

Incremental propensity score effects of buprenorphine retention on opioid-involved overdose events in pregnancy



Ashley I Naimi, PhD | Dept of Epidemiology
Professor | Emory University

✉ ashley.naimi@emory.edu
🌐 <https://ainaimi.github.io/>



EMORY

ROLLINS
SCHOOL OF
PUBLIC
HEALTH

My Research

The problems

Reproductive, perinatal, and maternal health.
Causal inference, Machine Learning, Study
Design, Statistical Inference, Reproducibility.

The questions

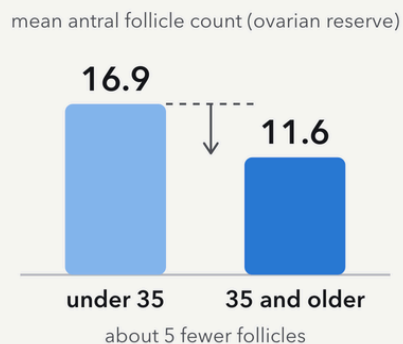
What would happen if something was different?
For whom? How do we estimate this credibly
from data (when possible)?

Methodological themes: **causal estimands** chosen to fit the data and the decision; **doubly robust estimation with machine learning**; **effect heterogeneity**; **reproducible, auditable workflows**.

Do Environmental Chemicals Speed Up Reproductive Aging?

Testing whether 16 endocrine-disrupting chemicals modify the age-related decline in ovarian reserve in 775 women.

1 The baseline decline



Ovarian reserve falls with age. Do chemical exposures change how steep that fall is?

2 The approach

16 candidate modifiers

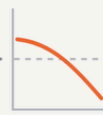
11 phthalates

3 parabens

bisphenol A

hair mercury

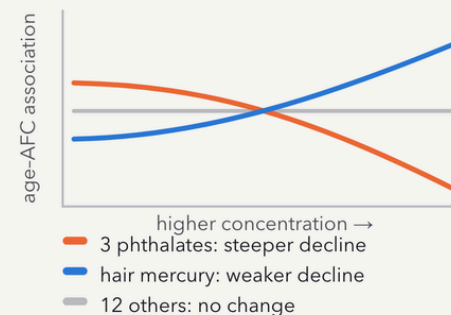
DR-learner
machine learning



age-AFC association by chemical concentration

A doubly robust learner flexibly estimates how the age-AFC association varies with each chemical's concentration.

3 What was found



MiBP, MCOP, and MEHP track a steeper, dose-dependent decline; hair mercury a weaker one, most likely a seafood proxy.

TAKEAWAY Select phthalates go with a steeper age-related loss of ovarian reserve: reproductive aging may be partly modifiable.

EARTH · The question

Does the age-related decline in ovarian reserve depend on a woman's exposure to **endocrine-disrupting chemicals**?

Background

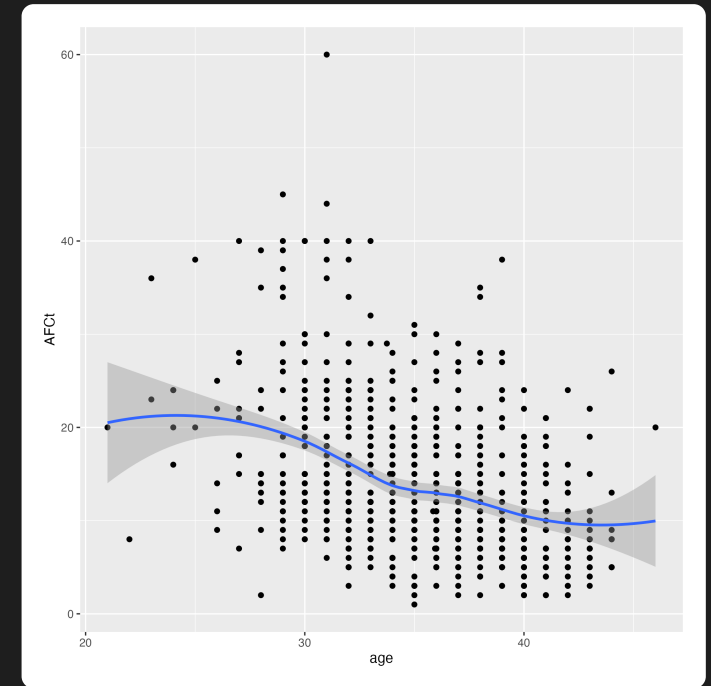
Antral follicle count falls with age: mean 16.9 under 35 versus 11.6 at 35 and older among 775 women at the MGH Fertility Center (EARTH Study, 2004–2019).

Question

Do any of 16 chemicals (11 urinary phthalate metabolites, 3 parabens, bisphenol A, hair mercury) change age associated declines in AFC?

Challenge

Sixteen candidate modifiers, skewed concentrations, possible interactions: standard regression / EMM is not built for this.



Antral follicle count by age, EARTH Study.

EARTH · Method and results

1 · Nuisance models

10-fold cross-validated Super Learner for $P(\text{age} \geq 35 \mid X)$ and $E(\text{AFC} \mid \text{age}, X)$: random forests, penalized regression, boosting, MARS, GLMs with screened interactions.

2 · Pseudo-outcome

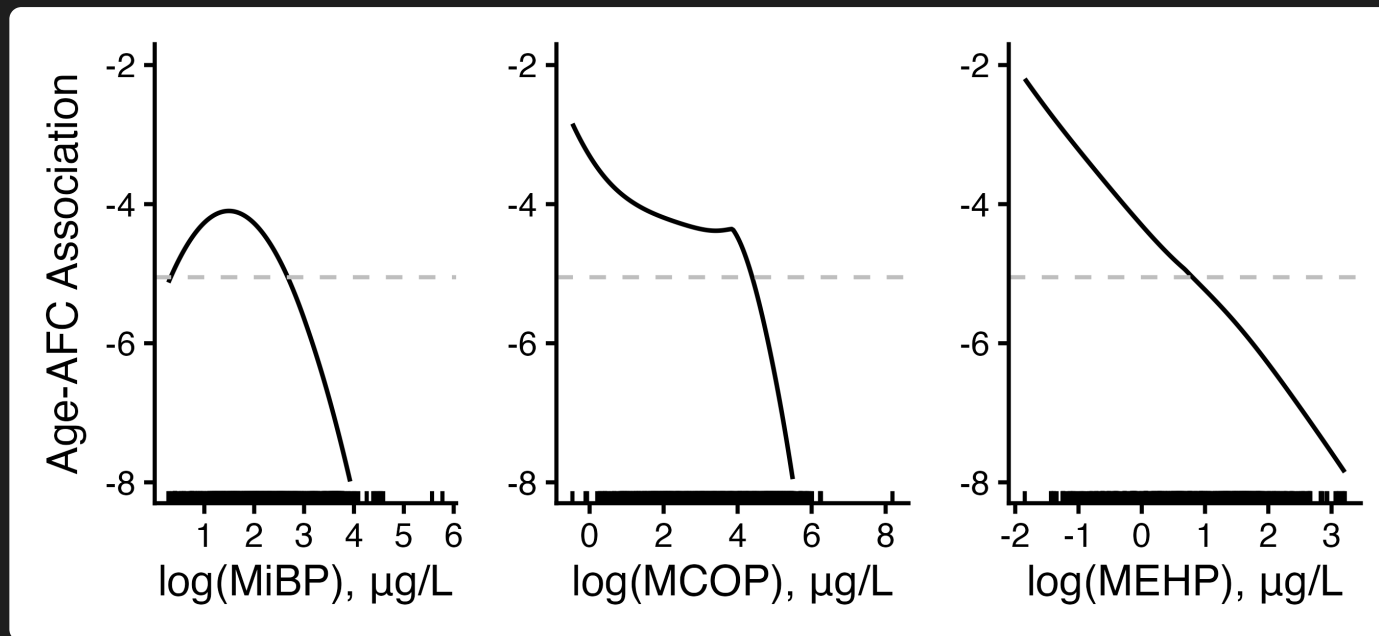
The AIPW score, the efficient influence function of the standardized mean difference. Its average: -5.1 follicles (95% CI $-6.2, -3.9$).

3 · Which chemicals modify?

Best linear projection of the score on the 16 log-concentrations; Wald statistics, conditional and unconditional.

4 · How?

Cross-validated LOESS of the score on the leading modifiers.

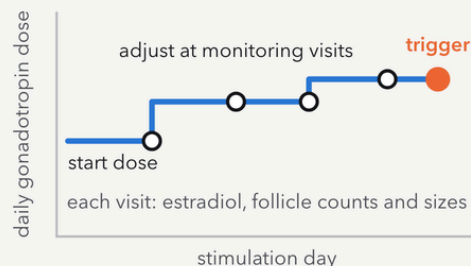


Age–AFC difference across MiBP, MCOP, and MEHP; dashed line: the overall -5.1 . The decline steepens beyond about $4.5 \mu\text{g/L}$ of MiBP and $50 \mu\text{g/L}$ of MCOP, and steadily with MEHP. Other chemicals: no modification.

Can IVF Stimulation Be Optimized One Decision at a Time?

Emulating sequential nested trials of daily gonadotropin dosing and trigger timing in about 4,000 patients at the Emory Reproductive Center.

1 The decisions

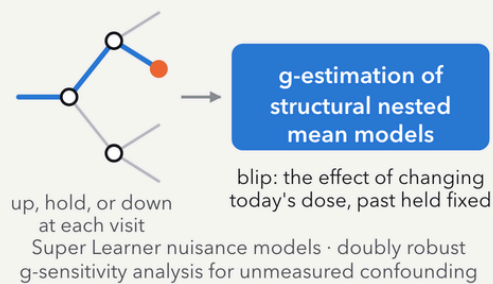


34% of cycles change dose mid-cycle
90% of visits end with no change

Doses are set from today's markers, which earlier doses shaped: time-varying confounding that regression cannot untangle.

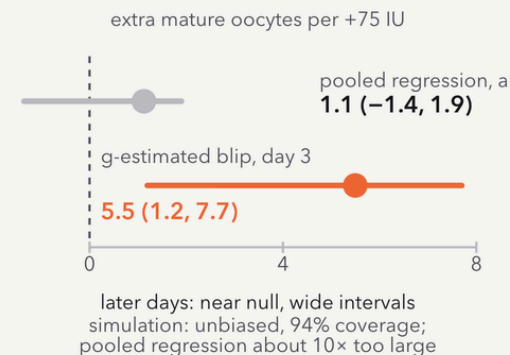
2 The approach

a trial nested at every decision day



Day-specific blip effects for dose, trigger timing, and deferring a monitoring visit, each conditional on the patient's history.

3 What the pilot shows



The start-dose decision matters; later titration shows little effect so far. Either answer changes practice: optimize doses, or monitor less.

TAKEAWAY

Estimate day-specific causal effects of dose decisions, then turn them into safety-constrained optimal regimens in a physician-in-the-loop tool, validated prospectively in shadow mode.

COESTAR-IVF · The question

During ovarian stimulation, does changing **today's** gonadotropin dose change the outcome, for which patients, and on which day?

The setting

About 435,000 ART cycles a year in the US; roughly half of cycles in women under 35 end in a live birth; a cycle costs \$15,000 to \$20,000.

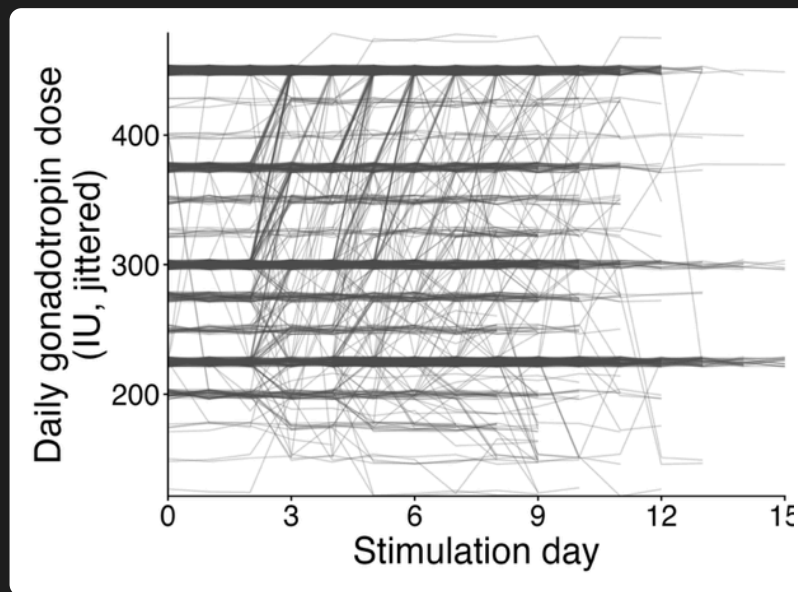
The decisions

A starting dose, adjustments at monitoring visits from estradiol and follicle growth, and the day of trigger, guided by heuristics that vary by physician. In the pilot, 34% of cycles change dose mid-cycle and 94% of changes fall on a visit day.

Why it is hard

Each day's dose responds to markers that earlier doses shaped: time-varying confounding by indication. The sequentially randomized trial that would settle it has never been run in IVF.

COESTAR-IVF: Causal Optimization with Emulated Sequential Trials for Assisted Reproduction. R01 research plan, 2026, with A. Gaskins and the Emory Reproductive Center.



Daily dose trajectories, pilot cohort: 1,485 autologous cycles, 1,142 women, 2024–2025.

COESTAR-IVF · Method and pilot results

Design

Sequential nested target trial emulation: a trial nested at every decision day, from EHR and SART-CORS data on about 4,000 patients (2014–2025).

Estimand

SNMM blip effects: the change in outcome if today's dose were raised 75 IU (or the trigger given today, or a visit deferred), holding history fixed.

Estimation

G-estimation by residual-on-residual regression; Super Learner nuisance models with cross-fitting; sandwich variance clustered on women; g-sensitivity analysis.

1.1

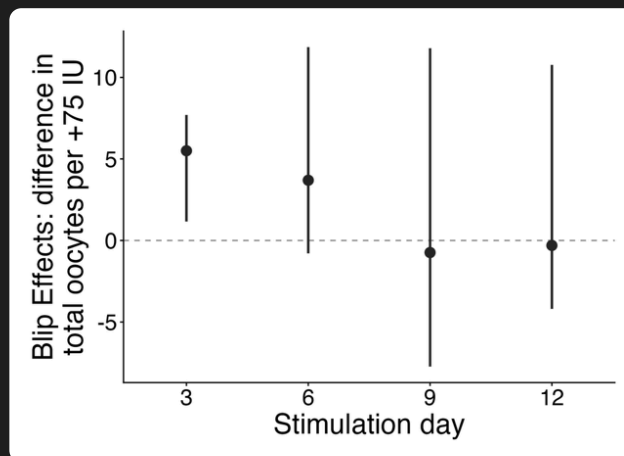
more mature oocytes per +75 IU from pooled regression on daily dose (95% CI -1.4 to 1.9)

5.5

g-estimated blip for +75 IU on day 3 (95% CI 1.2 to 7.7); days 6, 9, and 12 near null and imprecise

94%

interval coverage in a simulation calibrated to the cohort, bias -0.002 ; pooled regression about 10 times too large



Single-day blip effects of +75 IU on total oocytes retrieved, by stimulation day (pilot cohort, 1,485 cycles). The start-dose decision matters; later titration shows little effect so far.

Next: Aim 2, safety-constrained optimal regimens in a physician-in-the-loop tool; Aim 3, blinded shadow validation in 800 cycles.

Pilot: 2024–2025 EHR extract, 1,485 autologous cycles, 1,142 women. Full study: primary outcome cumulative live birth; safety outcome OHSS.

Acknowledgements

Manuscript

Rogers, Cooper, Kramer, Sullivan, Naimi. *Incremental propensity score effect of buprenorphine retention on opioid-involved overdose and acute hospital encounters during pregnancy and postpartum*. Under review.

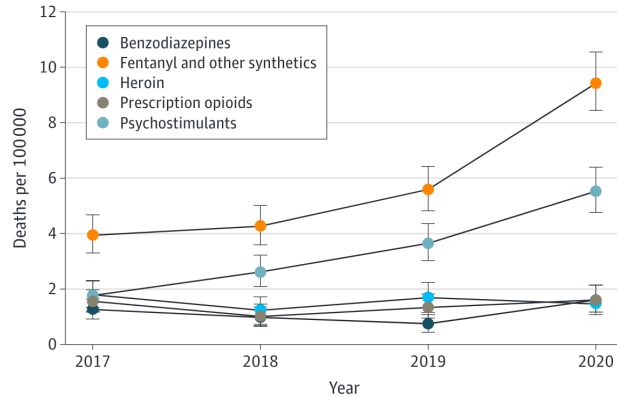
Support

NIH / NIDA training grant 5T32DA050552. Opinions are the authors' own.

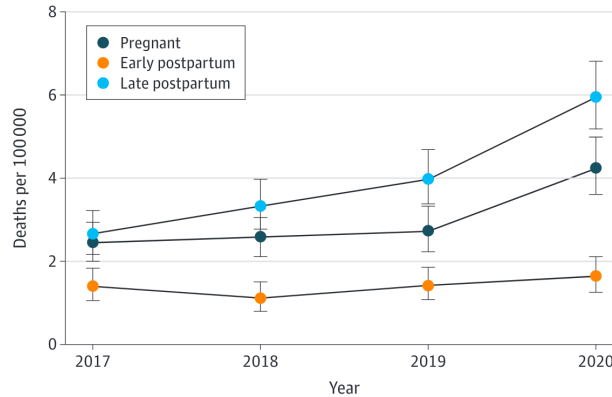


Overdose has become a leading cause of maternal death

A Drug types involved



B Pregnancy timing from 2017 to 2020



Drug types driving overdose deaths among pregnant and postpartum persons, United States, 2017–2020.

Opioid use disorder at delivery

More than quadrupled between 1999 and 2014: from 1.5 to 6.5 per 1,000 delivery hospitalizations.

Synthetic opioids

A near-exponential rise in fentanyl-involved deaths during pregnancy and postpartum, mirroring the broader epidemic.

The research gap

Most work on opioids in pregnancy studies birth outcomes. Far less asks what happens to the pregnant person.

Buprenorphine works, if people stay on it

Standard of care

Medication for opioid use disorder (buprenorphine or methadone) is first-line for pregnant people with OUD: better prenatal care attendance and birth outcomes.

How it works

Maintains low, stable opioid levels that suppress craving and withdrawal, reducing return to high-potency opioids and overdose.

Guidance

ACOG: universal screening and linkage to agonist therapy, rather than medically supervised withdrawal (detox), which carries high relapse rates.

The treatment gap

An estimated ~60% of pregnant people with OUD receive no treatment. Among those who do, many are underdosed.

The retention gap

Fear of provider stigma or custody loss, and providers tapering patients off medication, cut treatment short.

What we know about retention

Poor retention tracks with worse withdrawal symptoms and more emergency department visits. Its effect on overdose and acute opioid-related care across pregnancy and postpartum had not been studied.

The question

Among pregnant people who start buprenorphine, how does staying on treatment affect the risk of:

Opioid-involved overdose

Fatal or nonfatal; the outcome we most want to prevent.

Medically supervised withdrawal (detox)

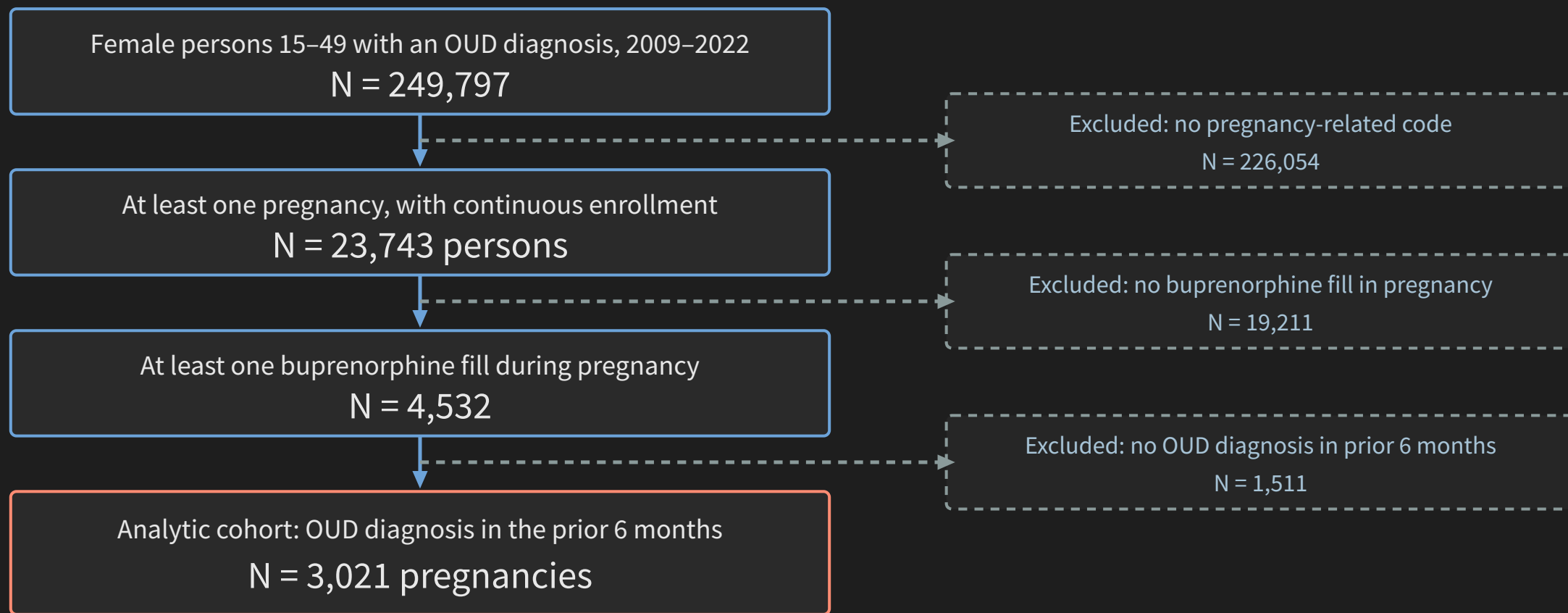
An acute encounter that signals discontinuation. Detox without ongoing medication lowers tolerance, and relapse afterward is common and dangerous.

Opioid-related emergency department visits

Acute care with an OUD diagnosis on record; an opportunity to re-engage, and a cost.

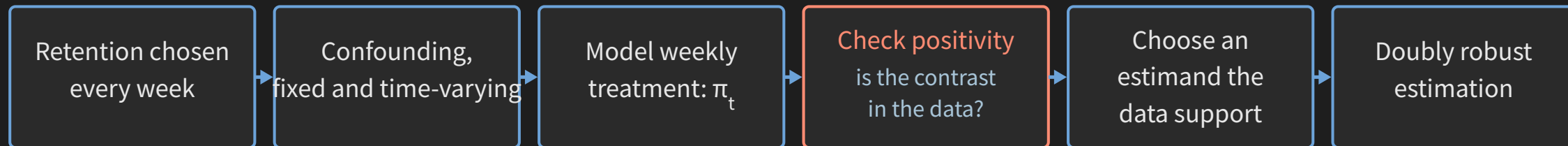
...during pregnancy and the first year after treatment starts, in a large commercially insured population.

Who is in the study



Additional eligibility: no methadone in the 6 months before the index fill (to avoid attributing effects to switching); enrollment from 6 months before pregnancy through the pregnancy outcome. Intermediate counts are person-level, as presented at CSEB 2025; the analytic sample is the manuscript's 3,021 pregnancies (Figure 1 reports pregnancy-level counts).

The logic of the analysis

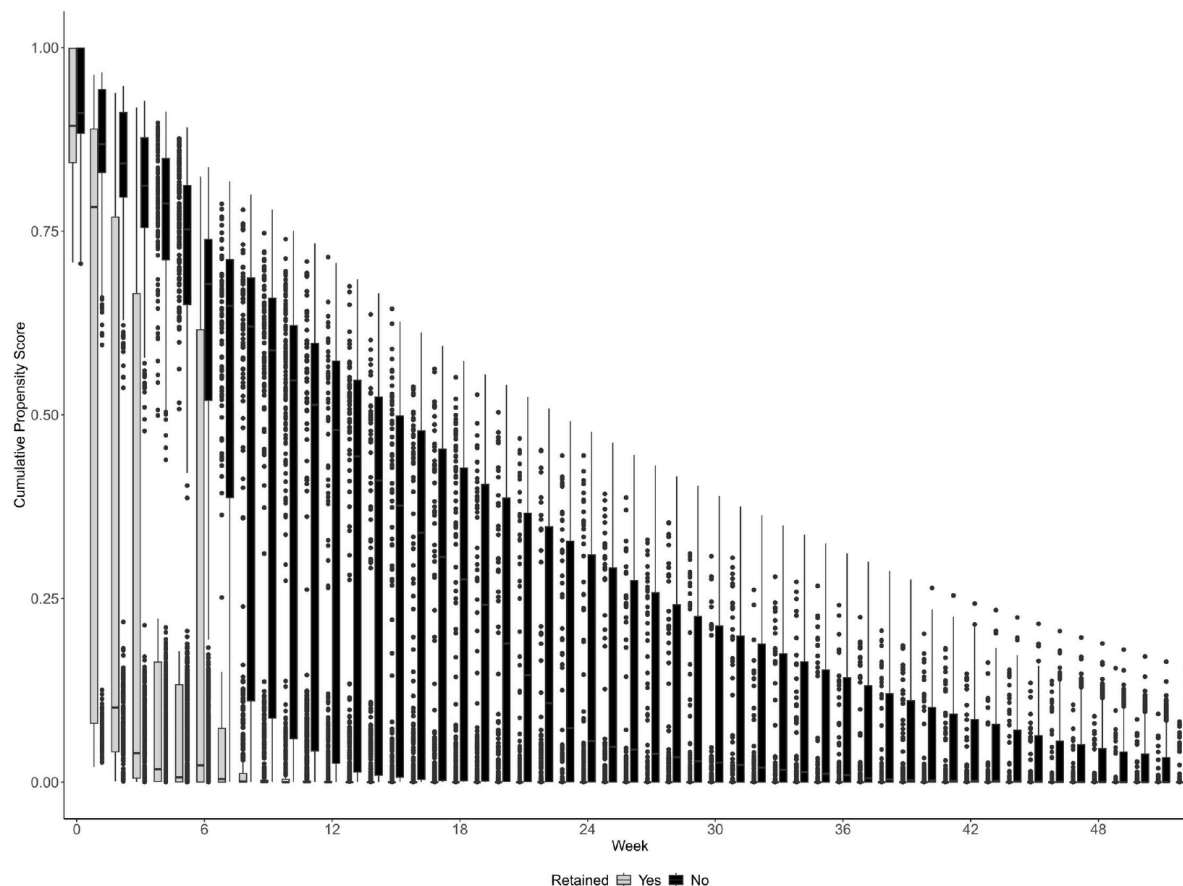


The comparison we set out to make: the longitudinal average treatment effect.

$$\psi = E(Y^{\bar{a}_{52}=1, D_{52}=1}) - E(Y^{\bar{a}_{52}=0, D_{52}=1})$$

Risk of the outcome by 52 weeks if everyone were retained every week (and stayed enrolled), minus the risk if no one were retained in any week. Identified under consistency, sequential exchangeability given measured history, no interference, and positivity.

Positivity, checked: the contrast is not in the data



Week 1

Cumulative propensity to have been retained so far: about 0.91 on average.

Week 12

About 0.07 (interquartile range below 0.001 to 0.55).

Week 36

Below 0.001.

"Retained for all 52 weeks" is essentially unobservable for most people. The ATE is not identified here, and it is not a realistic intervention either.

Cumulative propensity score by week, from the manuscript (Figure 4). Positivity violations of this kind grow with the number of time points.

The incremental propensity score intervention

Instead of setting treatment, shift each person's weekly odds of retention by a factor δ :

$$q_t(\delta) = \frac{\delta \pi_t}{\delta \pi_t + 1 - \pi_t}, \quad \delta = \frac{q_t / (1 - q_t)}{\pi_t / (1 - \pi_t)}$$

Example

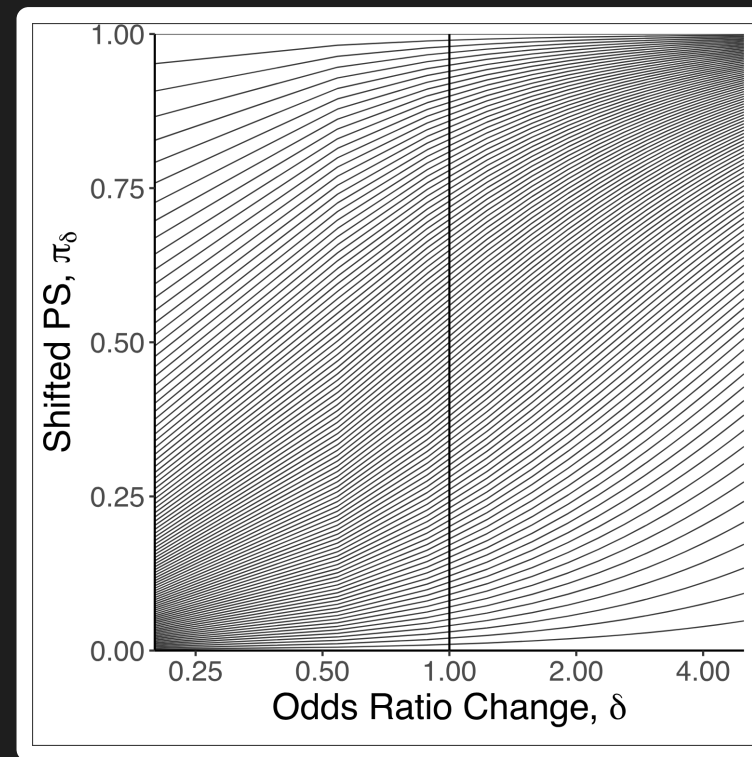
A person with $\pi = 0.35$ this week moves to 0.45 under $\delta = 1.5$, and one with $\pi = 0.5$ moves to 0.75 under $\delta = 3$.

$\delta = 1$

The observed world: no intervention on retention.

$\delta > 1 / \delta < 1$

Retention made more likely / less likely, for everyone, in proportion to where they start.



Each curve is one observed propensity score; follow it left or right to read $q(\delta)$. Vertical line: $\delta = 1$. Naimi et al., *Epidemiology* 2021.

The estimand: risk under a shifted retention process

$$\psi_{52}(\delta) = E\left[Y^{\bar{Q}_{52}(\delta), D_{52}=1}\right]$$

The risk of the outcome by 52 weeks if, each week, retention were drawn with the shifted probability $q_t(\delta)$ instead of the observed π_t , and no one dropped out.

Compare across δ , with $\delta = 1$ as the reference; δ ranged from 0.3 to 3.0.

Reading $\delta = 3$ against $\delta = 1$: what if we tripled everyone's weekly odds of staying on treatment?

What it buys

No positivity assumption; a contrast that corresponds to "make retention more likely," which is what programs do.

What it costs

The comparison is between distributions of treatment, not between "everyone treated" and "no one treated." Less familiar, and less useful for individual recommendations.

What it keeps

Consistency, sequential exchangeability given measured history, and no interference are still required.

Incremental effects, formally

Single time point (Kennedy 2019): the incremental effect is a weighted average of the two conditional risks, with weights set by the shifted propensity.

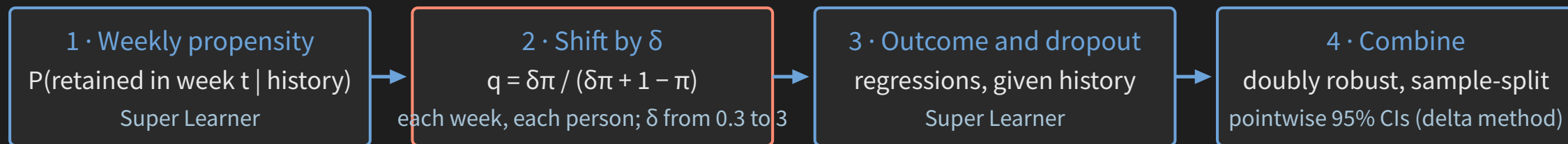
$$\varphi(\delta) = \mathbb{E} \left[\frac{\delta \pi(X) \mu(X, 1) + \{1 - \pi(X)\} \mu(X, 0)}{\delta \pi(X) + \{1 - \pi(X)\}} \right] = \mathbb{E}[\mu(X, A^*)], \quad A^* \mid X = x \sim \text{Bernoulli}\{q(x)\}$$

Longitudinal version with 52 weekly decisions (as shown at CSEB):

$$\varphi(\delta) = \sum_{\bar{a}_J \in \mathcal{A}^J} \int_{\mathcal{X}} \mu(x_J, a_J) \prod_{j=1}^J \frac{a_j \delta \pi_j(x_j) + (1 - a_j) \{1 - \pi_j(x_j)\}}{\delta \pi_j(x_j) + 1 - \pi_j(x_j)} d\mathbb{P}(x_j \mid x_{j-1}, a_{j-1})$$

Each week's treatment is drawn from the shifted propensity $q_j(\delta)$; the covariate process evolves as observed given the past. With dropout, the estimand is defined under $D_t = 1$ (no loss of enrollment), which adds dropout models to the nuisance set. Because no factor in the product requires π_j to be bounded away from 0 or 1, positivity is not needed.

Estimation



Why doubly robust

The estimator is built from the efficient influence function, so it is consistent if either the treatment/dropout models or the outcome models are right, and it reaches parametric rates when both converge at slow (machine-learning) rates.

Why Super Learner

A cross-validated ensemble of GLMs, random forests, and k-nearest neighbors for every nuisance model. No one has to believe a logistic regression with 52 weeks of history is correctly specified.

Why sample splitting

Fitting the nuisance models and evaluating the estimator on different folds keeps the machine learning from contaminating the confidence intervals.

What the causal interpretation still requires

Consistency

A week classified as "retained" corresponds to a well-defined treatment. Fills are not consumption; this is the assumption most at risk.

Exchangeability

Given measured history (baseline and weekly covariates), retention is as good as random each week. Claims lack stigma, provider practice, social support, and race/ethnicity.

No interference

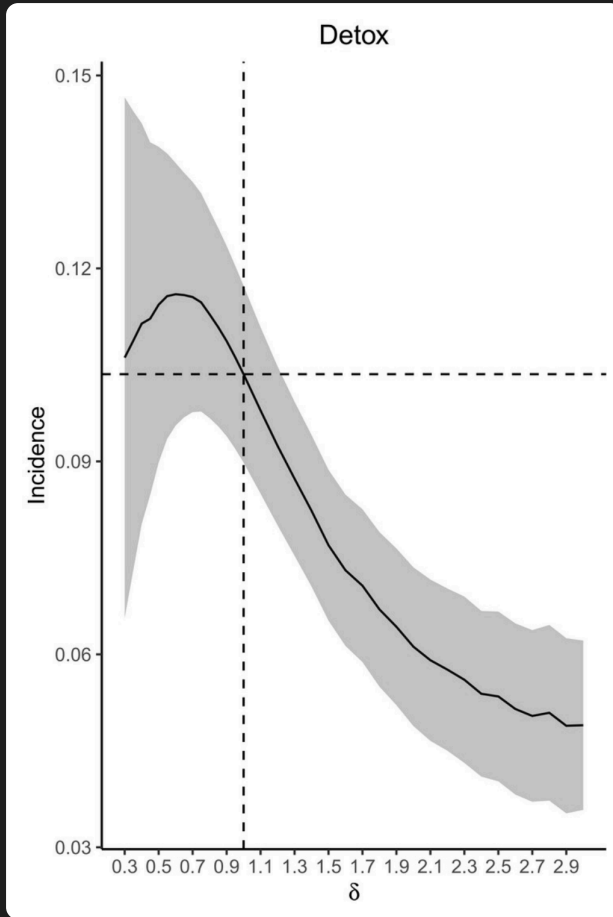
One person's retention does not affect another's outcome.

~~Positivity~~

Every person could be observed under every treatment history. Not needed for incremental effects.

Changing the estimand removed the assumption we could see failing. It did nothing about the two we cannot check.

Shifting retention upward lowers the risk of detox



10.4%

risk of detox by 52 weeks under the observed retention process ($\delta = 1$), had no one dropped out (95% CI 8.9 to 11.7)

4.9%

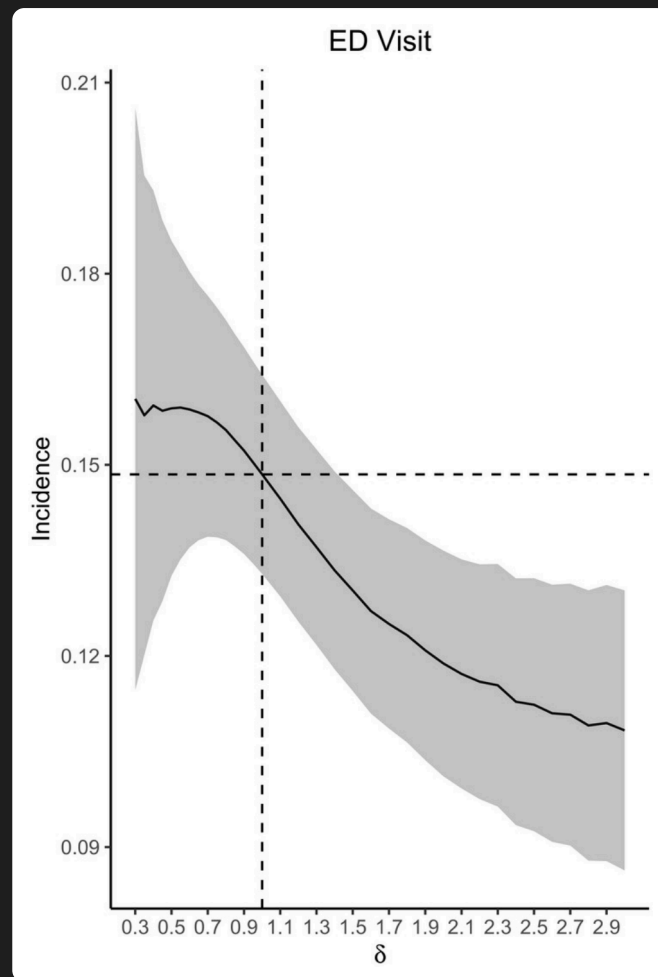
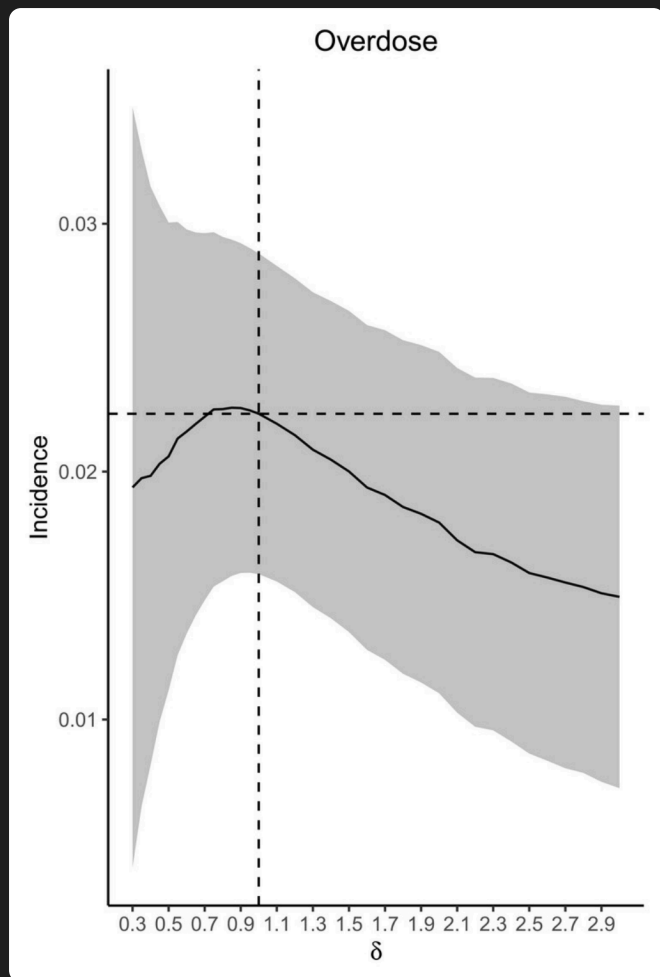
risk if everyone's weekly odds of retention were tripled ($\delta = 3$) (95% CI 3.6 to 6.2)

-5.5 per 100

risk difference, $\delta = 3$ versus $\delta = 1$ (95% CI -6.9 to -3.9): 55 fewer detox encounters per 1,000 pregnancies. At $\delta = 2$: -4.2 (-5.5 to -3.0).

Manuscript Figure 5 and Table 2 (retention: at least 5 of 7 days covered in the week; $N = 3,021$). Grey band: pointwise 95% confidence interval. Dashed lines: the observed world ($\delta = 1$). Making retention less likely ($\delta < 1$) produced no meaningful change.

Overdose and ED visits: the same direction



2.2% → 1.5%

overdose, $\delta = 1$ to $\delta = 3$; risk difference -0.7 per 100 (95% CI -1.5 to 0.03)

14.5% → 10.8%

ED visits with an OUD diagnosis, $\delta = 1$ to $\delta = 3$; risk difference -4.0 per 100 (95% CI -6.1 to -2.0)

Both fall as retention becomes more likely. Overdose is rare, so its interval is wide and just includes zero.

Summary of results

Outcome by 52 weeks	Observed	$\delta = 1$	$\delta = 2$	$\delta = 3$	RD, $\delta = 3$ vs 1 (95% CI)
Medically supervised withdrawal (detox)	8.8%	10.4%	7.7%	4.9%	-5.5 (-6.9, -3.9)
Opioid-related emergency department encounter	11.9%	14.5%	11.9%	10.8%	-4.0 (-6.1, -2.0)
Opioid-involved overdose	2.2%	2.2%	1.9%	1.5%	-0.7 (-1.5, 0.03)

Making retention on buprenorphine more likely, week by week, reduces acute opioid-related care during pregnancy and the postpartum year. The largest and most precise effect is on detox; overdose moves in the same direction with wide uncertainty.

Risks per 100 pregnancies under the shifted retention process with dropout removed (manuscript Table 2; N = 3,021). "Observed" is the crude incidence among all participants. Below $\delta = 1$ no outcome changed meaningfully. Threshold sensitivity analyses (PDC 0.5 to 0.9, CSEB 2025) show the same trajectories.

Limitations

Consistency: fills are not doses

Proportion of days covered is a billing proxy for medication taken. Dose changes, which matter in pregnancy, are not captured at all.

Generalizability

A commercially insured cohort. About a third of pregnant people with OUD; Medicaid populations face different barriers, especially postpartum.

Exchangeability: context is missing

No race or ethnicity, stigma, provider practices, incarceration, or social support. The estimator is doubly robust to model misspecification, not to unmeasured confounding.

Design limits

Left truncation (entry at the first fill in pregnancy, not at treatment start); first event only; no mortality linkage, so fatal overdoses are captured only if they generated a claim.

Incremental propensity score effects of buprenorphine retention on opioid-involved overdose events in pregnancy



Ashley I Naimi, PhD | Dept of Epidemiology
Professor | Emory University

✉ ashley.naimi@emory.edu
🌐 <https://ainaimi.github.io/>



EMORY

ROLLINS
SCHOOL OF
PUBLIC
HEALTH