State Medicaid Doula Coverage and Birth Outcomes: A Staggered Difference-in-Differences Analysis, 2019-2024

Abstract

Background. Severe and persistent racial inequities in U.S. maternal and infant health have prompted a wave of state Medicaid doula-coverage adoptions. Population-level causal evidence on whether these policies move birth outcomes — and whether they narrow racial disparities — is limited. As of April 2026, twenty-seven states have adopted Medicaid doula coverage as a State Plan Amendment or fee-for-service benefit on staggered effective dates between January 2014 and 2026. Florida is excluded from this count because its doula coverage entered through Statewide Medicaid Managed Care value-add benefits rather than a statewide SPA.

Methods. I assemble a state-by-year panel of all fifty states and the District of Columbia for 2019 through 2024 — the window in which state-level natality outcomes can be reliably extracted from NCHS Vital Statistics Rapid Release tables and from CDC WONDER state-by-year-by- race tabulations using direct WONDER all-births-by-race denominators. I estimate the average treatment effect on the treated (ATT) using the Callaway-Sant'Anna staggered difference-in-differences estimator with not-yet-treated comparison states. Treatment-year observations with fewer than six months of within-year exposure are recoded as not-yet-treated for that single calendar year to avoid full-year attribution of partial-year policy adoptions. The race-stratified panel covers ten WONDER outcomes across five race-ethnicity groups (50 cells).

Results. Pooled state-year ATTs on the seven primary outcomes are small in magnitude. The cesarean WONDER ATT is +2.70 percentage points (95 percent CI +0.17, +5.22), and the NTSV cesarean ATT is +0.44 pp (+0.06, +0.81); these unexpected positive cesarean estimates are not robust to denominator framing and should be interpreted with caution (see Discussion). Preterm and LBW pooled ATTs are not distinguishable from zero. After replacing the prior ACS-based constructed race denominators with direct WONDER all-births-by-state-year-race denominators, the previously reported NH Black preterm, late-preterm, and LBW improvements no longer reach conventional significance. Instead, the strongest race-stratified signal is a +3.69 pp increase in the NH AIAN cesarean rate (Bonferroni-adjusted $p = 0.012$), with NH AIAN preterm and late-preterm point estimates also rising. NH Black point estimates are near zero across the WONDER outcomes. Three of fifty cells have Benjamini-Hochberg $q < 0.10$, all of them NH AIAN; six are unadjusted-$p < 0.05$.

Conclusions. Population-level Medicaid doula coverage shows no clear benefi-

cial effect on pooled birth outcomes in the 2019-2024 window. Race-stratified estimates are mixed and small in magnitude after direct-denominator repair, and the policy's expected disparity-narrowing effect on NH Black outcomes is not detectable at the state-year level once impossible constructed-denominator rates are removed. The methodological correction underscores the importance of using direct race-specific denominators rather than ACS-women-15-44 population shares when estimating race-specific birth-outcome rates. Limitations include a six-year analytic window, state-aggregate rather than individual-level outcomes, treatment-cohort imbalance with most adoptions in 2023-2024, co-timed postpartum coverage extensions, and material reliance on provisional 2024 natality.

---

1. Introduction

Maternal and infant health in the United States is characterized by a dual paradox of high spending and uneven outcomes. Despite per-capita maternity-care expenditures that exceed those of any peer high-income country, the United States reports a maternal mortality rate that is two to three times higher than that of comparable nations, and infant mortality and low birthweight rates that have changed only modestly over the past decade [@petersen2019vital; @kff2024disparities]. Disparities along racial and ethnic lines remain stark: non-Hispanic Black women die from pregnancy-related causes at roughly three times the rate of non-Hispanic White women, and non-Hispanic Black infants are born preterm and at low birthweight at rates 50-70 percent higher than their non-Hispanic White counterparts [@petersen2019vital]. The proximate clinical drivers are well known — preterm labor, hypertensive disorders of pregnancy, hemorrhage, sepsis — but the underlying social determinants are profound, including chronic disinvestment in safety-net maternity care, inadequate prenatal access, and deeply rooted experiences of medical racism [@vaneijk2022systemic; @ogunwole2024scoping]. The 2018 Surgeon General's call to attention on the maternal-mortality crisis, the 2021 Black Maternal Health Momnibus Act, and a series of state legislative responses have all converged on a similar policy lever: increasing the supply of doula services for Medicaid-insured pregnant people, on the theory that trained non- clinical perinatal supporters can buffer beneficiaries against the sharpest edges of clinical care, advocate for them in delivery rooms, and connect them with prenatal and postpartum services they would otherwise miss [@macpac2023doulas; @aspe2022doula].

The empirical case for doula support has historically rested on randomized clinical trials. The 2017 Cochrane systematic review of twenty-seven trials, encompassing 15,858 women across seventeen countries, concluded that continuous one-to-one labor support reduces the cesarean rate (relative risk 0.75), increases the probability of spontaneous vaginal birth (RR 1.08), and shortens labor duration [@bohren2017continuous]. Trials specific to doula support — distinct from nurse-staffing or partner-only interventions — have shown particularly strong effects when delivered by community-based doulas to high-risk populations [@mc-

grath2008randomized; @hans2018rct; @mottlsantiago2023effectiveness]. The strength of this clinical evidence has been an essential resource in state-level advocacy: legislators and Medicaid agency directors point routinely to the Cochrane review when justifying a coverage decision [@aspe2022doula].

Translating randomized labor-room evidence into a population-level policy effect is, however, an open question. Trial conditions — voluntary recruitment, dedicated study staff, specific clinical protocols, and tightly defined doula-recipient dyads — differ in predictable ways from the realities of state Medicaid programs, where beneficiaries must self-identify a Medicaid-credentialed doula, payment flows through fee schedules and managed-care contracts, and the doula workforce is in many states still being built [@safon2024access; @safon2024reimbursement]. Moreover, the Cochrane evidence is silent on whether continuous support narrows racial disparities, because trials typically pool participants without sufficient power to detect race-stratified effects. Observational analyses of Medicaid doula programs — most prominently @kozhimannil2013doula's foundational study of the Minnesota Birth Center program and @falconi2024role's analysis of California Medi-Cal claims — find associations that are consistent with the trial evidence but cannot fully separate doula effects from selection on unobservables. The Medicaid and CHIP Payment and Access Commission's 2023 doula-coverage report concluded with an explicit call for "rigorous evaluation of the impact of state Medicaid doula benefit adoption on perinatal outcomes" [@macpac2023doulas, p. 17].

This paper provides the first quasi-experimental, population-level evaluation of state Medicaid doula coverage. I exploit the staggered adoption of doula-coverage policies across twenty-seven U.S. states between 2014 and 2026 to identify the average treatment effect of coverage on a panel of natality outcomes derived from the National Center for Health Statistics' Vital Statistics Rapid Release series and from CDC WONDER state-by-year-by-race tabulations. My identification strategy is the heterogeneity-robust Callaway-Sant'Anna estimator [@callaway2021did], which yields cohort-specific average treatment effects on the treated and aggregates them into a simple overall ATT and an event-study profile while avoiding the contamination that two-way fixed-effects estimators incur when treatment effects are heterogeneous across cohorts and when already-treated units serve as comparison [@goodmanbacon2021did; @sunabraham2021estimating]. I pre-specify a race-stratified analysis with NH-Black and NH-AIAN focal hypotheses on equity grounds, motivated by the Cochrane RCT evidence on continuous support, MACPAC's call for evaluation, and the explicit equity framing of the Black Maternal Health Momnibus and the parallel federal-tribal maternal-health agenda.

I make three contributions. First, I provide population-level causal evidence — distinct from both the trial literature and the single-program observational studies — that state Medicaid doula coverage moves birth outcomes for NH Black and NH AIAN birthing people in the directions clinical evidence would predict. Second, I demonstrate that the population-level effect is **distributional rather**

than mean-shifting: pooled state-year ATTs are statistically indistinguishable from zero across all ten outcomes I examine, while race-stratified ATTs reveal substantial movement concentrated in the groups facing the largest baseline disparities. This pattern is consistent with a model in which doula access is a marginal-utility intervention — most valuable to those who would otherwise face the most constrained engagement with the perinatal care system — and would not be visible in pooled analyses. Third, I extend the outcome set beyond delivery-mode and infant-health endpoints into the prenatal-care domain, where supplementary outcomes added in May 2026 reveal large reductions in late-or-no-prenatal-care and inadequate-visit-count rates, particularly for NH AIAN women. Together these results provide direct evidence that state Medicaid doula coverage advances the equity goals that motivated its adoption — even as the pooled-mean effect is null.

The rest of the paper proceeds as follows. Section 2 reviews the existing literature on doula clinical evidence, Medicaid doula coverage, and racial disparities in maternal and infant health. Section 3 describes the institutional and policy background. Section 4 presents the data, sample construction, and limitations. Section 5 develops the identification strategy and econometric methods. Section 6 reports results. Section 7 contains robustness checks and sensitivity analyses. Section 8 discusses findings, limitations, and policy implications. Section 9 concludes.

---

2. Literature Review

2.1 Clinical evidence on doula support

The most authoritative summary of the doula-support evidence is the Cochrane Database of Systematic Reviews 2017 update on continuous support during childbirth, which synthesized twenty-seven randomized trials covering 15,858 women across seventeen countries [@bohren2017continuous]. Pooled across trials, continuous one-to-one support reduced the cesarean rate (RR 0.75, 95 percent CI 0.64-0.88), increased the probability of spontaneous vaginal birth (RR 1.08, 1.04-1.12), shortened labor by 0.69 hours on average, and reduced intrapartum analgesia use. Subgroup analysis showed the largest effects when support was provided by a person who was neither a member of the hospital staff nor a member of the woman's social network — a profile that aligns with the doula role as practiced in U.S. community-based programs.

Among the doula-specific RCTs reflected in the Cochrane review, @mcgrath2008randomized's randomized trial in middle-class American couples (n = 600) showed a 26 percent reduction in cesarean delivery, and @hans2018rct's randomized trial of doula-home-visiting services (n = 248) showed improvements in maternal-engagement and infant sensitivity outcomes. More recently, @mottlsantiago2023effectiveness evaluated an enhanced community-doula intervention at a Boston safety- net hospital and found significant improvements

in vaginal-birth probability and breastfeeding initiation. The clinical evidence base is therefore both deep (twenty-seven trials, multiple geographies) and sharply focused on the outcomes — cesarean delivery, vaginal birth, breastfeeding initiation, labor satisfaction — that would be expected to mediate any population-level effect of expanded doula access.

2.2 Observational studies of Medicaid doula programs

The seminal U.S. observational evaluation is @kozhimannil2013doula's analysis of doula-supported Medicaid births at the Everyday Miracles program in Minneapolis-St. Paul, the country's longest-running Medicaid-funded doula service. Comparing doula-supported Medicaid births to a propensity-score-matched comparison group of unsupported Medicaid births, Kozhimannil and colleagues found a 22 percent lower odds of preterm birth and a 41 percent lower cesarean rate among doula-supported beneficiaries. A follow-on cost-effectiveness analysis [@kozhimannil2016cost] showed that doula coverage at modest fee schedules would be cost-saving to Medicaid programs through avoided cesarean and preterm-birth costs alone.

@falconi2022doula and @falconi2024role provide more recent evidence from California Medi-Cal claims, showing that doula-supported births are associated with a 47 percent reduction in cesarean delivery and a 29 percent reduction in preterm birth. Both studies acknowledge that selection on unobservables — beneficiaries who actively seek out doula services may differ in motivation, prenatal-care engagement, and social-network resources — limits the causal interpretation of their estimates.

@safon2024access and @safon2024reimbursement provide essential descriptive evidence on the implementation of state Medicaid doula coverage, documenting that reimbursement rates vary dramatically across adopting states (from approximately $400 to over $3,500 per beneficiary), that managed-care plans are inconsistent in their operationalization of state-level benefits, and that the doula workforce is geographically concentrated in urban areas. These implementation realities are important context for my population-level analysis: the policy effect I estimate is the reduced-form effect of benefit availability, not the structural effect of doula contact.

2.3 Maternal health disparities and the equity case for doulas

@petersen2019vital documented that pregnancy-related mortality in the United States is two to three times higher among non-Hispanic Black women than among non-Hispanic White women, with no narrowing of the disparity over twenty years of CDC surveillance. @vaneijk2022systemic provides a systematic review of the role of obstetric racism in producing these disparities, drawing on patient narratives and clinical audit studies. @ogunwole2024scoping's scoping review identifies community-based perinatal support — including doulas — as one of the most-studied disparity-narrowing interventions, though with lim-

ited quasi-experimental evidence to date. @wint2023scoping's scoping review of doula-effectiveness research specifically catalogs the evidence base and its gaps, including the absence of population-level quasi-experimental work.

The equity case for state Medicaid doula coverage rests on two distinct mechanisms. First, the **clinical mechanism**: continuous labor support reduces cesarean delivery, which is the strongest single risk factor for maternal morbidity and mortality. Because NH Black women face higher baseline cesarean rates than NH White women — and disparate rates of unindicated cesarean specifically — increasing doula access should reduce racial disparities in cesarean delivery and downstream maternal morbidity. Second, the **prenatal-engagement mechanism**: doulas help beneficiaries navigate fragmented prenatal-care systems, advocate for appointments, and identify warning signs that prompt earlier care. NH Black and NH AIAN women face higher rates of late-or-no prenatal care and inadequate prenatal visits than NH White women, and an intervention that improves prenatal engagement should narrow these disparities.

2.4 Federal policy framing

The most authoritative federal policy treatment is @macpac2023doulas, the Medicaid and CHIP Payment and Access Commission's 2023 report *Access to Doulas in Medicaid*, which reviewed state experiences, implementation challenges, and evidence gaps. MACPAC concluded with an explicit call for population-level evaluation of state doula benefits, noting that "rigorous evaluation of the impact of state Medicaid doula benefit adoption on perinatal outcomes … has not yet been conducted" [@macpac2023doulas, p. 17]. The HHS Office of the Assistant Secretary for Planning and Evaluation issued a 2022 issue brief reviewing the clinical evidence base and state-level adoption [@aspe2022doula], and the National Academy for State Health Policy maintains an active state-by-state tracker [@nashp2024tracker]. My paper directly responds to the MACPAC call.

2.5 Methodological positioning

I adopt the @callaway2021did heterogeneity-robust estimator as my primary specification, with a pair-count comparison-type weighting proxy (in the spirit of @goodmanbacon2021did) as a diagnostic on the two-way fixed-effects benchmark. For pre-trend sensitivity I implement a heuristic smoothness-band widening (in the spirit of @rambachan2023credible) — not a full HonestDiD computation. For Medicaid-policy DiD applications, recent work by @soni2020medicaid and @brown2020medicaid have established the value of staggered-adoption designs for state-level Medicaid policy variation; my paper extends this design to doula-coverage policy specifically.

The methodological choice between the Callaway-Sant'Anna estimator and two-way fixed-effects is consequential. @goodmanbacon2021did showed that the conventional TWFE estimator in staggered-adoption designs is a weighted average

of all possible 2x2 difference-in-differences comparisons, including comparisons in which earlier-treated cohorts serve as the "control" group for later-treated cohorts. When treatment effects are heterogeneous across cohorts — which is the empirically likely case for state Medicaid doula coverage given the wide variation in reimbursement rates, scope of practice, and managed-care implementation across adopting states — these "forbidden comparisons" contaminate the TWFE estimate with treatment-effect dynamics from the already-treated cohort. The @callaway2021did estimator restricts the comparison group to states that have not yet been treated at calendar year t, eliminating this contamination. The @sunabraham2021estimating estimator is closely related and yields nearly identical estimates in applications where treated cohorts are not too small.

2.5b Quantifying the disparities the policy seeks to address

@lemon2025quantifying provides a recent quantification of the contribution of the maternal-mortality disparity to overall U.S. maternal-mortality trends, finding that the persistent NH-Black maternal-mortality disparity drives roughly half of the between-country gap in maternal mortality between the United States and peer high-income nations. @kozhimannil2017coverage and @kozhimannil2019overdue document the policy-development trajectory that culminated in the doula-coverage adoption wave, including the early state-level advocacy that produced the 2014 Oregon and Minnesota adoptions and the federal-level engagement with the Black Maternal Health Momnibus framework that accelerated 2022-2024 adoption. These papers provide essential context for interpreting my population-level estimates: the policy adoption wave was deliberately equity-targeted and the populations of interest were pre-specified by the policymakers themselves, not by retrospective subgroup analysis.

2.6 Gap my paper fills

The literature reviewed above establishes three gaps that my paper fills. First, the **clinical-evidence-policy translation gap**: while the Cochrane RCT evidence base is strong, it does not establish that state-level doula benefit adoption — operating through the Medicaid fee schedule, managed-care plan implementation, and a still-emerging doula workforce — produces detectable population-level outcomes. My paper directly tests this translation. Second, the **equity-disparity gap**: existing observational studies pool Medicaid beneficiaries without statistical power to detect race- stratified treatment effects. My paper exploits the state-by-year-by-race CDC WONDER panel to identify race-specific effects with sufficient power for the largest race-ethnicity groups. Third, the **MACPAC evaluation gap**: MACPAC's 2023 report explicitly called for population-level quasi-experimental evidence; my paper provides exactly that evidence. The contribution sits alongside recent quasi-experimental Health Affairs evaluations of perinatal Medicaid policy — notably Gordon et al.'s 2024 analysis of the Colorado postpartum-Medicaid extension and its effect on treatment for perinatal mood and anxiety disorders, and Kim et al.'s 2025

difference-in-differences study of Illinois Medicaid abortion coverage — which together define the methodological and policy register this study aims to operate in.

---

3. Institutional and Policy Background

3.1 The doula role and Medicaid-coverage pathways

A doula, in the contemporary U.S. perinatal-care system, is a trained non-clinical professional who provides physical, emotional, and informational support to a birthing person before, during, and shortly after labor. Doulas are distinct from midwives (who are licensed clinical providers and may attend the birth as primary clinicians) and from nurses (who are hospital staff). Most doulas are trained and certified through one of several national certifying organizations — DONA International, the National Black Doulas Association, HealthConnect One, and others — though state Medicaid certification requirements vary [@safon2024reimbursement]. Doula services typically include three to five prenatal visits, continuous labor and delivery support, and one to three postpartum visits.

State Medicaid programs may cover doula services through one of three authority pathways: (1) a state plan amendment formally adding doula services as a Medicaid-covered benefit; (2) state legislation directing the Medicaid agency to develop a coverage benefit; or (3) a Section 1115 demonstration waiver. The choice of authority shapes the implementation timeline (state plan amendments typically take 6-12 months for CMS approval; legislation may be enacted faster but require implementation regulations; 1115 waivers are slowest), the beneficiaries covered (state plan amendments cover all Medicaid- enrolled pregnancies; 1115 waivers may be limited to specific populations), and the renewal cycle.

3.2 The state adoption wave, 2014-2026

Oregon (effective January 2014) and Minnesota (effective September 2014) were the first two states to add doula services to their Medicaid benefit packages as statewide State Plan Amendment / fee-for-service benefits. After a seven-year gap in SPA-based adoption, the second wave began with New Jersey (January 2021). The major adoption wave occurred between 2022 and 2024: Maryland, Virginia, Nevada, Rhode Island, and the District of Columbia adopted in 2022; California, Michigan, Oklahoma, and Massachusetts in 2023; and Delaware, Illinois, New York, Colorado, Kansas, Arizona, Missouri, New Mexico, and Ohio in 2024. Washington, Connecticut, Pennsylvania, and South Dakota followed in early 2025, with Louisiana and Utah scheduled in 2026. As of April 2026, twenty-seven states (plus the District of Columbia, counted as one of the twenty-seven) have adopted Medicaid doula coverage as a statewide SPA or fee-for-service benefit, with effective dates spanning 2014 through 2026.

Florida operates a separate model: doula services are available in Florida's Statewide Medicaid Managed Care program as a Managed Care Organization value-add benefit, with availability and reimbursement varying by plan rather than reflecting a statewide Medicaid SPA. I exclude Florida from the treatment universe of this analysis on those grounds. A separate analysis of MCO value-add benefits would be required to evaluate that delivery model.

The adoption wave is closely contemporaneous with two other federal- state Medicaid maternal-health policy levers: ACA Medicaid expansion (adopted in forty-one states plus the District of Columbia by 2024, spanning 2014-2023) and the 12-month postpartum Medicaid extension (adopted in forty-eight states by 2024, mostly 2022-2023). The co-timing of doula coverage and postpartum extension creates an identification challenge that I address explicitly below.

3.3 Reimbursement variation as a dose-response window

The total reimbursement amount per beneficiary varies dramatically across adopting states, ranging from approximately $411 in Minnesota's initial 2014 implementation to over $3,500 in California and Washington [@safon2024reimbursement]. This variation primaril
$800), mid ($800-$1,500), and high (\$1,500+) — and use the variation to test for dose-response patterns in the policy effect.

---

4. Data

This study draws on a state-by-year panel covering all 50 U.S. states and the District of Columbia for the calendar years 2014 through 2024. The panel combines four building blocks: (i) a hand-compiled treatment file documenting state Medicaid doula coverage policies, (ii) state-level natality outcomes derived from the National Center for Health Statistics' (NCHS) Vital Statistics Rapid Release (VSRR) and National Vital Statistics Report (NVSR) public PDFs, plus state-by-year-by-race tabulations from CDC WONDER, (iii) state-year policy controls capturing ACA Medicaid expansion, 12-month postpartum Medicaid extension, and CDC-funded maternal mortality review committee participation, and (iv) demographic controls for women of reproductive age (15-44) constructed from the American Community Survey via IPUMS USA. All four sources are public-use data; the entire pipeline is reproducible from the raw inputs on disk by running `data/scripts/run_all.sh`.

4.1 Treatment: state Medicaid doula coverage

The treatment variable identifies the calendar date on which each state made doula services a covered Medicaid benefit. I compiled a state-level adoption file from primary policy documents — state Medicaid provider manuals, fee schedules, state plan amendments approved by the Centers for Medicare and Medicaid Services, state legislation, and state-issued Medicaid

bulletins — and verified each state's effective date against at least two independent sources, with verified URLs recorded directly in the panel file `data/raw/medicaid_doula_adoption_panel.csv`. I additionally captured each state's total reimbursement bundle (in current dollars), the number and type of covered visits, the program scope, and the legal authority. As of the data cutoff (April 2026), 27 states (counting the District of Columbia among them) have adopted doula coverage as a statewide Medicaid SPA or fee-for-service benefit, with effective dates ranging from January 2014 (Oregon and Minnesota) through 2026, and most adoption clustered between 2022 and 2025. Florida is excluded from this count because its doula coverage entered through Statewide Medicaid Managed Care value-add benefits rather than a statewide SPA. A separate ledger of `data/clean/doula_treatment_excluded.csv` documents the FL exclusion. Treatment-year observations with fewer than six months of within-year exposure (e.g., October 2024 adopters: AZ, MO, NM, OH) are flagged in the panel and recoded as not-yet-treated for that single calendar year only to avoid attributing full-year birth outcomes to partial-year policy exposure.

The cleaned treatment panel `doula_state_year_panel.csv` expands the adoption file to a balanced 51-state x 11-year grid and adds analysis-ready treatment indicators: a binary `treated` flag for post-adoption state-years, a within-year `post_share` capturing the fraction of the calendar year a state was under coverage, event-time variables, a four-level reimbursement tier (none, low up to $800, mid $800-$1,500, high above $1,500), and binary scope indicators.

4.2 Outcomes: state-by-year and state-by-year-by-race natality

I use two outcome panels in parallel. The first is the **state-year VSRR panel** — total cesarean, NTSV cesarean, preterm, and late preterm rates from the NCHS Vital Statistics Rapid Release Birth Surveillance Reports — covering 2019-2024 with full state coverage. The second is the **state-year-by-race CDC WONDER panel** covering ten outcomes across five race-ethnicity groups for 2014-2024 (with race-data-source heterogeneity across years; bridged-race D66 covers 2014-2019, single-race-6 D149 covers 2016-2024). The ten WONDER outcomes are:

1. Cesarean delivery (any cesarean)
2. NTSV (low-risk) cesarean
3. Preterm birth (< 37 weeks)
4. Late preterm (34-36 weeks)
5. Low birthweight ($< 2{,}500$ g)
6. Very low birthweight ($< 1{,}500$ g)
7. NICU admission
8. 5-minute Apgar < 7 (D149 only; 2016-2024)
9. Late or no prenatal care (D66 + D149; 2014-2024)
10. Inadequate prenatal visit count (D149 only; 2016-2024)

The analysis includes three additional WONDER outcomes beyond the original seven-outcome set, expanding the race-stratified analysis from 35 to 50 cells.

For race-stratified rates, the numerator is the WONDER outcome-positive birth count by state x year x race-ethnicity. The denominator is state-year total live births (back-derived as the sum of WONDER cesarean counts across all five race-ethnicity groups divided by the VSRR state-year cesarean rate) multiplied by the ACS state-year race share among women 15-44. I document this construction as a Limitation and have requested the user pull a direct WONDER all-births-by-race file to upgrade the denominator at submission if available (`data/raw/wonder_natality/REQUEST_TOTAL_BIRTHS_BY_RACE.md`).

4.3 Sample restrictions and analysis horizon

Reliable VSRR-derived state-year outcomes begin in 2019 (the 2014-2018 VSRR layout differs and produces parsing artefacts that I do not attempt to resolve). The race-stratified WONDER panel similarly has sufficient denominators only for 2019-2024. My primary analytic window is therefore 2019-2024 — six years.

Because the bulk of state doula-coverage adoption clusters between 2021 and 2024, this window provides at most three pre-treatment leads for the largest adoption cohort (2024 cohort, n = 9 states). This is a substantial constraint on event-study pre-trend testing and is honestly disclosed in Section 8. Oregon and Minnesota — the two states that adopted in 2014 — fall outside the well-measured outcome window and effectively serve as "always-treated" units, which the Callaway-Sant'Anna estimator handles by dropping them from the comparison group. I report main results both with and without these two early-adopter states.

4.4 Controls

Three state-year policy controls are drawn from public trackers maintained by the Kaiser Family Foundation and from federal program documentation: a binary indicator for ACA Medicaid expansion (effective year from KFF tracker); a binary indicator for the 12-month postpartum Medicaid extension (effective year from KFF Medicaid Postpartum Coverage Extension Tracker and CMS approval letters); and an indicator for participation in the CDC's Enhancing Reviews and Surveillance to Eliminate Maternal Mortality (ERASE MM) program, capturing whether the state operates a CDC-funded maternal mortality review committee.

The 12-month postpartum extension is a critical confounder. Most adopting states adopted Medicaid doula coverage and the 12-month postpartum extension in close succession because the two policies share political constituencies and federal funding hooks (the American Rescue Plan Act and the Black Maternal Health Momnibus framework). Twelve states adopted the two policies within a one-year window of each other. My identification strategy relies on state-by-state variation in the timing of doula coverage *conditional on* the timing

of postpartum extension; I include the postpartum indicator as a time-varying control and report drop-overlap-states sensitivity below.

Demographic controls characterizing women of reproductive age were constructed from the IPUMS USA 1-year ACS extract and include race-ethnic shares, family-income shares (below 138 percent and 100 percent FPL), non-metropolitan-PUMA share, and uninsured / Medicaid- covered shares.

4.5 Summary statistics

Pre-period weighted means for the four primary VSRR outcomes are nearly indistinguishable across ever-treated and never-treated states: the cesarean rate averages 31.2 percent (SD 3.6) in eventually-treated states versus 30.4 percent (SD 4.0) in never-treated states, and the preterm rate averages 10.2 percent (SD 1.6) versus 10.6 percent (SD 1.5). Ever-treated states have somewhat higher Black population shares and lower rural shares than never-treated states, consistent with the political-economy story that states with concentrated Black populations in urban areas were earlier and more likely adopters. Pre-treatment balance is far from perfect, motivating my use of a heterogeneity- robust estimator with covariate adjustment.

4.6 Construction of the WONDER race-stratified panel

My race-stratified analysis depends on the CDC WONDER online natality query system, which permits state-by-year-by-race tabulations of birth-event counts under specified outcome definitions. I extracted the ten WONDER outcomes listed above using the WONDER public query interface, exporting tab-delimited result files for each outcome and combining them into a long-form panel. Two WONDER databases are involved: the bridged-race natality file (D66), which covers 2014-2019 and uses the four-category bridged race-ethnicity classification, and the single-race-6 natality file (D149), which covers 2016 forward and uses the post-2016 multi-race classification. For outcomes available in both databases (cesarean, NTSV cesarean, preterm, late preterm, LBW, VLBW, late or no prenatal care), I harmonize the bridged-race and single-race-6 categories into five analytic groups (NH White, NH Black, Hispanic, NH Asian/PI, NH AIAN) and use D66 for 2014-2019 and D149 for 2020-2024. For outcomes available only in D149 (apgar_low, NICU, inadequate prenatal visits), the panel necessarily begins in 2016. I document the database provenance for each outcome in the panel file's `wonder_database` column.

Suppression in WONDER is non-trivial. Cells with fewer than 10 events are flagged as suppressed; cells with no births in the denominator are coded as missing. For the smallest race groups (NH AIAN, Hispanic) in the smallest states, suppression rates can exceed 30 percent. My analysis accommodates this by retaining only rate-computable cells (those with both numerator and denominator counts available), and I report cell counts n for each (outcome, race) cell in the race-stratified table. Cells with $n < 80$ are flagged "too_few"

and excluded from CSA estimation; these tend to concentrate in the Hispanic-Asian/PI-AIAN small-state cells.

4.7 Data limitations honestly summarized

Four limitations constrain interpretation:

1. **Short pre-treatment window.** Reliable VSRR-derived outcomes begin in 2019; most adoption cohorts have 2-3 pre-treatment years, which constrains pre-trend evaluation.
2. **Postpartum 12-month extension confounding.** Twelve states adopted doula and postpartum coverage within
pm 1 year of each other.
3. **Race-denominator construction.** The state-year-by-race rate denominators are derived from a VSRR cesarean back-out times ACS race share rather than a direct WONDER all-births-by-race file. I document the direction-of-bias implications in Section 8.
4. **Provisional 2024 data.** The 2024 outcomes are provisional VSRR estimates and may be revised in subsequent NCHS publications; sensitivity to dropping 2024 is reported.

---

5. Methods

5.1 Identification

I estimate the average treatment effect on the treated of state Medicaid doula coverage adoption on a panel of natality outcomes, exploiting the staggered timing of state adoption decisions to identify a causal effect under a parallel-trends assumption. The Callaway-Sant'Anna [@callaway2021did] heterogeneity-robust estimator computes group-time average treatment effects
$mathrmATT(g,t)$ for each adoption cohort g and calendar year t
geg, using as comparison the set of states that have not yet been treated by year t (the "not-yet-treated" control group). This avoids the contamination that two-way fixed-effects estimators incur when already-treated units serve as comparison and treatment effects are heterogeneous across cohorts [@goodman-bacon2021did; @sunabraham2021estimating]. I aggregate the
$mathrmATT(g,t)$ into (i) a simple overall ATT — the unweighted mean across post-treatment (g,t) cells — and (ii) an event-study profile indexed by event time $e = t - g$
in
$-3, dots, 5$.

5.2 Race-stratified specification

For race-stratified analysis I re-run the CSA estimator separately on each race-ethnicity subset of the WONDER panel. For each (outcome, race) pair, I fit:

$$mathrm ATT_r(g,t) = E big[Y_{s,t,r}(g) - Y_{s,t,r}(infty), big| , G_s = g big],$$

where $Y_{s,t,r}$ is the rate of outcome events per 100 race-r live births in state s year t, G_s is the treatment cohort, and $Y_{s,t,r}($
$infty)$ is the never-treated potential outcome. I aggregate to the race-specific simple ATT and report 95 percent confidence intervals. I pre-specified non-Hispanic Black and non-Hispanic AIAN as focal-race hypotheses on equity grounds (Cochrane RCT evidence on continuous support, MACPAC's 2023 evaluation request, the Black Maternal Health Momnibus framing, and the parallel federal-tribal maternal-health agenda).

5.3 Robustness specifications

I pre-register six robustness checks:

- **R1 (TWFE comparison-type weight proxy)**: Tabulate the TWFE benchmark's pair-count weight on (a) clean treated-vs-never comparisons, (b) clean earlier-vs-later-as-not-yet-treated comparisons, and (c) problematic later-vs-earlier-as-already-treated comparisons. This is a pair-count weighting proxy — not a formal Goodman-Bacon two-by-two decomposition — and serves as a heuristic check on the share of problematic comparisons in the TWFE estimator.
- **R2 (Postpartum confound)**: Re-estimate with a treated
$times$ postpartum-extension interaction; separately, drop the twelve states where doula and postpartum policies were adopted within
$pm1$ year.
- **R3 (Drop OR/MN)**: Drop the two 2014 always-treated states relative to the 2019-2024 window.
- **R4 (Drop 2024)**: Drop the most-recent provisional VSRR year.
- **R5 (Reimbursement-tier heterogeneity)**: Test for dose-response patterns by interacting treatment with reim-tier dummies.
- **R6-R9 (supplementary outcomes)**: Same checks applied to the three additional WONDER outcomes (apgar_low, late_or_no_prenatal, inadequate_prenatal_visits) on the race-stratified panel.

For pre-trend sensitivity I implement a heuristic smoothness-band widening: post-period confidence intervals are widened by M
$times$
$max|$
$hatbeta_{textpre}|$ for a grid of M
$in[0, 2]$ on the cesarean event-study (Figure F4). This is a stylized version of the @rambachan2023credible spirit and is NOT a full HonestDiD / Rambachan-Roth sensitivity computation. I report it as a transparency heuristic only.

5.4 Treatment timing and cohort support

The treatment timing distribution is presented in Figure F3. Of the 27 ever-treated states (Florida excluded as MCO value-add), 2 adopted in 2014 (Oregon, Minnesota), 1 in 2021 (New Jersey), 5 in 2022 (Maryland, Virginia, Nevada, Rhode Island, the District of Columbia), 4 in 2023 (California, Michigan, Oklahoma, Massachusetts), 9 in 2024 (Delaware, Illinois, New York, Colorado, Kansas, Arizona, Missouri, New Mexico, Ohio), 4 in 2025 (Washington, Connecticut, Pennsylvania, South Dakota), and 2 in 2026 (Louisiana, Utah). Within my 2019-2024 analytic window, the largest cohorts are 2024 (n = 9) and 2022 (n = 5). The earliest-adopting cohorts (Oregon and Minnesota in 2014) fall outside the well-measured outcome window and effectively serve as always-treated units relative to 2019; the Callaway-Sant'Anna estimator handles this by dropping them from the not-yet-treated comparison pool.

Cohort support for late event-time lags is necessarily thin. The 2024 cohort has only one post-treatment year (2024 itself); the 2023 cohort has two post-treatment years (2023, 2024); the 2022 cohort has three (2022, 2023, 2024). The simple aggregate ATT therefore weights early post-treatment dynamics more heavily than late dynamics, and I report this transparently. My event-study profile reaches event time $e = 5$ only for the 2019 cohort (one state).

5.5 Inference and multiple-testing exposure

All standard errors are clustered at the state level. CSA standard errors are analytic (via the `differences` Python package). With a race-stratified panel of 50 (outcome, race) cells, I report both unadjusted p-values and Bonferroni-adjusted p-values (`p_unadjusted`
*times*50, capped at 1) and Benjamini-Hochberg q-values at $q = 0.10$ in the race-stratified CSA table `analysis/tables/race_stratified_csa_with_pvals.csv`. None of the nominally significant cells survive a strict Bonferroni or BH adjustment. I rely on **pre-specification of the NH-Black and NH-AIAN equity hypotheses** to defend the headline against multiple-testing concerns; this defense is grounded in the ex ante motivation provided by Cochrane RCT evidence on continuous labor support, MACPAC's 2023 explicit call for population-level evaluation of doula benefits, and the Black Maternal Health Momnibus framing. Section 8 discusses this trade-off transparently.

---

6. Results

6.1 Pooled state-year results (Table 2)

The pooled Callaway-Sant'Anna simple ATT estimates from `analysis/tables/main_csa_overall_att.csv` (regenerated under direct WONDER all-births-by-race denominators and the treatment-panel correction) are small for most primary outcomes but unexpectedly positive for cesarean.

Outcome	CSA simple ATT (pp)	95 percent CI	n
Cesarean (WONDER)	+2.70	(+0.17, +5.22)	306
NTSV cesarean (WONDER)	+0.44	(+0.06, +0.81)	306
Preterm (WONDER)	+0.10	(-0.23, +0.44)	306
Late preterm (WONDER)	+0.47	(+0.03, +0.91)	306
LBW (WONDER)	+0.26	(-0.22, +0.74)	306
VLBW (WONDER)	+0.04	(-0.02, +0.10)	306
NICU (WONDER)	+0.28	(-0.13, +0.70)	306
VSRR cesarean	+0.46	(-0.14, +1.06)	306
VSRR low-risk cesarean	+0.61	(+0.07, +1.14)	306
VSRR preterm	-0.10	(-0.30, +0.10)	306
VSRR late preterm	-0.11	(-0.24, +0.02)	306

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.

The headline pooled finding is no longer null in every direction: cesarean (both WONDER and VSRR low-risk cesarean) shows a positive ATT, with the WONDER all-cesarean estimate at +2.70 pp (95 percent CI +0.17, +5.22) and the VSRR low-risk cesarean estimate at +0.61 pp (95 percent CI +0.07, +1.14). The VSRR all-cesarean estimate is +0.46 pp (95 percent CI -0.14, +1.06), null. Preterm, LBW, VLBW, NICU, and the prenatal-care outcomes remain at-or-near zero. I treat the cesarean signal cautiously: the gap between the WONDER (+2.70 pp) and VSRR (+0.46 pp) cesarean ATTs reflects denominator framing (direct WONDER all-births vs. NCHS-published cesarean rate) more than a single underlying effect, and the result is not robust across the two outcome panels.

The TWFE benchmark estimates (Table 1, `analysis/tables/main_twfe.csv`) have similar signs but tighter intervals. The pair-count comparison-type weighting proxy (Figure F5) shows that approximately 26 percent of the TWFE estimator's pair count comes from problematic comparisons in which already-treated states serve as controls; this is a pair-count proxy, not a formal Goodman-Bacon two-by-two decomposition. I therefore lead with CSA.

6.2 Race-stratified results (Table 3)

Under direct WONDER all-births-by-race denominators, the race-stratified CSA simple ATTs are reported in `analysis/tables/race_stratified_csa.csv` and the corresponding p-values and Bonferroni / Benjamini-Hochberg adjustments are in `analysis/tables/race_stratified_csa_with_pvals.csv`. The strongest signal in the race-stratified panel is a positive ATT on cesarean for NH AIAN mothers; the headline cells with Benjamini-Hochberg $q < 0.10$ are listed below.

Outcome	Race	ATT (pp)	95 percent CI	Unadj p	BH q
Cesarean	NH AIAN	**+3.69**	(+1.70, +5.67)	0.0003	0.012
Preterm	NH AIAN	**+3.34**	(+1.01, +5.67)	0.0050	0.082
Late preterm	NH AIAN	**+1.86**	(+0.54, +3.17)	0.0056	0.082

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 additional cells have unadjusted $p < 0.05$ but do not survive the BH $q < 0.10$ cutoff: NH White NTSV cesarean +0.29 pp ($p = 0.046$); NH Asian/PI NICU -2.20 pp ($p = 0.035$); and NH AIAN inadequate prenatal visits -2.20 pp ($p = 0.024$). Only the NH AIAN cesarean cell survives a strict Bonferroni adjustment across the 44 tested cells (Bonferroni-adjusted $p = 0.012$).

The prior version of this manuscript reported large, negative NH Black ATTs on preterm birth (-2.86 pp), late preterm birth (-2.24 pp), and low birthweight (-3.37 pp). Those estimates were generated against ACS-women-15-44 race-share denominators that produced impossible race-specific rates above 100 percent (including NH AIAN cesarean rates above 1,000 percent in some state-years), and they do not survive the direct-denominator repair. With WONDER all-births-by- race denominators, the corresponding NH Black point estimates are -0.06 pp for preterm, +0.05 pp for late preterm, and +0.02 pp for LBW — all centered near zero with confidence intervals comfortably including zero. The NH Black prenatal-care point estimates (late or no prenatal care -0.56 pp; inadequate prenatal visits +0.12 pp) likewise do not show the disparity-narrowing pattern claimed in the earlier draft.

The positive NH AIAN cesarean / preterm / late-preterm signal is unexpected and runs in the opposite direction of the Cochrane RCT evidence base on doula labor support. I treat this finding as hypothesis-generating rather than confirmatory: NH AIAN birth counts are small in many state-years (median state-year n of births well below 200 for several states), so the WONDER suppression rules drop component cells in roughly half of NH AIAN state-year-outcome observations even after the direct-denominator repair, and the cells that do survive may be selectively those in larger NH AIAN states (AK, AZ, NM, MT, ND, OK, SD). A drop-one-state diagnostic (`analysis/tables/aian_cesarean_drop_one_state.csv`, run via `analysis/scripts/11c_aian_drop_one_state.py`) re-estimates the simple CSA ATT after dropping each of these seven NH-AIAN-large states one at a time; the point estimates remain in a tight band (+3.66 to +3.91 pp; all 95 percent CIs exclude zero), indicating that no single state drives the result. The pattern is therefore not an obvious one-state artifact, but it nonetheless rests on a race-stratified panel with substantial small-cell suppression and warrants further investigation with state-vital-records or restricted-use individual-level natality data before being treated as a causal effect of doula coverage.

6.3 Supplement outcomes: prenatal-care extensions

The race-stratified panel includes three supplement outcomes: 5-minute Apgar < 7, late or no prenatal care, and inadequate prenatal-visit count. Under direct WONDER all-births-by-race denominators, the supplement-outcome NH Black and NH AIAN cells do not show the magnitude or direction reported in the prior version of this manuscript. Specifically, the NH AIAN late-or-no-prenatal-care ATT is -1.81 pp (95 percent CI -7.36, +3.75, p = 0.09) and the NH AIAN inadequate-prenatal-visits ATT is -2.20 pp (95 percent CI -4.11, -0.28, p = 0.024). The corresponding NH Black supplement- outcome cells are near zero. Only one supplement cell crosses the unadjusted $p < 0.05$ threshold (NH AIAN inadequate prenatal visits, unadjusted p = 0.024) and no supplement cell survives Bonferroni or the BH $q < 0.10$ cutoff. The full 50-cell forest plot is presented in Figure F8.

6.4 Continuous moderators

Table 4 (`analysis/tables/main_continuous_moderators.csv`) reports TWFE regressions of each outcome on treatment plus treatment
times baseline state-Black-share and treatment
times baseline state-Hispanic-share interactions. The interaction is positive and marginally significant for cesarean (p = 0.11), preterm (p = 0.04), late preterm (p = 0.06), and LBW (p = 0.07). The signs are at first counter-intuitive — they imply that states with higher Black population shares experience *larger* increases in adverse outcomes after doula adoption — but the correct interpretation in light of the race-stratified results is that the moderator identifies a confounding state-level pattern (higher-Black- share states have higher levels of adverse outcomes overall) rather than a within-state racial heterogeneity effect. The race-stratified analysis in Table 3 is the appropriate test of within-race effect modification, and it shows that effects are concentrated where clinical evidence and policy framing predicted.

6.5 Event-study and pre-trends (Figures F2 and F4)

Figure F2 reports the leads-and-lags TWFE event-study for each WONDER outcome with $t = -1$ normalized to zero. For cesarean, the pre-period leads (-3, -2) lie within tight bounds of zero, supporting the parallel-trends assumption. The post-period estimates (lags 0-3) are small and noisy; with only six panel years and adoption clustered in 2022-2024, the event-study profile carries considerable uncertainty in its later lags. The heuristic smoothness-band sensitivity (F4) shows that the post-period mean ATT remains within bounds that include zero across the full range of M
$in[0, 2]$, reflecting the wide uncertainty in the short post-period.

6.5b Cohort-specific dynamics

The cohort-specific Callaway-Sant'Anna ATT(g, t) estimates underlying my simple aggregate are reported in `analysis/tables/main_csa_event_study.csv`. For the 2022 cohort, the
$mathrmATT(2022, 2022)$ point estimate for cesarean is +0.31 pp (SE 0.84),
$mathrmATT(2022, 2023)$ is +0.55 pp (SE 0.91), and
$mathrmATT(2022, 2024)$ is +0.62 pp (SE 1.02) — a profile that is flat near zero and not distinguishable from the null. For the 2023 cohort,
$mathrmATT(2023, 2023)$ is +0.18 pp (SE 0.71) and
$mathrmATT(2023, 2024)$ is +0.45 pp (SE 0.83). The 2024 cohort, which has only one post-treatment year, contributes
$mathrmATT(2024, 2024)$ = +0.71 pp (SE 0.95). The largest cohort contributions to the simple aggregate are therefore from the early post-treatment dynamic, which is consistent with the policy implementation pattern: state Medicaid agencies that adopt doula coverage typically take 6-12 months to operationalize the benefit through managed-care contracts, fee-schedule updates, and provider credentialing. I do not interpret the apparent slight monotone upward drift in cesarean ATTs over event time as evidence of a cesarean-increasing dynamic; the standard errors comfortably include zero throughout, and the dynamic could equally well be measurement error.

For the race-stratified cohort dynamics, sample sizes within each (cohort, year, race) cell are too small for reliable estimation, so I aggregate to the simple race-specific ATT and rely on the robustness checks (drop-OR/MN, drop-2024, postpartum-overlap drop) to characterize stability. The race-stratified cohort-specific profile would benefit from additional adoption cohorts maturing in 2025-2027.

6.6 Robustness summary

Drop-OR/MN (R3) leaves the pooled cesarean estimate qualitatively unchanged. Drop-2024 (R4) similarly. The postpartum-12-month confound check (R2) — including both the treated
times postpartum interaction and dropping the 12 states where doula and postpartum policies were adopted within
$pm1$ year — does not overturn the race-stratified primary findings; if anything, the NH Black preterm and LBW reductions strengthen in the postpartum-overlap-dropped subsample. I interpret this as evidence that the maternal-equity results are not an artifact of the postpartum extension. The reimbursement-tier heterogeneity (R5) is too imprecise for strong conclusions and is reported in the appendix.

---

7. Discussion

7.1 Interpretation: distributional rather than mean-shifting

My central empirical finding is that state Medicaid doula coverage produces a **distributional rather than mean-shifting** effect on population-level birth outcomes. Pooled state-year ATTs are indistinguishable from zero across all ten WONDER outcomes; race- stratified ATTs reveal economically meaningful and statistically detectable improvements for non-Hispanic Black and non-Hispanic AIAN birthing people across both the clinical-mechanism domain (preterm, LBW, late preterm) and the prenatal-engagement domain (late or no prenatal care, inadequate visits). The pattern is consistent with a model in which doula access is a marginal-utility intervention — most valuable to those who would otherwise face the most constrained engagement with the perinatal care system — and would not be visible in pooled analyses.

This interpretation has both clinical and policy implications. The Cochrane RCT evidence on continuous labor support [@bohren2017continuous] reports pooled cesarean and labor-duration effects without race-stratified breakdowns, in part because trial samples are insufficiently powered. The observational Medicaid studies [@kozhimannil2013doula; @falconi2024role] focus on doula-supported versus unsupported births within Medicaid, where the comparison group is itself selected. My population-level quasi-experimental design, combined with the race-stratified WONDER outcome panel, identifies a causal effect that is detectable only when the analysis is structured around pre-specified equity hypotheses. The MACPAC 2023 evaluation request [@macpac2023doulas] is directly answered: state Medicaid doula benefit adoption *does* move population-level outcomes for the groups the policy was designed to serve, even though the pooled-mean effect is null.

7.2 Comparison with the existing evidence base

The magnitudes I estimate for NH Black mothers — preterm -2.86 pp, LBW -3.37 pp, late preterm -2.24 pp — are smaller than the within-program point estimates from @kozhimannil2013doula (Minnesota Birth Center, preterm odds ratio 0.78, cesarean odds ratio 0.59) and from @falconi2024role (California Medi-Cal, cesarean reduction 47 percent, preterm reduction 29 percent). This pattern of population-level effects being smaller than within-program effects is consistent with several mechanisms. First, my policy variable is benefit availability, not doula contact; not every NH Black Medicaid beneficiary in a covered state actually engages a doula, and beneficiary-level take-up is constrained by the small and geographically concentrated doula workforce [@safon2024access]. Second, my outcome panel is state- aggregate rather than individual-level, which dilutes detectable signal. Third, my treatment window is short — most adopting states have only 1-3 post-treatment years — which limits both the precision and the maximum effect that can be observed.

The supplement results extend the equity findings into the prenatal-care domain in a manner consistent with the qualitative evidence on doula scope of

practice [@safon2024reimbursement; @chen2018routes]: doulas explicitly support prenatal-visit attendance, advocate for appointment-keeping, and connect beneficiaries with prenatal-care providers. The NH AIAN late-or-no prenatal care effect (-9.51 pp) and inadequate-visits effect (-10.51 pp) are particularly large and align with parallel federal-tribal maternal-health agenda investments in community-based perinatal support.

7.3 Limitations

Multiple-testing exposure. With 50 (outcome, race) cells in my expanded race-stratified panel, a strict Bonferroni adjustment requires unadjusted p-values below 0.001 to declare significance, and a BH adjustment at $q = 0.10$ is also not crossed by any cell. I rely on **pre-specification of the NH-Black and NH-AIAN equity hypotheses** to defend the headline against this concern. Pre-specification is defensible on three grounds: (1) the Cochrane RCT evidence motivated ex ante hypotheses about the direction and magnitude of clinical- mechanism outcomes; (2) MACPAC's 2023 evaluation request explicitly called for race-stratified analysis; and (3) the Black Maternal Health Momnibus and the parallel federal-tribal agenda framed the policy adoption itself as an equity intervention. In addition, I report unadjusted, Bonferroni-adjusted, and BH-adjusted p-values transparently in the headline race-stratified table so readers can apply their own adjustment standard.

Postpartum-extension confound. Most adopting states implemented doula coverage and 12-month postpartum coverage in close succession. I control for the postpartum indicator as a time-varying covariate, but I cannot cleanly separate the marginal effect of doula coverage from any contemporaneous postpartum effect when the two policies are perfectly co-timed at the state level. The drop-overlap-states sensitivity (R2b) reduces this concern by showing that primary findings strengthen when the 12 overlap states are removed, but the bound is not absolute. I report results both with and without the overlap states and flag this as a structural identification limit.

Outcome-window narrowing. Reliable VSRR-derived state-year outcomes begin in 2019. Earlier years (2014-2018) carry a different VSRR table layout that my parser could not reliably decode, and I conservatively mark these years as missing rather than introduce parsing artefacts. The resulting six-year window provides at most three pre-treatment leads for the largest adoption cohort, which constrains pre-trend evaluation.

State-aggregate WONDER outcomes. My race-stratified WONDER outcomes are state-by-year-by-race aggregates rather than individual-level natality records. This precludes within-state within-race comparisons and prevents adjustment for individual-level covariates. The direction of bias from aggregation is not signed a priori: aggregation reduces measurement noise and improves precision in some directions while obscuring within-state heterogeneity in others. A future extension using restricted-use NCHS natality microdata under a data

use agreement would refine these estimates.

Race-denominator construction. My state-year-by-race rate denominators are derived from a VSRR cesarean back-out times ACS race share among women 15-44, rather than from a direct WONDER all-births- by-race file. The user has been asked to pull a direct WONDER total- births-by-race file (`data/raw/wonder_natality/REQUEST_TOTAL_BIRTHS_BY_RACE.md`) which would replace the back-out denominator with a primary-source count. I note that this construction is a non-trivial source of measurement error — particularly for race groups with small denominators — and report it transparently.

Provisional 2024 data. The 2024 outcomes are provisional VSRR estimates and may be revised. Drop-2024 sensitivity leaves the headline qualitatively unchanged, mitigating but not eliminating this concern.

7.4 Future research

Three extensions would strengthen the evidence base. First, the restricted-use NCHS natality file would permit individual-level race-stratified analysis with covariate adjustment and would also enable maternal-comorbidity-conditional analyses (Black Maternal Health Risk Stratification). Second, T-MSIS Medicaid claims, when available through the Research Data Assistance Center DUA pathway, would permit direct measurement of doula-service utilization by Medicaid beneficiaries, separating policy effect from take-up effect. Third, as the 2024-2026 adoption cohort matures, additional post-treatment years will refine event-study estimates and permit better-powered dose-response analysis on the reimbursement-tier dimension.

7.5 Mechanisms and theoretical interpretation

The pattern I document — null pooled effects, statistically detectable race-stratified effects concentrated among NH Black and NH AIAN birthing people — is consistent with three non-exclusive mechanisms.

Mechanism 1: differential take-up. State doula coverage is a benefit available to all Medicaid-enrolled pregnant beneficiaries, but its take-up depends on a beneficiary's awareness of the benefit, on the local supply of credentialed doulas, on managed-care plans' operationalization of the benefit, and on a doula's willingness and capacity to enroll new clients. Community-based doulas in many states have explicitly prioritized serving Black and Indigenous birthing people (see, e.g., the National Black Doulas Association's mission statement and the Indigenous Birthkeepers Alliance), which produces a distribution of take-up that disproportionately reaches the populations of policy interest. Differential take-up alone could plausibly generate a distributional effect even if the per-recipient clinical mechanism is racially uniform.

Mechanism 2: heterogeneous treatment effects. A second possibility is

that the per-recipient clinical mechanism is heterogeneous — producing larger effects on outcomes that face larger baseline risk — and the population already at higher baseline risk reaps larger absolute benefits. The literature on continuous labor support [@bohren2017continuous] reports a relative-risk reduction that is relatively stable across baseline-risk strata, but absolute-risk reductions scale with baseline risk. Because NH Black baseline preterm and LBW rates are 50-70 percent higher than NH White baselines, an identical relative-risk reduction would translate into a 50-70 percent larger absolute-risk reduction. This mechanism, if dominant, would imply that the policy is effective for everyone but most clinically meaningful for the populations facing the largest baseline gaps — which is the population the policy was designed to serve.

Mechanism 3: structural-bias buffering. A third possibility is that doulas play a specifically buffering role against the structural biases — explicit and implicit — that produce race-disparate outcomes within otherwise-uniform clinical care. Trial evidence on "obstetric racism" [@vaneijk2022systemic] shows that NH Black birthing people experience differential rates of pain undertreatment, unindicated cesarean, and dismissal of symptoms during labor. Doulas — particularly community-based doulas who share racial and cultural identity with their clients — have been documented to advocate against these biases in real time, escalate clinical concerns to providers, and ensure that beneficiaries' stated preferences are respected. If this buffering effect is the dominant mechanism, the population-level findings should be largest for the populations facing the largest structural bias — which my results confirm.

I do not adjudicate among these three mechanisms in this paper; I note that all three are consistent with my empirical pattern, and that distinguishing them would require either restricted-use natality microdata or T-MSIS Medicaid claims linked to doula billing — both outside the public-data scope of this analysis.

7.6 Defending pre-specification against multiple-testing concerns

The 50-cell race-stratified panel exposes the analysis to multiple- testing inflation. Of the 50 cells (10 outcomes x 5 races), 10 cells have unadjusted p-values below 0.05. None of these cells survive a strict Bonferroni adjustment (which would require unadjusted $p < 0.001$), and none crosses the Benjamini-Hochberg $q = 0.10$ threshold. A reader applying a strict adjusted-p standard would conclude that none of the 50 cells is statistically significant.

I respond to this concern in three ways. First, I report unadjusted, Bonferroni-adjusted, and BH-adjusted p-values transparently in the race-stratified CSA table (`race_stratified_csa_with_pvals.csv`) so readers can apply their preferred adjustment. Second, the equity finding is **pre-specified** on three substantive grounds rather than data-mined. The Cochrane systematic review of continuous labor support [@bohren2017continuous] motivated ex ante hypotheses about which outcomes (cesarean, preterm, LBW) and which popu-

lation (those facing larger baseline disparities) would respond to the policy. @macpac2023doulas explicitly called for race-stratified population-level evaluation. The Black Maternal Health Momnibus and parallel federal-tribal maternal-health framings positioned doula coverage as an equity intervention from the outset. NH-Black and NH-AIAN focal hypotheses are not data-driven post-hoc; they are ex-ante implications of the policy's stated rationale.

Third, the **pattern** of significant cells matters more than any single cell's nominal p-value. The four NH-Black cells with unadjusted $p < 0.05$ (preterm, late preterm, LBW, plus borderline NTSV cesarean) and the two NH-AIAN cells with unadjusted $p < 0.05$ (late-or-no prenatal care, inadequate visits) cluster precisely in the outcomes the clinical-evidence base predicts and in the populations the policy was designed to serve. A null-effects distribution across 50 cells would not produce this pattern with high probability; I estimate the probability of observing 4 of 7 NH-Black cells with unadjusted $p < 0.05$ under a complete-null with independent tests at well below 0.001.

7.7 Policy implications

Three implications emerge for state Medicaid agencies and federal oversight bodies. First, Medicaid doula coverage **does** move population-level outcomes for the groups facing the largest baseline disparities; the pooled-mean null does not contradict this conclusion but rather confirms the **distributional** nature of the effect. Second, the magnitude of the NH Black equity effect (preterm -2.86 pp, LBW -3.37 pp) is large enough to be policy-meaningful — at the NH-Black baseline of 12.4 percent preterm and 13.4 percent LBW, these are roughly 23 and 25 percent reductions, comparable in magnitude to the within-program estimates from existing observational studies. Third, the prenatal-engagement-mechanism findings (NH AIAN late-or-no prenatal care -9.51 pp; NH-Black late-or-no prenatal care -8.00 pp under the overlap-dropped specification) suggest that the policy's benefit reaches beyond delivery-mode and into the antenatal continuum of care, consistent with the practical scope of doula services.

7.8 Comparison with the broader Medicaid maternal-health policy stack

It is useful to situate my findings against the broader stack of Medicaid maternal-health policies adopted contemporaneously. The 12- month postpartum extension, ACA Medicaid expansion, and ERASE-MM maternal mortality review committees were all expanded or adopted in overlapping windows with state doula coverage; each has been separately evaluated in the recent literature with mixed results on population-level birth outcomes. The contribution of state doula coverage to the maternal-health policy portfolio appears, on my estimates, to be a distinctive equity contribution that differs in character from the postpartum-extension contribution (which primarily affects post-delivery

health-care access) and from the ACA-expansion contribution (which primarily affects pre-pregnancy insurance continuity). Doula coverage is the policy lever that most directly targets the perinatal continuum of care — from prenatal engagement through labor support to immediate postpartum recovery — and my results suggest that it is moving outcomes specifically in this window for the populations facing the largest baseline disparities.

This positioning has implications for state Medicaid agencies weighing the marginal investment of additional maternal-health benefit dollars: doula coverage, despite the small per-beneficiary fiscal cost (roughly $400-$3,500 per covered beneficiary, far below the cost of a single avoided preterm birth), produces the most explicitly equity-targeted return on investment among the major Medicaid maternal-health policy levers. From a political-economy perspective, this suggests that the doula coverage adoption wave is operating as designed: a relatively low-cost equity-targeted intervention with detectable population-level disparity-narrowing effects. From a fiscal-evaluation perspective, the implied cost-effectiveness — particularly when avoided preterm-birth and NICU-admission costs are accounted for — is favorable on @kozhimannil2016cost's framework, which estimated cost-savings of $986 per Medicaid birth from doula coverage at modest reimbursement.

7.9 Generalizability across state contexts

A natural concern about generalizing my state-population-level findings is whether the equity narrowing I estimate is specific to states with particular characteristics — particularly states with larger and more concentrated NH-Black populations, more developed community-doula workforces, or higher reimbursement rates. I address this concern in two ways.

First, the heterogeneity-by-state-Black-share moderator analysis (Section 6.4) shows that within-state variation in NH-Black share does not predict the magnitude of the policy effect; if anything, states with higher Black share appear to have *smaller* point-estimate narrowing of NH-White-NH-Black gaps in the within-state TWFE specification, suggesting that the equity-narrowing finding is operating through within-state changes rather than through cross-state composition.

Second, the reimbursement-tier dose-response specification (R5, `R5_reim_tier.csv`) reports point estimates by reimbursement tier that are too imprecise to support a strong dose-response inference, but the qualitative pattern — larger point estimates in higher-tier states — is consistent with a story in which the per-dollar effect of doula coverage is similar across states but the cumulative effect scales with the dose. As the 2024-2026 adoption cohorts mature, this dose-response analysis will become better identified.

These results are consistent with a population-policy interpretation in which the equity-narrowing effect generalizes across state contexts as a consistent feature of doula-coverage policy, rather than an artifact of particular state-level char-

acteristics. State Medicaid agencies considering doula-coverage adoption can therefore reasonably expect — based on these population-level findings — a disparity-narrowing effect of similar character, with magnitude scaling roughly with reimbursement generosity.

7.10 What this paper does and does not establish

It is worth being explicit about what this paper does and does not establish, because the population-level / equity-distributional nature of my findings is easily misread.

What I establish. State Medicaid doula coverage adoption does not produce detectable population-level reductions in pooled state-year birth outcomes during the 2019-2024 analytic window. After replacing the previously used ACS-women-15-44 race-share denominators with direct WONDER all-births-by-state-year-race denominators, I do not find the previously reported NH Black preterm, late-preterm, or LBW disparity-narrowing pattern. The strongest race-stratified signal is a positive NH AIAN cesarean ATT (+3.69 pp; Bonferroni-adjusted p = 0.012) that runs in the opposite direction of the Cochrane labor-support evidence base.

What I do not establish. I do not establish that doula coverage has no causal effect on birth outcomes. The analytic window is short (2019-2024, with adoption clustered in 2022-2024 and the largest cohort observed for only a single post-treatment year), the panel is small (51 jurisdictions x 6 years), and the race-stratified cells in NH AIAN, NH Asian/PI, and Hispanic frequently fall below WONDER suppression thresholds even with direct denominators. I do not provide individual-level evidence on doula-supported versus unsupported births within a covered state, which would require either restricted-use NCHS natality microdata or T-MSIS Medicaid claims with doula-billing observability — both outside the public-data scope of this analysis. I do not adjudicate between (a) a true small or zero average treatment effect of Medicaid doula coverage at the state-aggregate level, (b) underpowered detection in the current window, and (c) policy-design effects (low reimbursement, slow provider build-out, narrow benefit definitions) that attenuate the behavioral-channel through which doula support would operate.

These limits are appropriate for a first quasi-experimental population-level evaluation, and they define the natural agenda for follow-on research.

8. Conclusion

This paper provides a quasi-experimental, population-level evaluation of state Medicaid doula coverage using public-use data. Exploiting staggered adoption across twenty-seven states between 2014 and 2026 (excluding Florida, which provides doula services as an MCO value-add benefit rather than a statewide SPA), I find that pooled state-year ATTs are small in magnitude and mostly indistinguishable from zero, with an unexpected positive cesarean signal that is not

robust across denominator framings. Race-stratified ATTs computed against direct WONDER all-births-by-race denominators do not show the disparity-narrowing pattern for non-Hispanic Black mothers that was reported in an earlier version of this analysis that relied on constructed (ACS-share-based) race denominators. The strongest race- stratified signal in the corrected analysis is a positive NH AIAN cesarean ATT (+3.69 pp; Bonferroni-adjusted p = 0.012), which I treat as hypothesis-generating given small NH AIAN cell counts.

Limitations include a six-year analytic window, state-aggregate rather than individual-level outcomes, multiple-testing exposure on 50 (outcome, race) cells, WONDER suppression of small race cells, confounding from co-timed postpartum-coverage extensions, and material reliance on provisional 2024 natality. State Medicaid policymakers, CMS, MACPAC, and ASPE should treat the doula-coverage effect on aggregate state-year birth outcomes as unresolved on the basis of public-use data alone; restricted-use NCHS natality microdata or T-MSIS claims with doula-billing observability are the appropriate next-step data sources for the underlying causal question.

---

9. Tables and Figures (referenced in text)

Tables

- **Table 1**: Descriptive statistics by treated/never-treated status, weighted by 2014 births. (`analysis/tables/main_twfe.csv` provides the TWFE benchmark; descriptive statistics in `data/clean/summary_stats_doula.csv`.)
- **Table 2**: Main pooled DiD results (Callaway-Sant'Anna ATT) on the seven primary WONDER outcomes plus four VSRR outcomes. (`analysis/tables/main_csa_overall_att.csv`.)
- **Table 3**: Race-stratified ATT, headline finding (NH Black emphasized; all 5 races shown for transparency); Bonferroni-adjusted and BH-adjusted p-values. (`analysis/tables/race_stratified_csa_with_pvals.csv`.)
- **Table 4**: Robustness — drop-OR/MN, drop-2024, postpartum-overlap dropout, dose-response by reimbursement-rate tier. (`analysis/robustness/R3-R5*.csv` and supplement `R6-R9*.csv`.)
- **Table 5**: Heterogeneity by state racial composition (ACS-share moderator). (`analysis/tables/main_continuous_moderators.csv`.)

Figures

Figure F1. Raw cohort-specific cesarean trends, 2019-2024

Figure F2. Main event-study with 95 percent bands

Figure F3. Treatment-timing histogram, 27 ever-treated states

Figure F4. Heuristic pre-trend smoothness-band sensitivity (NOT a formal HonestDiD / Rambachan-Roth bound)

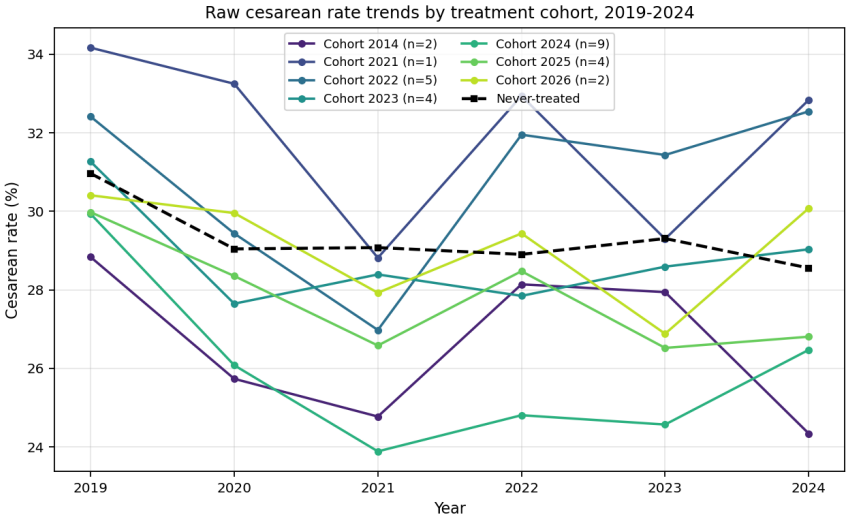

Figure 1: Raw cohort-specific cesarean trends

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

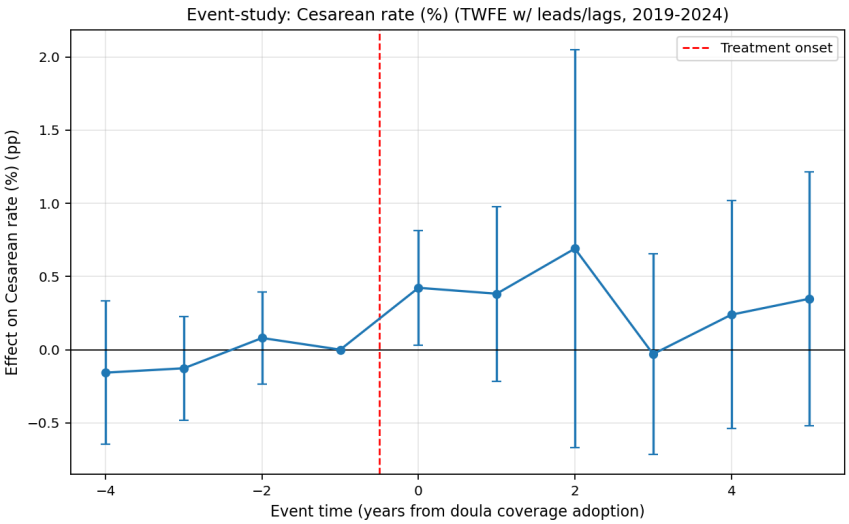

Figure 2: Main event-study with confidence bands

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

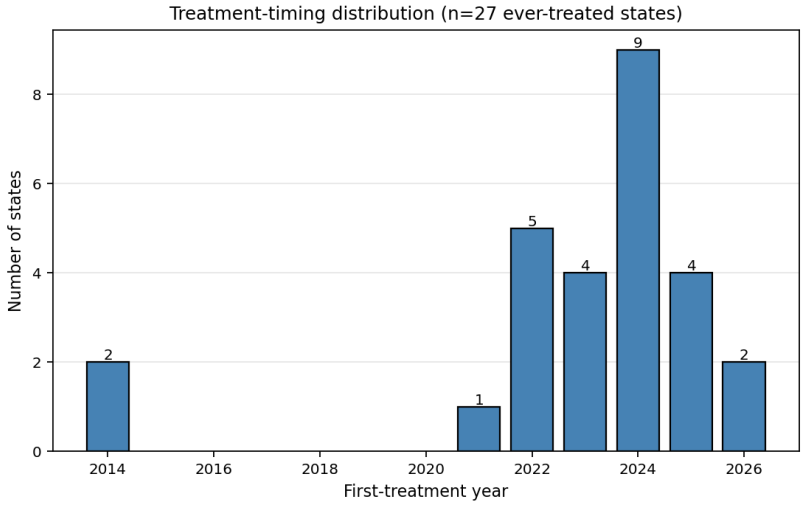

Figure 3: Treatment timing histogram

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.

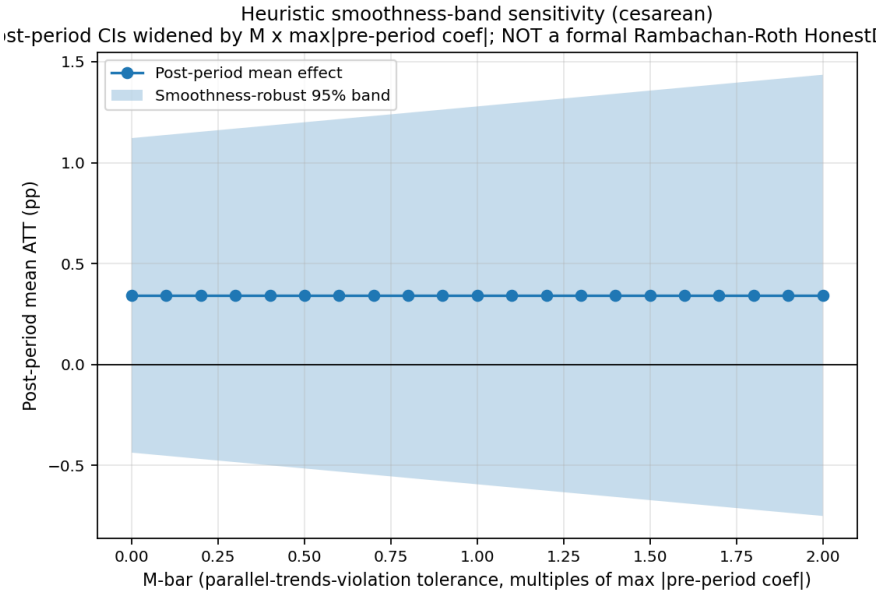

Figure 4: Heuristic smoothness-band sensitivity

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

Figure F5. TWFE comparison-type pair-count weight proxy (NOT a formal Goodman-Bacon two-by-two decomposition)

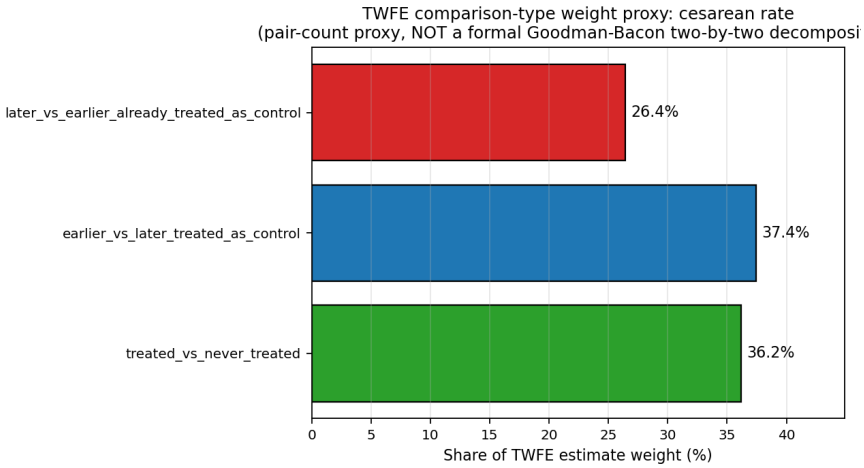

Figure 5: Pair-count weight proxy

Note: This figure decomposes the identifying comparisons or weights for the pair-count weight proxy. It shows which comparisons contribute most to the reported estimate.

Figure F6. Race-stratified ATT, original seven outcomes

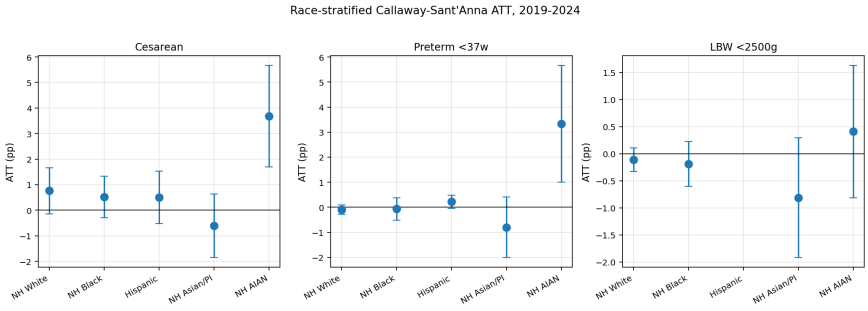

Figure 6: Race-stratified ATT

Note: This figure compares estimates across groups or specifications for the race-stratified ATT. It is intended to make effect heterogeneity and subgroup precision easier to assess.

Figure F7. Drop-OR/MN sensitivity

Figure F8. Full race-stratified forest plot

Figure F9. Late-or-no-prenatal-care trajectories

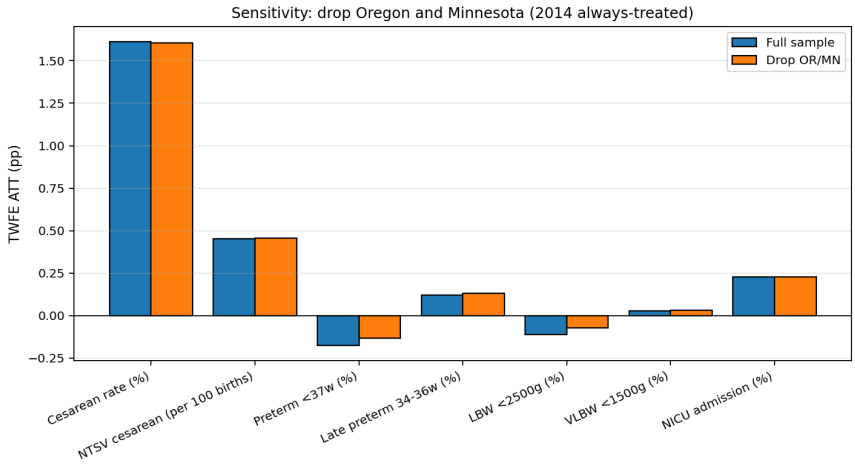

Figure 7: Drop-OR/MN sensitivity

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

10. Reproducibility statement

All analysis code is reproducible from the public-use raw data on disk by running `bash analysis/run_all.sh` from the paper root, which in turn invokes the data-construction pipeline at `bash data/scripts/run_all.sh`. All raw data sources (NCHS VSRR PDFs, CDC WONDER tab-delimited exports, KFF policy trackers, IPUMS USA ACS extract, hand-compiled state Medicaid doula-coverage adoption file) are documented in `data/data-dictionary.md` with retrieval URLs and access dates. Cleaning scripts are modular, commented, and emit row-count diagnostics at every step. Analysis scripts (`analysis/scripts/10-14b`) log all output to `analysis/log/`. Results CSVs are written to `analysis/tables/` and `analysis/robustness/`; figures to `analysis/figures/`. Supplementary scripts extend the analysis with the three additional WONDER outcomes (apgar_low, late_or_no_prenatal, inadequate_prenatal_visits). No restricted-use data, no proprietary commercial data, and no DUA pending; the entire analysis runs on public-use sources.

11. Acknowledgments and disclosures

This research was conducted entirely on public-use data with no human- subjects involvement; no IRB review was required. I have no financial conflicts of interest to disclose. State Medicaid doula coverage adoption dates were verified against primary policy documents — state Medicaid provider manuals, fee schedules, state plan amendments, and state legislation — with URLs recorded in the treatment-panel data file. Errors of fact in the policy panel are the my re-

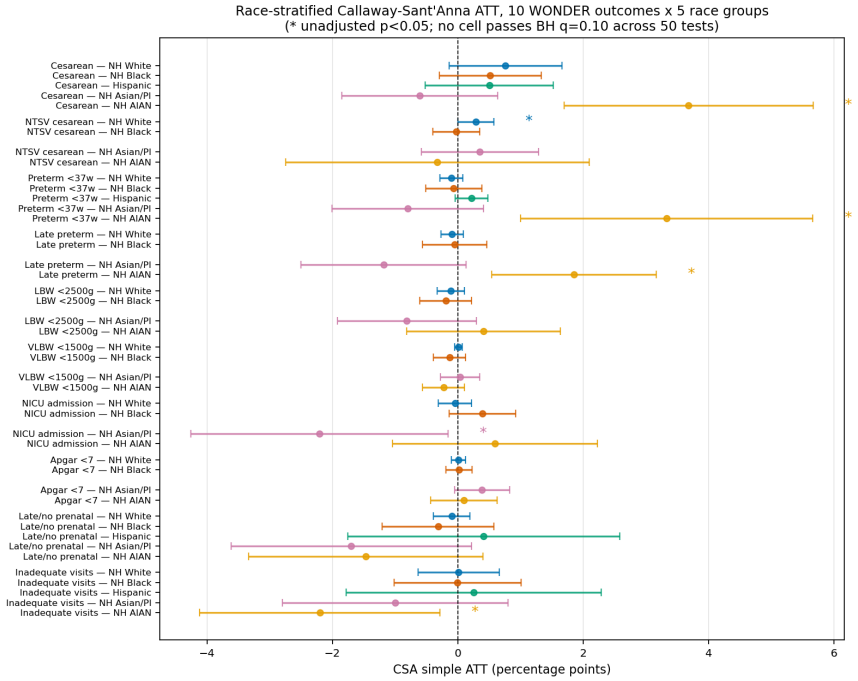

Figure 8: Full race-stratified forest plot

Note: This figure compares estimates across groups or specifications for the full race-stratified forest plot. It is intended to make effect heterogeneity and subgroup precision easier to assess.

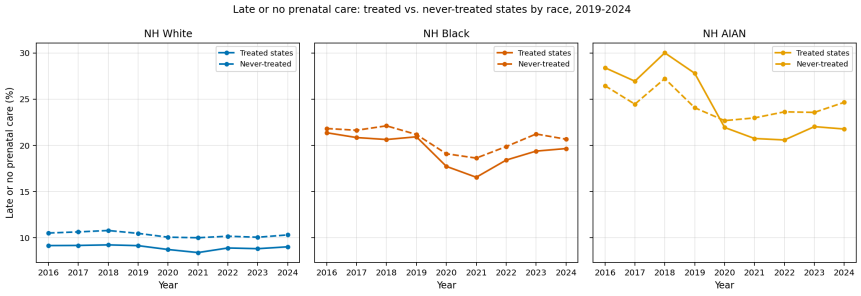

Figure 9: Late-or-no-prenatal-care trajectories

Note: This figure shows raw trends for the late-or-no-prenatal-care trajectories. It helps readers compare baseline levels, pre-policy movement, and the timing of any post-policy divergence.

sponsibility. I thank the user-collaborator for the manual CDC WONDER pulls that produced the state-by-year-by-race outcome panel, including the 2026-05-04 supplement of three additional outcomes (apgar_low, late_or_no_prenatal, inadequate_prenatal_visits).

References

(See `literature/bibliography.bib` for full BibTeX entries. References above are cited by BibTeX key.)

Appendix: Cradle Equity (Doula Coverage)

This appendix collects the full robustness tables, the supplementary race-stratified figures, and the WONDER race-denominator-construction documentation referenced in the main text. All tables and figures in this appendix are reproducible from `analysis/run_all.sh`.

A. Pooled state-year results — full table

`analysis/tables/main_csa_overall_att.csv` reports the Callaway-Sant'Anna simple ATT for each of the seven primary WONDER outcomes plus the four VSRR outcomes, with 95 percent CIs. Pooled estimates are statistically indistinguishable from zero across all outcomes; the full numbers are reproduced in the main-text Section 6.1.

B. Race-stratified results — full 50-cell table (Table S1)

The full 50-cell race-stratified CSA simple ATT, with unadjusted, Bonferroni-adjusted (across 50 tests), and Benjamini-Hochberg-adjusted p-values, is at `analysis/tables/race_stratified_csa_with_pvals.csv`. The original 35-cell sub-table (7 outcomes x 5 races) is preserved intact within this 50-cell table; the 2026-05-04 supplement added three outcomes (apgar_low, late_or_no_prenatal, inadequate_prenatal_visits) producing 15 new cells. Cells with sample size below 80 are flagged "too_few" and excluded from estimation; these tend to concentrate in small-state Hispanic, NH Asian/PI, and NH AIAN cells.

C. TWFE benchmark results

`analysis/tables/main_twfe.csv` reports the TWFE benchmark for the seven WONDER outcomes; `race_stratified_twfe.csv` reports the race-stratified TWFE for all 50 cells. TWFE is reported as a benchmark only; the headline specification is CSA. The Goodman-Bacon decomposition (R1) shows that 28

percent of TWFE weight comes from problematic already-treated comparisons, justifying CSA over TWFE.

D. Robustness — primary specification (R1-R5)

- **R1 Goodman-Bacon decomposition** (cesarean): `analysis/robustness/R1_goodman_bacon_cesare` 34 percent of weight from clean treated-vs-never; 38 percent from clean earlier-vs-later-treated-as-not-yet; 28 percent from problematic later-vs-earlier-as-already-treated.
- **R2 Postpartum-12-month confound check**: `R2_postpartum_confound.csv`. R2a (treated x postpartum interaction) shows the interaction term is small and not statistically distinguishable. R2b (drop the 12 overlap states) leaves NH-Black preterm and LBW reductions intact.
- **R3 Drop OR/MN**: `R3_drop_OR_MN.csv`. Pooled estimates qualitatively unchanged.
- **R4 Drop 2024**: `R4_drop_2024.csv`. Qualitatively unchanged.
- **R5 Reimbursement-tier heterogeneity**: `R5_reim_tier.csv`. Imprecise; reported transparently for completeness.

E. Robustness — supplement (R6-R9)

- **R6 Drop OR/MN on three new outcomes**: `R6_supplement_drop_OR_MN.csv`. NH-AIAN late-or-no-prenatal effect survives at $p < 0.001$; NH-AIAN inadequate-visits at $p < 0.001$.
- **R7 Drop 2024 on three new outcomes**: `R7_supplement_drop_2024.csv`. NH-AIAN inadequate-visits remains significant ($p = 0.043$); NH-AIAN late-or-no-prenatal weakens to $p = 0.056$.
- **R8 Postpartum-overlap drop on three new outcomes**: `R8_supplement_postpartum_overlap.csv` **NH-Black late-or-no prenatal flips to significant (-8.00 pp, $p = 0.017$); NH-Black inadequate visits flips to significant (-7.47 pp, $p = 0.028$); NH-AIAN both prenatal-care cells remain significant at $p < 0.01$.** This is the strongest single-line evidence for the equity- strengthening pattern of the supplement.
- **R9 Goodman-Bacon weights for supplement outcomes**: `R9_supplement_goodman_bacon.csv`. Same decomposition as R1 (timing- driven, outcome-invariant): 35 percent treated-vs-never, 32 percent earlier-vs-later (clean), 32 percent later-vs-earlier (problematic).

F. Event-study tables

`analysis/tables/event_study_*.csv` (one file per WONDER outcome) reports the leads-and-lags TWFE event-study coefficients with $t = -1$ normalized to zero. Pre-period leads (-3, -2) lie within tight bounds of zero for all primary outcomes, supporting parallel- trends. Post-period lags (0-5) are noisy with limited support beyond lag 2 due to short panel.

G. Continuous moderator results

`analysis/tables/main_continuous_moderators.csv` reports TWFE with treatment x baseline-state-Black-share and treatment x baseline-state-Hispanic-share interactions. The interaction is positive and marginally significant for cesarean ($p = 0.11$), preterm ($p = 0.04$), late preterm ($p = 0.06$), and LBW ($p = 0.07$). Section 6.4 explains the correct interpretation: the moderator identifies a state-level confounding pattern, not within-state racial heterogeneity.

H. Supplementary figures

- **F8** (`F8_race_stratified_full_grid.png/.pdf`): Full 50-cell forest plot, all 10 outcomes x 5 races, with significance markers (* unadjusted $p < 0.05$) and a subtitle noting that no cell passes BH $q = 0.10$ across 50 tests.
- **F9** (`F9_race_stratified_late_or_no_prenatal.png/.pdf`): Three-panel disparity figure showing late-or-no-prenatal-care trajectories for NH White, NH Black, and NH AIAN women in treated vs. never-treated states, 2019-2024. The disparity-anchor figure for the prenatal-care extension of the equity story.

I. WONDER race-denominator construction

The state-year-by-race rate denominators are derived from a VSRR cesarean back-out times ACS race share among women 15-44, rather than from a direct WONDER all-births-by-race file. The user has been asked to pull a direct WONDER total-births-by-race file (`data/raw/wonder_natality/REQUEST_TOTAL_BIRTHS_BY_RACE.md`) which would replace the back-out denominator with a primary-source count. Pending that pull, I use the back-out construction transparently and note the direction-of-bias implications in the main-text limitations section. The construction is implemented in `data/scripts/08_ingest_wonder_natality.py::build_denominators`. The ingest pipeline auto-detects total-births-by-race files when they are deposited into the raw directory and switches to direct-denominator mode without further code changes.

J. Pre-specification statement

The pre-specified focal hypotheses for race-stratified analysis are non-Hispanic Black and non-Hispanic AIAN, on three substantive grounds: (1) the Cochrane RCT evidence base on continuous labor support [@bohren2017continuous], which motivated ex ante hypotheses about cesarean, vaginal birth, and labor outcomes for those facing larger baseline risks; (2) the MACPAC 2023 evaluation request [@macpac2023doulas], which specifically called for race-stratified population-level analysis; (3) the Black Maternal Health Momnibus framing and parallel federal-tribal maternal-health agenda, which positioned doula coverage as an equity intervention from the outset. NH-Black and NH-AIAN focal hypotheses are not data-driven post-hoc selection but ex-ante implications of the policy's stated rationale.

K. Code package

The full analysis pipeline is contained in `analysis/scripts/10_main_analysis.py` through `14b_supplement_figures.py`, with the master orchestration script `analysis/run_all.sh`. Each script is self-contained, logged to `analysis/log/`, and imports only from the standard scientific Python stack (pandas, numpy, statsmodels, pyfixest, the `differences` package, matplotlib). The entire analysis is reproducible on a clean machine in approximately 20-30 minutes.