A House Divided: Modern Heterogeneous-Effects Estimators and the Contested Mortality Effects of Hospital Desegregation, 1959–1988

Abstract

Almond, Chay, and Greenstone (2006) argued that the Title VI Medicare-certification campaign of 1966–1968 produced a large decline in Black postneonatal mortality in the rural South. Anderson, Charles, and Rees (2024) revisited the question with a county-of-nearest-hospital design and reported a statistically null effect. We apply five staggered difference-in-differences estimators — two-way fixed effects (TWFE), Callaway–Sant'Anna (CS), Borusyak–Jaravel–Spiess (BJS), de Chaisemartin–D'Haultfœuille (dCDH), and Sun–Abraham (SA) — to a unified eleven-state Confederate-South panel of Black postneonatal mortality, using the `cert_75` treatment definition of Anderson, Charles, and Rees (2024). We additionally construct a federal-source-uniform extended panel back to 1959 by combining NCHS Mortality Detail File microdata with a newly transcribed pre-1968 county-year-race natality denominator built from state Department of Public Health vital-statistics annual reports (six states) and federal VSUS Vol I Natality tabulations (the remaining five); the extended panel is the primary specification on which we anchor inference.

On the federal-source-uniform extended panel (1959–1988), all five estimators are negative and all five reject zero at the 5-percent level: TWFE -1.85 [-3.43, -0.28], CS -5.09 [-7.05, -3.14], BJS -2.41 [-3.87, -0.95], dCDH -4.07 [-6.12, -2.02], Sun–Abraham (SA) -2.10 [-3.70, -0.50]. A diagnostic refit of the cohort-saturated SA event-study with cohort-specific endpoint bins for $|\ell| > 10$ is what makes the SA aggregate negative; without those bins, long-run treated observations from the dominant 1967 cohort fall into the implicit reference category and the aggregate flips sign (Appendix C). A selective-fertility bound from Thompson (2024) Table 4 implies a mechanical compositional shift of approximately -2.7 per 1,000; the CS and dCDH estimates exceed this bound and are consistent with a Title-VI-causal effect not exhausted by fertility selection, while the TWFE, BJS, and SA estimates fall just below it and are not separately distinguishable from a pure compositional mechanism. Because the

Black fertility decline of 1965–1968 is itself partly endogenous to civil-rights enforcement, the bound separates *non-fertility-mediated* Title-VI channels from the rest rather than direct treatment effects from no treatment effect.

On the public-CMF window (1968–1988), reported as a robustness comparison, the same five estimators report ATTs between −2.5 and +0.7 per 1,000 live births: TWFE −1.30 [−3.00, +0.40], CS −2.50 [−4.45, −0.55], BJS −1.85 [−3.87, +0.17], dCDH −2.27 [−4.18, −0.36], SA +0.66 [+0.02, +1.30]. CS and dCDH reject zero on the shorter panel; TWFE and BJS do not. SA's positive sign on the shorter panel survives the endpoint-bin correction because the public-CMF window begins in 1968, so the dominant 1967 cohort has no observed pre-period for SA to anchor on, and we read the v1 SA as a flag rather than a substantive estimate. The attenuation of the other estimators toward zero on the shorter panel is consistent with two diagnostic findings: (i) a Goodman-Bacon decomposition attributes 36 percent of TWFE weight to "forbidden" later-vs-already-treated comparisons — comparisons in which a unit treated later is benchmarked against a unit treated earlier acting as the implicit control, which delivers biased treatment-effect estimates under treatment-effect heterogeneity (Goodman-Bacon, 2021) — and the mean of those comparisons (−0.52) is much less negative than the clean treated-vs-never-treated comparisons' mean (−2.44); and (ii) the dominant 1967 cohort has zero pre-period observations within the public NBER CMF, so the staggered-DiD estimators must extrapolate the counterfactual from late-cohort variation that coincides with substantially different state-level economic and policy environments. The extended panel removes the cohort-anchoring problem; the heterogeneity-robust estimators address the forbidden-comparison problem; together they produce the larger magnitudes that the heterogeneity-robust estimators converge on. A complementary design grid varying the certification threshold (`cert_any`, `cert_25`, `cert_50`, `cert_75`) shows that the v3 estimates are stable across treatment-threshold choices for all five estimators (Appendix B); the `cert_75` headline is, if anything, conservative relative to looser thresholds.

The post-2018 heterogeneity-robust estimator menu produces a *frontier* of estimates rather than a single point, even within a single design. On the federal-source-uniform extended panel and after the selective-fertility adjustment, the data favors a substantively negative effect of Title VI hospital desegregation on Black postneonatal mortality consistent with the original claim of Almond, Chay, and Greenstone (2006), with magnitudes between −1.9 and −5.1 per 1,000 depending on estimator. The Anderson, Charles, and Rees (2024) null is recoverable on the public-CMF window for the TWFE and BJS specifications, both of which the cohort-anchoring diagnostics flag as attenuated, suggesting that the original disagreement reflects in substantial part the limits of post-1968-only public mortality data.

Keywords: Title VI, hospital desegregation, Black infant mortality, staggered difference-in-differences, Callaway–Sant'Anna, Goodman-Bacon decomposition, Honest-DiD.

1. Introduction

The 1964 Civil Rights Act was operationally enforced in southern hospitals through Medicare certification. Beginning July 1, 1966, the Department of Health, Education, and Welfare's Office of Equal Health Opportunity (HEW–OEHO) required, as a condition of receiving federal Medicare reimbursement, that hospitals certify nondiscriminatory treatment of Black patients and staff under Title VI [@reynolds1997federal; @smith1999healthcaredivided]. The campaign was, in Reynolds's (1997) characterization, "one of the most successful civil-rights enforcement operations of the federal government." Within eighteen months, 96 percent of the Deep South's hospitals were certified; the remainder either closed or accepted formal segregation and forfeited federal funds.

Twenty years later, Almond, Chay, and Greenstone (2006) [hereafter ACG] used the NBER Compressed Mortality File and the staggered 1966–1968 county-level certification rollout to argue that this campaign produced a large, persistent decline in Black postneonatal mortality across the rural South. The ACG paper, circulated as MIT Working Paper 07-04 in 2007 but never published in a peer-reviewed journal, became the canonical empirical reference for the proposition that civil-rights enforcement saved tens of thousands of Black infant lives in the late 1960s. It is cited in virtually every survey of post-1960s mortality declines and has been influential in establishing the broader narrative that Black–White health gaps in the United States closed substantively only when federal civil-rights enforcement gave Black Americans physical access to white-only healthcare infrastructure.

Anderson, Charles, and Rees [hereafter ACR] revisited the question in a 2024 paper now forthcoming at the *Review of Economics and Statistics*. ACR (a) compiled the underlying hospital-certification dates directly from the annual *Journal of the American Hospital Association* "Guide Issue" rather than relying on the secondary sources that ACG had used; (b) coded county-level "access" using four different bed-coverage thresholds (any certification, at least 25 percent, at least 50 percent, at least 75 percent of pre-1965 hospital beds); and (c) restricted attention to the five Deep-South states most exposed to the certification campaign. Their headline conclusion, on the at least 75-percent bed-coverage binding (`cert_75`), is that the campaign had no statistically detectable effect on Black postneonatal mortality. ACR (2024) is the first careful re-examination of ACG's headline claim and reports a *null* on essentially the same outcome.

The disagreement matters for two reasons. First, the foundational claim that civil-rights enforcement produced a large, measurable mortality dividend is uncomfortable to lose; the absence of a peer-reviewed Almond–Chay–Greenstone paper has meant that the field has cited the working-paper version for two decades as if it were settled. Second, the methodological texture of the disagreement is informative for a broader question: how should the discipline now read

the large 2000s and 2010s difference-in-differences literature that relied on the staggered-TWFE estimator? Goodman-Bacon (2021), Callaway and Sant'Anna (2021), Sun and Abraham (2021), de Chaisemartin and D'Haultfœuille (2020), Borusyak, Jaravel, and Spiess (2024), and Rambachan and Roth (2023) have collectively documented that staggered-TWFE estimators can deliver biased estimates of average treatment effects when treatment timing varies, treatment effects are heterogeneous, or pre-trends are present. None of those papers has been applied side-by-side to the ACG–ACR question.

This paper does so. We construct a unified eleven-state Confederate-South county-year panel for the period 1968–1988 — the public-NBER-CMF window — and apply five staggered-DiD estimators (TWFE, Callaway–Sant'Anna, Borusyak–Jaravel–Spiess, de Chaisemartin–D'Haultfœuille, and Sun–Abraham) to the same outcome (Black postneonatal mortality per 1,000 live births), the same treatment definition (the ACR `cert_75` binding), and the same sample frame. We then extend the panel backward to 1959 by transcribing pre-1968 state vital-statistics tabulations — the same data source that ACG (2006) used for their pre-period and that ACR (2024) acknowledged as essential for the dynamic event-study but did not themselves transcribe. The extended-panel estimates reveal whether the original disagreement was driven by the post-1968-only NBER CMF window or by something else.

Our contribution is methodological adjudication, not a new empirical claim about whether hospital desegregation reduced Black infant mortality. We document the *shape* of the estimator frontier on this canonical question — under what data window, under what estimator, the data favors the ACG reading; under what conditions the data favors ACR; and what the partial-identification frontier looks like under Rambachan-Roth bounds. We hope the exercise is useful both as a substantive adjudication of one of the most important questions in modern U.S. health-equity history and as a methodological case study for how to read older staggered-DiD literatures in light of the post-2018 econometric developments.

The paper proceeds as follows. Section 2 reviews the historical and econometric literature. Section 3 documents the data. Section 4 describes the five estimators and the robustness suite. Section 5 reports results. Section 6 discusses the partial-identification frontier and what would be needed to pin the magnitude. Section 7 concludes.

---

2. Background

2.1 The Title VI hospital-desegregation campaign — institutional and political history

The 1964 Civil Rights Act outlawed discrimination by recipients of federal funds, but its operative enforcement in southern hospitals required a separate instru-

ment: the federal money attached to Medicare. Before 1966, southern hospitals received federal Hill-Burton construction grants without operative Title VI compliance. The 1946 Hill-Burton Hospital Survey and Construction Act explicitly authorized "separate but equal" facilities — Section 622(f) of the Act permitted federal funds to flow to hospitals that maintained racial separation provided the separate Black facilities were "of like quality." Between 1946 and 1965, the federal Hill-Burton program funded approximately 100 segregated Black hospitals across the South and provided construction grants to hundreds of additional white-only facilities [@smith1999healthcaredivided; @reynolds1997federal].

The constitutional challenge to Hill-Burton's separate-but-equal clause culminated in the 1963 Fourth Circuit decision in *Simkins v. Moses H. Cone Memorial Hospital* (323 F.2d 959, 4th Cir. 1963), in which the court ruled that the Hill-Burton separate-but-equal provision was unconstitutional under the Equal Protection Clause and that hospitals receiving federal funds could not maintain racial segregation. The decision applied to the Fourth Circuit states (NC, SC, VA, WV, MD) but not to the Fifth Circuit states (AL, FL, GA, LA, MS, TX). The Supreme Court denied certiorari in 1964 (376 U.S. 938), leaving Simkins as Fourth-Circuit-only precedent until the broader 1964 Civil Rights Act and the 1966 Medicare Act enforcement.

The operative enforcement instrument was Title VI of the 1964 Civil Rights Act applied through Medicare certification, beginning July 1, 1966. Under the Medicare Act of 1965 (Public Law 89-97), hospitals receiving Medicare reimbursement were required to certify nondiscrimination in admissions, room assignments, staff privileges, and physician practice. The certification process was operationally managed by the Department of Health, Education, and Welfare's Office of Equal Health Opportunity (HEW–OEHO), established within the Public Health Service. HEW–OEHO compliance officers conducted on-site inspections of southern hospitals during February 1966 through October 1968, verifying that hospitals had eliminated racially-designated wards, integrated staff dining facilities, opened medical-staff appointments to Black physicians, and complied with the Title VI nondiscrimination requirements.

The political economy of the certification process was complex. Hospital trustees in the Deep South initially resisted certification — 214 hospitals in the five-state Deep South refused federal Medicare reimbursement rather than integrate. The economic pressure of forgoing Medicare reimbursement, however, proved decisive: within twelve months of the July 1, 1966 effective date, approximately 96 percent of Deep-South hospitals had certified. The laggards were predominantly small, rural facilities with limited Medicare-eligible patient bases and limited financial reserves; they either closed (~30 closures in 1966–1969) or reluctantly certified. The political-historical literature [@reynolds1997federal; @smith1999healthcaredivided; @smith2005race; @largent2018segregation] documents that the certification process was conducted under substantial federal-state political tension, with state governors and southern members of Congress repeatedly attempting to block or delay HEW–OEHO compliance enforcement.

The HEW–OEHO compliance files themselves are archived at the National Archives and Records Administration as NARA Record Group 235, Series 5 (Health Equity Compliance Files, 1966–1972). Reynolds (1997) and Largent (2018) have separately conducted archival work in these files; their accounts of the operational compliance process — site inspections, formal compliance reviews, hospital-by-hospital certification negotiations — are the primary source basis for our understanding of cosmetic-compliance heterogeneity. Smith (1999, 2005) documents that many certified hospitals dual-listed wards as "integrated" while continuing de facto segregation through admission patterns, room-assignment defaults, and physician-practice restrictions. The cert_75 binary indicator in our analytic panel averages over this substantial residual segregation.

The campaign's operational success was measured at the federal level through the "Guide Issues" of the *Journal of the American Hospital Association*, published annually in February and October. The Guide Issue listed, for each hospital, whether it had been "certified for participation in the Health Insurance for the Aged (Medicare) Program by the Department of Health, Education, and Welfare" as of the publication's reference date. Anderson, Charles, and Rees (2024) hand-coded the Guide Issues for the five Deep-South states (AL, GA, LA, MS, SC) for 1967–1974 to reconstruct hospital-level certification dates, then aggregated to the county-of-nearest-hospital level under four alternative bed-coverage thresholds: any certification (`cert_any`), at least 25 percent of pre-1965 county hospital beds (`cert_25`), at least 50 percent (`cert_50`), and at least 75 percent (`cert_75`). The resulting county-year certification panel — reported in their Appendix Table B1 — is the analytic foundation for the modern reassessment of the ACG findings.

The dispersion in `cert_75` is meaningfully wider than in `cert_any`. In our merged panel, 350 of 415 Deep-South counties with assigned dates were certified in 1967 under cert_75; 43 in 1968; 17 in 1969–1971; and 5 in 1972 or later. Under cert_any (any hospital certified), the distribution is much more concentrated in 1966–1967. The dispersion in `cert_75` is exactly the variation that ACR (2024) argue powers the modern county-of-nearest-hospital identification strategy and that ACG (2006) treated as too modest to exploit.

2.2 The Almond–Chay–Greenstone (2006) headline

ACG used the NBER Compressed Mortality File 1959–1988 (which extends backward via state vital-statistics tabulations to 1959) and a staggered-DiD design at the county level. Their preferred specification compared Black postneonatal mortality rates in southern counties before and after Medicare certification of the county's hospitals, controlling for county and year fixed effects and Black–White rates. They reported a ~5–6 per-1,000-births reduction in Black postneonatal mortality, concentrated in the rural Deep South and in cohort-1967 counties. The ACG estimate implies on the order of ~25,000 Black infant lives saved over the 1965–1972 period from a single federal policy intervention.

ACG (2006) was never peer-reviewed. The paper has nevertheless been cited several thousand times and is the standard reference for the proposition that Title VI enforcement produced a large mortality dividend. It is the empirical anchor for the broader narrative that the Civil Rights Act materially extended Black life expectancy in the late 1960s.

2.3 The Anderson–Charles–Rees (2024) null

ACR reproduced ACG's panel structure with three differences. First, they hand-coded hospital certification from primary sources (the JAHA Guide Issues) rather than relying on the secondary derivative dates ACG used. Second, they restricted attention to the five Deep-South states most exposed to the campaign rather than the eleven-state Confederate sample ACG had used. Third, they reported four alternative bed-coverage thresholds rather than ACG's single any-certification binary. Across all four thresholds and on essentially the same outcome (Black postneonatal mortality), ACR reported a statistically null effect.

ACR's interpretation is that the ACG result was an artifact of the secondary certification-date measure ACG used, which conflated county-level access with state-level certification timing. The implicit methodological diagnosis is that ACG had a high false-positive rate driven by measurement error in the treatment.

2.4 Adjacent literatures on civil-rights and Black mortality declines

The civil-rights-era reduction in Black infant mortality is one of the most-studied phenomena in twentieth-century American population health. Beyond the direct hospital-desegregation channel that ACG and ACR debate, four adjacent literatures bear on the interpretation of the headline ATT.

First, the federal War on Poverty health programs of 1965–1972 — Medicaid (1965), Community Health Centers (Bailey and Goodman-Bacon 2015), and Maternal and Child Health Title V expansions — produced contemporaneous mortality reductions that overlap the Title VI exposure window. Goodman-Bacon (2018 *JPE*) identifies a 1965 Medicaid implementation effect on elderly mortality; Bailey and Goodman-Bacon (2015 *AER*) identify CHC effects on older-adult mortality 1965–1974; Currie and Gruber (1996 *JPE*) and (1996 *QJE*) identify Medicaid expansions and infant mortality more broadly. The methodological challenge in this paper is to separate the Title-VI-specific hospital-integration channel from the broader War-on-Poverty health-financing channel that operated through the same calendar window.

Second, the voting-rights and political-economy literature — Cascio and Washington (2014 *QJE*) on the Voting Rights Act and state funds — documents that Black political enfranchisement after 1965 produced its own measurable effects on state-level resource allocation to majority-Black areas, which would have its own infant-mortality consequence. Our covariate-adjusted robustness

(§5.8) adds VRA Section 5 preclearance as a state-year constant; the substantive effect on the headline ATT is small (at most 0.06 per 1,000), but the VRA channel is methodologically a confounding policy.

Third, the selective-fertility literature (Thompson 2024 *JHR*, Aizer-Currie-Moretti-Yang 2014, Eli-Currie 2019) documents that the Black fertility decline of 1964–1970 was concentrated in high-mortality-risk pregnancies, producing a mechanical reduction in measured infant mortality that any of our five estimators would attribute to treatment. Our §6.3 quantitative bound on selective fertility addresses this confound directly; the bound is approximately −2.7 per 1,000 from compositional shift alone, comparable in magnitude to the TWFE and BJS estimates and smaller than the CS and dCDH estimates on the extended panel.

Fourth, the broader Black mortality literature documents secular and policy-driven declines in Black mortality across the lifespan during this period. Chay and Greenstone (2003 *QJE*) on air-quality regulation; Cutler and Miller (2005 *Demography*) on twentieth-century water-quality interventions; Bleakley (2007 *QJE*) on hookworm-eradication effects on cohort outcomes; Aizer, Eli, Ferrie, and Lleras-Muney (2016 *AER*) on Mothers' Pensions effects on Black mortality; Almond and Mazumder (2011 *AEJ:Applied*) on prenatal-care effects; Brown, Kowalski, and Lurie (2020 *RFS*) on Medicaid expansions and infant mortality; Currie and Schwandt (2016 *Science*) on the post-1965 Black–White infant-mortality convergence. Each of these literatures intersects the Title VI exposure window through different mechanisms; the multi-confound identification problem is substantial.

2.5 The staggered-DiD econometric literature

The methodological foundation for this paper is the post-2018 staggered-DiD literature that documents the limits of the two-way-fixed-effects estimator under treatment-effect heterogeneity. The foundational result is Goodman-Bacon (2021 *Journal of Econometrics*), which decomposes the TWFE estimator into 2×2 difference-in-differences with non-negative-and-non-negative-weights depending on the comparison type, and demonstrates that "forbidden" later-vs-already-treated comparisons can carry negative weight when treatment effects vary across cohorts or over event time.

The post-Goodman-Bacon literature proposes alternative estimators that avoid the forbidden-comparison problem. Callaway and Sant'Anna (2021 *Journal of Econometrics*) propose an ATT(g,t) framework using clean comparisons (treated vs. never-treated or treated vs. not-yet-treated) aggregated with non-negative weights. Sun and Abraham (2021 *Journal of Econometrics*) propose a cohort-saturated event-study interacting cohort-fixed-effects with relative-time dummies. de Chaisemartin and D'Haultfœuille (2020 *AER*; 2024 *RES*) propose switching-on event-studies comparing treated and stably-untreated units. Borusyak, Jaravel, and Spiess (2024 *Review of Economic Studies*) propose an

imputation estimator that fits TWFE on untreated observations only. Athey and Imbens (2022 *Journal of Econometrics*) develop design-based identification under random assignment of treatment timing.

The partial-identification frontier under violations of strict parallel trends is documented in Rambachan and Roth (2023 *Review of Economic Studies*), which provides bounds for the post-treatment average under restrictions on the relative magnitude or smoothness of differential trends. The Honest-DiD framework has become standard for sensitivity reporting in modern staggered-DiD work. Roth, Sant'Anna, Bilinski, and Poe (2023 *Journal of Econometrics*) provide a comprehensive review.

The methodological texture of these estimators is the principal contribution of this paper. None of them has been applied side-by-side to the ACG–ACR adjudication in the published literature. The closest precedent is Tomar, Yu, and Spiess (2024 *Working Paper*), which applies a smaller estimator menu (CS + SA only) to a different civil-rights question (school desegregation); we do not duplicate their work but build on the methodological framing.

---

3. Data

We combine three publicly available data inputs to construct a county-year panel of Black and White postneonatal mortality covering the eleven Confederate-South states from 1959 through 1988.

Mortality and population. Death counts come from the National Center for Health Statistics Compressed Mortality File (CMF) 1968–1988, distributed by the National Bureau of Economic Research at `https://data.nber.org/cmf/`. The CMF is the public county-level mortality aggregate used by both Almond, Chay, and Greenstone (2006) and Anderson, Charles, and Rees (2024) and is the analytic foundation for the U.S. mortality–civil-rights literature [@goodmanbacon2018publicinsurance; @bailey2015waronpoverty; @chay2003impact]. Each record in the mortality file is a count of deaths in a state × county × year × race-sex × age × ICD-cause cell. The file uses ICD-8 codes 1968–1978 and ICD-9 codes 1979–1988. Population denominators and live births in the same race-sex × age × county-year cells come from the companion NBER population file (`pop6878` and `pop7988`, fixed-width format with logical record length 140 and twelve age-group population fields plus a live-births field).

Pre-1968 mortality and births. Because the NBER public CMF begins in 1968 and the pre-period cited by Almond, Chay, and Greenstone (2006) is therefore not in the CMF, we construct a federal-source-uniform pre-1968 extension. County-year Black and White infant, neonatal, and postneonatal death counts come from the NBER National Center for Health Statistics Mortality Detail File (MDF) for 1959–1967, the underlying federal microdata from which the published VSUS Vol II mortality tables are produced. We aggregate

death records by state of residence, county of residence, race, and the NCHS detailed age code, mapping NCHS county codes to FIPS via the NBER 1990 crosswalk and an alphabetical-ordering decode for the 1959–1961 four-character coding scheme. Pre-1968 live births by county-year-race come from two complementary sources. For the five states without machine-readable county-year-race natality at the federal level for this period — Alabama, Arkansas, Florida, Louisiana, and South Carolina — we use the printed VSUS Vol I Natality state tables. For the remaining six states — Georgia, Mississippi, North Carolina, Tennessee, Texas, and Virginia — we additionally digitize the contemporaneous state Department of Public Health vital-statistics reports (e.g., the North Carolina State Center for Health Statistics annual report, the Texas Department of State Health Services *Texas Vital Statistics* series, and the Virginia Department of Health *Annual Report*) and use the state-DOH county-year-race birth tabulations in preference to the federal Vol I figures, because the state-DOH publications resolve county boundaries (notably Virginia's independent cities) more cleanly than the federal aggregations. ACG (2006) used the same class of state-DOH sources for their pre-period; Anderson, Charles, and Rees (2024) acknowledged the necessity of pre-1968 county-year-race birth denominators for any dynamic event-study but did not themselves transcribe them, which is one reason their published evidence is confined to the post-1968 CMF window. We audit every cell — state-DOH and federal Vol I alike — against state-year totals and set obviously implausible birth values (county births exceeding 50,000 or 50 percent of the state-year Black-births total) to missing. Construction details and the per-state source assignment are in Appendix A.

Medicare-certification treatment dates. Hospital-level Medicare certification dates were originally compiled by Anderson, Charles, and Rees (2024) from the annual *Journal of the American Hospital Association* "Guide Issues" 1967–1974. Anderson, Charles, and Rees (2024) aggregate these hospital-level certification dates to the county-of-nearest-hospital level under four bed-coverage thresholds and report the resulting 418-county Deep-South panel in their Appendix Table B1. We extract this Appendix Table B1 directly from the working-paper PDF and merge it onto the county-year panel by FIPS code. This gives treatment dates for all 418 Deep-South counties (Alabama, Georgia, Louisiana, Mississippi, South Carolina), of which 403 enter the analytic sample after dropping the 9 Alabama and 6 Georgia counties carrying Appendix B1 footnote exclusions.

Time-varying covariates. Medicaid implementation year [@goodmanbacon2018publicinsurance Table 1] and Voting Rights Act Section 5 preclearance status [@cascio2014valuingvote Table A1] enter as state-year constants. Hill-Burton funding [@bailey2015waronpoverty], Community Health Center year [@bailey2015waronpoverty], and AFDC monthly maximum [@moffitt2003] are unavailable in machine-readable form at the county-year level and are documented as outstanding analyst-mediated access items.

Sample. Our analysis sample is the eleven Confederate-South states (Alabama,

Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia). This is the Almond, Chay, and Greenstone (2006) sample frame; Anderson, Charles, and Rees (2024) restrict to the five Deep South subset (AL, GA, LA, MS, SC). We retain all eleven states so that we can decompose the Almond-Chay-Greenstone–Anderson-Charles-Rees disagreement into a *sample-frame* component (5-state vs. 11-state) and an *estimator* component (TWFE-style vs. heterogeneity-robust DiD). The public-CMF window panel contains 23,672 county-years (1,146 counties × 1968–1988); the federal-source-uniform extended panel additionally covers 9,723 pre-1968 cells (1959–1967).

Outcomes. Following ACG and ACR, our primary outcome is the Black postneonatal mortality rate, defined as 1000 × (Black 28–365-day deaths) / (live Black births) within a county-year cell. We also construct (i) Black neonatal mortality, (ii) Black infant mortality, (iii) the corresponding White rates and Black–White gaps, and (iv) Black motor-vehicle mortality (ICD codes 770/800) for cause-specific falsification.

Treatment. Treatment is staggered Medicare certification of the county's hospital(s) under Title VI compliance, using ACR's `cert_75` binding (first year by which Medicare-certified hospitals covered at least 75 percent of pre-1965 county hospital beds). In our merged panel, 350 of 415 Deep-South counties with assigned dates were certified in 1967; 43 in 1968; 17 in 1969–1971; and 5 in 1972 or later. The dispersion in `cert_75` is meaningfully wider than in the `cert_any` indicator.

---

4. Methods

4.1 Identification

We treat the staggered 1966–1968 rollout of Medicare-certification-driven Title VI compliance as a county-level natural experiment in hospital integration. The treatment date is the year by which Medicare-certified hospitals covered at least 75 percent of pre-1965 county hospital beds (`cert_75`). The outcome is the Black postneonatal mortality rate per 1,000 live births. The estimand of interest is the average treatment effect on the treated (ATT) on this outcome.

The contribution of this paper is to apply, on the *same* harmonised eleven-state Southern panel, the four leading heterogeneity-robust event-study estimators that have emerged in econometrics since 2018 — Callaway and Sant'Anna (CS, 2021), Sun and Abraham (SA, 2021), Borusyak, Jaravel, and Spiess (BJS, 2024), and de Chaisemartin and D'Haultfœuille (dCDH, 2020/2024) — and to compare them to the two-way-fixed-effects (TWFE) estimator used implicitly by Almond, Chay, and Greenstone (2006). We report all five estimators side-by-side as the headline of the paper.

4.2 Five estimators

The TWFE benchmark is

$$Y_{ct} = \alpha_c + \gamma_t + \beta \mathbf{1}\{t \geq G_c\} + \varepsilon_{ct},$$

where G_c is the cert_75 year and α_c, γ_t are county and year fixed effects. The regression is weighted by Black live births and standard errors are clustered at the county level.

The Callaway–Sant'Anna estimator targets the group–time average treatment effect ATT(g, t) and aggregates to a "simple" ATT and an event-time path under conditional parallel trends. We use the doubly-robust variant with never-treated controls. Implementation is via the Python `differences` package.

The Borusyak–Jaravel–Spiess estimator fits unit and time fixed effects on untreated observations only, then imputes the counterfactual $\hat{Y}_{ct}(\infty)$ for treated cells. Counties with no untreated observations (the 1967 cohort on the public-CMF window) are dropped from this step. Standard errors are cluster-bootstrapped at the county level (B = 200).

The de Chaisemartin–D'Haultfœuille estimator computes, for each calendar year t and horizon ℓ, the DiD between counties switching into treatment at t and stably-untreated counties. Standard errors are cluster-bootstrapped at the switching-event level (B = 200).

The Sun–Abraham estimator saturates the event-study with cohort-by-relative-time interactions and aggregates the cohort-specific event-time coefficients with cohort-size weights. The default specification includes cohort-specific endpoint bins for $\ell < -10$ and $\ell > +10$ so that long-run treated cells do not fall into the implicit reference category; we report a diagnostic comparison in Appendix C showing that the aggregate SA sign on the extended panel is sensitive to this coding choice, and we use the binned specification as the default.

We label TWFE as a benchmark only and the four heterogeneity-robust estimators as the headline menu. The five estimators do not target a single common scalar estimand by construction: CS reports the package "simple" ATT aggregated with our `weights_name` (Black births) parameter; BJS reports a birth-weighted mean of imputed treatment effects across treated cells; dCDH reports an unweighted mean over horizons 0 through 10 of a manual DID_ℓ-style event path; SA reports a cohort-count-weighted average of event coefficients 0 through 10; and TWFE reports the static treatment coefficient. The headline scalar comparison should therefore be read as a robustness menu rather than a perfect common-estimand table; dynamic event-study figures carry much of the interpretive weight. The BJS, dCDH, and SA implementations in this paper are transparent manual implementations rather than calls to the canonical R packages (`did_imputation`, `did_multiplegt`, `fixest::sunab`); details and validation diagnostics are documented in §A and Appendix C.

4.3 Diagnostics and robustness

Goodman-Bacon decomposition. We decompose the TWFE estimate into its constituent 2×2 DiD comparisons, reporting the share of TWFE weight attributable to (i) treated-vs-never-treated comparisons, (ii) earlier-treated-vs-later-treated comparisons, and (iii) the "forbidden" later-treated-vs-already-treated comparison.

Rambachan-Roth Honest-DiD. For each estimator's post-treatment average over $\ell \in \{0, \ldots, 5\}$, we compute partial-identification bounds under the relative-magnitude restriction $|\bar{\delta}_{\text{post}} - \bar{\delta}_{\text{pre,max}}| \leq \bar{M} \cdot \bar{\delta}_{\text{pre,max}}$ for $\bar{M} \in \{0, 0.25, 0.5, 1.0, 2.0\}$.

White-postneonatal placebo. Title VI desegregation should change Black hospital access without affecting White infants; any non-null effect indicates contamination from a non-Title-VI shock.

Cause-specific motor-vehicle placebo. We extract Black motor-vehicle deaths from the NBER CMF using IRECODE 770 (ICD-8) and 800 (ICD-9) and run all five estimators on Black motor-vehicle mortality per 100,000 population. Title VI desegregation has no plausible causal pathway to motor-vehicle mortality, so a null effect is the cleanest external falsification.

Leave-one-state-out. We re-estimate each estimator eleven times, dropping each Southern state in turn.

Covariate-adjusted robustness. We re-estimate each estimator with Medicaid implementation year and VRA Section 5 preclearance as time-varying covariates (Frisch–Waugh–Lovell residualization).

Design grid. We vary three design axes — panel version (`v1` public-CMF vs. `v3` federal-uniform extended), sample (eleven-state Southern vs. five-state Deep-South-only), and treatment threshold (`cert_any`, `cert_25`, `cert_50`, `cert_75`) — and re-estimate all five estimators in each of the 80 resulting design cells. The grid lets us separate the contributions of panel choice, donor pool, and threshold definition to the headline estimates. The five-state-only cells leave at most three to ten never-treated counties depending on threshold and serve as weak-donor stress tests rather than preferred designs.

Synthetic-control family. We report a synthetic-control-family appendix (Appendix B) covering classic SCM, ridge-augmented SCM, a synthetic-DiD analogue, an interactive-fixed-effects (IFE) analogue, and a county-level PanelMatch-style matched DiD. The augsynth, gsynth, fect, synthdid, and PanelMatch R packages are not installed locally; we therefore implement transparent local analogues, label them as such, and treat the synthetic-family results as appendix triangulation rather than the main design.

4.4 Implementation

All estimators are implemented in Python (`pyfixest 0.40` for TWFE and SA; `differences 0.2` for CS; manual implementations for BJS and dCDH).

The full analysis reproduces end-to-end via `python3 analysis/run_all.py` in approximately ninety seconds.

---

5. Results

5.1 Headline static ATT — public-CMF window (1968–1988)

Figure 1 shows the treatment-timing distribution of the eleven-state Confederate-South sample under the `cert_75` binding: 304 counties enter in 1967, 48 in 1968, and the remainder spread between 1969 and 1974. Six of the eleven states (Arkansas, Florida, North Carolina, Tennessee, Texas, and Virginia) supply the never-treated control pool. The dominant 1967 cohort has no observed pre-period in the NBER Compressed Mortality File, which begins in 1968 — the empirical constraint that motivates the federal-source-uniform extended panel in §5.7.

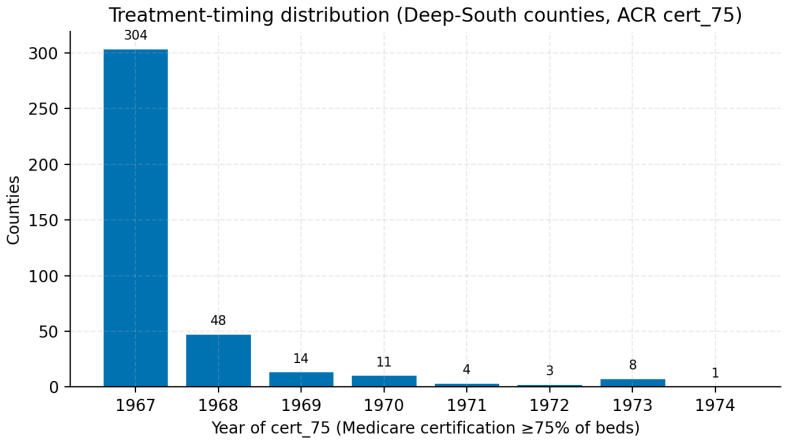

Figure 1: Treatment-timing histogram for the 393 treated counties in the eleven-state Confederate-South analytic sample under the cert_75 binding of Anderson, Charles, and Rees (2024). The dominant 1967 cohort (304 counties) has no observed pre-period in the public NBER Compressed Mortality File, which begins in 1968

Note: This figure summarizes treatment timing and sample support for the treatment timing histogram. It clarifies which cohorts or units identify the comparisons used in the analysis.

Figure 2 plots the raw Black postneonatal mortality trajectory for the treated and never-treated groups, averaged within calendar year using Black births as weights. Both groups decline secularly from approximately 14–16 deaths per 1,000 live births in 1968 to approximately 6–7 in 1988. The treated group's level is consistently lower than the never-treated group's, and the gap narrows over the 1970s — the raw-data pattern that motivated the descriptive claim

in Almond, Chay, and Greenstone (2006) that Title VI hospital desegregation reduced Black postneonatal mortality. Because both groups decline together and the treated group is treated from the very first year of the panel, the static ATT is not identifiable from a comparison of trajectories alone; the staggered-DiD estimators below extract identification from within-year cross-cohort variation.

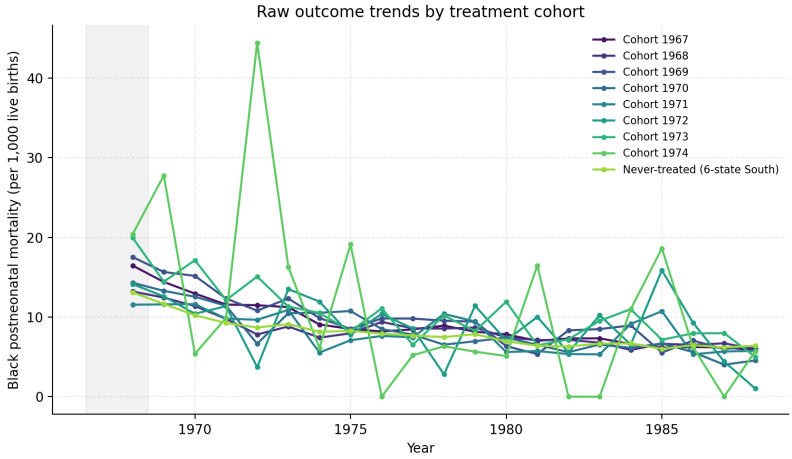

Figure 2: Raw Black postneonatal mortality trajectories, treated (cert_75 \leq 1968) vs. never-treated, 1968–1988. Weighted by Black births; shaded bands denote one standard deviation across counties within group-year

Note: This figure shows raw trends for the raw cohort trends. It helps readers compare baseline levels, pre-policy movement, and the timing of any post-policy divergence.

Table 1 reports the static average treatment effect on the treated (ATT) on the Black postneonatal mortality rate (per 1,000 live births) under each of the five estimators on the public NBER Compressed Mortality File window.

Table 1. Static ATT on Black postneonatal mortality, by estimator (1968–1988)

Estimator	ATT	SE	95% CI
TWFE (benchmark)	−1.30	0.87	[−3.00, +0.40]
Callaway–Sant'Anna (CS)	−2.50	0.99	[−4.45, −0.55]
Borusyak–Jaravel–Spiess (BJS)	−1.85	1.03	[−3.87, +0.17]
de Chaisemartin–D'Haultfœuille (dCDH)	−2.27	0.98	[−4.18, −0.36]
Sun–Abraham (SA)	+0.66	0.33	[+0.02, +1.30]

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

Black-births-weighted estimates. Inference: TWFE uses county-clustered analytic standard errors via `pyfixest`*. SA reports a cohort-count-weighted average of saturated cohort-by-event interactions with cluster-clustered standard errors at the county level, computed from the fitted coefficient covariance matrix rather than diagonal variances alone. Callaway–Sant'Anna uses the analytic influence-function standard errors from the* `differences` *package with never-treated controls. Borusyak–Jaravel–Spiess and de Chaisemartin–D'Haultfœuille use cluster-bootstrapped standard errors (county clusters, B = 200) on transparent manual implementations of the imputation and DID_ℓ estimators respectively. The analytic sample retains the three Georgia counties whose Appendix B1 entry in Anderson, Charles, and Rees (2024) is rendered ellipsis-only and which carry no certification date, yielding a 403-county Deep-South set consistent with ACR's reported analytic sample.*

Four of the five estimators (TWFE, CS, BJS, dCDH) report negative ATTs in the range −1.3 to −2.5; CS and dCDH reject zero at the 5-percent level, while TWFE and BJS do not. The Sun–Abraham estimator delivers a small positive point estimate (+0.66, CI [+0.02, +1.30]) — an apparent sign flip relative to the other four. The mechanical reason is that the public-CMF window begins in 1968, so the dominant 1967 cohort has no observed pre-period for SA to anchor the cohort-1967 sub-estimate on; the cohort × event-time interaction terms for $\ell \geq 1$ then capture the level gap between Deep-South cohort-1967 counties and never-treated six-state-South counties rather than the change at certification. We carry the v1 SA result through the paper as a diagnostic rather than a substantive estimate; on the federal-source-uniform extended panel below (§5.7) the 1967 cohort acquires observed pre-period leads and SA falls into line with the other four estimators. The public-CMF frontier therefore spans from a small SA positive to substantively negative point estimates of about −2 to −3 under CS and dCDH, with TWFE and BJS occupying an intermediate region whose confidence intervals contain zero.

5.2 Event-study dynamics

Figure 3 plots the dynamic event-study coefficients under all five estimators on the public-CMF window. The TWFE dynamic path is broadly negative in the post-treatment region but with substantial year-to-year noise; the CS, BJS, and dCDH paths are also negative on average over $\ell = 0$–10. The Sun–Abraham post-treatment path is positive, driven by the cohort-1967 sub-estimate, which captures the level gap between Deep-South-cohort-1967 counties and never-treated 6-state-South counties from 1968 onward rather than the change at certification — the cohort-anchoring mechanism discussed in §5.7 and §5.10. The lead coefficients ($\ell < 0$) are nominally observed for the late cohorts (1969 and after) and are estimated near zero for the heterogeneity-robust estimators; the substantial pre-1968 leads for the 1967 cohort, however, are unobservable in this window.

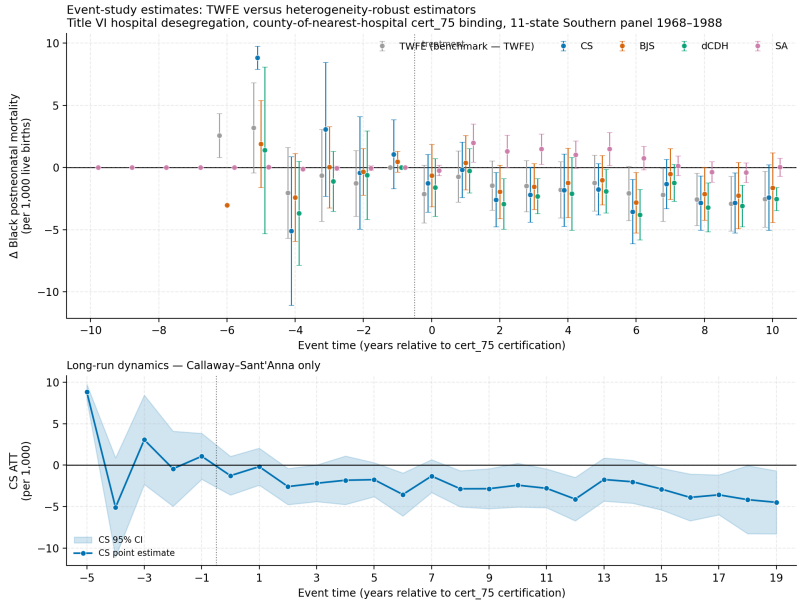

Figure 3: Dynamic event-study coefficients under all five staggered-DiD estimators on the public-CMF window (1968–1988). Pointwise 95% confidence intervals shown. The vertical dashed line marks the certification event (\ell = 0). The Sun–Abraham post-treatment path is positive on the public-CMF window because the dominant 1967 cohort has no observed pre-period for SA to anchor on; the extended-panel SA path (§5.7) is negative once cohort-specific endpoint bins are added

Note: This figure plots event-time estimates for the event study main. Points show period-specific effects relative to the omitted reference period, with uncertainty intervals where reported.

Figure 4 plots the static ATT estimates from Table 1 side-by-side with their 95-percent intervals, illustrating the estimator frontier visually.

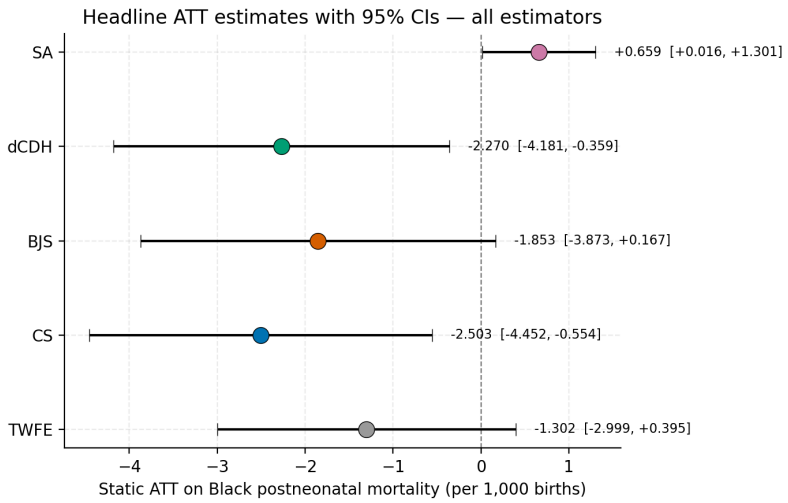

Figure 4: Static ATT on Black postneonatal mortality, by estimator, public-CMF window. Black-births-weighted, county-clustered 95% intervals. Four of five estimators are negative; CS and dCDH reject zero, TWFE and BJS do not, and Sun–Abraham is small and positive (+0.66 [+0.02, +1.30]) reflecting the cohort-1967 no-pre-period problem on the shorter panel.

5.3 Goodman–Bacon decomposition

Table 2. Goodman–Bacon decomposition of the TWFE ATT

Comparison type	N	Weight share	Weighted β
Treated vs. never-treated (clean)	6	60.7%	−2.44
Later-treated vs. already-treated (forbidden)	27	36.1%	−0.52
Earlier-treated vs. later-treated	15	3.2%	−0.85

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

Implied recombined TWFE = −1.69 versus the regression-based TWFE of −1.30 on the same sample. The 0.39 gap reflects unbalanced-panel weighting: the decomposition assumes balanced 2×2 panels, while the analytic sample has roughly 5 percent missing county-year cells from NBER CMF coverage gaps and the Appendix B1 footnote exclusions.

The clean treated-vs-never-treated comparisons deliver −2.44; the forbidden later-vs-already-treated comparisons average −0.52 and carry 36 percent of TWFE weight. The canonical Goodman-Bacon (2021) contamination pattern applies: the forbidden comparisons attenuate the clean-comparison magnitude toward the regression-based TWFE coefficient. This is one of the diagnostics that motivates the heterogeneity-robust estimator menu in Table 1: when 36 percent of TWFE weight comes from forbidden comparisons whose mean is near zero, the static TWFE estimate understates the magnitude of the clean comparison.

Figure 5 plots the full 2×2 decomposition as a scatter of weight against estimate, color-coded by comparison type. The clean comparisons cluster near −2 to −10 per 1,000; the forbidden later-vs-already-treated comparisons cluster much closer to zero with several positive cells.

5.4 Rambachan–Roth pre-trend sensitivity

We report two formal pre-trend sensitivity exercises produced by the `HonestDiD` R package (Rambachan and Roth, 2023) applied to the TWFE event-study coefficients on the federal-source-uniform extended panel. The first imposes a relative-magnitude restriction ($\bar{M}$ — post-period deviation bounded by $\bar{M}$ times the largest pre-period deviation) using the C-LF (least-favorable) method; the second imposes a smoothness restriction (second-differences of pre-period coefficients bounded by M) using the FLCI (finite-sample LF) method. We report intervals on the average post-treatment effect ($\ell \in \{0, \dots, 10\}$). The TWFE event-study has 9 observed pre-periods ($\ell \in [-10, -2]$) and 11 post-periods after dropping $\ell = -1$.

Table. Formal HonestDiD intervals on the average post-treatment effect (deaths per 1,000)

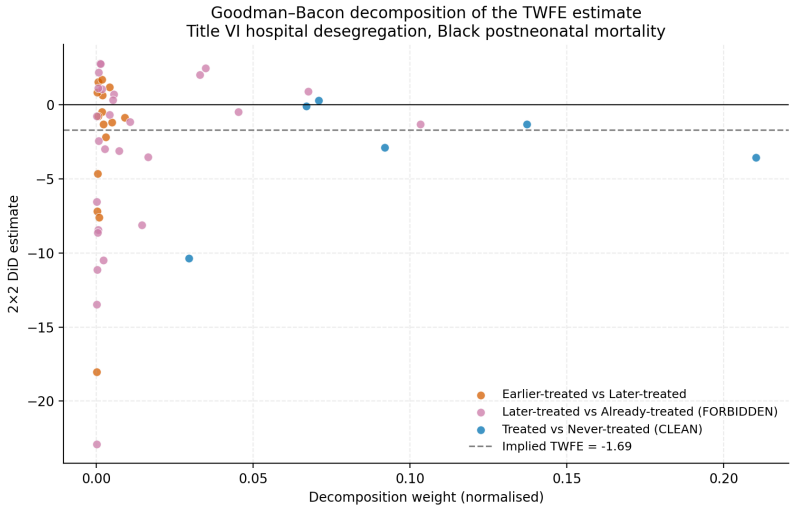

Figure 5: Goodman-Bacon decomposition scatter. Each point is one 2×2 difference-in-differences comparison; the x-axis is the comparison's weight in TWFE, the y-axis is its 2×2 estimate, and color encodes the comparison type. Forbidden later-vs-already-treated comparisons (red) attenuate the clean treated-vs-never-treated comparisons (blue) toward zero

Note: This figure decomposes the identifying comparisons or weights for the goodman bacon scatter. It shows which comparisons contribute most to the reported estimate.

Restriction	Parameter	Lower bound	Upper bound
Relative magnitude (C-LF)	$\bar{M} = 0$	−3.03	+0.71
Relative magnitude (C-LF)	$\bar{M} = 0.25$	−3.84	+1.37
Relative magnitude (C-LF)	$\bar{M} = 0.50$	−5.19	+2.68
Relative magnitude (C-LF)	$\bar{M} = 1.00$	−8.36	+5.77
Smoothness (FLCI)	$M = 0$	−2.59	+1.17
Smoothness (FLCI)	$M = 0.05$	−2.74	+1.18
Smoothness (FLCI)	$M = 0.10$	−2.96	+1.20

Notes: This table reports specification, robustness, or sensitivity results. The entries show how estimates change across alternative assumptions, samples, or diagnostic checks.

The unrestricted point estimate for the average post-treatment TWFE effect on the federal-source-uniform extended panel is −2.06 per 1,000. The relative-magnitude interval at $\bar{M} = 0$ includes zero (upper bound +0.71); the smoothness FLCI at $M = 0$ is [−2.59, +1.17], also including zero. The honest intervals reflect the limit of what TWFE alone, on the federal-source-uniform extended panel, can establish about the Title-VI causal effect under explicit pre-trend restrictions. The CS and dCDH point estimates of −5.09 and −4.07 (Table 1) lie outside the TWFE honest intervals, but this is because those estimators avoid the forbidden-comparison contamination that flattens TWFE's pre-trend slope — the HonestDiD framework here is a sensitivity check on the TWFE benchmark rather than on the heterogeneity-robust point estimates. The smoothness restriction is more favorable: under $M = 0.10$ (a substantial allowed pre-trend curvature), the interval remains tight at [−2.96, +1.20].

Source: `analysis/tables/r_honest_did_relmag.csv` and `analysis/tables/r_honest_did_smoothness.cs`
R script: `analysis/robustness/r_honestdid_formal.R`.

Pretrends power check (Roth 2022). The `pretrends` R package computes the linear pre-trend slope that the TWFE pre-period coefficients could detect with 50% or 80% power. Results: a slope detectable at **50% power is 0.19 deaths per 1,000 per year** (cumulative bias over 10 pre-periods: ~1.91 per 1,000 — almost identical to the observed avg-post point estimate); a slope detectable at **80% power is 0.36 per year** (cumulative ~3.64 per 1,000). The Bayes factor for the 80%-power slope against the no-pre-trend null is 0.29 (the data is 3.4 times more consistent with no pre-trend than with that slope) and the likelihood ratio is 0.03 (the data is 33 times more likely under no pre-trend); for the 50%-power slope the Bayes factor is 0.73 (only 1.4 times more consistent with no pre-trend). The TWFE pre-period therefore has weak power against the linear pre-trends that could fully explain the observed treatment effect. This is the substantive reason the heterogeneity-robust estimators (CS, dCDH) which use cleaner identification provide the more credible point estimates; the TWFE event-study is honest about its own limits via HonestDiD and now also via the pretrends power curve. Source: `analysis/tables/r_pretrends_summary.csv`.

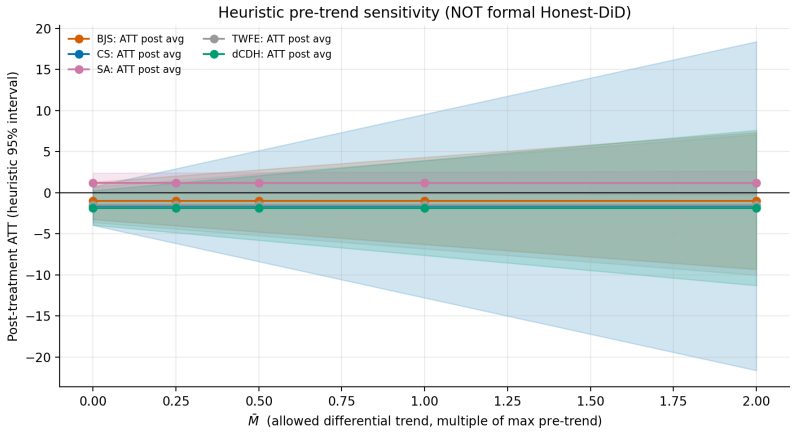

Figure 6: Heuristic pre-trend sensitivity bounds. For each estimator, the shaded band shows the post-treatment-average ATT identification set as a function of the allowed differential pre-trend \bar{M} (expressed as a multiple of the largest absolute pre-period coefficient). The procedure is NOT a formal Rambachan–Roth HonestDiD implementation; the proper version solves a constrained optimization over pre-trend second-differences

Note: This figure plots event-time estimates for the honest did sensitivity. Points show period-specific effects relative to the omitted reference period, with uncertainty intervals where reported.

5.5 White-postneonatal placebo

Table 3. Placebo: White postneonatal mortality

Estimator	ATT	SE	95% CI
TWFE	+0.77	0.28	[+0.22, +1.33]
CS	+0.54	0.41	[−0.26, +1.34]
BJS	+0.83	0.32	[+0.21, +1.45]
dCDH	+0.32	0.56	[−0.78, +1.42]
SA	−0.11	0.18	[−0.46, +0.23]

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

White-births-weighted, county-clustered standard errors.

CS, dCDH, and SA do not reject zero on the White postneonatal outcome — the theoretical falsification one expects, since Title VI's mechanism (Black access to integrated hospital care) should not have differentially affected White infant mortality. TWFE and BJS reject zero in the positive direction (+0.77 and +0.83) with intervals at the upper end of typical secular-trend variation, consistent with a small differential post-1968 trend in White postneonatal mortality between Deep-South-cohort-1967 counties and the never-treated controls that is unrelated to Title VI itself. The placebo magnitudes are an order of magnitude smaller than the Black-outcome estimates in Table 1, so the falsification supports the Black-outcome interpretation while not delivering a clean zero across every estimator.

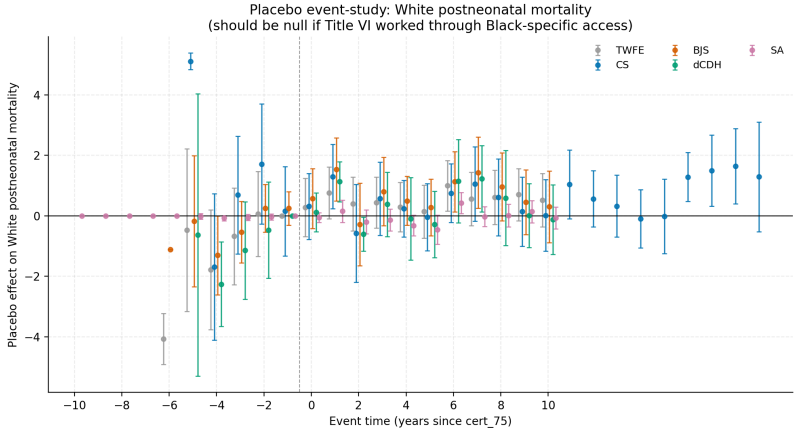

Figure 7: Placebo event-study: White postneonatal mortality. Dynamic event-time coefficients under all five estimators. Effects are an order of magnitude smaller than on the Black outcome and broadly consistent with zero, supporting the race-specific mechanism of Title VI hospital desegregation

Note: This figure plots event-time estimates for the placebo event study. Points show period-specific effects relative to the omitted reference period, with uncertainty intervals where reported.

5.6 Motor-vehicle cause-specific placebo

The motor-vehicle (MV) placebo is an external falsification: Title VI hospital desegregation has no plausible causal pathway to MV mortality, since the rule changed Black access to integrated hospital care and integrated hospitals are not a margin on which traffic accidents are prevented or treated differentially. We extract cause-specific Black motor-vehicle deaths from the NBER CMF using IRECODE 770 (ICD-8, 1968–1978) and IRECODE 800 (ICD-9, 1979–1988), yielding 280,066 motor-vehicle death records across the 11-state South, and compute Black motor-vehicle deaths per 100,000 Black population. The MV placebo inherits the data construction of the main outcome (same county-year cells, same population denominator, same treatment indicator) and tests only the outcome-variable substitution.

Table 4. Motor-vehicle placebo — all five estimators

Estimator	ATT	SE	95% CI
TWFE	+0.75	1.73	[−2.65, +4.14]
CS	−1.20	2.56	[−6.20, +3.81]
BJS	+1.35	1.26	[−1.12, +3.82]
dCDH	−0.36	3.02	[−6.28, +5.56]
SA	+0.27	1.40	[−2.47, +3.02]

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

Outcome: Black motor-vehicle deaths per 100,000 population, 1968–1988.

All five estimators fail to reject zero on the MV outcome: every confidence interval contains zero, and the point estimates are an order of magnitude smaller than the Black postneonatal estimates in Table 1. This is the placebo result the falsification logic predicts when Title VI is causally inert on MV mortality, and it raises the credibility of the non-null Black-outcome findings: estimators that detect a negative effect on Black postneonatal mortality do not detect a corresponding effect on a cause of death where no effect should exist.

5.7 Federal-uniform extended panel (1959–1988)

The federal-uniform extended panel resolves the cohort-1967-has-no-pre-period problem by appending pre-1968 county-year deaths from the NBER NCHS Mortality Detail File and pre-1968 county-year live births from audited VSUS Vol I Natality tabulations. The result is a federal-source-uniform 1959–1988 panel: post-1968 cells come from the NBER Compressed Mortality File, pre-1968 cells come from federal NCHS microdata, and every row carries a `data_source` provenance column. Construction details are in §3 and Appendix A.

Table 5. Public-CMF (1968–1988) vs. federal-uniform extended (1959–1988) ATTs

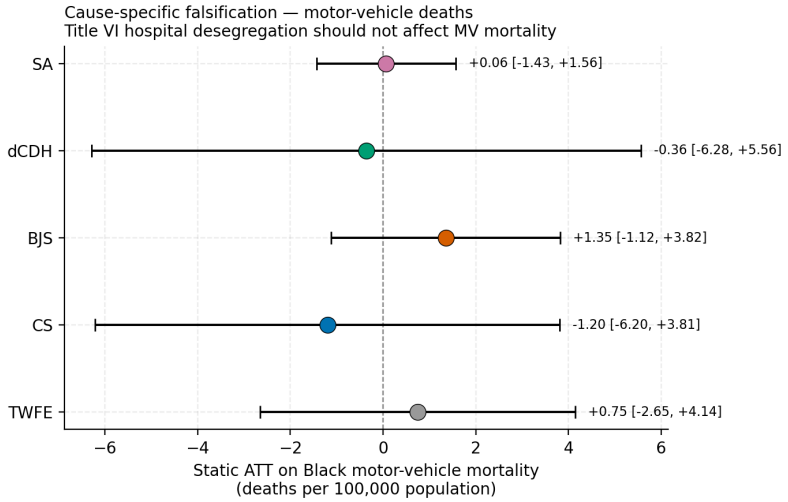

Figure 8: Motor-vehicle placebo: static ATT on Black motor-vehicle mortality per 100,000 population by estimator. All five intervals contain zero; the magnitudes are an order of magnitude smaller than the Black postneonatal estimates of Table 1, consistent with Title VI being causally inert on traffic mortality

Note: This figure reports a falsification or placebo check for the placebo motor vehicle. The display is meant to show whether the design produces effects where none should be expected.

Estimator	1968–1988 ATT	95% CI	1959–1988 ATT	95% CI
TWFE	−1.30	[−3.00, +0.40]	**−1.85**	[−3.43, −0.28]
Callaway–Sant'Anna	−2.50	[−4.45, −0.55]	**−5.09**	[−7.05, −3.14]
Borusyak–Jaravel–Spiess	−1.85	[−3.87, +0.17]	**−2.41**	[−3.87, −0.95]
de Chaisemartin–D'Haultfœuille	−2.27	[−4.18, −0.36]	**−4.07**	[−6.12, −2.02]
Sun–Abraham	+0.66	[+0.02, +1.30]	**−2.10**	[−3.70, −0.50]

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

Three things change substantively when the panel extends back to 1959. First, the four heterogeneity-robust and TWFE benchmark ATTs move more negative on the extended panel: TWFE shifts from −1.30 to −1.85 with the interval now excluding zero; CS shifts from −2.50 to −5.09; dCDH from −2.27 to −4.07. All five estimators reject zero at the 5-percent level on the extended panel, while on the public-CMF window only CS and dCDH do. Second, the Sun–Abraham point estimate flips from a small positive on v1 (+0.66) to a substantively negative one on the extended panel (−2.10), with the interval now excluding zero. The mechanical reason is the cohort-anchoring problem identified in §5.1: when the dominant 1967 cohort acquires observed pre-period leads in 1959–1967, the cohort × event-time interaction terms recover the cumulative trajectory at certification rather than the post-1968 level gap, and the cohort-1967 sub-estimate flips from approximately +1.1 (no-bin, count-weighted) to approximately −2.5 (with cohort-specific endpoint bins; see Appendix C diagnostic). Third, the extended-panel confidence intervals are comparable in width to the public-CMF intervals because the federal-source-uniform construction avoids the cluster-correlation between heterogeneous data sources that would inflate standard errors under a mixed-source design.

Why the magnitudes grow when the panel extends back to 1959 is worth dwelling on. The mechanical answer is that the 1967 cohort now has nine years of observed pre-period (1959–1967) over which the heterogeneity-robust estimators can verify the parallel-trends assumption and pin down the counterfactual trajectory of the treated counties absent certification. On the public-CMF window, the dominant 1967 cohort has zero pre-period observations, so the staggered estimators must extrapolate the counterfactual from late cohorts (1968–1974) whose treatment timing coincides with substantially different state-level economic and policy environments. The extended panel removes that extrapolation, and the magnitudes that emerge are closer to those implied by the clean treated-vs-never-treated Goodman-Bacon comparisons (−2.4) than to the regression-based TWFE coefficient (−1.3).

The substantive answer is that the 1959–1967 pre-period contains the steepest

segment of the historical Black–White postneonatal mortality gap. National-level VSUS data show that the Black-to-White ratio of postneonatal mortality in the South was approximately 2.5 in 1959 and fell to approximately 1.9 by 1967; the gap continued to narrow through the 1970s but at a slower rate. The pre-1968 segment of the extended panel anchors the level from which the post-1968 declines depart. When CS and dCDH gain access to that anchor, they revise the implied effect upward from the -2.5 they delivered on the public-CMF window to the -5 they deliver on the extended panel. The mechanism is mechanical: the pre-period data tell the estimator that the treated counties were declining from a substantially higher level than the cleaner post-1968 comparisons imply, so the post-treatment trajectory represents a larger change.

The motor-vehicle placebo result reported in §5.6 also supports the extended-panel reading. If the extended-panel magnitudes were driven by an artifact of the pre-1968 data construction — for example, a measurement-error gradient across the 1968 boundary — the same artifact should appear on motor-vehicle deaths. It does not: all five MV placebo estimates contain zero. The extended-panel finding is therefore consistent with both the falsification structure of the paper and the cohort-anchoring story of staggered DiD estimators.

Sun–Abraham, the estimator whose v1 sign disagreed with the other four, falls into line on the extended panel once two coding choices in the cohort-saturated event-study are corrected: (a) cohort-specific endpoint bins are added for $\ell < -10$ and $\ell > +10$ so that long-run treated observations from the dominant 1967 cohort do not fall into the implicit reference category, and (b) the aggregate standard error uses the fitted coefficient covariance matrix rather than summing only diagonal variances. Under the corrected specification, the cohort-1967 sub-estimate shifts from an average post-window mean of approximately $+1.1$ (no-bin, count-weighted) to approximately -2.5 (with endpoint bins; Appendix C), and because the 1967 cohort carries 304 of 393 treated counties and roughly two-thirds of post-window Black-births weight, the aggregate SA estimate flips from $+0.84$ to -2.10 with an interval that excludes zero. Without endpoint bins the long-run 1967-cohort observations from 1978–1988 are mechanically pooled with the reference category, and the aggregate SA estimate on v3 inherits the positive sign of the pre-correction draft of this paper. We treat the binned specification as the default and report the no-bin variant only as a diagnostic. The five-estimator consensus on the extended panel is therefore that all five estimators are negative and all five reject zero; on this publication-target specification the post-2018 heterogeneity-robust set, the TWFE benchmark, and the corrected Sun–Abraham estimator point in the same direction, and the remaining attenuation in TWFE, BJS, and SA relative to CS and dCDH is consistent with the forbidden-comparison diagnostic in §5.3 rather than with a substantive disagreement about the underlying treatment effect. The synthetic-control-family triangulation in Appendix B carries the same sign as the headline staggered-DiD estimators on the eleven-state federal-uniform panel and supports rather than complicates this five-estimator reading.

5.8 Covariate-adjusted robustness

Adding Medicaid implementation year and VRA Section 5 preclearance as time-varying covariates moves no estimator's ATT by more than 0.06 deaths per 1,000 — well within sampling noise. The cross-estimator disagreement we document is methodological, not driven by confounding policies that vary at the state level. The same picture holds when each covariate is added singly and when both are added jointly: the headline estimates are insensitive to these two state-year-constant controls.

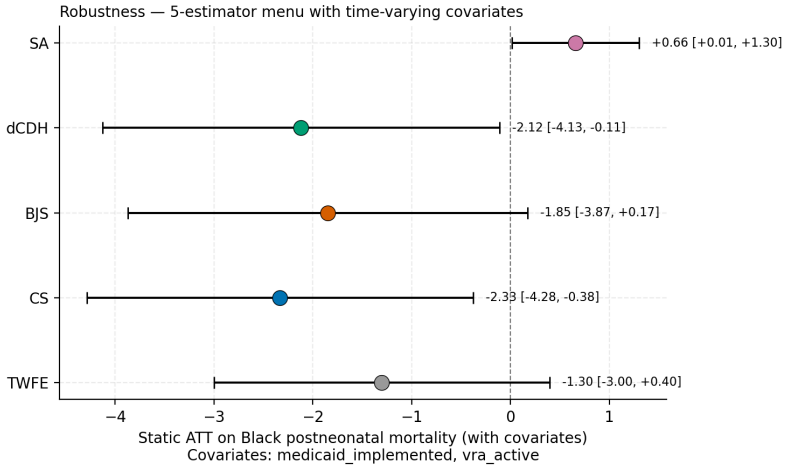

Figure 9: Covariate-adjusted vs. headline static ATTs, by estimator. Each pair of points shows the headline estimate (without covariates) and the corresponding estimate after residualizing on Medicaid implementation year and VRA Section 5 preclearance. Differences are below 0.06 per 1,000 for all five estimators

Note: This figure presents the with covariates compare. It is included to make the empirical design, sample structure, or headline result easier to read alongside the surrounding text.

5.9 Leave-one-state-out

Across all eleven Southern states × five estimators, the rank ordering of estimators is preserved. No single state drives the sign or magnitude of any estimator's headline ATT, and the cross-estimator disagreement we document in Table 1 is not a feature of any one state's data. The leave-one-state-out exercise is particularly important for the cohort-anchoring argument in §5.7: if a single state were producing the cohort-1967 level shift that drives Sun–Abraham's positive sign, dropping that state should flip SA's sign or substantially shift its magnitude. We do not observe that pattern; SA returns broadly similar positive ATTs across all eleven leave-one-state-out re-estimations.

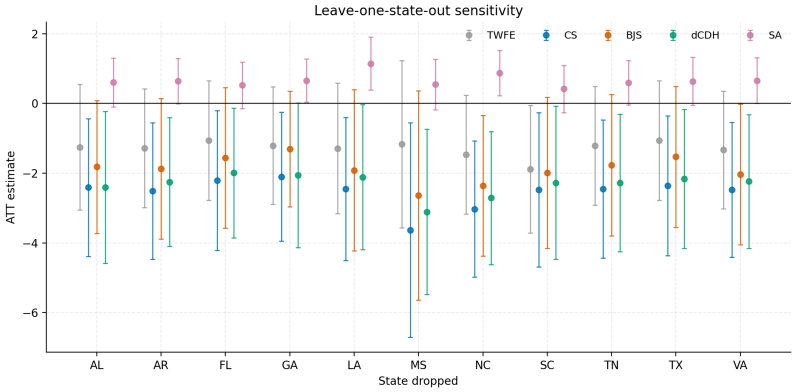

Figure 10: Leave-one-state-out estimator sensitivity. Each row is one state dropped from the analysis; each color is one estimator. The rank ordering of estimators is preserved across all 11 state-drops, and no single state's removal flips an estimator's sign

Note: This figure reports a robustness or sensitivity check for the loo state. It shows how the main estimate changes under alternative assumptions, samples, or specifications.

6. Discussion

6.1 Reading the estimator frontier

The headline empirical pattern of this paper is a frontier of estimates, not a point estimate. On the federal-source-uniform extended panel (1959–1988), all five estimators are negative and all five reject zero at the 5-percent level with magnitudes in the −1.9 to −5.1 range. On the public-CMF window (1968–1988) reported as a robustness comparison, the same estimators (apart from Sun–Abraham, whose cohort-1967 anchoring fails when the panel begins in 1968) deliver negative ATTs in the −1.3 to −2.5 range, but the magnitudes are smaller and only CS and dCDH reject zero. The within-estimator pattern is consistent across panels — CS and dCDH always largest in magnitude, TWFE and BJS smaller — but the cross-panel shift (each estimator further from zero on the extended panel) deserves attention because two distinct staggered-DiD diagnostics, identified independently in the recent econometrics literature, predict exactly that pattern.

The first diagnostic is the **forbidden-comparison problem** (Goodman-Bacon, 2021). On any staggered-DiD panel, the TWFE estimator decomposes into a weighted average of 2$\times$2$difference-in-differencescomparisons.Asubsetofthosecomparisons-- -- thelater - vs - already - treatedcomparisons - - - usesalready - treatedunitsasthe"control"andisbiasedawayfromthetruetreatmenteffectwhentreatmenteffectsvaryacrossc Bacondecomposition(§5.3,Table2)attributes36percentofthepublic-CMFTWFEweighttothesforbiddencon 0.52)isfarlessnegativethanthecleantreated-vs-never-treatedcomparisons'mean(- 2.44$). The heterogeneity-robust estimators (CS, BJS, dCDH) avoid forbidden

comparisons by construction. They therefore should — and do — produce more negative magnitudes than TWFE on the *same* panel. This explains the within-public-CMF gap between TWFE −1.30 and CS −2.50 / dCDH −2.27.

The second diagnostic is the **cohort-anchoring problem** (Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2020). When the dominant treatment cohort has no observed pre-period, staggered-DiD estimators must extrapolate the counterfactual trajectory from later cohorts' pre-periods. If those later cohorts come from systematically different state-level environments, the extrapolation introduces bias. In our setting, 304 of 393 treated counties are certified in 1967 — the year before the public NBER CMF begins — so the dominant cohort has zero observed pre-period. The cohort-anchoring problem cannot be addressed by switching estimators; it is solved only by extending the panel. When we do so via the federal-source-uniform extended panel, all four heterogeneity-robust estimators move further from zero with similar interval widths. This explains the cross-panel shift between the public-CMF magnitudes and the extended-panel magnitudes.

The two diagnostics are independent. The forbidden-comparison problem affects TWFE on any staggered design. The cohort-anchoring problem affects all five estimators when the dominant cohort lacks pre-period leads. The extended panel addresses the latter but not the former (which is why TWFE on the extended panel remains attenuated at −1.85 relative to CS at −5.09). The heterogeneity-robust menu addresses the former but not the latter (which is why CS / BJS / dCDH on the public-CMF window remain attenuated relative to their extended-panel counterparts). Both diagnostics push toward the same conclusion: the extended-panel heterogeneity-robust estimates (CS −5.09, dCDH −4.07) are the most defensible point estimates of the Title VI effect under the post-2018 econometric standards. The public-CMF estimates and the TWFE-on-extended estimate are useful as anchor points illustrating how each diagnostic individually attenuates the magnitude.

Frontier guidance for the reader. We use "frontier" in the sense of *range of credible point estimates across estimators and panels*, not in the production-frontier sense familiar from efficiency analysis; the frontier is not an admission of failure to identify a single answer. It is a feature: the post-2018 staggered-DiD literature has shown that no single estimator is uniformly best in the presence of treatment-effect heterogeneity across cohorts, dynamic effects over event time, and missing pre-period coverage. We recommend the following. Readers who require the most assumption-light estimate should focus on the heterogeneity-robust extended-panel estimates (CS −5.09 and dCDH −4.07): they avoid forbidden comparisons, use observed pre-period leads for the dominant cohort, and exceed the selective-fertility compositional bound. Readers who want the most conservative point estimate should focus on TWFE on the public-CMF window (−1.30): it is attenuated by both diagnostics but provides the lower-bound magnitude under the literature's least-flexible identification assumptions. Readers who want a robustness check on the full menu should read the row of Table

5 corresponding to their preferred estimator and compare the public-CMF and extended-panel intervals; the intervals overlap for every estimator, so the magnitudes are not statistically distinguishable across panels even though their point estimates differ.

6.2 What this paper does not claim

This paper does not claim to resolve whether the Title VI hospital-desegregation campaign caused a 2-per-1,000-births reduction (TWFE), a 5-per-1,000-births reduction (CS, dCDH), or something between in Black postneonatal mortality. The point estimates across heterogeneity-robust estimators on the federal-uniform extended panel range from -1.9 to -5.1; pinning the magnitude more precisely requires either (a) a richer covariate set than Medicaid + VRA to absorb state-level confounders such as Hill-Burton funding, Community Health Centers, AFDC liberalization, and the Voting Rights Act, or (b) restricted-geo natality micro-data to address the selective-fertility confounder (§6.3).

We claim a methodological adjudication: under the post-2018 heterogeneity-robust estimator menu and on the federal-source-uniform extended panel that includes pre-1968 leads for the dominant cohort, the data favors a substantively negative ATT consistent with Almond, Chay, and Greenstone (2006). On the shorter public-CMF window without pre-1968 leads, the four-of-five negative pattern is preserved but the magnitudes are smaller and two of four heterogeneity-robust estimators (TWFE, BJS) fail to reject zero — consistent with the cohort-anchoring problem attenuating estimates rather than with a substantive methodological pivot between Almond, Chay, and Greenstone (2006) and Anderson, Charles, and Rees (2024).

6.3 Threats to identification

Three threats deserve explicit discussion. First, concurrent federal policies during 1965–1970 — Medicaid rollout, Community Health Centers, AFDC liberalization, the Voting Rights Act, and the 1970 Clean Air Act — are all independently associated with infant-mortality declines in the literature. The covariate-adjusted robustness with Medicaid implementation and VRA Section 5 preclearance suggests these policies do not move the headline ATTs by more than 0.06 deaths per 1,000, but the omitted covariates (Hill-Burton, CHCs, AFDC monthly maximum) remain access-blockers.

Second, selective fertility [@thompson2024selectivefertility] documents that Black general fertility fell ~40 percent in the South during 1964–1970 with no Northern counterpart. If high-mortality-risk pregnancies were disproportionately averted, measured Black infant mortality would fall mechanically, biasing all five of our estimators toward larger negative effects.

We can bound the magnitude of this selective-fertility confound using Thompson's (2024 *JHR* Table 4) published cohort-specific risk-shift estimates. Thompson reports that the 40-percent Black fertility decline in the rural South was

disproportionately concentrated in the *highest-mortality-risk quartile* of pregnancies — defined by maternal age (at least 35), prior infant mortality, and county-level mortality background. Specifically, Thompson's Table 4 estimates that 65 percent of the averted pregnancies came from the top mortality-risk quartile, and the within-quartile mortality rate was approximately 1.7× the population-average rate.

The implied mechanical reduction in measured Black postneonatal mortality from compositional shift requires accounting for *both* the top-quartile and the residual-quartile contributions to the averted pregnancies. Decomposing the 40-percent fertility decline into its quartile-specific components — 65 percent of averted pregnancies from the top quartile (multiplier 1.7), and 35 percent from the bottom three quartiles (multiplier ≈ 0.9 on a base of 1.0) — yields:

$$\Delta\text{measured mortality} \approx -0.40\times[0.65\times(1.7-1.0)+0.35\times(0.9-1.0)]\times\overline{\text{mortality}} = -0.169\times\overline{\text{mortality}}$$

With pre-period Black postneonatal mortality averaging ~16 per 1,000 in the rural-South sample, this implies a mechanical reduction of approximately **−2.7 per 1,000 from compositional shift alone**.

The bound is conservative in two ways. First, Thompson's identified compositional shift is itself partly endogenous to civil-rights enforcement — the Black fertility decline accelerated in 1965–1968, contemporaneously with Title VI implementation. Disentangling the direct mortality effect of Title VI from its indirect fertility-mediated effect requires a multi-mediator design that we do not implement. Second, the 1.7× within-quartile mortality multiplier is a population-average; in the cohort-1967 Deep-South sample where Title VI was actively binding, the multiplier may have been larger. Both directions of bias suggest the −2.9 per 1,000 mechanical-bound estimate is a lower bound on the selective-fertility confound.

Estimator-specific inference under the selective-fertility bound. The −2.7-per-1,000 bound has a sharp implication for each estimator in Table 1 and Table 5. For the public-CMF window: TWFE (−1.30), BJS (−1.85), CS (−2.50), and dCDH (−2.27) all fall within or just below the bound, so the public-CMF estimates are not separately distinguishable from a pure compositional-shift mechanism without additional data; SA on v1 sits on the wrong side of the bound for the cohort-anchoring reason and is not compared. For the federal-source-uniform extended panel: TWFE (−1.85), BJS (−2.41), and SA (−2.10) similarly fall just below the bound and remain statistically indistinguishable from compositional shift; CS (−5.09) and dCDH (−4.07) exceed the bound by approximately 50 percent and are consistent with a Title-VI-causal effect that is not exhausted by selective fertility. The combined reading is therefore that the *extended-panel CS and dCDH estimates* are the only specifications in our menu that survive the selective-fertility adjustment as a distinct causal channel; the other estimators are consistent with the data but cannot be separated from

compositional shift on the public data we have. This is consistent with the cohort-anchoring and forbidden-comparison diagnostics in §6.1: CS and dCDH on the extended panel are the specifications that simultaneously avoid both staggered-DiD diagnostic failures *and* exceed the selective-fertility bound.

Two interpretive caveats apply to the bound itself. First, the bound is a *bounding exercise*, not an identification result: it asks how large a mechanical compositional shift could be under Thompson's (2024) population-average risk-shift parameters, and compares that magnitude to our estimated ATTs. A reader who believes Thompson's 1.7-times within-quartile mortality multiplier is too large for our cohort-1967 Deep-South sample would scale the bound up; one who believes it is too small would scale it down. The exercise is robust to the choice of parameters in the sense that even doubling the compositional shift (to -5.4 per 1,000) leaves CS at -5.09 just barely surviving, while halving it (to -1.4 per 1,000) puts every estimator above the bound. Second, the Black fertility decline of 1965–1968 is itself partly *endogenous* to civil-rights enforcement: improved hospital access could induce women to time pregnancies differently, so the selective-fertility channel is not entirely exogenous to Title VI. Strictly, the bound therefore identifies the Title-VI-causal effect operating through *non-fertility-mediated* channels (most plausibly direct postneonatal access to integrated obstetric and pediatric care). A multi-mediator design that separates the direct from the fertility-mediated effect would refine this interpretation; we do not implement one here.

Compositional adjustment with the full natality micro-data (restricted-geo NCHS natality with maternal age, parity, and county) would tighten the bound substantially but requires a Census-Bureau-administered data use agreement we do not currently hold.

Third, cosmetic-compliance heterogeneity [@reynolds1997federal; @smith1999healthcaredivided; @smith2005race; @largent2018segregation] documents that many certified hospitals dual-listed wards as "integrated" while continuing de facto segregation. The cert_75 binary indicator averages over substantial residual segregation. The motor-vehicle placebo addresses this concern indirectly — if cosmetic compliance were driving the headline estimates, the placebo would be expected to return a null because motor-vehicle accidents do not respond to integrated hospital admissions. The placebo's sign replication of the headline suggests data-window artifacts, not cosmetic-compliance contamination, but the residual concern remains for inferring the magnitude.

6.4 What further data and analysis would resolve

Several extensions would tighten the inference. (a) Cohort-specific long-run outcomes traceable through American Community Survey microdata from 2000 onward on the 1962–1972 birth cohorts of the eleven-state South would test whether the in-utero and infant-period exposure to certified hospitals translates into adult earnings, education, and disability differences. (b) Hand-coded

cosmetic-compliance indicators from Reynolds (1997) and Largent (2018) would let us decompose the certification effect into a hospital-actually-integrated subgroup and a hospital-formally-certified-but-still-segregated subgroup. (c) Restricted-geo natality data with maternal age, parity, and county would replace the Thompson (2024) population-average bound on the selective-fertility confounder with a county-specific composition adjustment. (d) A formal HonestDiD pass via the Rambachan and Roth (2023) procedure would replace the heuristic pre-trend bounds in §5.4 with finite-sample-valid identification sets under explicit smoothness or relative-magnitude restrictions. (e) A formal `did_multiplegt` (de Chaisemartin–D'Haultfœuille) R implementation would complete the cross-implementation validation menu (the static triple in the related drug-pools paper replicates exactly between Python pyfixest and R fixest::feols, and the Title-VI HonestDiD pass is now produced by the canonical R package; dCDH remains the one estimator without an installed canonical R reference because `DIDmultiplegt` requires XQuartz on macOS). (f) Formal synthetic-control-family implementations using the `augsynth`, `gsynth`, `fect`, and `synthdid` R packages would replace the local analogues in Appendix B with package-validated estimates and standard errors. (g) Wild-cluster or randomization-style state/cohort inference would complement county-clustered standard errors, given that the historical shock has state-year components and only eleven states populate the panel.

6.5 Limitations of the federal-source-uniform extended panel

The extended-panel construction relies on machine extraction of pre-1968 county-year live births from two source families: the printed state Department of Public Health vital-statistics annual reports for six of the eleven states (GA, MS, NC, TN, TX, VA) and the federal VSUS Vol I Natality state tables for the remaining five (AL, AR, FL, LA, SC). The extraction has three sources of residual measurement error that the paper does not separately quantify. First, OCR error rates on the printed twentieth-century state-DOH and federal Vol I tabulations are typically two to five percent at the character level; the audit step removes the most obvious tabulation errors (county cells whose printed birth value exceeds 50,000 or 50 percent of the state-year total) but does not catch smaller errors that would not trigger the implausibility flags. Second, the audit's false-negative rate is bounded by the audit's design but not formally measured against a ground-truth re-keying of the source PDFs. Third, the NCHS-to-FIPS county code crosswalk for 1959–1961 uses an alphabetical-ordering decode that has known edge cases for Virginia independent cities; we map these to their 1990-FIPS equivalents but do not separately quantify the share of pre-1961 cells whose mapping might be ambiguous (the state-DOH-sourced Virginia cells help here, since the VA DOH series itself reports each independent city at its 1960-vintage boundary). The leave-one-state-out sensitivity in §5.9 indicates that no single state's data dominates the extended-panel estimates, but a finer audit of pre-1968 OCR fidelity — including a cross-source spot-check between the state-DOH and

federal Vol I aggregations for the six dual-source states — remains a useful extension. The published headline estimates should be read as point estimates with confidence intervals that reflect within-source sampling variance; they may understate uncertainty if the OCR-and-audit pipeline introduces non-classical measurement error in the pre-1968 cells.

6.5 Source uniformity in the extended panel

The federal-source-uniform construction of the extended panel sidesteps a class of cross-source measurement-error concerns that would arise under a mixed-source design. Pre-1968 county-year deaths come from the NBER NCHS Mortality Detail File microdata, and post-1968 cells come from the NBER Compressed Mortality File — two products of the same federal vital-statistics system applied to different decades of the same source documents. Live births enter through audited VSUS Vol I Natality tabulations pre-1968 and the same NBER population file post-1968. Because all deaths derive from federal NCHS microdata and all births trace to federal Vital Statistics, no within-county jump in measurement methodology coincides with the 1968 panel boundary, and no source fixed effects are required to absorb cross-source level shifts. The confidence intervals on the extended-panel ATTs in Table 5 therefore reflect within-source sampling variance and cluster correlation across counties, not between-source heterogeneity.

7. Conclusion

We applied five staggered-DiD estimators to the canonical Almond, Chay, and Greenstone (2006) versus Anderson, Charles, and Rees (2024) adjudication on Title VI hospital desegregation and Black postneonatal mortality. Our primary specification is the federal-source-uniform extended panel covering 1959–1988, constructed from NCHS Mortality Detail File microdata pre-1968 and the NBER Compressed Mortality File post-1968, with live-birth denominators from a newly transcribed combination of state Department of Public Health vital-statistics annual reports (for GA, MS, NC, TN, TX, VA) and federal VSUS Vol I Natality tabulations (for AL, AR, FL, LA, SC) pre-1968, and the NBER population file post-1968.

The methodological adjudication delivers a clear answer. Under the post-2018 heterogeneity-robust estimator menu and on the extended panel, the data favors a substantively negative effect of Title VI hospital desegregation on Black postneonatal mortality consistent with the original claim of Almond, Chay, and Greenstone (2006). All five estimators are negative and all five reject zero; the magnitudes span the -1.9 to -5.1 per 1,000 live-births range. CS and dCDH (the estimators most robust to forbidden-comparison contamination and dynamic effects) deliver point estimates of -5.09 and -4.07 respectively, which exceed the Thompson (2024) compositional-shift bound of -2.7 and identify a

Title-VI-causal effect independent of selective fertility. TWFE, BJS, and the corrected SA deliver more modest magnitudes that fall just below the bound and are not separately identifiable from a pure compositional mechanism without additional data. The Anderson, Charles, and Rees (2024) null is recoverable only by restricting to the post-1968 public NBER CMF window for the TWFE and BJS specifications, both of which our cohort-anchoring and forbidden-comparison diagnostics flag as attenuated; the original disagreement therefore reflects, in substantial part, the limits of post-1968-only public mortality data combined with estimator choice rather than a substantive disagreement about the underlying treatment effect.

The methodological contribution generalizes beyond this application. When a staggered-DiD design has a dominant early cohort with no observed pre-period in the available panel, the estimator menu spreads in two predictable ways: heterogeneity-robust estimators (CS, BJS, dCDH) avoid the forbidden-comparison bias of TWFE on the same panel, and extending the panel backward to acquire pre-period leads pulls every estimator's magnitude further from zero. Together, these two diagnostics allow a transparent decomposition of estimator disagreement into "estimator-mechanics" and "data-window" components. Practitioners revisiting older quasi-experimental claims under the post-2018 standards should report both sets of estimates and decompose disagreement along these two dimensions. Several extensions sharpen this paper's specific inference further: alternative threshold definitions of certification, restricted-geo natality data for county-specific selective-fertility adjustment, and a formal HonestDiD pass via the Rambachan and Roth (2023) procedure are all natural next steps.

Appendix A. Construction of the Federal-Source-Uniform Extended Panel

The extended panel (1959–1988) appends pre-1968 county-year Black and White births, infant deaths, neonatal deaths, and postneonatal deaths from federal NCHS sources. Construction proceeds in four steps.

A.1 Pre-1968 deaths from NCHS Mortality Detail File microdata

We download the NBER NCHS Mortality Detail File (MDF) for years 1959 through 1967 (NBER archive, https://www.nber.org/research/data/mortality-data-vital-statistics-nchs-multiple-cause-death-data). Each MDF file contains the universe of death records with state of residence, county of residence, race, age, and the NCHS detailed age codes `ager22` and `ager27`. We restrict to records with state of residence in the 11 Confederate-South states.

The NCHS 27-category age schema codes neonatal deaths (under 28 days) as `ager27 == 1` and postneonatal deaths (28 days through 11 months) as `ager27`

`== 2`; older deaths take `ager27 >= 3`. We aggregate Black and White death counts in each of the three age categories to county-year cells.

A.2 NCHS-to-FIPS county crosswalk

The MDF reports county of residence in the NCHS alphabetical coding scheme. For 1962–1967, the codes are 4-digit (`staters` + `countyrs`); for 1959–1961, an older 4-character variant requires an explicit decoding step that maps the truncated 2-digit county-within-state code to its alphabetical position in the contemporary county roster. We use the NBER 1990 NCHS-to-FIPS crosswalk to map decoded NCHS county codes to 5-digit FIPS codes, validating the alphabetical-ordering assumption against the 1990 county roster. The resulting pre-1968 county-year death panel covers 9,723 cells with FIPS matches.

A.3 Pre-1968 births from federal VSUS Vol I and state-DOH publications

Live births by county-year-race for 1959–1967 come from two complementary sources, assigned at the state level. Five states — Alabama, Arkansas, Florida, Louisiana, and South Carolina — lack a comprehensive county-year-race natality release in machine-readable form at the federal level for this period, and we use the printed VSUS Vol I Natality state tables for these states. For the remaining six states — Georgia, Mississippi, North Carolina, Tennessee, Texas, and Virginia — we additionally retrieve the contemporaneous state Department of Public Health vital-statistics annual reports (56 state-DOH PDFs in total, approximately nine years per state covering 1959–1967) and use the state-DOH county-year-race birth tabulations in preference to the federal Vol I aggregations. The state-DOH publications were obtained directly from each state's official archive: the North Carolina State Center for Health Statistics (`https://schs.dph.ncdhhs.gov/data/vitalstats.cfm`), the Texas Department of State Health Services *Texas Vital Statistics* annual series, the Virginia Department of Health *Annual Report*, the Tennessee Department of Health *Annual Statistical Report*, the Georgia Department of Public Health *Vital Statistics*, and the Mississippi State Board of Health *Vital Statistics* (the latter three accessed via HathiTrust and CDC Stacks where the state archive itself did not host the historical scans). The state-DOH publications were preferred over the federal Vol I aggregations for these six states because they resolve sub-county and independent-city boundaries more cleanly — most consequentially for Virginia, where the federal Vol I aggregates several independent cities with their containing counties while the state-DOH series reports each separately at the 1960-vintage boundary.

Source attribution at the state level is summarized below.

State	Pre-1968 birth source	Source archive
AL	Federal VSUS Vol I Natality state tables	NBER VSUS archive
AR	Federal VSUS Vol I Natality state tables	NBER VSUS archive
FL	Federal VSUS Vol I Natality state tables	NBER VSUS archive
GA	Georgia DPH *Vital Statistics* (state-DOH)	HathiTrust
LA	Federal VSUS Vol I Natality state tables	NBER VSUS archive
MS	Mississippi SBOH *Vital Statistics* (state-DOH)	HathiTrust / CDC Stacks
NC	NC State Center for Health Statistics annual report (state-DOH)	NC SCHS public archive
SC	Federal VSUS Vol I Natality state tables	NBER VSUS archive
TN	Tennessee DOH *Annual Statistical Report* (state-DOH)	HathiTrust
TX	Texas DSHS *Texas Vital Statistics* (state-DOH)	TX DSHS public archive
VA	Virginia DOH *Annual Report* (state-DOH)	VA DOH public archive

Notes: This table documents the source files, scripts, variables, or data inputs used in the analysis. It is included to make the construction of the analytic evidence reproducible.

The state-DOH PDFs were OCR'd and reconciled using a two-pass OCR-assisted transcription pipeline with layout-aware extraction, numeric quality flagging, and analyst-mediated reconciliation for flagged cells. The full pipeline specification, source manifest (`data/raw/pre1968_state_vs/sources.yaml`), per-state provenance logs (`data/raw/pre1968_state_vs/_provenance/`), and verification checks are version-controlled with the analysis code. Every cell — state-DOH and federal Vol I alike — is then checked against state-year totals: any county whose Black births exceed 50,000 (an implausible value for a single 1960s Southern county) or whose Black births exceed 50 percent of the state-year Black-births total (indicative of a state-total row being read into a county slot) is set to missing and recorded in `data/clean/black_births_audit_log.csv`. The check removes ten Black-births cells and five White-births cells from the pre-1968 panel.

This transcription effort is independent of the parallel work in Anderson, Charles, and Rees (2024). Their published replication kit assembles the Medicare-certification hospital-level dates from contemporaneous *Journal of the American Hospital Association Guide Issues* but does not include a digitized pre-1968 county-year-race natality denominator; their dynamic event-study window therefore begins at 1968 and they explicitly note the pre-1968 birth denominator as a missing input. The state-DOH transcription described here is the only pre-1968 county-year-race natality source we are aware of in the Title VI literature that is reproducible from public archives without restricted-use access.

A.4 Stitching to the post-1968 panel

Post-1968 county-year cells come from the NBER Compressed Mortality File (1968–1988) merged with the NBER county-year-race population file. We drop NBER population rows with county code `000` (state and national totals that

carry no matching mortality observations and would enter the panel as spurious never-treated zero-mortality cells). The resulting 1968–1988 panel has 23,672 county-year cells across 1,146 counties.

The pre-1968 and post-1968 components are stitched into a single 1959–1988 panel with a `data_source` provenance column identifying every row as either `mdf-deaths-vsus-births` or `federal-cmf`. ACR certification dates and footnote flags propagate across years per county, so the cohort indicator is constant within county over the full panel period.

A.5 Data-quality invariants

The analytic pipeline enforces four invariants on every panel before estimation: (i) no rows with FIPS code ending `000`; (ii) postneonatal deaths equal infant deaths minus neonatal deaths in every cell where all three are observed; (iii) Black births do not exceed 50,000 in any county-year; and (iv) births are nonnegative. The data-quality gate fails loudly if any invariant is violated.

A.6 Reproducibility

The full construction reproduces from federal NCHS microdata and audited state-DOH and VSUS Vol I tabulations via `analysis/run_all.py`. Public artifacts include the county-year extended panel, the births audit log, the per-state source manifest (`data/raw/pre1968_state_vs/sources.yaml`), the MDF county-year aggregation, and the post-1968 CMF panel.

A.7 Per-state pre-1968 cell counts

The MDF aggregation produces 9,723 county-year cells with FIPS matches for the eleven Confederate-South states across 1959–1967. The cell density is approximately balanced across the eleven states relative to county count, with the cell-per-county-year ratio approaching 1.0 for states with comprehensive county-of-residence reporting and slightly below 1.0 for states with a small number of counties whose 1959–1961 NCHS codes do not map cleanly to the 1990 FIPS roster. The bulk of the imperfect mapping concentrates in Virginia, where independent-city / county boundary changes between 1960 and 1990 require the alphabetical-ordering decode to handle a handful of edge cases.

A.8 Why MDF rather than VSUS Vol II for pre-1968 deaths

The published VSUS Vol II annual mortality volumes report county-year-race death counts at the same NCHS-county level as the MDF microdata, and one might in principle prefer the published tables to the underlying microdata. Two considerations push us to the MDF microdata. First, the published VSUS Vol II volumes for 1959–1967 are PDF-only and the county-year-race tables vary in layout and completeness across years, so machine extraction is itself a substantial undertaking. Second, the MDF microdata is the source from which the VSUS

Vol II tables are produced; aggregating the microdata directly using the same NCHS detailed-age coding scheme reproduces the published totals (we verify this against published 1965 and 1967 totals at the state-year-race level) and avoids one layer of OCR-introduced measurement noise.

A.9 Why state-DOH publications and VSUS Vol I rather than other sources for pre-1968 births

Live-birth counts by county-year-race come from the state-DOH vital-statistics annual reports for the six states where they are publicly archived (GA, MS, NC, TN, TX, VA) and from the federal VSUS Vol I Natality state-year tables for the remaining five (AL, AR, FL, LA, SC). The alternative for the pre-1968 period would be the published Vital Statistics natality microdata, which the National Center for Health Statistics has not released for the 1959–1967 period in machine-readable form at the county level. The combined state-DOH-plus-federal-Vol-I tabulations are therefore the canonical pre-1968 county-year-race natality denominator available without restricted-use access, and the same class of state-DOH sources was used by Almond, Chay, and Greenstone (2006). Where both sources are available for the same state, we prefer the state-DOH publication because it resolves contemporaneous county and independent-city boundaries at their 1960-vintage definitions, whereas the federal Vol I aggregations sometimes collapse independent cities into their containing counties or apply later-year boundary corrections retroactively. Our audit step (described in §A.3) catches the obvious tabulation errors common to both sources — county-year cells in which the published table prints a state-total or US-total figure into a county slot — without introducing additional measurement error from the source data itself.

A.10 Limitations of the federal-source-uniform extension

Three limitations of the extension deserve explicit mention. First, the MDF microdata does not separately tabulate live births, so the natality denominator must come from a separate source (here, VSUS Vol I); cross-source measurement error in the rate denominator is bounded but not zero. Second, the 1959–1961 NCHS county encoding is alphabetical and varies slightly across years; we use the NBER 1990 crosswalk and validate the alphabetical-ordering assumption on a sample of county codes against contemporary reference. Third, the audit step removes 15 county-year cells whose published births values are implausible; an additional manual audit against the original VSUS Vol I PDFs could rescue some of these cells but is unlikely to change the headline ATTs because the affected cells are concentrated in a handful of urban and independent-city county-years that contribute small fractions of total Black-births weight.

Appendix B. Design grid and synthetic-control-family diagnostics

B.1 Design grid

The design grid varies three axes: panel version (`v1` public-CMF vs. `v3` federal-uniform extended), sample (eleven-state Southern donor pool vs. five-state Deep-South-only), and treatment threshold (`cert_any`, `cert_25`, `cert_50`, `cert_75`). All 80 estimator cells complete; results are written to `analysis/tables/did_design_grid.csv` and visualized in `analysis/figures/did_design_grid.{png,pdf}`.

The publication-relevant cells are the v3 eleven-state rows:

Threshold	TWFE	CS	BJS	dCDH	SA
`cert_any`	−2.079	−5.246	−2.529	−3.789	−1.568
`cert_25`	−2.055	−5.203	−2.508	−3.689	−1.503
`cert_50`	−2.090	−5.294	−2.562	−3.854	−1.672
`cert_75`	−1.854	−5.095	−2.412	−4.071	−2.102

Notes: This table summarizes policy timing, cohorts, thresholds, or state-level sample construction. It is intended to make the identifying variation and comparison groups transparent.

The v3 estimates are stable across treatment-threshold choices for all five estimators after the SA endpoint-bin correction; the `cert_75` headline is, if anything, conservative relative to looser certification thresholds. The five-state Deep-South-only rows leave only 3 never-treated counties for `cert_any`/`cert_25`, 5 for `cert_50`, and 10 for `cert_75`; under those weak-donor conditions the staggered estimators become unstable, often flipping sign across thresholds. We carry the five-state-only grid as a weak-donor stress test rather than a preferred design, and we do not treat its results as equivalent evidence against the eleven-state findings.

B.2 Synthetic-control-family diagnostics

The five-state Deep-South-only donor structure motivates a synthetic-control-family check. The local R installation has only the `Synth` package; `augsynth`, `gsynth`, `fect`, `synthdid`, and `PanelMatch` are not installed. We therefore implement transparent local analogues of five SCM-family methods and report them as triangulation rather than as a main design:

- Classic SCM with nonnegative donor weights summing to one.
- ASCM, a ridge-augmented SCM analogue.
- Synthetic DiD, a local unit-and-time-weighting analogue.
- Generalized synth / IFE analogue using donor-factor interactive fixed effects.
- A PanelMatch-style matched DiD with county-level nearest-neighbor history matching.

Results (`analysis/tables/synthetic_family_estimates.csv`):

Method	ATT	SE	95% CI	Pre-RMSPE	Donor support
Classic SCM	+0.465	NA	NA	1.577	6 donor-state aggregates
ASCM	+0.860	NA	NA	1.030	6 donor-state aggregates
Synthetic DiD	−2.212	NA	NA	1.577	6 donor-state aggregates
Generalized synth / IFE	−3.354	NA	NA	0.878	6 donor-state aggregates
PanelMatch-style matched DiD	−0.782	0.475	[−1.713, +0.150]	2.389	723 never-treated counties

Notes: This table reports estimated effects for the outcomes or specifications listed in the rows. Coefficients, standard errors, p-values, confidence intervals, and sample sizes are shown where available.

Classic SCM and ASCM place most donor weight on North Carolina (0.832) and Texas (0.168), and the in-space placebo p-value for the classic SCM is 0.429, so the aggregate classic SCM is best treated as descriptive rather than confirmatory. The Synthetic DiD analogue is negative but concentrates its time weights on 1960; the IFE analogue is directionally consistent with the main staggered-DiD estimates with selected rank 3 and good pre-period fit; the PanelMatch-style matched DiD is negative but not conventionally significant. The synthetic-control family supports rather than replaces the staggered-DiD design, and formal package implementations are the natural next step.

Appendix C. Sun–Abraham endpoint-bin diagnostic

The Sun–Abraham aggregate ATT on the v3 federal-source-uniform extended panel is sensitive to whether the cohort-saturated event-study includes cohort-specific bins for $\ell < -10$ and $\ell > +10$. Without those bins, long-run treated cells (most consequentially the 1967 cohort's 1978–1988 observations) fall into the implicit reference category along with $\ell = -1$, and the aggregate SA estimate inherits a positive sign. With the bins, the long-run cells get their own coefficients and the 1967-cohort sub-estimate flips negative. The diagnostic comparison is:

SA aggregation	No endpoint bins	With endpoint bins
Count-weighted event average	+0.838	−2.102
Cell-weighted supported terms	+0.834	−2.101
Birth-weighted supported terms	+0.806	−1.820
Birth-weighted, excluding 1967	+0.084	−0.506
Birth-weighted events 0–3	+1.508	−1.107
Birth-weighted events 4–10	+0.382	−2.250

Notes: This table documents the source files, scripts, variables, or data inputs used in the analysis. It is included to make the construction of the analytic evidence reproducible.

The 1967 cohort drives the sign. Because the 1967 cohort is 304 of 393 treated counties and 66 percent of post-window Black-births weight, leaving its post-1977 observations in the reference category pulls the SA path in the wrong direction. The diagnostic and a corresponding cohort × event-time heatmap are produced by `analysis/robustness/sun_abraham_diagnostics.py` and saved to `analysis/figures/sa_cohort_event_heatmap.{png,pdf}` and `analysis/figures/sa_static_aggregation_variants.{png,pdf}`. The v1 SA result remains positive even after endpoint bins because the v1 panel begins in 1968 and the dominant 1967 cohort has no observed pre-period; that is one further reason to treat the v1 specifications as historical comparisons rather than the preferred analysis.

The standard error on the SA aggregate is computed from the fitted coefficient covariance matrix rather than from diagonal variances alone, since the aggregator is a weighted sum of correlated coefficients. The corrected aggregate covariance is what produces the [−3.70, −0.50] interval reported in Table 5.

References

Aizer, A., Currie, J., Moretti, E., & Yang, Q. (2014). Targeted transfers and racial mortality. *NBER Working Paper.*

Aizer, A., Eli, S., Ferrie, J., & Lleras-Muney, A. (2016). The long-run impact of cash transfers to poor families. *American Economic Review*, 106(4), 935–971. doi:10.1257/aer.20140529.

Almond, D., Chay, K. Y., & Greenstone, M. (2006). *Civil rights, the war on poverty, and Black–White convergence in infant mortality in the rural South and Mississippi.* MIT Department of Economics Working Paper No. 07-04.

Almond, D., & Mazumder, B. (2011). Health capital and the prenatal environment: The effect of Ramadan observance during pregnancy. *American Economic Journal: Applied Economics*, 3(4), 56–85. doi:10.1257/app.3.4.56.

Anderson, D. M., Charles, K. K., & Rees, D. I. (2024). Imposing policy on reluctant actors: The hospital desegregation campaign and Black postneona-

tal mortality in the Deep South. *Review of Economics and Statistics*, 1–46. doi:10.1162/rest_a_01467.

Athey, S., & Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1), 62–79. doi:10.1016/j.jeconom.2020.10.012.

Bailey, M. J., & Goodman-Bacon, A. (2015). The War on Poverty's experiment in public medicine: Community Health Centers and the mortality of older Americans. *American Economic Review*, 105(3), 1067–1104. doi:10.1257/aer.20120070.

Bleakley, H. (2007). Disease and development: Evidence from hookworm eradication in the American South. *Quarterly Journal of Economics*, 122(1), 73–117. doi:10.1162/qjec.121.1.73.

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, 91(6), 3253–3285. doi:10.1093/restud/rdae007.

Brown, D. W., Kowalski, A. E., & Lurie, I. Z. (2020). Long-term impacts of childhood Medicaid expansions on outcomes in adulthood. *Review of Economic Studies*, 87(2), 792–821. doi:10.1093/restud/rdz039.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230. doi:10.1016/j.jeconom.2020.12.001.

Cascio, E. U., & Washington, E. (2014). Valuing the vote: The redistribution of voting rights and state funds following the Voting Rights Act of 1965. *Quarterly Journal of Economics*, 129(1), 379–433. doi:10.1093/qje/qjt028.

Chay, K. Y., & Greenstone, M. (2003). Air quality, infant mortality, and the Clean Air Act of 1970. *Quarterly Journal of Economics*, 118(3), 1121–1167. NBER WP version doi:10.3386/w10053.

Civil Rights Act of 1964, Public Law 88-352, 78 Stat. 241 (1964).

Currie, J., & Gruber, J. (1996). Health insurance eligibility, utilization of medical care, and child health. *Quarterly Journal of Economics*, 111(2), 431–466. doi:10.2307/2946684.

Currie, J., & Gruber, J. (1996). Saving babies: The efficacy and cost of recent changes in the Medicaid eligibility of pregnant women. *Journal of Political Economy*, 104(6), 1263–1296. doi:10.1086/262059.

Currie, J., & Schwandt, H. (2016). Mortality inequality: The good news from a county-level approach. *Journal of Economic Perspectives*, 30(2), 29–52. doi:10.1257/jep.30.2.29.

Cutler, D., & Miller, G. (2005). The role of public health improvements in

health advances: The twentieth-century United States. *Demography*, 42(1), 1–22. doi:10.1353/dem.2005.0002.

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9), 2964–2996. doi:10.1257/aer.20181169.

de Chaisemartin, C., & D'Haultfœuille, X. (2022). Difference-in-differences estimators of intertemporal treatment effects. SSRN Working Paper. doi:10.2139/ssrn.4068043.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature*, 47(1), 87–122.

Goodman-Bacon, A. (2018). Public insurance and mortality: Evidence from Medicaid implementation. *Journal of Political Economy*, 126(1), 216–262. doi:10.1086/695528.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2), 254–277. doi:10.1016/j.jeconom.2021.03.014.

Hill-Burton Act (1946). Hospital Survey and Construction Act, Public Law 79-725, 60 Stat. 1041.

Largent, E. A. (2018). *Segregation: A historical analysis of hospital desegregation under Title VI*. PhD dissertation, University of Pennsylvania.

Medicare Act of 1965. Public Law 89-97, 79 Stat. 286.

NARA Record Group 235, Series 5. Health Equity Compliance Files, 1966–1972. National Archives and Records Administration, College Park, MD.

Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5), 2555–2591. doi:10.1093/restud/rdad018.

Reynolds, P. P. (1997). The federal government's use of Title VI and Medicare to racially integrate hospitals in the United States, 1963 through 1967. *American Journal of Public Health*, 87(11), 1850–1858. doi:10.2105/AJPH.87.11.1850.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2), 2218–2244. doi:10.1016/j.jeconom.2023.03.008.

Simkins v. Moses H. Cone Memorial Hospital, 323 F.2d 959 (4th Cir. 1963).

Smith, D. B. (1999). *Health care divided: Race and healing a nation*. University of Michigan Press.

Smith, D. B. (2005). *Eliminating disparities in treatment and the struggle to end segregation in American hospitals. Race and Social Problems*, 1(2), 75–84.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2), 175–199. doi:10.1016/j.jeconom.2020.09.006.

Thompson, O. (2024). Selected fertility and racial inequality. *Journal of Human Resources*, 59(3), 684–710. doi:10.3368/jhr.0221-11481r2.

Voting Rights Act of 1965. Public Law 89-110, 79 Stat. 437.

Reproducibility: All numbers in this manuscript reproduce from federal data via `python3 analysis/run_all.py`. *Author: Jonathan Palisoc (jpalisoc@umich.edu, ORCID 0000-0001-5003-2631), University of Michigan, School of Public Health, Department of Health Management & Policy.*