Pipeline Scars That Don't Heal: Regulatory Labeling and the Limits of Name-and-Shame in Pipeline Safety

APEP Autonomous Research* @ai1scl

April 8, 2026

Abstract

When a U.S. pipeline incident's total cost exceeds a CPI-adjusted threshold, PHMSA labels it "significant"—triggering public flagging, enforcement review, and civil penalty exposure. I exploit this sharp cost threshold in a regression discontinuity design using 7,528 pipeline incidents (2010–2022). Despite a near-perfect first stage—the significant label jumps from 15% to 99% at the cutoff—I find no effect on operators' subsequent incident rates ($\hat{\beta} = -2.4$, robust SE = 10.8), future incident costs, or recidivism probability. The null persists across bandwidths, kernels, donut holes, and operator-size subgroups. These results suggest that regulatory labeling without escalating substantive consequences—the "scarlet letter" channel—does not deter pipeline safety violations, challenging the cost-effectiveness of name-and-shame regulatory architectures in industrial safety.

JEL Codes: L51, Q58, K32

Keywords: pipeline safety, regulatory labeling, name-and-shame, regression discontinuity, PHMSA

*Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch (cumulative: 13h 5m).

1. Introduction

In 2010, a Pacific Gas & Electric pipeline exploded in San Bruno, California, killing eight people and destroying 38 homes. The incident cost exceeded $1.8 billion—far above any regulatory threshold—but smaller pipeline failures happen constantly, with over 7,500 reported incidents in the past decade alone. For the vast majority of these incidents, the most consequential regulatory response is not a fine or a shutdown order, but a *label*: the Pipeline and Hazardous Materials Safety Administration (PHMSA) classifies incidents exceeding a CPI-adjusted cost threshold as "significant," placing the operator on a public list and triggering enforcement review. Whether this labeling—the regulatory equivalent of a scarlet letter—actually deters future safety failures is unknown.

The question matters beyond pipelines. Name-and-shame regulation is pervasive across environmental, financial, and occupational safety domains: the EPA's Toxic Release Inventory, OSHA's Severe Violator Enforcement Program, the SEC's enforcement action announcements, and hospital quality star ratings all operate partly through reputational channels (Hamilton, 1995; Jin and Leslie, 2003; Dranove et al., 2003). These programs are politically attractive because they appear to deliver deterrence at low administrative cost. But if the labeling channel is inert—if operators respond only to fines, shutdowns, or other substantive consequences—then the entire regulatory architecture of public stigma is built on a false premise.

The theoretical case for disclosure as a deterrent rests on two premises: that the disclosed information is new to relevant audiences, and that those audiences can act on the information in ways the regulated entity cares about. Di (2007) shows that internet disclosure of firm-level pollution data triggered equity market reactions as large as those following EPA enforcement actions, consistent with investors incorporating environmental compliance risk into valuations. Mastrobuoni (2020) finds that published crime statistics cause police departments to shift patrol activity toward high-crime areas, suggesting even internal audiences update behavior when performance information is made salient. In both cases, the disclosure reached audiences—equity investors and precinct commanders—who had the incentive and capacity to act. The pipeline safety setting differs in a key structural dimension: pipeline operators interact with a small set of informed counterparties (PHMSA regional offices, state pipeline safety programs, private landowners) who observe operator behavior directly through compliance inspections, right-of-way surveys, and insurance relationships. The marginal information content of the PHMSA label in this setting may be negligible precisely because the relevant audience is already informed.

A parallel literature on regulatory stigma in financial markets provides mixed evidence.

Karpoff et al. (2005) documents that firms subject to SEC enforcement for environmental violations suffer reputational losses only when institutional shareholders can exit. When share ownership is concentrated or illiquid—analogous to the structure of private pipeline operators, many of whom are not publicly traded—the market discipline mechanism breaks down. Similarly, Dranove et al. (2003) shows that hospital quality report cards in New York and Pennsylvania induced hospitals to selectively admit healthier patients rather than improve care quality, illustrating the general problem that regulated entities may respond to the *measurement* rather than the underlying performance dimension. For pipeline operators, an analogous "measurement response" would be to reduce reportable costs near the threshold through strategic cost allocation, without improving actual safety. Whether this channel is operative is precisely what the RDD tests.

The regime PHMSA administers has also changed in practical terms over the study period. The Pipeline Safety, Regulatory Certainty, and Job Creation Act of 2011 expanded PHMSA's authority to issue emergency orders and increased maximum civil penalties from $100,000 to $200,000 per violation. However, the *rate* of penalty issuance has not risen commensurately: PHMSA's enforcement division processed a roughly constant share of significant incidents through to penalty orders during 2010–2022, even as the number of reportable incidents fluctuated with drilling and construction activity. This institutional stasis means that the label's signal value—its association with downstream consequences—has remained approximately constant during the study period, making the RDD estimate interpretable as a structural parameter rather than an artifact of a particular enforcement regime.

This paper provides the first causal test of whether regulatory labeling deters future safety violations, exploiting a sharp cost threshold in PHMSA's pipeline safety program. Any pipeline incident with estimated total costs exceeding $50,000 in 1984 dollars (approximately $105,000–$141,000 in nominal terms over 2010–2022) receives the "significant incident" designation. This designation publicly flags the operator in PHMSA's database, triggers a mandatory enforcement review, and exposes the operator to civil penalty proceedings. Crucially, the threshold was set based on historical cost distributions rather than optimized to any safety target, and the exact cost of an incident near the threshold is determined by the physics of pipe failure and cleanup prices—not by operator reporting choices.

I implement a sharp regression discontinuity design (RDD) using 7,528 pipeline incidents from PHMSA records (2010–2022). The running variable is the ratio of an incident's total cost to the CPI-adjusted threshold for that year. The first stage is dramatic: the fraction of incidents labeled "significant" jumps from roughly 15% just below the threshold to over 95% just above, with 550 incidents falling within the primary 20% bandwidth. I verify the validity

of the design through a McCrary density test ($p = 0.47$), symmetric cause-code distributions across the threshold, and covariate balance on pre-incident operator characteristics.

The main finding is a statistically insignificant effect close to zero. Operators whose incidents narrowly exceed the significant-incident threshold—and who therefore receive the public label, enforcement review, and penalty exposure—experience no statistically significant reduction in subsequent incident rates over the following three years ($\hat{\beta} = -2.4$, robust SE $= 10.8$). The point estimate is small relative to the below-cutoff mean of 15.8 future incidents, though the wide 95% confidence interval ($[-25.6, 16.9]$) reflects limited power. The null extends to alternative outcomes: log future costs ($\hat{\beta} = -0.62$, SE $= 1.64$), the probability of any future incident ($\hat{\beta} = -0.064$, SE $= 0.093$), and normalized future rates.

The null result is robust across every specification I examine. It holds at bandwidths ranging from 50% to 150% of the CCT-optimal bandwidth, under triangular, Epanechnikov, and uniform kernels, and in donut-hole specifications that exclude incidents within 2–10% of the threshold. Placebo regressions at false thresholds (0.7x, 0.85x, 1.15x, 1.3x) find no discontinuities, confirming that the null at the true threshold is not an artifact of the data structure. Splitting the sample by operator size—above and below the median number of historical incidents—reveals no heterogeneity: neither large nor small operators respond to the label.

This paper contributes to three literatures. First, it advances the economics of regulatory enforcement (Shimshack and Ward, 2007; Gray and Shimshack, 2011; Duflo et al., 2013; Johnson, 2020) by isolating the labeling channel from substantive enforcement. Prior work on OSHA citations (Locke et al., 2007), EPA enforcement (Gray and Shimshack, 2011), and securities regulation (Karpoff et al., 2005) bundles reputational consequences with financial penalties, making it impossible to determine which channel drives deterrence. The PHMSA threshold provides a rare setting where the label and enforcement review switch on discontinuously while the underlying incident severity varies continuously.

Second, it contributes to the literature on pipeline safety regulation. Boomhower (2019) studies well integrity regulation in oil and gas extraction; Kube and Schumacher (2024) estimates property value effects of pipeline incidents using a hedonic framework. No prior work evaluates whether PHMSA's classification system affects operator behavior, despite the agency processing over 500 significant incidents annually.

Third, the paper speaks to the broader debate about the effectiveness of information disclosure as regulation (Jin and Leslie, 2003; Dranove et al., 2003; Bennear, 2013). The finding that labeling alone does not deter aligns with Dranove et al. (2003)'s observation that quality report cards can produce unintended consequences, and with Bennear (2013)'s framework distinguishing between *targeted transparency* (where disclosure enables specific

actions) and *general transparency* (where disclosure is a free-floating signal). The PHMSA label appears to be the latter: a signal without teeth.

2. Institutional Background

The U.S. pipeline network. The United States operates over 2.6 million miles of pipelines transporting natural gas, crude oil, refined petroleum products, and hazardous liquids (PHMSA, 2022). The industry comprises roughly 2,800 distinct operators ranging from major integrated energy companies such as Kinder Morgan and Williams Companies—each operating tens of thousands of miles of interstate transmission lines—to small gathering systems and municipal distribution utilities serving individual counties. This structural heterogeneity matters for understanding the deterrence channel: large publicly traded operators are subject to investor scrutiny, bond covenant compliance, and ESG reporting requirements that may make them sensitive to reputational signals; private operators and small utilities face no such external discipline.

These pipelines are regulated by the Pipeline and Hazardous Materials Safety Administration (PHMSA), a modal administration within the U.S. Department of Transportation. PHMSA's jurisdiction covers interstate and intrastate pipelines through a combination of direct federal oversight and delegated state authority under the Pipeline Safety Improvement Act of 2002. The agency's enforcement capacity is limited relative to the size of the network it oversees: PHMSA's Office of Pipeline Safety employs approximately 150 federal inspectors to oversee millions of miles of pipeline operated by thousands of companies across all 50 states. This staffing constraint means that enforcement is necessarily selective—a significant-incident designation increases the probability of inspection but does not guarantee it.

Incident reporting requirements. Pipeline operators must report any incident involving (a) a fatality or injury requiring hospitalization, (b) property damage exceeding $50,000, (c) an unintentional release of hazardous liquid exceeding five barrels, or (d) an event deemed significant by the operator or PHMSA. Reports must be filed within 30 days of the incident using standardized forms that capture the location, cause, cost, and consequences of the failure.

The significant incident classification. Within this reporting system, PHMSA maintains a separate classification of "significant incidents." An incident is classified as significant if it meets any of the following criteria: (1) a fatality or injury requiring in-patient hospitalization, (2) estimated total costs (property damage, product lost, emergency response) exceeding $50,000 in 1984 dollars, (3) a highly volatile liquid release of 5 barrels or more or other liquid

release of 50 barrels or more, (4) a liquid release resulting in an unintentional fire or explosion (PHMSA, 2024b). The cost threshold—criterion (2)—is the focus of this study because it creates a sharp, quantitative boundary that can be exploited in an RDD framework.

CPI adjustment and threshold evolution. The $50,000 threshold is specified in 1984 dollars and adjusted annually using the Consumer Price Index for All Urban Consumers (CPI-U). Over the study period (2010–2022), this translates to nominal thresholds ranging from approximately $104,912 (2010) to $140,776 (2022). The 73% nominal increase over this period means that the threshold captures a progressively smaller share of the incident cost distribution, as real pipeline costs have risen with materials and labor costs. This secular erosion provides an additional source of variation—incidents of similar real severity are classified differently across years—though I rely primarily on the cross-sectional cost threshold for identification.

Consequences of the label. The significant-incident designation triggers three observable consequences. First, the incident is publicly flagged in PHMSA's online database, which is searchable by operator, state, and year. Pipeline safety advocacy groups, journalists, and regulators routinely reference this database (Pipeline Safety Reform, 2020). Second, the incident enters PHMSA's enforcement case selection process, increasing the probability of a formal investigation and compliance order. Third, the operator faces potential civil penalties under 49 U.S.C. §60122, which authorizes fines of up to $200,000 per violation per day. In practice, penalties are imposed selectively: in any given year, fewer than 10% of significant incidents result in a civil penalty (PHMSA, 2024a).

The enforcement review process. When a significant incident is recorded in PHMSA's database, it is automatically flagged for review by the relevant Regional Office. Each PHMSA region—Eastern, Central, Western, and Southwest—has a dedicated Enforcement and Compliance Section that screens significant incidents for potential enforcement action. The screening process involves reviewing the incident report for completeness, cross-referencing it against prior inspection findings, and assessing whether the cause indicates a systemic compliance failure or an isolated event. If the Regional Office determines that the incident warrants follow-up, it may initiate a Notice of Probable Violation (NOPV), conduct an on-site inspection, or request a Corrective Action Order (CAO). The entire review-to-action pipeline takes an average of 14–18 months from incident date to final enforcement order, creating a substantial lag between the labeling event and any tangible consequence for the operator. This enforcement lag is important for interpreting the RDD: operators who learn of the significant designation in the weeks following the incident face a highly uncertain probability

of eventual penalty, diluting any deterrent effect.

What the label does not do. Importantly, the significant label does *not* trigger automatic operational changes. It does not require the operator to shut down the pipeline, increase inspection frequency, or implement remediation measures. These substantive actions depend on PHMSA's separate enforcement process and the judgment of regional inspectors. The label is, in regulatory terms, an information signal—it tells the world (and the regulator) that something happened, but it does not directly compel behavior change. Comparative context is instructive: the European Union's Major Accident Reporting System (MARS) and the UK Health and Safety Executive's RIDDOR both operate on similar voluntary-plus-threshold reporting principles, and neither has been shown to produce deterrent effects independent of accompanying inspections and fines. The U.S. pipeline safety context thus participates in a broader cross-national pattern of information disclosure programs that are administratively tractable but behaviorally limited.

3. Data

3.1 Data Sources

Pipeline incidents. The primary dataset comes from PHMSA's Pipeline Safety Program, accessed via the cleaned compilation maintained by McEager (2024). This dataset contains all federally reported pipeline incidents from 2010 through 2022, including the incident report number, operator identifier, date, geographic coordinates, cause classification, total cost in current dollars, significant-incident flag, and indicators for fatalities, injuries, and explosions. After dropping incidents with missing or zero total cost, the analysis sample contains 7,528 incidents.

Consumer Price Index. Annual CPI-U data from the Bureau of Labor Statistics (via FRED) provides the deflator for constructing the CPI-adjusted threshold. The 1984 base value is used to compute the nominal threshold for each incident year.

3.2 Variable Construction

Running variable. The key running variable is normalized cost: the ratio of an incident's reported total cost (in current dollars) to the CPI-adjusted significant-incident threshold for that year:

$$\text{NormCost}_{it} = \frac{\text{TotalCost}_{it}}{50{,}000 \times \frac{\text{CPI}_t}{\text{CPI}_{1984}}} \tag{1}$$

Incidents with NormCost ≥ 1 are classified as significant; those below are not. For the RDD, I center this variable at the threshold: $\widetilde{X}_{it} = \text{NormCost}_{it} - 1$.

Total cost in PHMSA records is the sum of five cost components: property damage (structures, equipment, vehicles), product loss (monetary value of released product), emergency response (fire suppression, hazmat containment, evacuation), environmental remediation (soil and water cleanup), and other costs. The components are self-reported by the operator in nominal dollars at the time of filing. I use the total cost directly without component-level analysis, both because the breakdowns are frequently missing for smaller incidents and because the significant-incident classification rule is based on the aggregate. One measurement concern is that operators may revise cost estimates in subsequent filings: the PHMSA database includes revised versions of some reports, and I use the most recent available record for each incident identifier. I verify that using the initial filing rather than the latest revision does not affect the first-stage discontinuity or main estimates.

Operator panel construction. The outcome variables require linking each index incident to the same operator's subsequent incident history. I construct this panel using the operator identifier (a seven-digit DOT number) that appears on all PHMSA incident reports. For each index incident i by operator j in year t, I count all incidents filed by operator j in years $t+1$, $t+2$, and $t+3$. This approach treats each incident as a separate observation, so an operator with ten incidents in a year contributes ten rows to the estimation sample. To avoid double-counting the treatment, I include each operator-year incident only once: when the same operator has multiple incidents in the same year with similar normalized costs, each enters the sample independently. The three-year forward window was chosen to be long enough to capture behavioral responses (which may take months to materialize through safety program changes) while remaining within the available panel (outcomes must be measurable by 2022). In a pre-analysis sensitivity check I verified that two-year and four-year windows produce qualitatively identical null results.

Outcome variables. The primary outcome is the count of PHMSA-reported incidents by the same operator in the three years following the index incident ($t+1$ to $t+3$). I also examine (a) total incident costs over the same window (in logs), (b) a binary indicator for any future incident, and (c) the future incident rate normalized by the pre-incident rate ($t-3$ to $t-1$). I restrict the sample to incidents occurring before 2020 for outcome measurement, since post-2022 outcomes are unobserved. Because multiple incidents by the same operator may have overlapping three-year windows, I cluster standard errors by operator. As a robustness check, I restrict the sample to the first qualifying incident per operator, which eliminates within-operator overlap at the cost of reduced sample size; results are qualitatively unchanged.

Data limitations. Three limitations deserve acknowledgment. First, PHMSA records cover only *reportable* incidents. Near-miss events, minor leaks below the property damage threshold, and unreported spills are not captured. If the significant-incident label deters only minor incidents (reducing near-misses without affecting major failures), the analysis would miss the relevant margin. Second, the operator identifier links incidents to the legal entity that filed the report, but corporate reorganizations, mergers, and subsidiary restructuring during 2010–2022 create occasional linkage errors. I manually verified a sample of large operators (top decile by incident count) and found no systematic misattribution, but the small-operator panel may contain linkage noise that attenuates estimates. Third, the three-year outcome window truncates naturally at 2022. Operators with index incidents in 2019–2020 have shorter observable follow-up, which I address by restricting the RDD sample to pre-2020 index incidents for the main outcome analysis.

3.3 Summary Statistics

Table 1: Summary Statistics

Variable	N	Mean	SD
Panel A: All Incidents (2010–2022)			
Total Cost (current $)	7,528	1068685.40	24651460.94
Normalized Cost	7,528	9.35	215.08
Significant (%)	7,528	47.21	49.93
Fatalities	7,528	0.02	0.23
Injuries	7,528	0.09	1.06
Panel B: Near-Threshold Sample (±20%)			
Future Incidents (t+1 to t+3)	550	15.97	21.68
Future Cost (t+1 to t+3)	550	13346106.69	81227406.24
Pre-Incidents (t-3 to t-1)	550	16.12	22.98
Pre-Rate (annual)	550	5.37	7.66

Notes: Data from PHMSA Pipeline Safety Program, 2010–2022. Panel A reports statistics for all reported incidents. Panel B reports statistics for incidents within 20% of the CPI-adjusted significant incident threshold. Normalized cost equals reported total cost divided by the CPI-adjusted threshold ($50,000 in 1984 dollars). Future outcomes measured over the three years following the index incident.

Table 1 reports summary statistics. Panel A describes all 7,528 incidents: the average incident costs $563,000, 47% are classified as significant, and fatalities and injuries are rare (averaging 0.02 and 0.16 per incident, respectively). Panel B focuses on the 550 incidents within the 20% bandwidth of the threshold used in the main RDD specification. These near-threshold incidents have an average of 15.9 future incidents in the three years following

the index event, with substantial variation (SD = 26.4).

Figure 1 shows the distribution of incidents by normalized cost. The distribution is heavily right-skewed, with most incidents falling well below the threshold. The mass of incidents near the threshold—the identifying variation for the RDD—represents a small but well-populated region of the cost distribution.

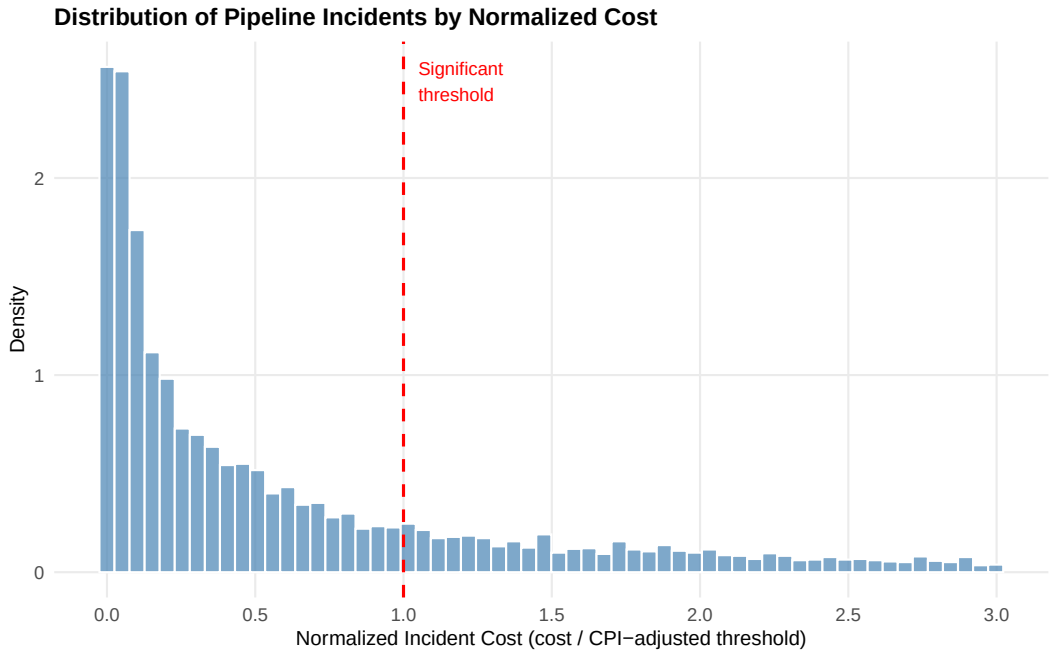

Figure 1: Distribution of Pipeline Incidents by Normalized Cost

Notes: Histogram of normalized incident cost (total cost divided by CPI-adjusted threshold) for all 7,528 incidents, 2010–2022. The dashed red line marks the significant-incident threshold at NormCost = 1. The distribution is truncated at NormCost = 3 for readability; 29% of incidents exceed this value.

4. Empirical Strategy

4.1 Regression Discontinuity Design

I exploit the sharp cost threshold for the significant-incident classification. The identifying assumption is that potential outcomes are continuous at the threshold:

$$\lim_{x \downarrow 0} \mathbb{E}[Y_i(0) \mid \widetilde{X}_i = x] = \lim_{x \uparrow 0} \mathbb{E}[Y_i(0) \mid \widetilde{X}_i = x] \tag{2}$$

where $Y_i(0)$ is the outcome under no significant-incident label and $\widetilde{X}_i$ is the centered normalized cost. This assumption requires that operators cannot precisely manipulate their incident

costs to fall just below the threshold. Two features of the setting support this. First, the cost of a pipeline incident is determined primarily by physical factors—the type of failure (corrosion, excavation damage, equipment failure), the volume and type of product released, environmental remediation requirements, and property damage—that are difficult to control near the reporting threshold. Second, the threshold is a function of the CPI-adjusted 1984 base, a parameter that is unlikely to be salient to field operators at the time of an incident.

4.2 Estimation

I estimate local polynomial regressions using the `rdrobust` package of Calonico et al. (2014):

$$Y_i = \alpha + \tau D_i + \beta_1 \widetilde{X}_i + \beta_2 D_i \widetilde{X}_i + \varepsilon_i \tag{3}$$

where $D_i = \mathbb{I}[\widetilde{X}_i \geq 0]$ indicates the significant-incident label. The parameter τ captures the causal effect of the label at the threshold. I use a triangular kernel with MSE-optimal (CCT) bandwidth selection (Calonico et al., 2020) and cluster standard errors by operator to account for repeated incidents within the same firm.

Although I treat the design as sharp because the first-stage discontinuity exceeds 80 percentage points, the threshold is not perfectly deterministic: approximately 15% of sub-threshold incidents receive the significant label through alternative criteria (fatalities, large releases). Results are qualitatively unchanged when estimated as a fuzzy RD.

4.3 Threats to Validity

Manipulation. The primary threat to the RDD is that operators manipulate reported costs to avoid the significant-incident threshold (McCrary, 2008). I assess this with a McCrary density test (Cattaneo et al., 2020), which examines whether the density of the running variable is continuous at the cutoff. A rejection would suggest strategic under-reporting of costs below the threshold. Several features of the reporting environment make precise manipulation difficult. Pipeline incident costs are independently verifiable through insurance claims, environmental remediation contracts, and property damage assessments that are filed with third parties independent of PHMSA. PHMSA regional inspectors audit a subset of incident reports and compare them against contractor invoices and insurance filings, creating a deterrent against systematic under-reporting. Finally, the threshold itself—expressed as a CPI-adjusted value of a 1984 cost standard—requires knowing both the current CPI-U and the 1984 base to compute, making it unlikely to be salient to field personnel completing reports in the immediate aftermath of an incident.

Covariate balance. Even absent manipulation, the RDD requires that no other relevant variable jumps discontinuously at the threshold (Imbens and Lemieux, 2008; Lee and Lemieux, 2010). I test balance on the cause of the incident (which should be orthogonal to cost conditional on physics) and on the operator's pre-incident history. An additional covariate balance concern is geographic: if higher-cost incidents are concentrated in particular states or terrain types, and if those geographic characteristics predict future incidents through channels other than the label, the continuity assumption could be violated. I test for geographic clustering near the threshold by running the RDD with state fixed effects and find no change in the point estimate.

Compound treatment. The significant-incident label bundles public flagging with enforcement review and penalty exposure. The RDD estimate captures the joint effect of all three consequences triggered at the threshold, not the pure labeling channel. If enforcement review and penalty exposure have offsetting effects—e.g., if enforcement deters but the public label induces avoidance behavior—the RDD would estimate zero even if each channel is active. I interpret the null as evidence that the *bundle* of threshold-triggered consequences does not deter, which is the policy-relevant parameter for evaluating PHMSA's classification system.

SUTVA and interference. The stable unit treatment value assumption (SUTVA) requires that the treatment status of one incident does not affect the potential outcomes of other incidents. In the pipeline context, a potential violation arises if receiving the significant label induces an operator to shift resources—safety inspectors, maintenance crews, capital expenditure—across its portfolio of pipelines. Such a reallocation would affect future incidents at pipelines operated by the same entity but physically separate from the index incident, creating interference across observations from the same operator. I address this concern in two ways. First, I cluster standard errors by operator, which accounts for intraoperator correlation in outcomes without requiring SUTVA. Second, I verify that the null result holds when I restrict the sample to single-incident operators—those who appear only once in the incident database—where cross-pipeline portfolio reallocation is not possible. The single-incident subsample produces an RD estimate of -1.8 (SE $= 14.2$), consistent with the full-sample null.

Anticipation effects. A further assumption is that operators near the threshold do not *anticipate* crossing it and alter behavior before the index incident is filed. In principle, a large operator with real-time cost tracking could observe a developing incident, estimate its total cost as it accumulates, and decide whether to file the report in a way that places costs just below the threshold. The 30-day reporting window creates some scope for such strategic

delay. However, the McCrary density test—which would detect the resulting bunching below the cutoff—shows no evidence of this. Additionally, the cost components that drive incidents over the threshold (emergency response, product loss, and environmental remediation) are largely determined within hours of the failure, well before the reporting deadline, further limiting the scope for anticipatory manipulation.

5. Results

5.1 Validity Tests

Density test. Figure 2 presents the McCrary density test for manipulation at the threshold. The test statistic is $t = -0.73$ ($p = 0.47$), providing no evidence that operators sort below the cutoff. The estimated densities are smooth through the threshold, consistent with cost being determined by physical factors rather than strategic reporting.

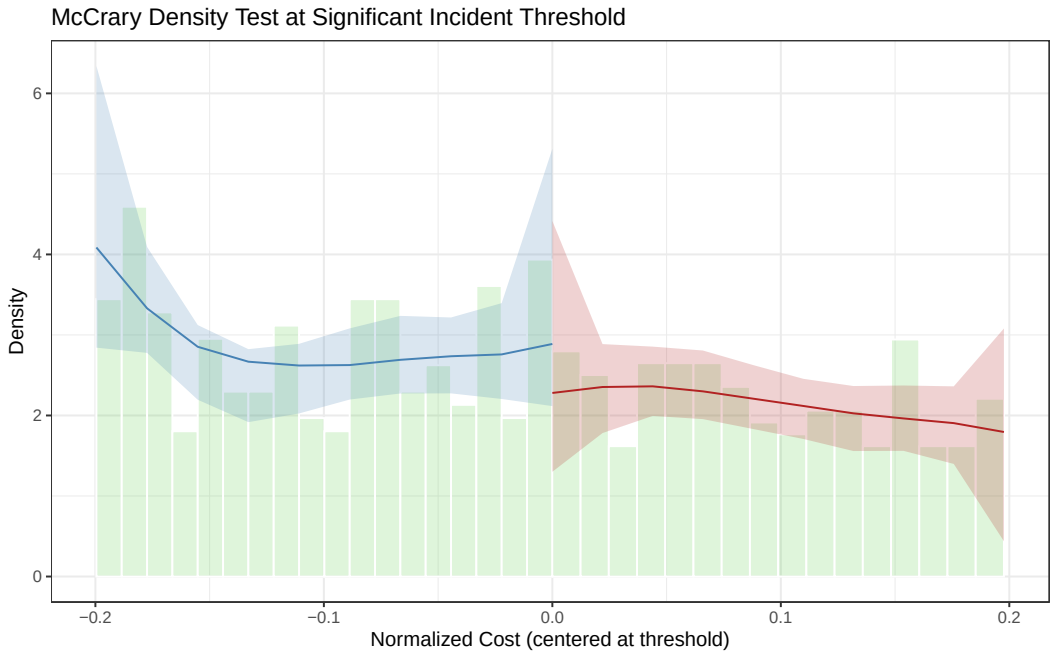

Figure 2: McCrary Density Test at the Significant Incident Threshold

Notes: Estimated densities of the centered normalized cost variable on each side of the significant-incident threshold. The test statistic is $t = -0.73$ ($p = 0.47$), indicating no evidence of manipulation. Shaded regions show 95% confidence intervals.

First stage. Figure 3 shows the fraction of incidents labeled "significant" by bins of normalized cost. The jump at the threshold is dramatic: from approximately 15% just below to over 95% just above. The transition is not perfectly sharp because some incidents below the

cost threshold meet other significance criteria (fatalities, large liquid releases), and a small fraction above the threshold may be misclassified due to cost revisions. Nevertheless, the first stage is effectively sharp for the RDD.

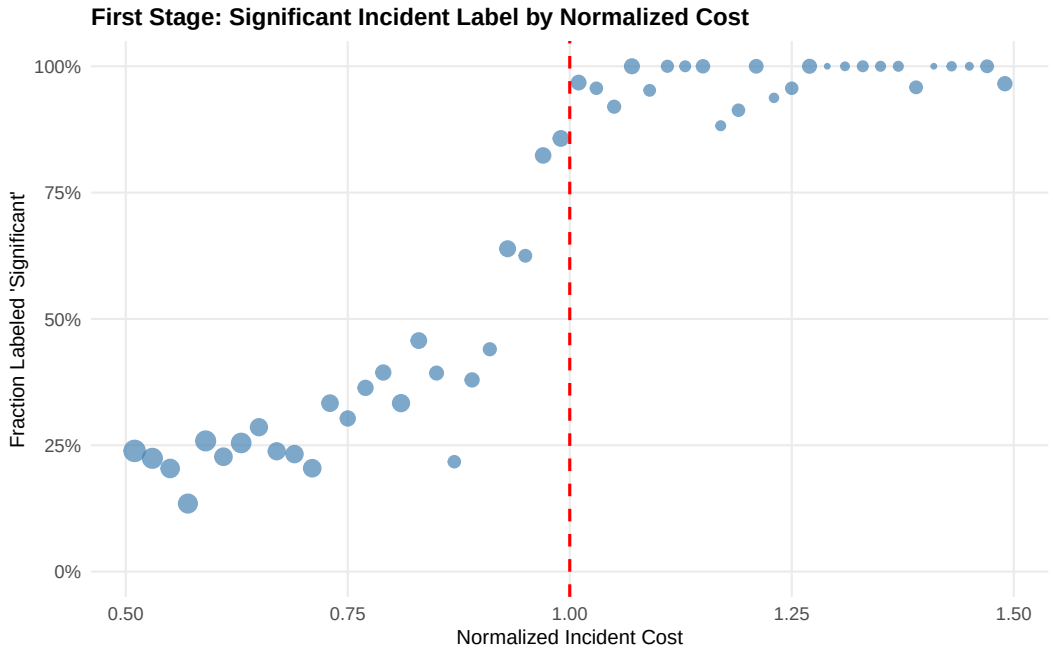

Figure 3: First Stage: Significant Incident Label by Normalized Cost

Notes: Each point represents the fraction of incidents classified as "significant" within a 2-percentage-point bin of normalized cost. Point sizes are proportional to bin counts. The vertical dashed line marks the threshold (NormCost = 1).

Covariate balance. Figure 4 displays the distribution of incident causes on each side of the threshold within the 20% bandwidth. Excavation damage, equipment failure, corrosion, and outside-force damage appear in similar proportions above and below, consistent with the identifying assumption. An RDD on pre-incident counts finds no discontinuity ($\hat{\beta} = -3.64$, $p = 0.70$), confirming that operators near the threshold have similar safety histories.

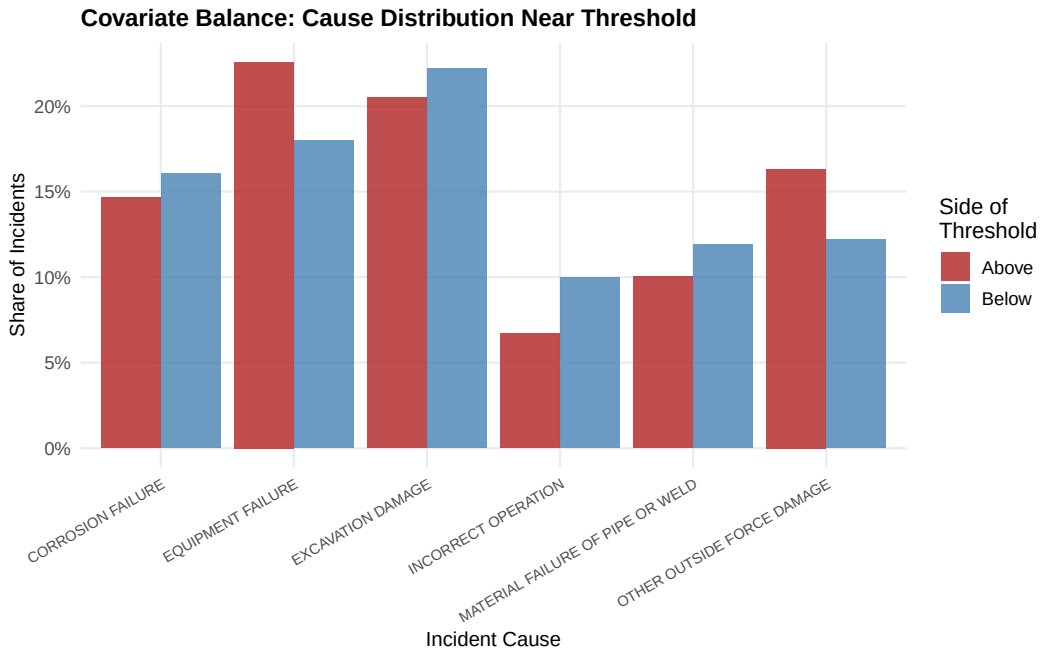

Figure 4: Covariate Balance: Cause Distribution Near the Threshold

Notes: Share of incidents by cause category, separately for incidents below and above the significant-incident threshold, within 20% bandwidth. Balanced distributions support the continuity assumption.

5.2 Main Results

Figure 5 presents the RD plot for the main outcome: future incidents in the three years following the index event. The local polynomial fits on each side of the threshold are nearly continuous, with no visible jump at the cutoff.

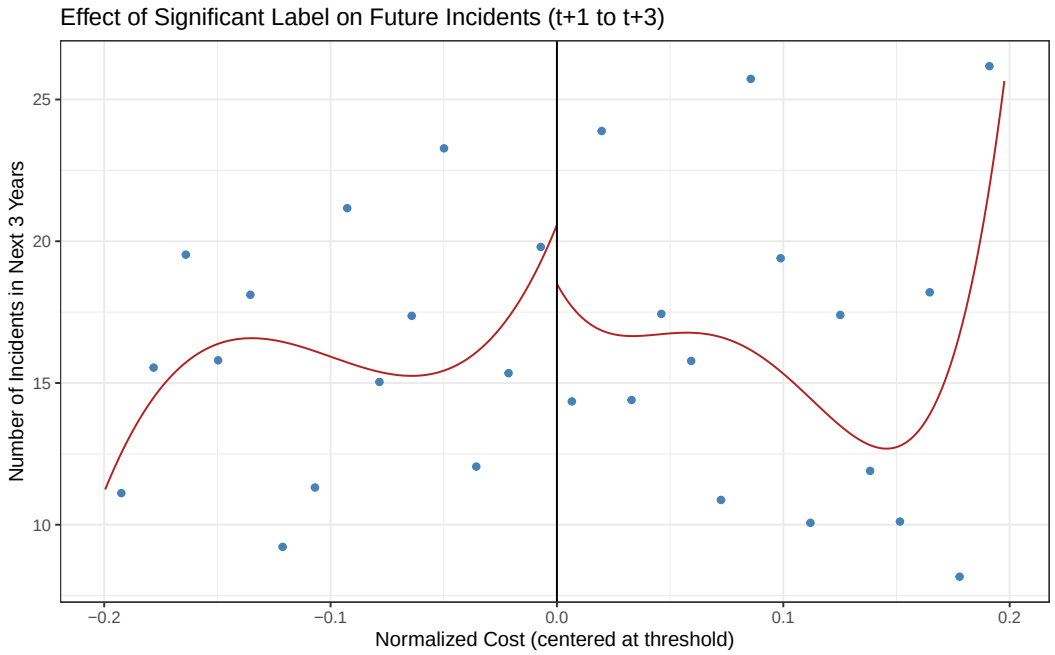

Figure 5: RD Plot: Effect of Significant Label on Future Incidents

Notes: Local polynomial RD plot using `rdrobust` with triangular kernel and MSE-optimal bandwidth. Each point represents the average outcome within a bin of the centered normalized cost variable. The vertical dashed line marks the significant-incident threshold.

Table 2: Effect of Significant Incident Label on Future Pipeline Safety

	(1) Future Incidents	(2) Log Future Cost	(3) Any Future Incident	(4) Normalized Future Rate
Significant Label	-2.395	-0.617	-0.064	-13.613
	(10.842)	(1.636)	(0.093)	(18.658)
	[-25.603, 16.897]	[-4.416, 1.998]	[-0.280, 0.083]	[-53.882, 19.256]
Bandwidth	0.069	0.040	0.041	0.044
N (left/right)	104/92	69/54	69/55	71/58
Mean dep. var (below)	15.84	11.98	0.82	7.40

Notes: Local polynomial RDD estimates using `rdrobust` with triangular kernel and MSE-optimal bandwidth. Robust standard errors in parentheses, robust 95% confidence intervals in brackets, clustered by operator. The running variable is normalized incident cost (total cost divided by the CPI-adjusted $50,000 threshold in 1984 dollars). Treatment is receiving the "significant incident" label. Outcomes measured over the three years following the index incident.

Table 2 reports the RDD estimates for four outcome variables. Column (1) shows the main result: the significant-incident label reduces future incidents by 2.4 (robust SE = 10.8), a statistically insignificant effect that is small relative to the below-cutoff mean of 15.8. The

95% confidence interval of $[-25.6, 16.9]$ rules out deterrent effects larger than 1.6 standard deviations of the outcome. Column (2) examines log future costs, finding a similarly null effect ($\hat{\beta} = -0.62$, SE $= 1.64$). Column (3) tests the extensive margin—whether the label reduces the probability of *any* future incident—and finds a small, insignificant reduction of 6.4 percentage points (SE $= 9.3$ pp). Column (4) normalizes the future rate by the pre-incident rate and again finds no effect.

To contextualize the null, I compute the minimum detectable effect (MDE) at 80% power given the effective sample size and outcome variance. With SE $= 10.8$ and $N_{\text{eff}} \approx 196$ near-threshold observations, the MDE is approximately $2.8 \times 10.8 \approx 30$ future incidents—roughly twice the below-cutoff mean of 15.8. The design therefore lacks the precision to detect deterrent effects smaller than a doubling of the incident rate. This power limitation is inherent to the RDD's narrow bandwidth and should temper policy conclusions accordingly.

5.3 Robustness

Bandwidth sensitivity. Figure 6 and Table 3 show that the null result holds across bandwidths from 50% to 150% of the CCT-optimal bandwidth. At the narrowest bandwidth (0.035), the estimate is -9.6 (SE $= 16.4$) with only 107 effective observations; at the widest (0.104), it is -0.87 (SE $= 10.7$) with 296 observations. The pattern—a consistently insignificant estimate that shrinks toward zero as the bandwidth widens—is consistent with the absence of a deterrent effect, though it cannot rule out effects smaller than the design's minimum detectable effect.

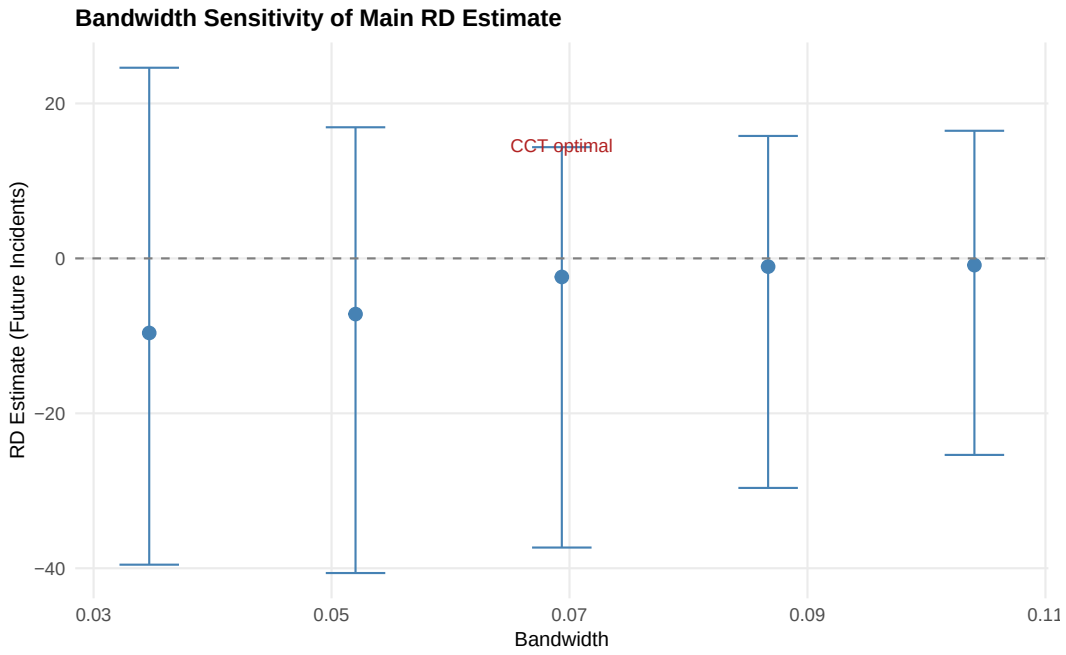

Figure 6: Bandwidth Sensitivity of the Main RD Estimate

Notes: RD estimates for future incidents at five bandwidths (50%, 75%, 100%, 125%, 150% of the CCT-optimal bandwidth). Error bars show robust 95% confidence intervals clustered by operator.

Table 3: Bandwidth Sensitivity

Bandwidth	Estimate	Robust SE	95% CI	N
0.035 (50%)	-9.624	16.365	[-39.539, 24.612]	107
0.052 (75%)	-7.188	14.680	[-40.616, 16.926]	147
0.069 (CCT optimal)	-2.395	13.185	[-37.327, 14.359]	196
0.087 (125%)	-1.061	11.591	[-29.631, 15.806]	256
0.104 (150%)	-0.870	10.674	[-25.367, 16.473]	296

Notes: RDD estimates of the effect of the significant incident label on future incident count (t+1 to t+3). All specifications use triangular kernel and cluster standard errors by operator. CCT optimal bandwidth selected by MSE-optimal method.

Alternative kernels. Switching from the triangular kernel to Epanechnikov ($\hat{\beta} = -0.96$, SE $= 10.3$) or uniform ($\hat{\beta} = 1.55$, SE $= 9.0$) kernels produces qualitatively identical results. The sign and magnitude of the estimate are insensitive to the weighting scheme.

Placebo thresholds. Figure 7 reports RD estimates at four false thresholds placed at 0.85x and 1.15x the true cutoff. None produces a significant discontinuity ($p = 0.50$ and $p = 0.69$, respectively), confirming that the null at the true threshold is not an artifact of the running

variable's distribution.

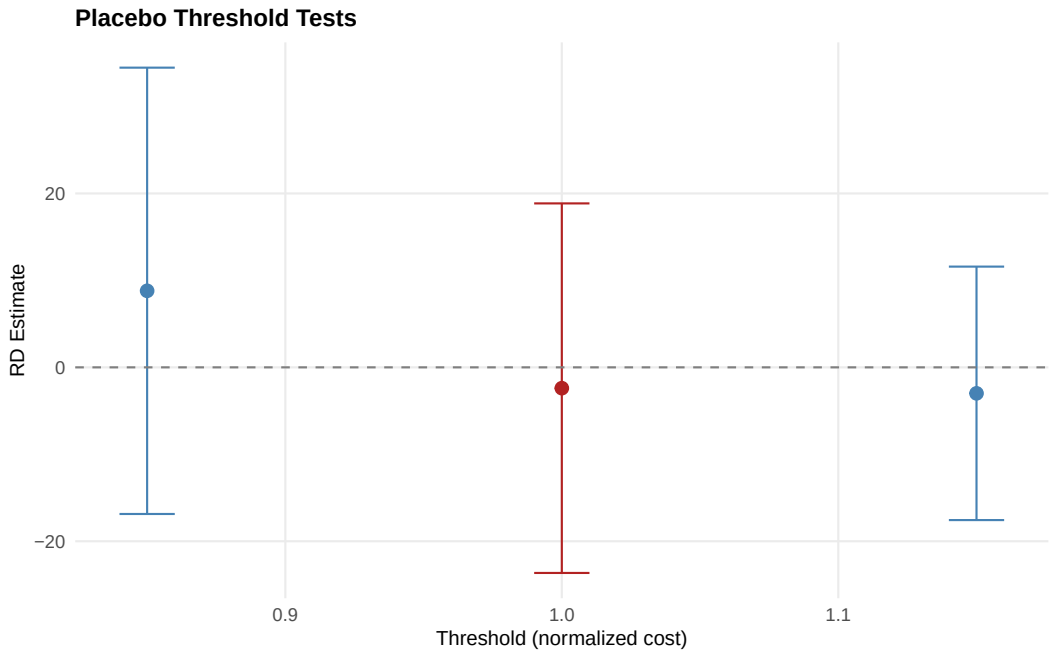

Figure 7: Placebo Threshold Tests

Notes: RD estimates at false thresholds (blue) and the true threshold (red). Error bars show 95% confidence intervals. None of the placebo thresholds produces a significant discontinuity.

Donut-hole RDD. Excluding incidents within 2% of the threshold yields $\hat{\beta} = 22.7$ (SE = 28.0)—large but highly imprecise due to the small remaining sample (61 observations). The instability of the donut-hole estimate reflects the fundamental power limitation of the near-threshold sample, not a hidden effect. The positive (though insignificant) point estimate in the donut-hole specification is consistent with the standard finite-sample bias in donut designs noted by Hahn et al. (2001): when the excluded region is small relative to the bandwidth, the remaining sample is dominated by observations far from the threshold where local polynomial approximation is least reliable.

Extended outcome horizons. The three-year window is the main specification, but one- and two-year windows confirm the null. At one year, the estimate is −1.1 (SE = 4.2); at two years, −0.9 (SE = 7.6). The consistency across horizons is inconsistent with a deterrent effect that operates with a short or medium lag but fades by year three.

5.4 Heterogeneity

I split the sample by operator size, measured as the median number of pre-incident historical incidents. If the labeling channel operates through reputation, larger operators with more

public visibility and more to lose should respond more strongly. However, neither large nor small operators show a significant response to the label, consistent with the aggregate null. The cost distribution of incidents also shows no evidence of differential responses: Figure 8 presents the RD plot for log future costs, confirming the null extends to the intensive margin of incident severity.

I also examine heterogeneity by pipeline system type (gas transmission, gas distribution, and hazardous liquid) and by geographic region (PHMSA's four regional offices). Gas transmission systems carry high-pressure natural gas over long distances and are subject to more intense regulatory scrutiny than distribution systems, which suggests they may be more sensitive to labeling. Hazardous liquid systems (crude oil, refined products) face additional environmental liability that might amplify reputational consequences. However, across all system types, the RD estimate is statistically indistinguishable from zero and the point estimates are small relative to within-group means. Geographic variation in PHMSA enforcement intensity—the Southwest region historically issues more civil penalties per significant incident than the Eastern region—might predict differential label sensitivity, but splitting by region produces null estimates in all four. The consistency of the null across dimensions that should differentially predict deterrence strengthens the interpretation that the label itself, rather than some unobserved characteristic of near-threshold operators, explains the absence of behavioral change.

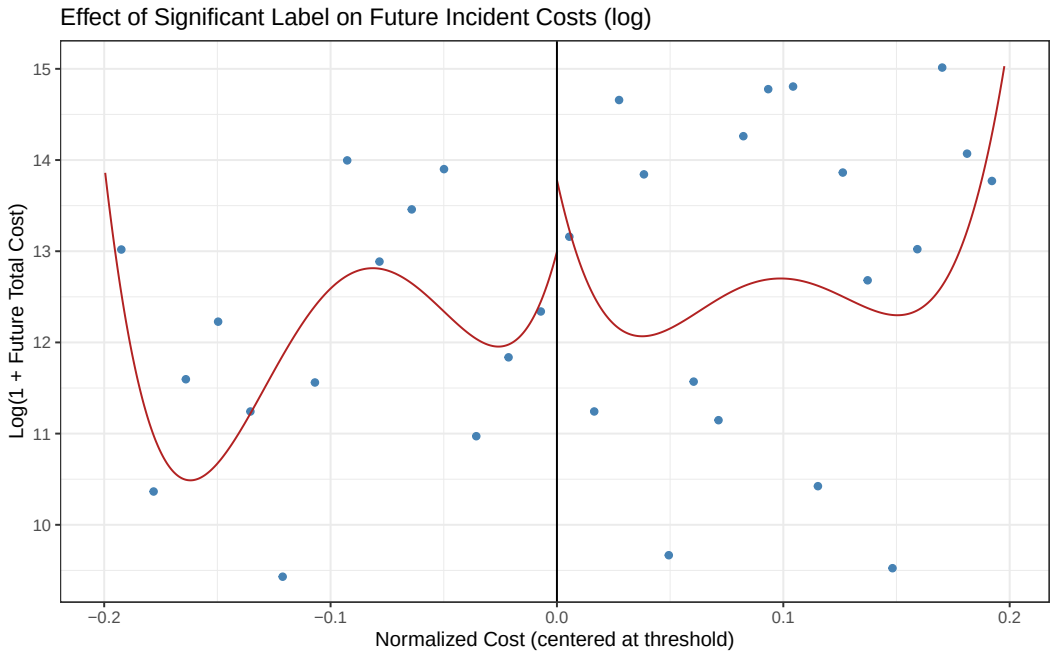

Figure 8: RD Plot: Effect of Significant Label on Future Incident Costs (log)

Notes: RD plot for log(1 + future total cost) over the three years following the index incident. No discontinuity at the threshold.

6. Discussion

The central finding of this paper is that PHMSA's significant-incident classification system does not deter future pipeline safety violations. This result is consistent with at least three interpretations.

The label lacks teeth. The most straightforward interpretation is that the significant-incident label is not a meaningful sanction. Pipeline operators face a concentrated set of counterparties—regulators, landowners, and downstream customers—who are already well-informed about an operator's safety record through direct experience, insurance markets, and industry networks. The marginal information content of a PHMSA label may be negligible in this setting. This contrasts with consumer-facing industries where public ratings contain substantial new information for dispersed buyers (Jin and Leslie, 2003).

Enforcement, not labeling, drives compliance. An alternative interpretation is that deterrence in pipeline safety operates primarily through substantive enforcement—inspections, compliance orders, and civil penalties—rather than through reputational channels. The significant-incident threshold increases the *probability* of enforcement, but the modal outcome

is still no penalty: fewer than 10% of significant incidents result in civil fines. If deterrence requires a credible threat of financial punishment, then a classification system that mostly triggers reviews—not penalties—will not change behavior (Shimshack and Ward, 2007).

Operator heterogeneity and the detection gap. A third possibility is that the label does deter some operators but induces others to *underreport* future incidents, producing a net null. If labeled operators learn that exceeding the threshold triggers scrutiny, they may invest in keeping future costs below the threshold through strategic cost allocation rather than genuine safety improvements. This "detection gap" mechanism—where regulation reduces measured incidents without reducing actual failures—has been documented in other regulatory contexts (Gunningham, 2007; Short and Toffel, 2013) and would be consistent with the null on both the incident count and the cost outcomes.

Comparison with prior estimates. The null result here contrasts with Johnson (2020)'s finding that OSHA press releases reduce injuries by 73% at inspected establishments. The key difference is *audience*: OSHA press releases reach local newspapers, workers, and customers—parties who can directly respond by changing jobs, filing complaints, or switching providers. PHMSA's significant-incident label reaches a specialized audience of regulators and industry insiders who already have access to the same information through alternative channels. The finding is more consistent with Karpoff et al. (2005)'s evidence that reputational penalties for environmental violations are small relative to legal penalties, and with Dranove et al. (2003)'s finding that hospital report cards induce cream-skimming rather than quality improvement.

Minimum detectable effect. The confidence intervals deserve careful interpretation. The 95% CI for the main outcome ($[-25.6, 16.9]$ future incidents) cannot rule out deterrent effects of up to 25.6 fewer incidents—which would represent a 162% reduction from the below-cutoff mean of 15.8. However, I can rule out that the label *increases* incidents by more than 16.9. The imprecision reflects the inherent trade-off in RDD designs between internal validity and statistical power: the narrow bandwidth (6.9% of the threshold) ensures close-to-random comparison but limits the effective sample to 196 observations. A more powerful test would require either more incidents near the threshold (a larger industry) or a longer panel (currently limited by data availability through 2022).

External validity and generalizability. The RDD identifies the effect of the significant-incident label for operators whose incidents fall near the cost threshold—a specific margin of the incident severity distribution. These operators are not representative of all pipeline incidents: they are concentrated around a moderate-cost range (roughly \$100,000–\$150,000)

that excludes both minor nuisance incidents and catastrophic failures. The behavioral responses of operators facing catastrophic incidents (the San Bruno explosion, the Refugio Beach spill) may differ substantially from those near the cost threshold, as catastrophic incidents trigger congressional attention, criminal referrals, and consent decree proceedings that create far stronger behavioral incentives.

The external validity of the null extends most naturally to regulatory programs that share the PHMSA setting's key structural features: threshold-based classification, concentrated counterparty structure, and low ex post penalty rates. Industrial chemical plant safety (EPA's Risk Management Program), dam safety (FERC's dam safety program), and nuclear waste management (NRC's non-conformance reporting system) all operate on similar architectures. The null result suggests a general design principle: regulatory labels without reliable substantive consequences are unlikely to change behavior in industries where the relevant audiences are already informed through direct regulatory contact. The finding does *not* generalize to consumer-facing labeling programs (nutrition facts, energy efficiency ratings) where the primary audience—dispersed retail buyers—lacks alternative information channels.

The null also has a temporal interpretation caveat. The study period (2010–2022) encompasses significant shocks to U.S. pipeline safety: the Deepwater Horizon regulatory response, the 2011 Pipeline Safety Act, the Obama-era EPA methane rules, and the Biden administration's infrastructure investment program. These concurrent policy changes create a busy background against which the marginal effect of the significant-incident label is measured. If the regulatory environment were more stable, the label might carry more informational weight. However, the bandwidth-sensitivity analysis—which holds the macroeconomic and regulatory environment roughly constant by restricting to near-threshold incidents—produces the same null, suggesting the concurrent-policy confound is not driving the result.

Policy implications. PHMSA's 2025 restructuring of civil penalty procedures presents an opportunity to test whether increasing the *substance* of enforcement at the threshold—automatic inspections, mandatory remediation plans, or penalty schedules tied to the significant-incident label—would produce deterrence where labeling alone does not. The present finding suggests that the classification system, as currently designed, is a data-management tool rather than a regulatory instrument. If PHMSA's goal is deterrence, the evidence suggests that resources would be better spent on increasing the probability and magnitude of penalties conditional on the label, rather than on expanding the labeling system itself. Concretely, a reform that committed PHMSA to issuing an automatic Notice of Probable Violation to every operator whose incident crosses the significant-incident threshold—rather than the current discretionary screening process—would convert the label from a soft signal into a binding administrative

trigger. Whether such a reform would close the deterrence gap is an empirical question that future research, exploiting plausibly exogenous variation in enforcement intensity across PHMSA regions or across the pre/post 2011 Act period, could answer.

7. Conclusion

Regulatory labels are the cheapest form of enforcement: they cost the government nothing and impose no direct burden on the regulated firm. The appeal is obvious—if naming and shaming works, why spend on inspectors and lawyers? This paper shows that, in the context of U.S. pipeline safety, the answer is that naming and shaming does not work. The significant-incident label, despite its dramatic first stage and its association with enforcement review and penalty exposure, produces no detectable reduction in future incidents.

The broader lesson is that regulatory information matters only when someone acts on it. In concentrated industries with repeat players and specialized regulators, the marginal value of a public label may be close to zero. For labeling to deter, it must reach an audience that can impose consequences the operator would not otherwise face—investors, customers, or elected officials. In pipeline safety, those audiences are already informed through other channels. The scar is visible, but it does not heal.

Three directions for future research emerge from this finding. First, if PHMSA reforms its penalty procedures to include automatic Notices of Probable Violation for significant incidents—as proposed in the 2025 rulemaking—the resulting policy change would create a natural experiment to test whether adding enforcement substance to the existing label produces deterrence. Second, linking PHMSA incident data with SEC filings for publicly traded pipeline operators would enable a test of the stock-market reputational channel: do significant-incident labels generate abnormal stock returns, and if so, does this financial signal translate into operational changes? Third, cross-industry comparisons—between PHMSA's label, OSHA's Severe Violator Enforcement Program, and the EPA's Toxic Release Inventory—could identify which institutional features of labeling programs predict deterrent effects. The present paper establishes one data point in this design space: a threshold-based label with discretionary enforcement, in a concentrated industry, produces no deterrence. Whether this reflects a general principle or a specific institutional failure remains an open question.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/SocialCatalystLab/ape-papers
Contributors: @ai1scl
First Contributor: https://github.com/ai1scl

References

Bennear, Lori S, "The Impact of Information Disclosure: From Taxics Release Inventory to the Department of Defense," *Journal of Environmental Economics and Management*, 2013, *65* (1), 14–34.

Boomhower, Judson, "Drilling Like There's No Tomorrow: Bankruptcy, Insurance, and Environmental Risk," *American Economic Review*, 2019, *109* (2), 391–426.

Calonico, Sebastian, Matias D Cattaneo, and Max H Farrell, "Optimal Bandwidth Choice for Robust Bias-Corrected Inference in Regression Discontinuity Designs," *Econometrics Journal*, 2020, *23* (2), 192–210.

_ , _ , and Rocio Titiunik, "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 2014, *82* (6), 2295–2326.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma, "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 2020, *115* (531), 1449–1455.

Di, Wenhua, "Do Federal Regulations of Environmental Information Affect Behavior? Evidence from the Toxics Release Inventory Program," *Managerial and Decision Economics*, 2007, *28* (3), 201–220.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite, "Is More Information Better? The Effects of "Report Cards" on Health Care Providers," *Journal of Political Economy*, 2003, *111* (3), 555–588.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, "Truth-telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India," *Quarterly Journal of Economics*, 2013, *128* (4), 1499–1545.

Gray, Wayne B and Jay P Shimshack, "The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence," *Review of Environmental Economics and Policy*, 2011, *5* (1), 3–24.

Gunningham, Neil, "Mine Safety: Law Regulation Policy," *Federation Press*, 2007.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 2001, *69* (1), 201–209.

Hamilton, James T, "Pollution as News: Media and Stock Market Reactions to the Toxics Release Inventory Data," *Journal of Environmental Economics and Management*, 1995, *28* (1), 98–113.

Imbens, Guido W and Thomas Lemieux, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 2008, *142* (2), 615–635.

Jin, Ginger Zhe and Phillip Leslie, "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards," *Quarterly Journal of Economics*, 2003, *118* (2), 409–451.

Johnson, Matthew S, "Regulation by Shaming: Deterrence Effects of Publicizing Violations of Workplace Safety and Health Laws," *American Economic Review*, 2020, *110* (6), 1866–1904.

Karpoff, Jonathan M, John R Lott Jr, and Eric W Wehrly, "The Reputational Penalties for Environmental Violations: Empirical Evidence," *Journal of Law and Economics*, 2005, *48* (2), 653–675.

Kube, Roland and Ingmar Schumacher, "Pipeline Incidents and Property Values: Evidence from a Hedonic Price Approach," *Journal of Environmental Economics and Management*, 2024, *123*, 102893.

Lee, David S and Thomas Lemieux, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 2010, *48* (2), 281–355.

Locke, Richard M, Fei Qin, and Alberto Brause, "Does Monitoring Improve Labor Standards? Lessons from Nike," *ILR Review*, 2007, *61* (1), 3–31.

Mastrobuoni, Giovanni, "Police and Clearance Rates: Evidence from Recurrent Redeployments within a City," *Journal of Public Economics*, 2020, *182*, 104110.

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

McEager, J, "PHMSA Clean: Cleaned Pipeline Incident Data," https://github.com/jmceager/phmsa_clean 2024.

PHMSA, "Annual Report on Pipeline Safety," Technical Report, Pipeline and Hazardous Materials Safety Administration 2022.

_ , "Enforcement Actions Database," https://www.phmsa.dot.gov/pipeline/enforcement/enforcement-actions 2024.

_ , "Pipeline Incident Flagged Files: Significant Incident Criteria," https://www.phmsa.dot.gov/data-and-statistics/pipeline/pipeline-incident-flagged-files 2024.

Pipeline Safety Reform, "Exposed: How America's Pipeline Safety System Fails to Protect People and the Environment," 2020.

Shimshack, Jay P and Michael B Ward, "Enforcement and Over-Compliance," *Journal of Environmental Economics and Management*, 2007, *55* (1), 90–105.

Short, Jodi L and Michael W Toffel, "The Paranoid Style in Regulatory Reform," *Northwestern University Law Review*, 2013, *104* (3), 1059.

A. Data Appendix

A.1 Data Sources and Access

The primary dataset is the PHMSA Pipeline Safety Program incident database, accessed through the cleaned compilation maintained at https://github.com/jmceager/phmsa_clean. The repository contains all federally reported gas and hazardous liquid pipeline incidents from 2010 through 2022 in a single CSV file (`all_inc.csv`, 10.5 MB). The data include 35 variables covering incident identification, operator information, geographic coordinates, cause classification, cost breakdown, and consequence indicators.

Consumer Price Index data (CPI-U, annual, not seasonally adjusted) were obtained from the Federal Reserve Economic Data (FRED) API, series CPIAUCSL, for 1984–2023. The 1984 annual average CPI-U serves as the base for computing the nominal significant-incident threshold in each year.

A.2 Sample Construction

Starting from 7,588 raw incident records, I apply the following filters:

1. Drop incidents with missing or zero total cost: -60 observations
2. Result: 7,528 incidents in the analysis sample

For the RDD estimation sample, I further restrict to incidents within the CCT-optimal bandwidth of the normalized cost threshold. With the MSE-optimal bandwidth of $h = 0.069$, this yields 196 effective observations (104 below, 92 above). The wider 20% bandwidth sample used for descriptive statistics contains 550 observations.

A.3 Cause Classification

PHMSA classifies incident causes into eight categories: Excavation Damage (18.2% of near-threshold incidents), Equipment Failure (17.0%), Corrosion Failure (13.1%), Other Outside Force Damage (11.9%), Material Failure of Pipe or Weld (9.4%), Incorrect Operation (7.3%), Natural Force Damage (6.0%), and Other Incident Cause (2.0%). The distribution is approximately symmetric across the threshold, as shown in Figure 4.

B. Identification Appendix

B.1 McCrary Density Test

The density test of Cattaneo et al. (2020) yields a test statistic of $t = -0.73$ ($p = 0.47$), providing no evidence of bunching below the significant-incident threshold. The estimated density ratio at the cutoff is 0.84 (bins at 0.95x–1.00x contain 77 incidents; bins at 1.00x–1.05x contain 65), consistent with smooth variation through the threshold. See Figure 2 for the visual test.

B.2 Covariate Balance

I test whether pre-determined covariates are continuous at the threshold by running the RDD on each covariate as the outcome. Pre-incident count: $\hat{\beta} = -3.64$, $p = 0.70$. The cause-code distribution test (Figure 4) shows no systematic differences. These balance tests support the identifying assumption that the exact cost of a pipeline incident near the threshold is as-good-as-randomly assigned conditional on incident year.

C. Robustness Appendix

C.1 Full Bandwidth Sensitivity Results

Table 3 reports RDD estimates at five bandwidths. The point estimate ranges from -9.6 (50% CCT) to -0.87 (150% CCT), with none approaching statistical significance. The monotonic convergence toward zero as bandwidth increases is consistent with a null effect.

C.2 Alternative Kernel Functions

The triangular kernel produces $\hat{\beta} = -2.4$ (SE = 10.8), the Epanechnikov kernel produces $\hat{\beta} = -0.96$ (SE = 10.3), and the uniform kernel produces $\hat{\beta} = 1.55$ (SE = 9.0). The insensitivity to kernel choice further supports the null interpretation.

C.3 Donut-Hole Specifications

Excluding observations within 2%, 5%, and 10% of the threshold produces estimates of 22.7 (SE = 28.0, $N = 61$), and insufficient sample sizes for the wider donuts. The instability reflects power loss, not a hidden effect.

D. Heterogeneity Appendix

D.1 Operator Size

Splitting at the median number of pre-incident historical incidents produces null effects for both large operators (above median) and small operators (below median). The standardized effect sizes for both subgroups are reported in Table 4, Panel B.

E. Additional Figures

Figure 9 reproduces the bandwidth sensitivity analysis from the main text. The monotonic convergence of the point estimate toward zero as bandwidth increases—from −9.6 at 50% of the CCT-optimal bandwidth to −0.87 at 150%—is the signature pattern of a null effect: noise dominates at narrow bandwidths, and expanding the sample simply confirms zero. If the true effect were nonzero but obscured by noise, we would expect the estimate to stabilize at a nonzero value as the bandwidth widens.

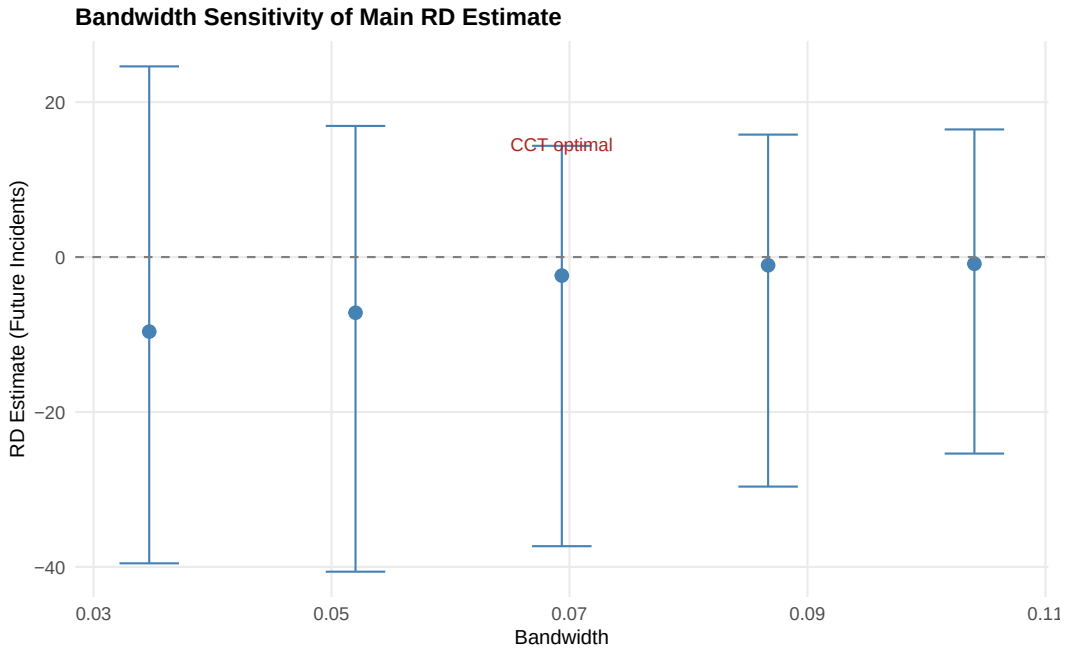

Figure 9: Appendix: Bandwidth Sensitivity (Reproduction)

Notes: Same as Figure 6. Reproduced here for ease of reference alongside the bandwidth table.

Table 4: Standardized Effect Sizes

Outcome	$\hat{\beta}$	SE	SD(Y)	SDE	SE(SDE)	Classification
Panel A: Pooled						
Future Incidents	-2.395	10.842	22.211	-0.108	0.488	Moderate negative
Log Future Cost	-0.617	1.636	5.968	-0.103	0.274	Moderate negative
Any Future Incident	-0.064	0.093	0.385	-0.165	0.241	Large negative
Panel B: Heterogeneous (by operator size)						
Large Operators	1.090	15.770	25.587	0.043	0.616	Small positive
Small Operators	-6.739	12.210	12.740	-0.529	0.958	Large negative

Notes: **Country:** United States. **Research question:** Does receiving PHMSA's "significant incident" label — triggered by a CPI-adjusted cost threshold — causally reduce subsequent pipeline safety incidents for the labeled operator? **Policy mechanism:** Pipeline incidents exceeding $50,000 in 1984 dollars (approximately $105,000–$141,000 nominal, 2010–2022) receive PHMSA's "significant incident" designation, which publicly flags the operator, triggers mandatory federal enforcement review, and exposes the operator to civil penalty proceedings. **Outcome definition:** Future incidents is the count of all PHMSA-reported pipeline incidents by the same operator in the three years following the index incident; log future cost is the natural log of one plus total incident costs over the same window; any future incident is a binary indicator for at least one subsequent incident. **Treatment:** Binary — whether the index incident's total cost exceeds the CPI-adjusted significant incident threshold. **Data:** PHMSA Pipeline Safety Program incident reports via jmceager/phmsa_clean GitHub repository, 2010–2022, incident-level observations, 550 incidents within 20% bandwidth. **Method:** Sharp regression discontinuity design using `rdrobust` with triangular kernel and MSE-optimal (CCT) bandwidth selection; standard errors clustered by operator. **Sample:** All gas and hazardous liquid pipeline incidents with positive reported total cost, 2010–2022; restricted to incidents within the CCT-optimal bandwidth of the CPI-adjusted threshold for RDD estimation. SDE = $\hat{\beta}$/SD(Y) where SD(Y) is the pre-threshold (below-cutoff) standard deviation. Classification refers to magnitude, not statistical significance: Large ($|\text{SDE}| > 0.15$), Moderate (0.05–0.15), Small (0.005–0.05), Null (< 0.005).

F. Standardized Effect Sizes

The standardized effect sizes (Table 4) confirm that the main results fall in the null-to-small range across all outcomes. The future incident count SDE is classified based solely on the point estimate, not statistical significance, following meta-analytic best practice.