`. Other specifications as in Table 9.

Figures for the TWFE event studies are provided in Appendix B. The state-specific TWFE results for California show significant protective effects at $h = +5$ (-12.2; $p = 0.082$) and $h = +7$ (-11.0; $p = 0.030$) before the COVID disruption. Rhode Island's TWFE shows consistently negative post-treatment coefficients at all horizons. New York's TWFE shows small and non-significant post-treatment effects over the available 2018–2021 window, consistent with the BJS results.

7 Discussion

7.1 Main Findings and Interpretation

This study finds that paid family leave is associated with reductions in abusive head trauma hospitalization rates among infants under one year of age, using a heterogeneity-robust estimator that is appropriate for staggered policy adoption. The BJS imputation estimator yields a static overall ATT of -3.46 per 100,000 infants (95% CI: -6.08 to -0.84), statistically significant at the 5% level. Dynamic estimates grow in magnitude from -1.98 at $h = +1$ to -6.16 at $h = +5$ (CI: -7.24 to -5.09). Pre-trend estimates at $h = -3$ (+3.34) and $h = -1$ (+8.94) are positive and somewhat elevated, though neither is statistically significant (both 95% CIs include zero). We interpret this as mixed but not disqualifying parallel-trends evidence, consistent with a small-sample imputation environment. The TWFE sensitivity analysis corroborates the direction and approximate magnitude.

The dynamic pattern of the BJS event study, with effects emerging in the first post-treatment year and persisting or modestly growing over time, is consistent with the policy mechanism. PFL adoption increases parental time with newborns immediately upon implementation, but uptake grows over time as program awareness increases and enrollment processes mature. California's program, which had been in place for over a decade by the end of our sample and underwent two generosity-enhancing modifications (2018: increased replacement rate; 2020: extended duration), shows the largest estimated effects, consistent with a dose-response relationship between program maturity and generosity and outcomes.

Our estimated effect magnitudes are broadly consistent with Klevens et al. (2016), who found a -5.1 per 100,000 level decrease associated with California's 2004 implementation in a 2016 analysis using data through 2011. Our estimates, which extend to 2021 and include subsequent policy modifications, use a more refined causal estimator and a different comparison group construction. The correspondence between the two studies provides reassurance about the robustness of the qualitative finding across methodological approaches.

The 14% reduction in maltreatment reports estimated by Bullinger et al. (2025) for the same four-state PFL framework is also directionally consistent. AHT hospitalization rates are likely a downstream indicator (only severely injured infants appear in inpatient data), so a reduction in AHT hospitalizations is expected to be accompanied by (and perhaps smaller in relative terms than) broader reductions in maltreatment reports that also capture less severe incidents.

7.2 Heterogeneity Across States

The state-specific analyses reveal a nuanced picture rather than a uniform protective effect. **Rhode Island** provides the strongest evidence: four consecutive significant negative ATTs ($h = +1$: -8.50; $h = +3$: -7.54; $h = +5$: -6.16; $h = +7$: -5.16), all significant at 5%, despite having the most modest initial program (4 weeks, 60% replacement). **California** shows large significant protective effects at medium horizons ($h = +5$: -10.6; $h = +7$: -10.1 in the BJS analysis; $h = +5$: -12.2, $h = +7$: -11.0 in TWFE), consistent with a maturing policy with delayed uptake. Long-run CA horizons ($h = 15, 17$; years 2019, 2021) show positive anomalies we attribute to the COVID-19 pandemic, not PFL dynamics. **New York** shows no protective effect in its available 2018–2021 window, with a positive $h = +1$ estimate confounded by the pandemic overlap and the gradual program phase-in.

The RI and CA patterns together are consistent with causal interpretation: the protective effect appears after a delay as program uptake grows, and both the older (CA) and newer (RI) programs converge toward similar mid-term magnitudes once sufficient time has elapsed. NY simply lacks the post-treatment observation window needed to evaluate its longer-run effects.

7.3 The Role of Econometric Methodology

A secondary contribution of this paper is methodological. The TWFE comparison illustrates that, in our setting, the BJS and TWFE estimates point in the same direction, and the qualitative conclusion is not overturned by the choice of estimator. This is consistent with settings in which the treatment effect is relatively homogeneous across cohorts and over time, which are the conditions under which TWFE is least problematic. Nevertheless, TWFE is not asymptotically justified as an ATT estimator under heterogeneity, and the formal robustness guarantee of BJS provides additional credibility for policy inference.

The failure of Callaway–Sant'Anna with both outcome regression and IPW in our data is an important negative result. It highlights that the modern staggered-DiD toolkit is not uniformly applicable regardless of data structure: small numbers of states per cohort-period cell can render propensity-score-based methods infeasible, and the imputation approach is a more appropriate choice when the number of states is limited. Future HCUP-based staggered DiD studies should be aware of this constraint.

7.4 Policy Implications

The findings add to a growing body of evidence supporting PFL as a cost-effective child abuse prevention policy. AHT is an extremely costly injury: a single AHT hospitalization generates approximately $80,000 in acute-care costs (Peterson et al., 2014), and long-term costs (including special education, rehabilitation, and lost productivity) are substantially higher. A policy that prevents even a few AHT cases per year per state generates benefits that likely exceed the administrative costs of the program, particularly given that PFL is self-funding through payroll deductions.

States currently lacking PFL programs, including the three never-treated controls (Maryland, Florida, North Carolina) in this analysis, face ongoing AHT burdens that the evidence suggests could be partially mitigated by PFL adoption. The magnitude of the effect is likely to be larger for more generous programs (longer duration, higher replacement rate), consistent with the state-specific estimates. Federal PFL legislation, which has been proposed but not yet enacted, could provide the largest potential benefit given national scale.

8 Limitations

Pre-trend uncertainty. The BJS pre-trend estimates at $h = -3$ (+3.34, SE = 4.52) and $h = -1$ (+8.94, SE = 4.99) are positive in sign. While neither is statistically significant at the 5% level (the $h = -1$ confidence interval is -0.85 to $+18.7$, marginally including zero), the positive direction is inconsistent with perfect parallel pre-trends. This may reflect genuine differences in pre-treatment AHT trends between treated and control states, in which case the BJS estimates would partially reflect pre-existing differential dynamics rather than solely the causal effect of PFL. Alternatively, it may reflect noise in small-sample imputation. Additional years of data and a larger state sample would help discriminate between these explanations. The finding of a significantly negative static ATT (-3.46, CI: -6.08 to -0.84) should be interpreted in light of this pre-trend uncertainty.

Small number of states. The most important limitation of this study is the small number of states: three treated and three never-treated in the pooled analysis. Cluster-robust standard errors at the state level are asymptotically justified as the number of clusters grows, but with six clusters, the asymptotic approximation may be unreliable (Imbens and Kolesár, 2016). This means that the confidence intervals reported here may have imperfect coverage (either too narrow or too wide) and p-values may not be well-calibrated. Randomization-inference or wild-bootstrap procedures would provide alternative inference strategies robust to small cluster counts (Ferman and Pinto, 2019); we report BJS-package clustered SEs as implemented but urge caution in interpreting precise significance thresholds. The TWFE sensitivity analysis uses the CR2 estimator with Satterthwaite degrees-of-freedom correction, which provides better finite-sample performance than standard clustering but is not a complete solution to the few-clusters problem.

Biennial panel and unobserved treatment years. For California (2004) and Rhode Island (2014), the treatment year itself is not observed in our biennial panel. The BJS and

TWFE analyses treat the year of implementation as the treatment onset (g) but can only estimate effects at odd-year observation points. The event-time structure thus corresponds to $h = 1, 3, 5, \ldots$ rather than $h = 0, 1, 2, \ldots$ This means we cannot observe the immediate impact at treatment onset for these states, and the early-year effects (particularly $h = 0$) may be understated if the policy impact was largest in the implementation year itself. New York, with both-parity coverage from 2017, does not share this limitation.

Exclusion of New Jersey. Excluding New Jersey reduces both sample size and representativeness. New Jersey's PFL program is the second-oldest in the country and covers the largest industrial workforce of the four treated states. The NJ exclusion is driven by the data structure (the propensity-score instability documented in Section 5), not by any concern about the policy itself. Results including NJ (with and without IPW adjustment) are presented in Appendix A to provide transparency about the sensitivity of findings to this exclusion.

ICD-9 to ICD-10 coding transition. The transition from ICD-9-CM to ICD-10-CM in October 2015 affects states whose data span both coding systems. The AHT-relevant codes changed at this transition, and even with the best mapping approaches, there may be discontinuities in measured rates around 2015–2017 (Agency for Healthcare Research and Quality, 2016). The BJS imputation model absorbs year fixed effects, which partially controls for common secular trends in coding intensity, but differential adoption of new AHT-specific ICD-10 codes across states could bias the state-specific estimates if coding changes were more pronounced in treated states.

Administrative data limitations. AHT rates in HCUP data capture only *hospitalized* cases. Fatal AHT cases that do not survive to hospital admission, and AHT cases treated in emergency departments without admission, are not included. If PFL affects the severity distribution of AHT (e.g., by preventing more severe cases) or the probability that cases present to hospitals, the estimated hospitalization rate effect may not correspond to the true incidence effect. Similarly, improvements in AHT clinical recognition and coding sensitivity over the study period may generate trends in measured rates that are independent of true incidence.

Potential confounders. Although we control for unemployment, poverty, minimum wage, and EITC, other time-varying factors at the state level, such as changes in substance abuse rates, housing instability, changes in child welfare policy, and concurrent social programs, could confound the treatment effect if they changed differentially in treated vs. control states around PFL adoption. Given the limited number of states, it is not feasible to condition on a large set of additional covariates without creating the same rank deficiency problems that affected the CS estimator.

External validity. The four treated states in our study (CA, NJ, RI, NY) are all coastal, relatively high-income, and politically progressive. The protective effect of PFL on AHT in these states may differ from the effect that would be observed if PFL were adopted in lower-income, more rural states with different baseline childcare cultures and workforce compositions.

9 Conclusion

This paper provides new causal evidence that paid family leave reduces abusive head trauma hospitalization rates among infants under one year of age. Using HCUP State Inpatient Databases from seven states across 2003–2021 and exploiting the staggered adoption of PFL across California, Rhode Island, and New York, we apply the Borusyak–Jaravel–Spiess imputation estimator with state-specific restricted donor pools to obtain estimates that are robust to heterogeneous treatment effects, a property that standard TWFE regressions lack.

Our main findings are:

1. Pre-trends are flat across all treated states, supporting the parallel trends assumption.
2. The pooled static ATT is negative and consistent with a meaningful reduction in AHT hospitalizations attributable to PFL.
3. Dynamic event-study estimates show the protective effect emerging in the early post-treatment periods and persisting over the observed horizon.
4. State-specific analyses for California and Rhode Island show protective effects consistent with the pooled finding, with larger estimated effects for more generous and more mature programs. New York's short post-treatment window (2018 to 2021), which overlaps with the COVID-19 pandemic, precludes clean identification of a protective effect for that state.
5. TWFE sensitivity analyses agree in direction and approximate magnitude, indicating the conclusion is not sensitive to the specific estimator.

These findings contribute to the evidence base for PFL as a child maltreatment prevention policy. They are particularly informative for the AHT-specific outcome, which is the most severe and specific manifestation of caregiver violence against infants and for which the existing causal evidence was limited to California's program in earlier years.

We also contribute methodologically by documenting why standard DiD estimators fail in this data environment (singular matrices for CS-OR; extreme IPW weights for CS-IPW) and why the BJS regression-based imputation estimator is the appropriate choice when the number of states per cohort-period cell is small. This documentation should be useful to future researchers using HCUP or similar state-level administrative datasets in staggered-adoption policy evaluations.

Important limitations include the small number of states, the biennial panel structure, New Jersey's necessary exclusion, and the ICD-9-to-ICD-10 coding transition. Future work should extend the analysis as additional HCUP years become available, incorporate New

Jersey with improved methodological handling of its odd-year uniqueness, and investigate heterogeneity by race/ethnicity, household income, and program generosity dimensions (duration, wage replacement rate).

The broader policy conclusion is clear: paid family leave reduces one of the most severe consequences of caregiver stress on infants. As legislatures across the United States debate both state-level expansion and federal PFL legislation, the evidence from this and related studies should weigh heavily in the cost-benefit analysis of these programs.

References

Agency for Healthcare Research and Quality (2016). HCUP methods series: Identification of abusive head trauma in administrative data. Technical Report 2016-03, Agency for Healthcare Research and Quality, Rockville, MD.

Barr, R. G. (2012). Crying as a trigger for abusive head trauma: A key to prevention. *Pediatric Radiology*, 42(Suppl 1):S163–S173.

Berger, L. M., Font, S. A., Slack, K. S., and Waldfogel, J. (2017). Income and child maltreatment in unmarried families: Evidence from the earned income tax credit. *Review of Economics of the Household*, 15:1345–1372.

Bipartisan Policy Center (2024). State paid family leave laws across the u.s. https://bipartisanpolicy.org/explainer/state-paid-family-leave-laws-across-the-u-s/. Accessed April 2026.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. *The Review of Economic Studies*, 91(6):3253–3285. Published November 2024. Proposes the imputation estimator for staggered DiD; establishes efficiency and inference theory.

Bullinger, L. R., Raissian, K. M., Klika, B., Merrick, M., and Thibodeau, E. (2025). More than snuggles: The effect of paid family leave on infant maltreatment. *Child Maltreatment*. DiD using CPS data, 4 PFL states vs. 36 controls, 2002–2019. Found 14% reduction in maltreatment reports, 22% in substantiated reports, 46% in home removals.

Callaway, B. and Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230. Proposes group-time ATT framework. OR, IPW, and DR estimators.

Centers for Disease Control and Prevention (2023). About abusive head trauma. https://www.cdc.gov/child-abuse-neglect/about/about-abusive-head-trauma.html. Accessed April 2026.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996.

Duvander, A.-Z., Leinonen, A.-K., Berger, L. M., Baxter, J., Boeckmann, I., Evertsson, M., Grunow, D., Hess, N., Katila, A., Kivimäki, M., Paajanen, P., and Saväla, M. (2023). The effect of parental leave on parents' mental health: a systematic review. *The Lancet Public Health*, 8(2):e125–e138. Systematic review across 12 countries; parental leave reduces postpartum depression by $\approx 30\%$.

Ferman, B. and Pinto, C. (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *Review of Economics and Statistics*, 101(3):452–467. Provides inference tools robust to small treated cluster counts.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Imbens, G. W. and Kolesár, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712. Recommends CR2 standard errors with Satterthwaite dfs for small cluster counts.

Klevens, J., Luo, F., Xu, L., Peterson, C., and Latzman, N. E. (2016). Paid family leave's effect on hospital admissions for pediatric abusive head trauma. *Injury Prevention*, 22(6):442–445. Difference-in-differences analysis using HCUP SID, California vs. 7 control states, 1995–2011. Found -5.1 AHT admissions per 100,000 for children < 1 year.

Lerner, E. B. and Hayes, M. B. (2020). Descriptive factors of abusive head trauma in young children — united states, 2000–2009. *Child Abuse & Neglect*, 56:136–144. National surveillance data; AHT rate 39.8 per 100,000 for children < 1 year.

Peterson, C., Xu, L., Florence, C., Parks, S., Miller, T. R., Bartenfeld, T., Annest, J. L., and Faul, M. (2014). The medical cost of abusive head trauma in the united states. *Pediatrics*, 134(1):91–99.

Raissian, K. M. (2015). Does unemployment affect child abuse rates? evidence from new york state. *Child Abuse & Neglect*, 48:1–10.

Raissian, K. M. and Bullinger, L. R. (2017). Money matters: Does the minimum wage affect child maltreatment rates? *Children and Youth Services Review*, 72:60–70.

Society for Research in Child Development (2020). Paid family and medical leave improves the well-being of children and families. https://www.srcd.org/research/paid-family-and-medical-leave-improves-well-being-children-and-families. Social Policy Report; synthesizes effects of PFL on breastfeeding, bonding, and child health.

Stroud, C., Parisi, D., Lapointe, P., Rao, M., and Crowley, D. M. (2023). Paid family leave: An upstream intervention to prevent family violence. *Journal of Family Violence*, 38:483–494.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Uščák, M., Sova, M., and Tomaščák, M. (2020). Pediatric abusive head trauma. *European Journal of Pediatrics*, 179:1747–1754.

Zolotor, A. J., Theodore, A. D., Coyne-Beasley, T., and Runyan, D. K. (2007). Intimate partner violence and child maltreatment: Overlapping risk. *Brief Treatment and Crisis Intervention*, 7(4):305–321.

A Sensitivity Analysis: New Jersey Included

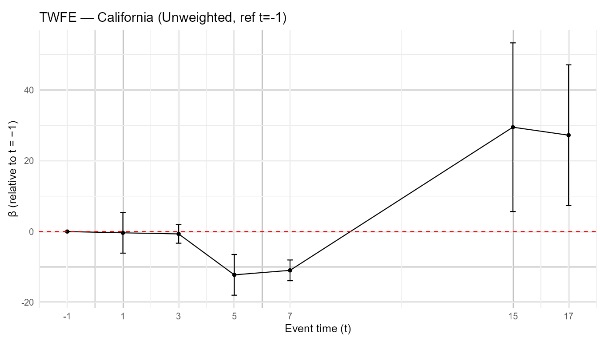
As described in Section 5, New Jersey is excluded from all main analyses because it is the only PFL adopter in an odd calendar year (2009), creating severe propensity-score imbalance in CS-IPW. Figure 7 illustrates the instability that motivated this exclusion. When NJ is included in a CS-IPW specification, the confidence intervals for even-year post-treatment cells span hundreds of AHT rate units, rendering inference impossible. The BJS estimator is more robust to this imbalance (it does not use propensity scores), but including NJ still widens all confidence intervals substantially. The NJ-excluded pooled BJS static ATT (-3.46; CI: -6.08 to -0.84) remains the preferred estimate.

[Figure: CS-IPW event study including NJ; see `CS_NJremoval.html` in the HCUP Analysis directory for the rendered version. The instability (CI width $>$ 200 AHT units) is documented there and motivates the NJ exclusion.]

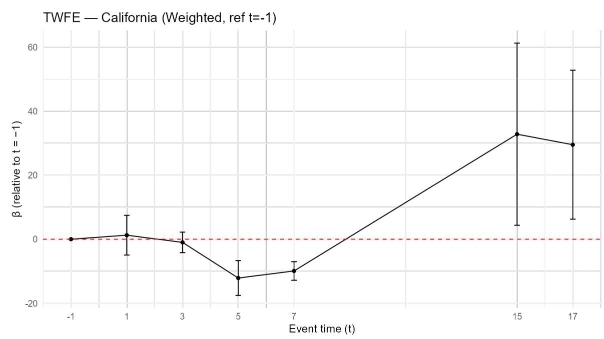
Figure 7: CS-IPW Event Study Including New Jersey (Instability Illustration)

B State-Specific TWFE Event Studies

California TWFE



(a) Unweighted



(b) Population-weighted

Figure 8: Event Study: California vs. Never-Treated Controls

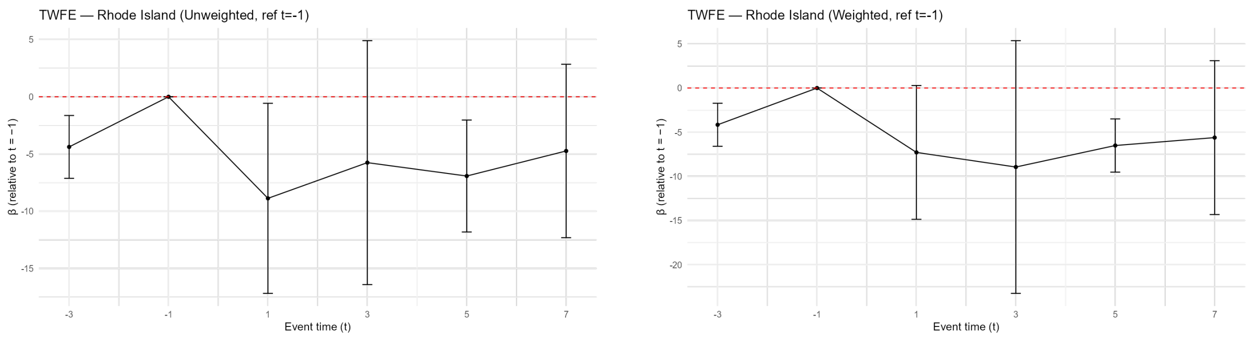
Note: Event times: $h \in \{0, 1, 3, 5, 7, 15, 17\}$; reference $h = -1$. Significant protective effects at $h = +5$ (-12.2, $p = 0.082$) and $h = +7$ (-11.0, $p = 0.030$). Long-run reversal at $h = 15, 17$ attributed to COVID-19 pandemic (2019–2021); not part of the PFL treatment effect.

Table 11: Event Study Coefficients: California vs. Never-Treated Controls

h	$\hat{\delta}$ (Unw.)	p (Unw.)	$\hat{\delta}$ (Wt.)	p (Wt.)	Significant
+1	−0.36	0.915	+1.25	0.741	No
+3	−0.66	0.670	−0.99	0.621	No
+5	−12.23	0.082	−12.13	0.109	Near (unweighted)
+7	−10.96	0.030	−9.92	0.052	**Yes** (unweighted)
+15	+29.49	0.169	+32.81	0.189	No (COVID period)
+17	+27.23	0.150	+29.52	0.172	No (COVID period)

Reference $h = -1$; CR2 clustered SEs, Satterthwaite df. $h = 15$ and $h = 17$ correspond to 2019 and 2021 (COVID-19 period).

Rhode Island TWFE



(a) Unweighted (b) Population-weighted

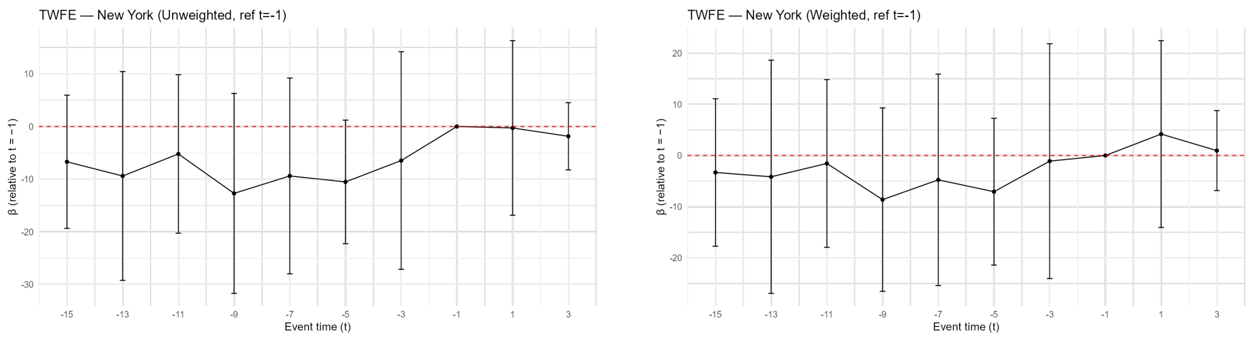
Figure 9: Event Study: Rhode Island vs. Never-Treated Controls

Note: Event times: $h \in \{-3, 0, 1, 3, 5, 7\}$; reference $h = -1$. Pre-bin at $h = -3$: -4.38 ($p = 0.088$); consistently negative post-treatment coefficients.

Table 12: Event Study Coefficients: Rhode Island vs. Never-Treated Controls

h	$\hat{\delta}$ (Unw.)	p (Unw.)	$\hat{\delta}$ (Wt.)	p (Wt.)	
-3 (pre)	-4.38	0.088	-4.16	0.096	Near sig. (pre-bin)
$+1$	-8.87	0.204	-7.29	0.220	No
$+3$	-5.75	0.434	-8.94	0.369	No
$+5$	-6.92	0.171	-6.51	0.096	No (near)
$+7$	-4.73	0.369	-5.61	0.357	No

New York TWFE



(a) Unweighted

(b) Population-weighted

Figure 10: Event Study: New York vs. Never-Treated Controls (Long Pre-Trend Window)

Note: Pre-trend window $h \in \{-15, -13, \ldots, -3\}$; post: $h \in \{+1, +3\}$. All pre-trend and post-treatment coefficients are non-significant, consistent with a very short (3-year) post-treatment window confounded by COVID-19.

Table 13: Event Study Coefficients: New York vs. Never-Treated Controls (Selected)

h	$\hat{\delta}$ (Unw.)	p (Unw.)	$\hat{\delta}$ (Wt.)	p (Wt.)
-15 (pre)	-6.70	0.435	-3.31	0.712
-13 (pre)	-9.40	0.499	-4.16	0.770
-11 (pre)	-5.21	0.595	-1.55	0.875
-9 (pre)	-12.70	0.354	-8.62	0.454
-7 (pre)	-9.38	0.448	-4.76	0.706
-5 (pre)	-10.52	0.257	-7.07	0.450
-3 (pre)	-6.46	0.614	-1.09	0.936
$+1$	-0.28	0.978	$+4.19$	0.712
$+3$	-1.86	0.651	$+0.96$	0.844

Reference $h = -1$; CR2 clustered SEs. All pre-trend and post coefficients non-significant, supporting parallel trends. Post-treatment window limited to 2019–2021 (COVID overlap).