NBER WORKING PAPER SERIES

POLICY EVALUATION BEYOND SIGNIFICANCE TESTING: EVIDENCE FROM CHICAGO'S SHOTSPOTTER CONTRACT

Terence Chau
Philip J. Cook
Javier Lopez
Jens Ludwig
Alejandro Roemer
Adam K. Soliman

Working Paper 36005
http://www.nber.org/papers/w36005

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2026

This project was supported by Arnold Ventures Grant 24-12760 and University of Chicago Crime Lab general operating funds; the project was reviewed and approved by the University of Chicago's Institutional Review Board (IRB25-1416). We have no financial or other conflicts of interest to report. Thanks to Aaron Chalfin, David Drukker, John Deke, John MacDonald, Derek Neal, John Pepper, Greg Ridgeway, Evan Rose, Rosanna Smart, Chris Walters and seminar participants at the NBER and the annual American Society of Evidence-Based Policing conference and the LSE / University of Chicago economics of crime conference for helpful comments. Thanks to Natalie Lu for excellent assistance with the data analysis. This work was carried out while Terence Chau was at the University of Chicago Crime Lab, and does not relate to his position at Amazon. Any errors and all opinions are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Terence Chau, Philip J. Cook, Javier Lopez, Jens Ludwig, Alejandro Roemer, and Adam K. Soliman. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Policy Evaluation Beyond Significance Testing: Evidence from Chicago's ShotSpotter Contract
Terence Chau, Philip J. Cook, Javier Lopez, Jens Ludwig, Alejandro Roemer, and Adam K. Soliman
NBER Working Paper No. 36005
November 2026
JEL No. C10, C12, C44, K42

ABSTRACT

Program impact estimates that are not statistically significant are often treated as providing little guidance for policy choice, even when economically important effects remain plausible. In such cases, the decision-relevant task is not simply to test whether the effect differs from zero, but to estimate the probability distribution of programmatic effects. The importance of going beyond hypothesis testing is especially clear when evaluations are underpowered relative to the effect needed for benefits to exceed costs. We illustrate this point using Chicago's contract for ShotSpotter, an acoustical gunshot detection system that reduces police response times and may therefore have improved survival among shooting victims. Using a geographic RD design, we estimate that ShotSpotter reduced the shooting fatality rate by 2.5 percentage points, equivalent to roughly 49 fewer deaths per year. Although the estimate is not statistically significant, its implied benefit far exceeds the cost of the service. To have enough statistical power under the hypothesis testing approach to detect an effect of this size would require 87 years' worth of Chicago shooting data. We instead implement a decision-relevant perspective using Bayesian analysis, treating the life-saving effect as a random variable. We find positive expected net benefits, with posterior probabilities that benefits exceed costs ranging from 70% under a flat prior and a $0.5 million value of statistical life to 99.6% under a no-harm prior and a $13.6 million value. The results illustrate how reliance on statistical significance can discard decision-relevant evidence when economically meaningful effects are difficult to detect precisely.

Terence Chau
Amazon, Inc.
terence.chau.c@icloud.com

Philip J. Cook
Duke University
Sanford School of Public Policy
and NBER
pcook@duke.edu

Javier Lopez
The University of Chicago
jlopez75@uchicago.edu

Jens Ludwig
The University of Chicago
Harris School of Public Policy
and NBER
jludwig@uchicago.edu

Alejandro Roemer
The University of Chicago
aleroemer@uchicago.edu

Adam K. Soliman
Clemson University
Wilbur O. and Ann Powers College of Business
John E. Walker Department of Economics
asolima@clemson.edu

I Introduction

Policy evaluations are often summarized by whether an estimated effect is statistically distinguishable from zero, but the policy question is different: Is the program worth adopting or continuing? As Imbens (2021), Manski (2019), and Wasserstein and Lazar (2016) (among many others) have noted, answering that question requires information about expected net benefits, the probability that benefits exceed costs, and the distribution of possible gains and losses.[1] Whether the estimated impact is statistically different from zero is, by itself, largely irrelevant to the decision of whether a program should be adopted. The American Statistical Association called for "a world beyond $p < 0.05$" (Wasserstein et al., 2019). Yet in practice, hypothesis testing continues to play a central role in applied economics. When an estimated impact is not statistically significant by the usual standard, evaluations are often treated as uninformative about whether the program is worthwhile. To the extent that "not statistically significant" is interpreted as "no effect," decisionmakers may therefore discontinue, or fail to adopt, worthwhile programs.

Since appeals to principle within the econometrics literature have not changed applied practice much, we focus here on a more pragmatic rationale for doing what econometricians have been suggesting: Many of our studies in economics are underpowered, often severely so, relative to the effect size required for a program to be worthwhile. To see why this is such a problem, note that the common applied practice can be represented schematically as follows: If an impact estimate is statistically significant, consider its economic magnitude as measured by, for example, a benefit-cost ratio; if not, treat the evaluation as providing little evidence of benefit. This common practice means that a Type 2 error in program-impact evaluation can conceal evidence that the program is in fact worthwhile. The "significance filter" in reporting evaluation results is particularly distorting for evaluations that are under-powered relative to the break-even effect.

We illustrate the magnitude of the distortion that can arise in practice with Chicago's contract for ShotSpotter, an acoustical gunshot detection system that alerts police to gunfire. Existing evaluations of ShotSpotter have focused largely on police activity and gun violence, generally finding improvements in some operational outcomes. We instead examine whether faster notification and police response reduce the probability that a shooting victim dies. We use a geographic regression discontinuity (RD) design that capitalizes on the fact that some police districts, but not others, adopted the ShotSpotter technology. While sound waves obviously travel across district boundaries, the company basically only sent notifications to the Chicago Police Department (CPD) for gunshots that happen within the 'treatment' districts. We estimate that ShotSpotter reduced the shooting fatality rate by 2.5 percentage points, equivalent to roughly 49 fewer deaths per year. If accurate, the monetized benefit far exceeded the cost of the service, but the estimate is not statistically significant and thus would often be treated in applied work as uninformative about whether the policy was worthwhile. Even though Chicago is the third-largest city in the country with the highest number of annual shootings (by far), to have adequate statistical power to detect a 2.5 pp effect

[1] "In many cases, researchers are interested in a point estimate and the degree of uncertainty associated with that point estimate as the precursor to making a decision or recommendation to implement a new policy. In such cases, the absence or presence of statistical significance (in the sense of being able to reject the null hypothesis of zero effect at conventional levels) is not relevant, and the all-too-common singular focus on that indicator is inappropriate." (Imbens, 2021, p. 158)

under the hypothesis testing approach would require 87 years' worth of Chicago data.

The key issue is not simply that our point estimate is imprecise, but that the effect required for ShotSpotter to be worthwhile is far smaller than the effect a feasible evaluation could reliably detect. At an annual contract cost of roughly $9 million and a value of statistical life of $13.6 million, ShotSpotter would pass a benefit-cost test if it saved even just one life per year. Among 1,966 outdoor shootings in ShotSpotter districts, that corresponds to a reduction in the fatality rate of only about 0.05 percentage points (around one-twentieth of one percentage point), compared with a standard error of roughly 3 percentage points. Achieving 80% power to detect an effect as small as this would require on the order of 200,000 years' worth of Chicago shooting data. In cases like this, no feasible evaluation can precisely distinguish economically relevant effects from zero.

This mismatch is not unique to ShotSpotter: As Ioannidis et al. (2017) argue (and we discuss in more detail below), the overwhelming share of studies within empirical economics are underpowered. Those authors argue that the solution is for economists to simply increase power through use of both bigger samples and meta-analyses. But even some of the most famous large-scale randomized controlled trials (RCTs) in the social sciences like the Tennessee STAR class-size experiment (Krueger, 1999, 2003) are ex ante underpowered to detect the break-even effect; how exactly should we get funders to support even larger RCTs? Similarly, some of the most famous nationwide "natural experiment" studies of key policy questions like the effects on crime from more prisons (Levitt, 1996) or more police (Evans and Owens, 2007) are underpowered. How do economists expand sample when there are only so many U.S. cities and states? How do we accumulate enough studies in our meta-analyses when there are only so many convincing sources of exogenous variation?

Further compounding the difficulty of "sample sizing" our way out of this challenge is how far we can be from the sample sizes required to be powered to detect the break-even effect, as our ShotSpotter application illustrates. Where would the needed inputs for a useful meta-analysis on ShotSpotter ever come from? This problem is only likely to grow over time as "policy R&D" opens the door to new types of interventions with low marginal costs, including algorithmic and other technological interventions (Ludwig et al., 2024).

A Bayesian decision framework provides a different approach that enables us to use whatever impact-estimation evidence we have more constructively, as it asks a different question. Rather than reducing the evaluation to whether an effect can be distinguished from zero, it provides an estimate of the probability distribution of the effect with associated decision-relevant quantities, such as expected net benefits and the probability that benefits exceed costs. It also provides a transparent way to incorporate prior information (in cases where that is available). Applying this framework to ShotSpotter, we find positive expected net benefits and posterior probabilities that benefits exceed costs ranging from 70% under an uninformative prior and a $0.5 million VSL to 99.6% under a no-harm prior and a $13.6 million VSL.

Despite its advantages, Bayesian methods remain uncommon in contemporary applied economics. Among the 760 articles published in the *American Economic Journal: Applied Economics* and *American Economic Journal: Economic Policy* from 2020 through 2025, only 31 (4.1%) made use of Bayesian methods or reasoning, and only three (0.4%) connected Bayesian analysis directly to policy choice. By contrast, full-document searches over the same articles returned "null" in

41.6%, "hypothesis" in 59.2%, and "significant" in 87.1%.[2]

The contribution of this paper is therefore not the well-known observation that statistical significance isn't itself a decision rule, but rather to emphasize that the limits of the hypothesis testing approach are especially problematic when the evaluation is underpowered relative to the benefit-cost break-even effect—as often seems to be the case. When the effect required for a policy to be worthwhile lies below what the evaluation can reliably detect, treating statistical insignificance as evidence of no benefit discards the information relevant to the policy choice. A decision-relevant analysis, in contrast, takes full advantage of the available information.

II Illustrating the Problem: ShotSpotter Case Study

A ShotSpotter and Existing Evidence

ShotSpotter is an acoustic gunshot-detection system that uses networks of outdoor microphones to identify likely gunfire and transmit an alert and estimated location to police. The technology can therefore affect police response in two ways: by notifying police of shootings that otherwise generate no 911 call, and by providing information more quickly and precisely when a 911 call is also made. In a one-year pilot in Durham, North Carolina, Cook and Soliman (2024) find that ShotSpotter substantially increased the number of shooting incidents reported to police and reduced response times by about a minute.[3]

Existing evaluations have focused primarily on police activity and gun violence, with mixed results. In Chicago, Topper and Ferrazares (2024) find that ShotSpotter adoption slowed response to other high-priority calls and reduced arrests associated with those calls, while finding no detectable reduction in gun victimizations. Connealy et al. (2025) find increased gun confiscations but no statistically detectable effects on fatal or nonfatal shootings. Studies in other cities similarly suggest improvements in some operational outcomes, including response times and evidence recovery, but provide limited evidence of reductions in aggregate gun violence.[4]

Less attention has been devoted to a different potential outcome: whether faster police response reduces the probability that a shooting victim dies. Correlational evidence suggests that mortality from gunshot wounds rises with the delay until emergency medical services arrive (Nasser et al., 2020). In Chicago specifically, opening a Level I trauma center in 2018 shortened travel time by 9.5 minutes and was associated with a 4% decrease in firearms mortality in its service area (Poulson et al., 2026). Chicago police officers are trained and equipped to provide first aid, including tourniquets and other bleeding-control measures, and may arrive before emergency medical personnel.[5] ShotSpotter could therefore reduce shooting fatality rates even if it has little effect on the number of shootings: faster notification and more precise location information may bring officers to seriously injured victims sooner. This mechanism motivates our focus on shooting fatality rather than

[2] Appendix Table A1 reports the full search and classification of all 760 papers.

[3] During the pilot, 57% of gunfire notifications were ShotSpotter alerts without a corresponding 911 call, 28% were 911 calls without a ShotSpotter alert, and 15% involved both.

[4] See, for example, Koren (2018); Goldenberg et al. (2019); Mares and Blackburn (2021); Mares and Buettner (2024); Cheng (2023); Piza et al. (2024); Doucette et al. (2021).

[5] See Banks (2018) and Chicago Police Department (2023).

shooting incidence.

B Data

We rely on publicly available victimization data from the Violence Reduction Dashboard (VRD), hosted by the Chicago Mayor's Office and developed with technical assistance from our research center, the University of Chicago Crime Lab.[6] The VRD also includes a dataset of all ShotSpotter alerts recorded in Chicago since the system's introduction in 2017. Using timestamp and location information, we match alerts to shooting incidents to determine which victimizations were detected by ShotSpotter, utilizing precise geocodes and timestamps from both datasets.[7]

Between January 1, 2010 and September 22, 2024, the data include 41,347 victims who were shot outdoors, of which 22,927 occurred after ShotSpotter was first implemented. Outdoor shootings, which comprise 88.8% of all incidents, constitute our main analysis sample because they are the most likely to be detected by ShotSpotter.

The main outcome variable for our analysis is the shooting fatality rate, defined as the ratio of victims who were shot and killed to the total number of shooting victims. The fatality rate for outdoor shootings in Chicago during our study period is 16.1%, but with a strong upward trend from 2010 to 2024 (Appendix Figure A.2). Given these time trends, our research design will need to control for any factors that may have changed shooting fatality rates over time.

The likelihood of death is influenced by a number of factors, such as the location and number of gunshot wounds they suffer, and the caliber of the bullet(s) (Braga and Cook, 2018). Of course there are many cases in which the victim survives without any medical assistance in the field, and others where the victim dies immediately. It is the intermediate cases—serious injuries in which the victim survives but would die within minutes in the absence of first aid—in which there is a possibility that the police response time matters.

C ShotSpotter in Chicago

Chicago adopted ShotSpotter in 2017 following a sharp increase in gun violence in 2016 (Appendix Figure A.3). Since the contract with ShotSpotter set prices proportional to the land area covered, deployment was concentrated in the police districts within the South and West sides of Chicago where shootings were most frequent (Appendix Figure A.4). Those districts account for something like 50% of Chicago's total land area but (depending on the year) around 80% of the city's shootings. Chicago implemented a staggered rollout across districts in 2017, but several years later its contract was not renewed. ShotSpotter was terminated citywide on September 22, 2024.

Our research design (explained further below) uses the cross-sectional geographic discontinuity created by the use of ShotSpotter in some districts but not others. A number of alternative designs are possible, but are all limited in important ways. We could, for example, exploit the over-time variation introduced by the staggered roll-out and then the abrupt termination of the ShotSpotter contract, but such an analysis is complicated by apparent violations of the parallel trends assumption during the "pre" period and the strong upward trend in shooting fatality rates citywide.

[6] The underlying data are available through the Chicago Data Portal, and the VRD provides visualization tools.
[7] Appendix D assesses robustness to looser or tighter spatial and temporal tolerances.

Another alternative, a simple cross-section comparison of fatality rates between ShotSpotter and non-ShotSpotter districts with respect to fatality rates, is not valid, since a number of location-based factors influence the fatality rate other than police response time. For example, the relatively high-violence areas served by ShotSpotter could experience shootings that are intrinsically more lethal on average (due to the prevalence of automatic weapons or high-caliber weapons), or could have less adequate emergency medical services. As shown in Appendix Table A2, a simple mean comparison suggests higher lethality in ShotSpotter districts, but this difference cannot be interpreted causally.

D Research Design and Results

Our research design exploits a geographic discontinuity: shootings occurring just inside and just outside ShotSpotter district boundaries are plausibly comparable. Following Black (1999), we assume that in the absence of treatment, expected fatality rates vary smoothly across district borders, so any discontinuous jump in outcomes can be attributed to ShotSpotter.

Even though sound travels, so ShotSpotter microphones in the "treatment" districts can detect gunshots in the "control" districts as a matter of physics, ShotSpotter only alerts the city to gunshots within the treatment areas (with a small buffer). Empirically, about 60.9% of shootings in ShotSpotter districts are linked to an alert compared with 5.0% in control districts, with a sharp rise in alert probability at the boundary (Figure 1).[8]

Focusing on outdoor shootings, which ShotSpotter is most likely to detect, we estimate the following sharp RD specification:

$$Y_i = \beta_1 + \beta_2 r_i + \beta_3 Z_i + \beta_4 f_h(r_i) + \epsilon_i, \tag{1}$$

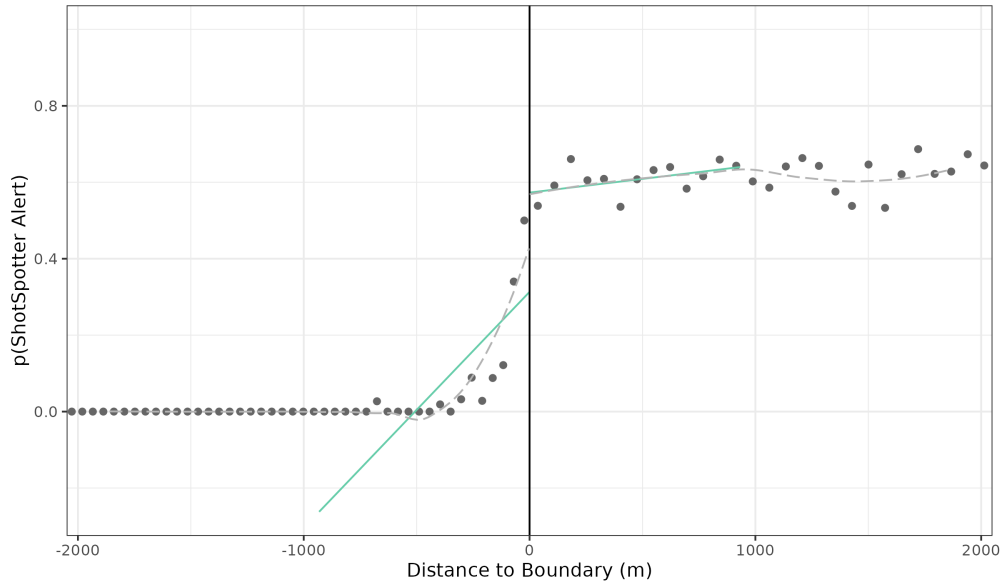
where Y_i indicates whether shooting i was fatal, r_i is the normalized distance to the district boundary, Z_i equals one if the incident occurred within a ShotSpotter district, and h defines the local bandwidth. The coefficient β_3 captures the intent-to-treat (ITT) effect at the boundary, $E(Y_{1i} - Y_{0i} \mid r_i = 0)$. We follow best practices in RD estimation, including local linear regressions with a triangular kernel, data-driven bandwidth selection (Imbens and Kalyanaraman, 2012), and robust inference procedures from Calonico et al. (2014) and Cattaneo et al. (2020). The results are robust to alternative bandwidths, kernels, and polynomial orders, as discussed below.

Identification of β_3 in Equation (1) requires that potential outcomes be continuous at the boundary in the absence of treatment. This assumption would be violated if, for example, police or medical response protocols changed discretely across district lines for unrelated reasons (Topper and Ferrazares, 2024).

We find no detectable evidence of such discontinuities in the pre-ShotSpotter period (January 1, 2010 - January 13, 2017). Fatality rates appear to be smooth across district borders: estimated

[8] The City of Chicago randomly jitters victimization coordinates within a circle roughly one city block in radius (≈200 meters) to preserve anonymity. See University of Chicago (2020) and Chicago Department of Transportation (2007). This spatial randomization may slightly attenuate RD estimates by blurring the boundary, so we assess potential bias from spatial measurement error. Using police-district identifiers (which are not jittered), we estimate that only 250 shootings (1.7%) within 2 km of the relevant boundaries are misclassified across treatment status.

Figure 1: Probability of Receiving a ShotSpotter Alert



Notes: Optimal-bandwidth local linear regressions in teal (coefficient = 0.265, robust $p < 0.001$), with local linear smoothers fitted separately on each side of the discontinuity in dashed gray. Bins are constructed using the mimicking-variance evenly spaced method (Calonico et al., 2015).

differences are small (1-2 pp) and statistically indistinguishable from zero ($p = 0.63$), consistent with continuity of potential outcomes (Table 1 and Appendix Figure A.5). Observable victim characteristics are likewise balanced across boundaries (Appendix Tables A3 and A4).

In principle, we might also worry about "bunching" of observations around the district boundary cutoffs if, for example, offenders know about ShotSpotter and follow their victims until they traverse a district boundary into a place without the technology. To test for this hypothesis, we use the density tests of Cattaneo et al. (2020) and McCrary (2008). As reported in Appendix Figures A.6 and A.7, we find no evidence of manipulation or bunching near the cutoff.

Our primary results are presented in Table 1. We interpret the post-implementation geographic discontinuity at the boundary of ShotSpotter and non-ShotSpotter districts as evidence of ShotSpotter's effect (see also Figure 2). Our preferred specification implies a -2.5 percentage-point effect, equal to about 15% of the control mean, with a 95% confidence interval of $[-0.091, 0.027]$. Across bandwidths and specifications, point estimates remain consistently negative but not statistically significant at conventional levels. Specifically, in the appendix, we show that the results are robust to the choice of alternative bandwidths (100 m - 2 km) and polynomial order; see Appendix Figures A.8, A.9, and A.10, as well as Appendix Table A5.

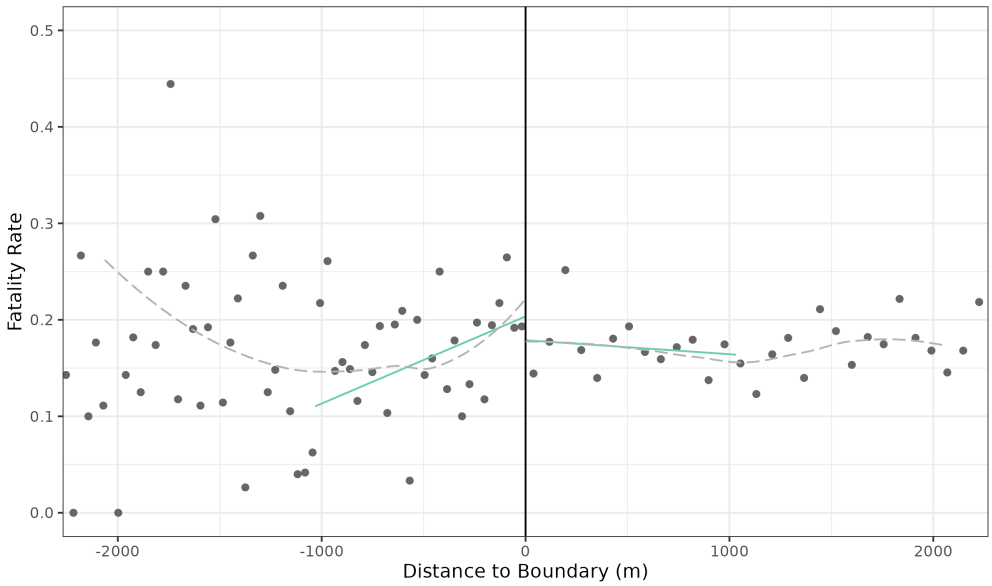
We can also carry out a series of placebo geographic tests following Cattaneo et al. (2020), in which we select "pseudo-thresholds" that divide areas in ways that the two sides of the threshold do not actually differ in their receipt of ShotSpotter services. In placebo locations without treatment, estimated discontinuities are consistently small and statistically insignificant, suggesting the main results are not artifacts of finite-sample noise or model misspecification (see Appendix C).

Table 1: Sharp Regression Discontinuity Estimates

	Outcome: Fatal Shooting	
	Treatment Period	Pre-Treatment Period
In ShotSpotter Area	-0.025	0.015
	[-0.091, 0.027]	[-0.039, 0.064]
	(0.286)	(0.634)
Control Mean in Bandwidth	0.17	0.128
Optimal Bandwidth	1032.105	1147.441
Observations	4021	3926

Notes: Bandwidth-selection robust 95% confidence intervals are reported in brackets and robust p-values in parentheses. Estimates are from local linear regressions using a triangular kernel.

Figure 2: Shooting Fatality Rate at the ShotSpotter Boundary



Notes: Optimal-bandwidth local linear regressions in teal (coefficient = -0.025, robust $p = 0.286$), with local linear smoothers fitted separately on each side of the discontinuity in dashed gray. Bins are constructed using the mimicking-variance evenly spaced method (Calonico et al., 2015). Each observation is a shooting victimization.

What to make of these results? The outcome of interest (shooting fatality rates) is important and the policy has a plausible mechanism of action on that outcome. The research design and resulting estimates seem plausible under the usual standards of evidence. Yet since the confidence interval around our impact estimates includes zero, under the standard hypothesis-testing framework, these results would in practice typically be treated as either uninformative, or as demonstration of zero effectiveness. Decisionmakers would be free to conclude that "ShotSpotter doesn't work."

E Power Considerations

If our point estimate (a 2.5 percentage point reduction in the fatality rate) were accurate, then the ShotSpotter contract would pass a cost-benefit test. Under our benchmark assumption that the RD effect applies to the 1,966 outdoor shootings in Chicago's ShotSpotter districts in 2024, the implication is roughly 49 fewer deaths in that year. Given the 2025 federal standard for the value of a statistical life (VSL), $13.6 million, those lives would be valued at about $666 million. Even at a much lower VSL, this benefit greatly exceeds the annual contract cost ($9 million) added to the opportunity costs of the increase in police deployments from shooting alerts (estimated to be in the range of $2-7 million by Topper and Ferrazares (2024)).

The problem is that our estimation procedure is underpowered relative to this considerable beneficial effect. How much sample would be required to have 80% power to detect an effect that reduced the shooting fatality rate by 2.5 percentage points? That would require sufficient statistical power such that 2.5 percentage points was 2.8 times the standard error; to shrink our actual standard error of 3.0 percentage points down to (2.5 pp / 2.8) = 0.9 pp would require roughly 11 times our current sample (of about 7.7 years of "treatment data"), which would correspond to about 87 years' worth of data.[9]

The analysis would be still more underpowered with respect to smaller effect sizes. In fact, using the 2025 federal VSL, ShotSpotter would be worthwhile if just one life were saved per year, equivalent to a 0.05 percentage point reduction. To be able to detect an effect at this one-life benchmark—that is, to get our standard error down to (0.05 pp / 2.8) = 0.018 pp (one-fifty-sixth of a percentage point)—would require roughly 28,000 times our current sample, or on the order of 200,000 years of data.[10] The point is that for an underpowered evaluation, it is inappropriate to interpret a null hypothesis test as demonstration that a program is not worthwhile.

As Imbens (2021) notes, the conservatism of this hypothesis testing approach might make sense for the purposes of guiding scientific progress, assuming the null hypothesis represents the current consensus of scientists. But for policy decisions, the null hypothesis has no such privileged position. The results of a hypothesis test are essentially irrelevant to the choice.

[9] This calculation uses the conventional approximation that the standard error shrinks with the square root of the sample size, so that 80% power at the 5% significance level requires $|\tau|/SE \geq 1.96 + 0.84$. With $SE = 0.030$, the required sample is $(2.80 \times 0.030/|\tau|)^2$ times our current sample, which covers roughly 7.7 years of post-implementation data (January 2017 through September 2024). The conversion to years is therefore specific to this design; the multiples depend only on the standard error and the effect size.

[10] The exact break-even effect is $\frac{9}{13.6 \times 1{,}966} \approx 0.034$ percentage points. Achieving 80% power would therefore require a target standard error of approximately $0.034/2.8 = 0.012$ percentage points, which implies a sample-size multiplier of approximately $(3.0/0.012)^2 \approx 62{,}500$, or about 480,000 years of data.

III Illustrating the Solution: Decision-Relevant Analysis

We now use the ShotSpotter estimates to calculate expected net benefits and the probability that the program's benefits exceed its costs (Section A).[11] We then examine how robust those conclusions are to concerns about model misspecification (Section B).

A Expected-Value Decision Analysis

The problem with hypothesis testing for policy evaluation is that it answers the wrong question, namely, "is the effect distinguishable from zero?" Policymakers, however, must decide whether to adopt a policy, which is fundamentally a decision problem under uncertainty. Therefore, the relevant question is not whether the effect differs from zero, but whether the expected benefits exceed the costs, with attention to both statistical uncertainty and the consequences of incorrect decisions.

Following Claxton (1999), we formalize the decision problem by treating net benefit as the primitive object:

$$NB(\tau) = B(\tau) - C = V \times N_{\text{exposed}} \times (-\tau) - C,$$

where τ denotes the causal effect on shooting fatality risk, V is the value of a statistical life, N_{exposed} is the number of shootings affected by the policy, and C is the cost of implementation.

Under risk neutrality, the policy choice is governed by the expected-value decision rule:

$$\text{Adopt if } \mathbb{E}[NB \mid \text{data}] > 0.$$

Under this rule, statistical significance is neither necessary nor sufficient for policy adoption. The problem arises when statistical significance is used as a decision rule, so that an estimate failing to reject zero is treated as providing no evidence of benefit—Appendix Figure A.1 highlights that the benefit-cost break-even effect can lie well below the effect size required for conventional statistical significance.

To implement the decision rule above, we need to characterize uncertainty about the treatment effect τ. We begin with a flat prior over the real line as a transparent benchmark. Using our preferred estimates from Table 1, the RD point estimate is -0.025 with a standard error of 0.030. Under our Normal approximation and a flat prior, the posterior is therefore

$$\tau \mid \text{data} \sim \mathcal{N}(-0.025, 0.030^2).$$

Given the net-benefit function defined above, the benefit-cost break-even effect is

$$\tau_{BE} = -\frac{C}{V \times N_{\text{exposed}}},$$

the reduction in the shooting fatality rate required for benefits to equal costs. Using baseline values of $C = \$9$ million, $V = \$13.6$ million, and $N_{\text{exposed}} = 1966$, this yields $\tau_{BE} \approx -0.000337$.

[11] Appendix Table A6 summarizes the different approaches for policy evaluation.

For each posterior draw $\tau^{(s)}$, the corresponding annual net benefit is $NB^{(s)} = V \times N_{\text{exposed}} \times (-\tau^{(s)}) - C$.[12] Under the flat-prior benchmark, expected annual net benefits are approximately \$659 million (\$13.6M $\times$ 1966 $\times$ 0.025 $-$ \$9M). ShotSpotter would therefore be adopted under the expected-value decision rule. A complementary summary of uncertainty is the posterior probability that benefits exceed costs,

$$P(B > C) = P(\tau < \tau_{BE}),$$

which is approximately 0.796 under the benchmark VSL.

The flat prior leaves the treatment effect unrestricted. We therefore also consider a simple ShotSpotter-specific "no-harm" benchmark, $\pi(\tau) \propto \mathbf{1}\{\tau \leq 0\}$, which allows for no effect but rules out an increase in shooting fatality risk. The sign restriction is plausible here because the outcome is death due to a shooting, and the proposed mechanism is faster detection and police response, but we do not recommend such a restriction for interventions generally. Combining the Normal likelihood with the sign restriction yields a truncated-Normal posterior with mean approximately -0.036 and standard deviation 0.023. At the benchmark VSL and exposure level, $P(B > C)$ is approximately 0.996 under the no-harm prior.

Figure 3 summarizes the posterior distribution of annual net benefits under these two benchmark priors and two values of a statistical life. The circles report posterior means, the horizontal bars characterize the distribution of possible net benefits, and the boxed labels report $P(B > C)$. The main conclusion is the same under both priors: expected net benefits are positive, and most posterior mass lies above the benefit-cost threshold.

We report $P(B > C)$ to summarize uncertainty, not as a separate adoption rule. Under risk neutrality, the relevant criterion remains expected net benefit. Policymakers with different objectives or greater aversion to losses may instead place more weight on the distribution of possible outcomes, which Figure 3 also reports.

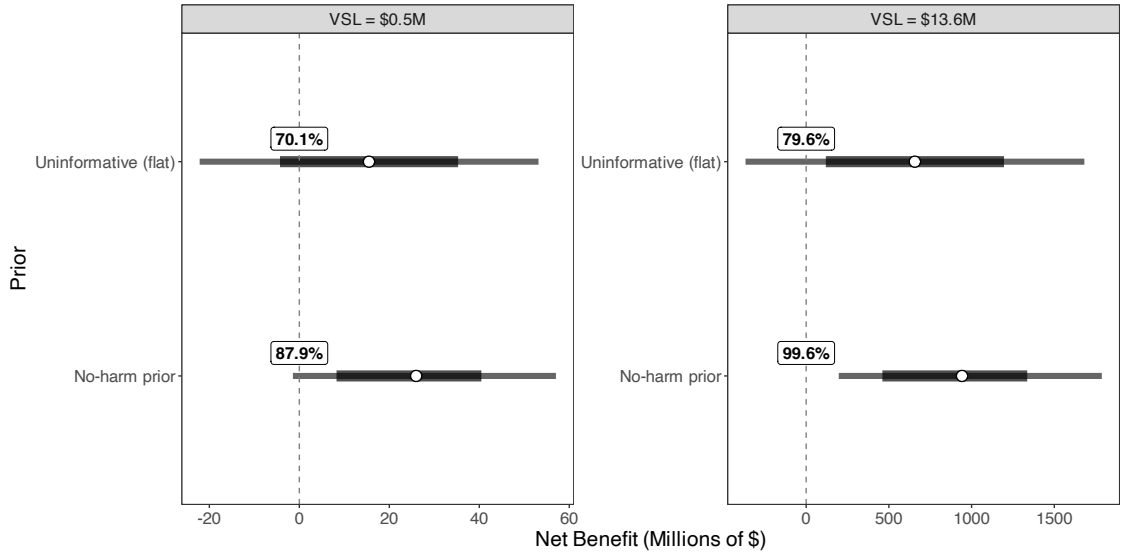
Sensitivity and Robustness

The two benchmark priors are not meant to capture the full range of plausible treatment effects. In theory, external evidence on the fraction of gunshot victims whose survival could be affected by faster response would help discipline the magnitude of the prior, but we have not found the necessary information to calibrate such bounds. The appendix therefore reports sensitivity exercises that relax the no-harm restriction or place greater prior weight on effects near zero. We view these as sensitivity checks rather than calibrated priors. Appendix Figure A.11 summarizes the results, with additional detail in Appendix Figures A.12, A.13, and A.14.

We next examine sensitivity to the geographic reach of the RD effect. Our baseline benefit-cost calculation applies the local RD estimate to all 1,966 outdoor shootings in ShotSpotter districts. As a more conservative alternative, we restrict the exposure base to shootings within the geographic footprint used to estimate the RD effect, which yields $N_{\text{exposed}} \approx 110$ shootings per year; Appendix Figure A.15 reports results across the full range of exposure assumptions. Even under this much

[12] The corresponding benefit-cost ratio is $BCR^{(s)} = \frac{V \times N_{\text{exposed}} \times (-\tau^{(s)})}{C}$. Because $BCR^{(s)}$ is a monotone transformation of $NB^{(s)}$, the posterior distribution of net benefits directly provides the corresponding distribution of benefit-cost ratios.

Figure 3: Posterior Net Benefits under Benchmark Priors and VSL Assumptions



Notes: Boxed labels report the posterior probability that program benefits exceed costs, $P(B > C)$. White circles report posterior mean annual net benefits. Lighter horizontal segments span the 10th–90th percentiles and darker segments span the interquartile range. Results are shown for the uninformative flat prior and the no-harm prior at VSL values of $0.5 million and $13.6 million. All calculations assume $N_{\text{exposed}} = 1966$ and an annual program cost of $C = \$9$ million.

narrower exposure definition, the posterior probability that benefits exceed costs remains about 0.74 under the flat prior and above 0.93 under the no-harm prior.

Lastly, we account for the enforcement spillovers documented by Topper and Ferrazares (2024). They find no detectable effect of ShotSpotter on gun victimizations, but they do estimate roughly 70 fewer Priority-1 arrests per year. Valuing that loss at $2–7 million annually lowers estimated net benefits, but not enough to change the adoption decision under either benchmark prior.

B Robust Control under Model Misspecification

The analysis in Section A treats the posterior distribution as our description of uncertainty about ShotSpotter's effect. However, the assumptions linking the estimated effect to the policy decision may be wrong. We therefore use the robust control framework of Hansen and Sargent (2001) to examine how the policy conclusion changes when greater weight is placed on unfavorable outcomes. Rather than specifying an alternative model, this approach allows the distribution to deviate from the baseline posterior, subject to a penalty that increases with the size of the deviation.

Model Perturbations

Let P denote the posterior distribution of the treatment effect τ, which summarizes uncertainty based on the data and prior assumptions (as in Section A). Robust control considers alternative distributions Q that place more probability on unfavorable outcomes while remaining close to P.

We measure that distance using relative entropy (the Kullback-Leibler divergence),

$$\mathrm{KL}(Q\|P) = \mathbb{E}_Q\left[\log\left(\frac{dQ}{dP}\right)\right].$$

Larger deviations from the posterior carry a larger penalty.

Robust Evaluation of Net Benefits

Under robust control, policy evaluation starts by asking how net benefits would change if probability mass were shifted toward less favorable outcomes. The robust certainty-equivalent net benefit is

$$NB_{\mathrm{rob}}(\theta) = \inf_{Q \ll P}\left\{\mathbb{E}_Q[NB(\tau)] + \theta\,\mathrm{KL}(Q\|P)\right\},$$

where the parameter θ governs the tradeoff between expected net benefits and concern about misspecification. Smaller values of θ allow for greater reweighting toward adverse outcomes, while larger values keep the evaluation closer to the posterior benchmark. The relevant question is therefore how small θ must become before the policy conclusion changes.

For the static adoption decision considered here, robust net benefit has the closed-form representation

$$NB_{\mathrm{rob}}(\theta) = -\theta\,\log\mathbb{E}_P\left[\exp\left(-\frac{NB(\tau)}{\theta}\right)\right],$$

and we evaluate this expression using posterior draws of τ. For each draw, we calculate the corresponding net benefit and use a numerically stable log-sum-exp transformation. We can therefore calculate robust net benefits across values of θ without having to solve a separate optimization each time.

Interpretation

We summarize robustness using the break-even level θ^*, defined as the value of θ at which robust net benefits equal zero. Appendix Table A7 reports these thresholds across priors and alternative VSL assumptions. Under the flat prior and benchmark VSL, the adoption decision remains favorable unless θ falls below roughly \$485 million. Under the no-harm prior, the corresponding threshold is approximately \$1.2 million.

However, the value of θ^* is difficult to interpret on its own, so we also report how the posterior is reweighted at the break-even threshold. For a given θ, robust control assigns weights to posterior draws proportional to $\exp(-NB/\theta)$, so lower-net-benefit realizations receive more weight as θ falls. Appendix Table A8 reports the resulting concentration of decision weight in the lower tail across priors and VSL assumptions.[13] At these break-even thresholds, the flat-prior evaluation places approximately 25% of total decision weight on the worst 1% of posterior outcomes and about 50% on the worst 5%. Under the no-harm prior, essentially all of the decision weight falls on the worst 1% of outcomes.

[13] The worst 1% and 5% are defined by ranking posterior draws from lowest to highest net benefits.

IV ShotSpotter as an Illustration of a Larger Problem

The limits of hypothesis testing for policy decisions for our ShotSpotter application illustrate a much larger problem: Empirical economics studies are often underpowered, and the problem can be especially severe relative to the effects that matter for policy decisions.

For example, Ioannidis et al. (2017) examine 159 meta-analyses of different economic literatures, capturing 64,076 separate point estimates from 6,700 different studies. The areas they examine cover international, macro, finance, labor, development and other areas of micro-economics. Their test for statistical power is to ask how many studies would have 80% power to detect the "true effect" in a literature, defined for example by the average effect in the meta-analysis. Put differently, they ask: how many studies have a standard error smaller than the true effect divided by 2.8? They note that the median statistical power is 18% - very far from the 80% conventional target. They note that half the research areas they examine have at least 90% of all studies underpowered under this definition.

Ioannidis et al. (2017) suggest the solution is for economists to increase the statistical power of their individual studies and make greater use of meta-analyses. But even a little bit of reflection on how empirical economics works in practice is enough to realize the limits of this suggestion.

For starters, sample sizes in RCTs are rarely set by what researchers would need to achieve adequate power; in practice, they are largely determined by available research budgets. In Table 2, for example, we review several canonical policy experiments for which a benefit-cost estimate is available and derives largely from a single outcome, making our illustrative exercise simpler to carry out. These include the Tennessee STAR class-size reduction experiment (Krueger, 1999, 2003), high-dosage tutoring (Guryan et al., 2023), and Youth Guidance's Becoming a Man violence-prevention program (Heller et al., 2017). In every case, the break-even effect is less than the minimum effect the study can detect with 80% power (2.8 times the standard error). Power to detect the break-even effect is often 20% or less. Note that these are some of the larger RCTs we have in economics, at least in the domestic U.S. context. Closing these power gaps would often require sample sizes (and research budgets) many times larger than those in even some of economics' most prominent policy experiments. Nor will meta-analysis solve this given how few RCTs there are in these policy areas.

Similarly, the bottom panel of the table shows that canonical 'natural experiment' evaluations of key policies like increasing prison populations (Levitt, 1996) or police staffing (Evans and Owens, 2007) are also ex ante underpowered to detect the break-even effect where the policy would just pass a benefit-cost test. It is hard to imagine how to improve sample size here; Levitt (1996) already uses data on the universe of states that had a prison-overcrowding lawsuit filed, while Evans and Owens (2007) examine every city that received COPS funding grants from the federal government. And here again, meta-analysis seems an unlikely solution; there are so few credible sources of identifying variation in many of these literatures that it is hard to imagine accumulating enough sample through additional credible studies.

Table 2: Power Calculations for Selected Policy Evaluations

Study	Point estimate	Benefit-cost ratio	Break-even effect	MDE at 80% power ($2.8 \times SE$)	Power at break-even
Panel A: Canonical policy evaluation RCTs					
Tennessee STAR class-size reduction (Krueger, 1999, 2003)	5.4 percentile points	1.98–4.10	1.317–2.727	3.332	20–63%
High-dosage tutoring, Study 1 (Guryan et al., 2023)	0.18 SD	2.4–3.6	0.050–0.075 SD	0.185 SD	12–21%
High-dosage tutoring, Study 2 (Guryan et al., 2023)	0.40 SD	5.4–8.0	0.050–0.074 SD	0.294 SD	8–11%
Becoming a Man, Study 1 (Heller et al., 2017)	0.0345 fewer violent crimes per participant	5–30	0.001–0.007	0.046	5–7%
Panel B: Canonical policy evaluation natural experiments					
Effects of prison on crime (Levitt, 1996)	Violent-crime elasticity: −0.379	2.0	0.190	0.504	18%
Effects of police on crime (Evans and Owens, 2007)	All-crime effect: −2.246	1.77	1.269	1.938	45%

Notes: The break-even effect is the treatment effect at which estimated program benefits equal program costs. For studies reporting a benefit-cost ratio, we calculate the implied break-even effect by dividing the magnitude of the point estimate by the benefit-cost ratio, assuming benefits scale proportionally with the reported effect measure. The 80% MDE is the approximate minimum detectable effect for 80% power using a two-sided 5% test, calculated as $2.8 \times SE$. The final column reports the probability of rejecting a zero treatment effect at the 5% level if the true treatment effect were equal to the benefit-cost break-even effect. Where the benefit-cost ratio is reported as a range, we report the corresponding range of break-even effects and power. Break-even effects are reported in absolute value. For Guryan et al. (2023), we report results separately for Studies 1 and 2 because the point estimates and benefit-cost ratios differ substantially across the two trials and are reported separately in the paper. For Heller et al. (2017), we report Study 1 because that is the trial for which the authors conduct a benefit-cost analysis, reflecting the more complete outcome data available for that experiment. For Levitt (1996), we report the elasticity of violent crime with respect to imprisonment because the study reports effects separately for violent and property crime, and violent crime accounts for the overwhelming share of the estimated social costs of crime (see Levitt, 1996, Table 8).

V Conclusion

The primary task of a program-impact evaluation is to provide a decisionmaker with the evidence relevant to the choice of whether to adopt (or continue) the program. In a simple go/no-go decision about a particular program, that should involve, for starters, a measure of impact valued appropriately net of cost. The relevant statistical information is also characterized by the posterior probability distribution of this impact. For the most part, however, applied economists continue to assess economic magnitude only if the impact estimate is statistically significantly different from zero (in the "right" direction). That approach is conducive to misleading conclusions and, hence, faulty decisions, especially given that underpowered evaluations are commonplace. The ShotSpotter case study in Chicago serves as an illustration of this problem.

Our argument, in short, is that in a world of underpowered evaluations, avoiding the shortchanging of potentially worthwhile programs requires that program analysts accept the econometrician's advice and adopt a decision-relevant analysis.

References

Banks, P. (2018, January). Re: First aid, Policy, training, and equipment. Policy Report, Civilian Office of Police Accountability, City of Chicago.

Black, S. E. (1999, May). Do Better Schools Matter? Parental Valuation of Elementary Education. *The Quarterly Journal of Economics 114*(2), 577–599.

Braga, A. A. and P. J. Cook (2018, July). The Association of Firearm Caliber With Likelihood of Death From Gunshot Injury in Criminal Assaults. *JAMA Network Open 1*(3), e180833.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014, November). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs: Robust Nonparametric Confidence Intervals. *Econometrica 82*(6), 2295–2326.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2015, October). Optimal Data-Driven Regression Discontinuity Plots. *Journal of the American Statistical Association 110*(512), 1753–1769.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2020). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge Elements: Quantitative and Computational Methods for Social Science. Cambridge: Cambridge University Press.

Cattaneo, M. D., M. Jansson, and X. Ma (2020, July). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association 115*(531), 1449–1455.

Cheng, Y. (2023, July). Houston's gunshot alert system isn't curbing violence but delays police response times, data shows.

Chicago Department of Transportation (2007, April). Street and Site Plan Design Standards. Design Standards, City of Chicago.

Chicago Police Department (2023). 2023 annual report. Technical report, Chicago Police Department, Chicago, IL.

City of Chicago (2024a, October). Violence reduction - ShotSpotter alerts - historical. https://data.cityofchicago.org/Public-Safety/Violence-Reduction-Shotspotter-Alerts-Historical/3h7q-7mdb/about_data.

City of Chicago (2024b). Violence reduction - victims of homicides and non-fatal shootings. https://data.cityofchicago.org/Public-Safety/Violence-Reduction-Victims-of-Homicides-and-Non-Fa/gumc-mgzr/about_data. Dataset containing individual-level homicide and non-fatal shooting victimizations from 1991 to present.

Claxton, K. (1999). The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of health economics 18*(3), 341–364.

Connealy, N. T., E. L. Piza, R. A. Arietti, G. O. Mohler, and J. G. Carter (2025). Staggered deployment of gunshot detection technology in chicago, il: a matched quasi-experiment of gun violence outcomes. *Journal of experimental criminology 21*(3), 1027–1053.

Cook, P. J. and A. Soliman (2024). Evaluation of Durham's ShotSpotter Installation: Results of a 12-Month Pilot Project. *SSRN Electronic Journal*.

Doucette, M. L., C. Green, J. Necci Dineen, D. Shapiro, and K. M. Raissian (2021, October). Impact of ShotSpotter Technology on Firearm Homicides and Arrests Among Large Metropolitan Counties: A Longitudinal Analysis, 1999–2016. *Journal of Urban Health 98*(5), 609–621.

Evans, W. N. and E. G. Owens (2007). Cops and crime. *Journal of public Economics 91*(1-2), 181–201.

Goldenberg, A., D. Rattigan, M. Dalton, J. P. Gaughan, J. S. Thomson, K. Remick, C. Butts, and J. P. Hazelton (2019, December). Use of ShotSpotter detection technology decreases prehospital time for patients sustaining gunshot wounds. *Journal of Trauma and Acute Care Surgery 87*(6), 1253–1259.

Guryan, J., J. Ludwig, M. P. Bhatt, P. J. Cook, J. M. V. Davis, K. Dodge, G. Farkas, R. G. Fryer, S. Mayer, H. Pollack, L. Steinberg, and G. Stoddard (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review 113*(3), 738–765.

Hansen, L. P. and T. J. Sargent (2001). Robust control and model uncertainty. *American Economic Review 91*(2), 60–66.

Heller, S. B., A. K. Shah, J. Guryan, J. Ludwig, S. Mullainathan, and H. A. Pollack (2017). Thinking, fast and slow? some field experiments to reduce crime and dropout in chicago. *The Quarterly Journal of Economics 132*(1), 1–54.

Imbens, G. and K. Kalyanaraman (2012, July). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies 79*(3), 933–959.

Imbens, G. W. (2021, August). Statistical Significance, p -Values, and the Reporting of Uncertainty. *Journal of Economic Perspectives 35*(3), 157–174.

Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos (2017). The power of bias in economics research. *The Economic Journal*, F236–F265.

Koren, D. (2018, October). ShotSpotter pilot assessment.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The quarterly journal of economics 114*(2), 497–532.

Krueger, A. B. (2003). Economic considerations and class size. *The economic journal 113*(485), F34–F63.

Levitt, S. D. (1996). The effect of prison population size on crime rates: Evidence from prison overcrowding litigation. *The quarterly journal of economics 111*(2), 319–351.

Lopez, J. and T. Ballard (2024, May). Summer safety analysis: Chicago crime trends. Technical report, University of Chicago Crime Lab.

Ludwig, J., S. Mullainathan, and A. Rambachan (2024, May). The unreasonable effectiveness of algorithms. *AEA Papers & Proceedings 114*, 623–627.

Manski, C. F. (2019). Treatment choice with trial data: Statistical decision theory should supplant hypothesis testing. *The American Statistician 73*(sup1), 296–304.

Mares, D. and E. Blackburn (2021, June). Acoustic gunshot detection systems: A quasi-experimental evaluation in St. Louis, MO. *Journal of Experimental Criminology 17*(2), 193–215.

Mares, D. and A. Buettner (2021 (updated 2024)). The Impact of Shotspotter Deployment in Winston-Salem, NC. Technical report, Center for Crime Science and Violence Prevention.

McCrary, J. (2008, February). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698–714.

Nasser, A. A., C. Nederpelt, M. El Hechi, A. Mendoza, N. Saillant, P. Fagenholz, G. Velmahos, and H. M. Kaafarani (2020). Every minute counts: the impact of pre-hospital response time and scene time on mortality of penetrating trauma patients. *The American Journal of Surgery 220*(1), 240–244.

Piza, E. L., G. O. Mohler, J. G. Carter, D. N. Hatten, N. T. Connealy, R. Arietti, J. Cho, and E. Castillo (2024, January). The impact of gunshot detection technology on gun violence in Kansas City and Chicago: A multi-pronged evaluation. Technical Report 308357, National Criminal Justice Reference Service / U.S. Department of Justice, Office of Justice Programs.

Poulson, M. R., A. Benjamin, and D. R. Scantling (2026). Firearm mortality and equitable access to trauma care in chicago. *JAMA surgery 161*(4), 406–411.

SoundThinking (2024). Gunshot Detection. https://www.soundthinking.com/law-enforcement/leading-gunshot-detection-system/.

Topper, M. and T. Ferrazares (2024). The Unintended Consequences of Policing Technology: Evidence from Shotspotter.

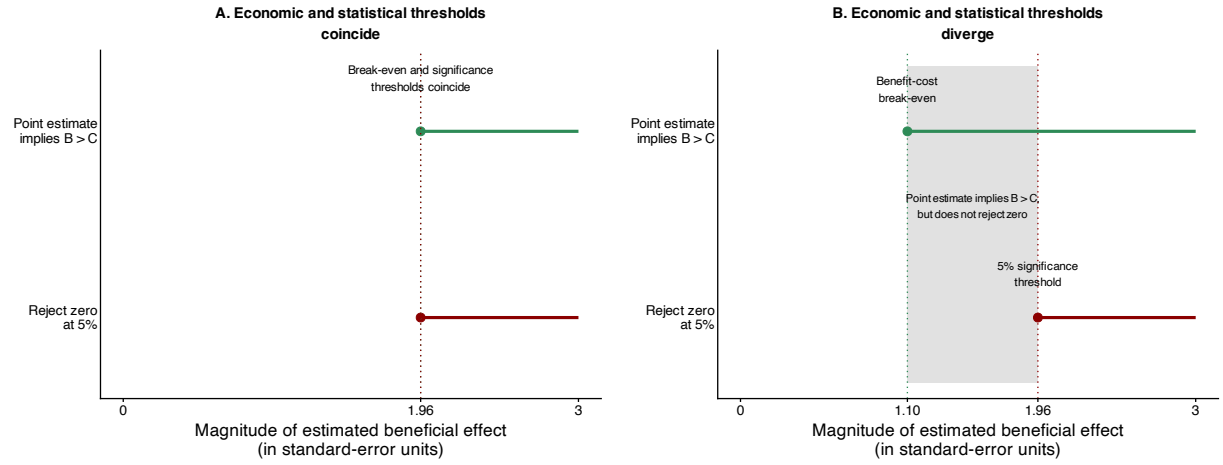
University of Chicago (2020, August). Chicago's Grid System. https://chicagostudies.uchicago.edu/grid.

Wasserstein, R. L. and N. A. Lazar (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond "p < 0.05".

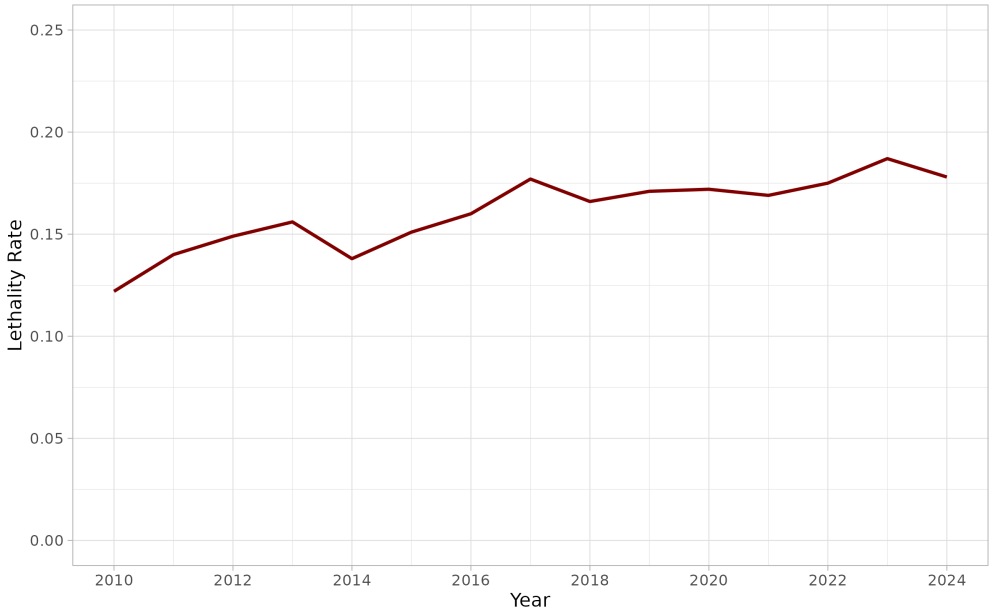
A Appendix Figures

Figure A.1: When Economic Relevance and Statistical Detectability Diverge



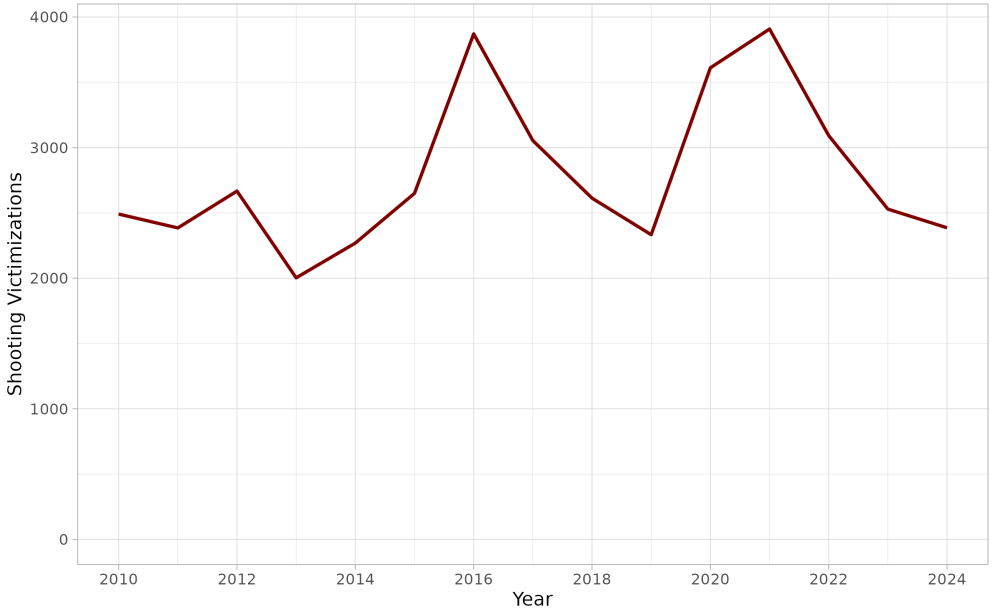
Notes: The horizontal axis measures the magnitude of the estimated beneficial effect in standard-error units. The benefit-cost threshold is the effect size at which the point estimate implies benefits equal costs; the statistical threshold is the effect size required to reject a zero effect at the 5% level. Panel A illustrates a case in which these thresholds coincide. Panel B illustrates a case in which the benefit-cost threshold is smaller than the conventional significance threshold. The shaded region therefore contains estimates whose implied benefits exceed costs based on the point estimate, but that are not statistically distinguishable from zero. The figure is intended to illustrate the distinction between economic relevance and statistical detectability, rather than to treat either threshold as a complete policy decision rule.

Figure A.2: Yearly Fatality Trends in Chicago, 2010-2024



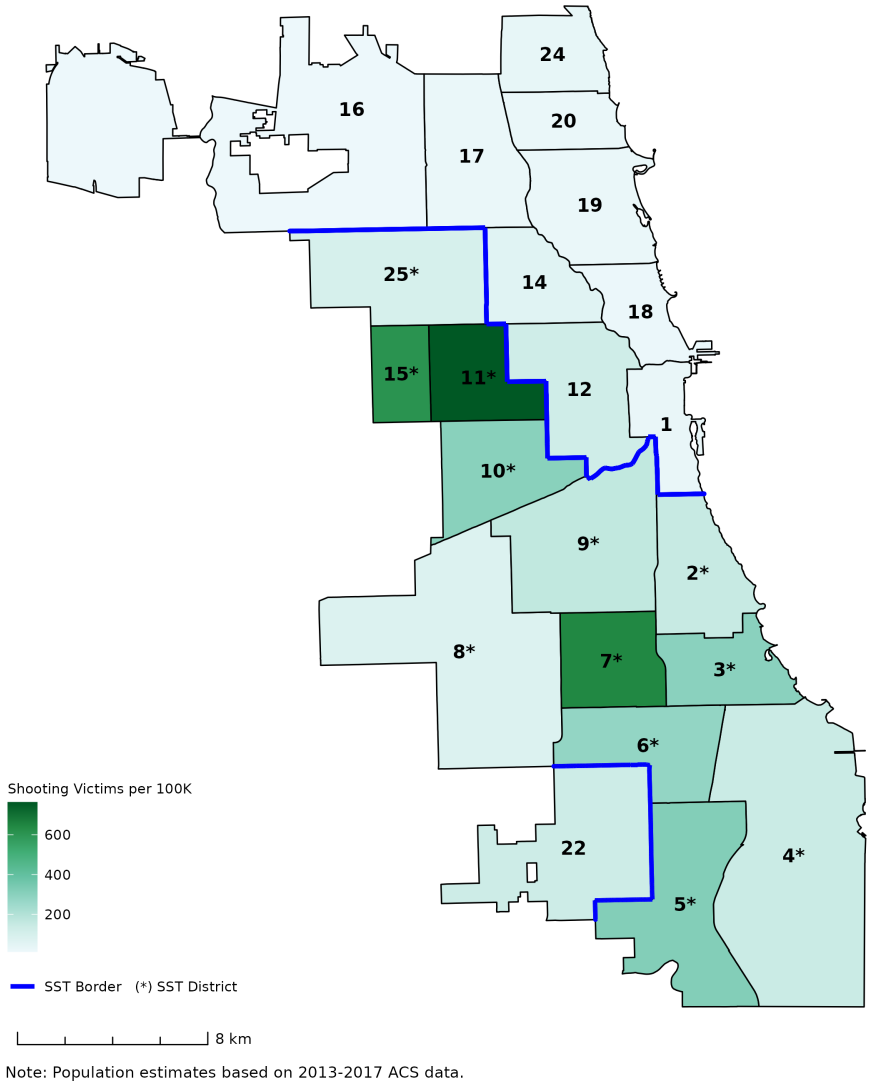
Notes: The figure plots the annual fatality rate among outdoor shooting victims in Chicago from 2010 through 2024. The fatality rate is defined as the share of shooting victims who died from their injuries.

Figure A.3: Yearly Gunshot Injury Trends in Chicago, 2010-2024



Notes: The figure plots the annual number of shooting victimizations in Chicago from 2010 through 2024. ShotSpotter was first implemented in Chicago in 2017, following the sharp increase in shootings in 2016.

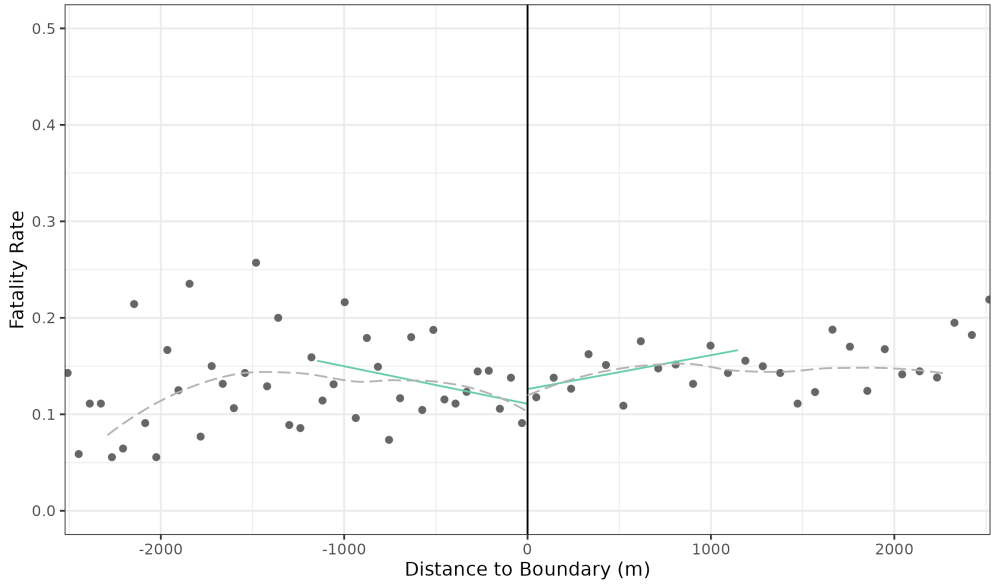
Figure A.4: Shooting Victimizations per 100K by CPD District, 2016



Note: Population estimates based on 2013-2017 ACS data.

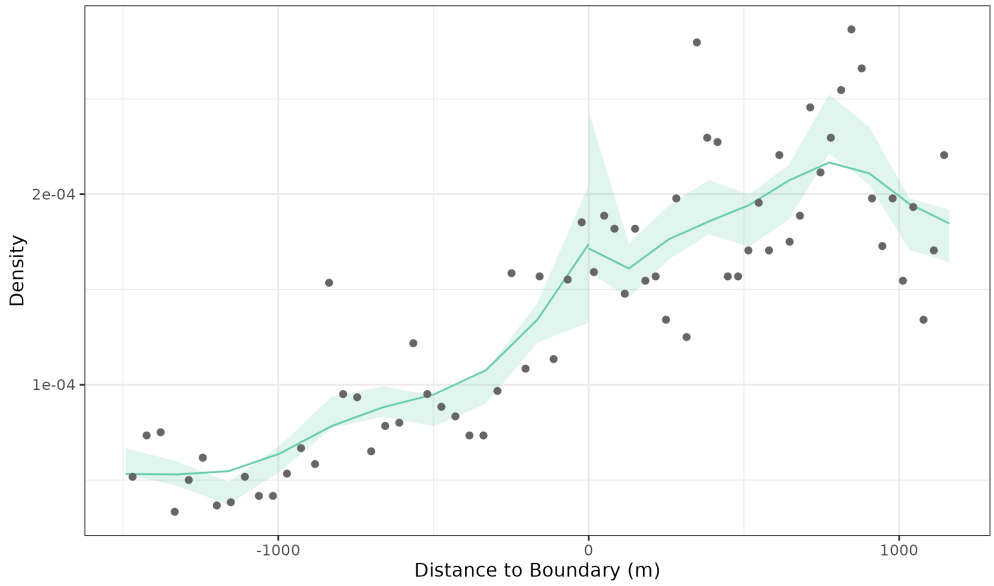
Notes: The figure maps shooting victimizations per 100,000 residents by Chicago Police Department district in 2016. The blue boundary denotes the area subsequently covered by ShotSpotter. Population estimates are based on 2013–2017 ACS data.

Figure A.5: Placebo Regression Discontinuity Estimates in the *Pre*-ShotSpotter Period



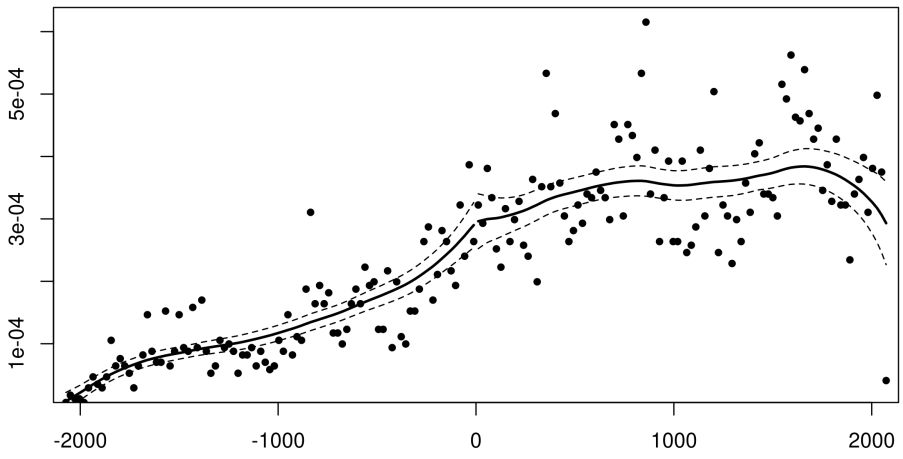
Notes: Optimal bandwidth local linear regressions in teal (coefficient = 0.015, robust p-value = 0.634), local linear smoother fitted on each side of the discontinuity in dashed gray. Bins created using mimicking variance evenly-spaced method (Calonico et al., 2015).

Figure A.6: Running Variable Density Test



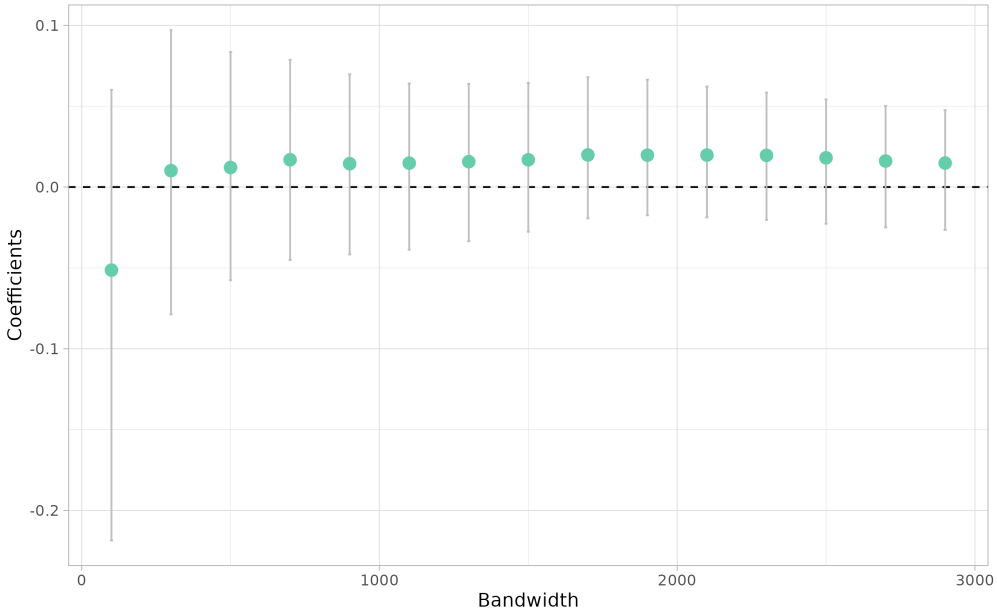
Notes: p-value = 0.253. Data includes only observations within two kilometers of the district boundary.

Figure A.7: McCrary (2008) Density Plot



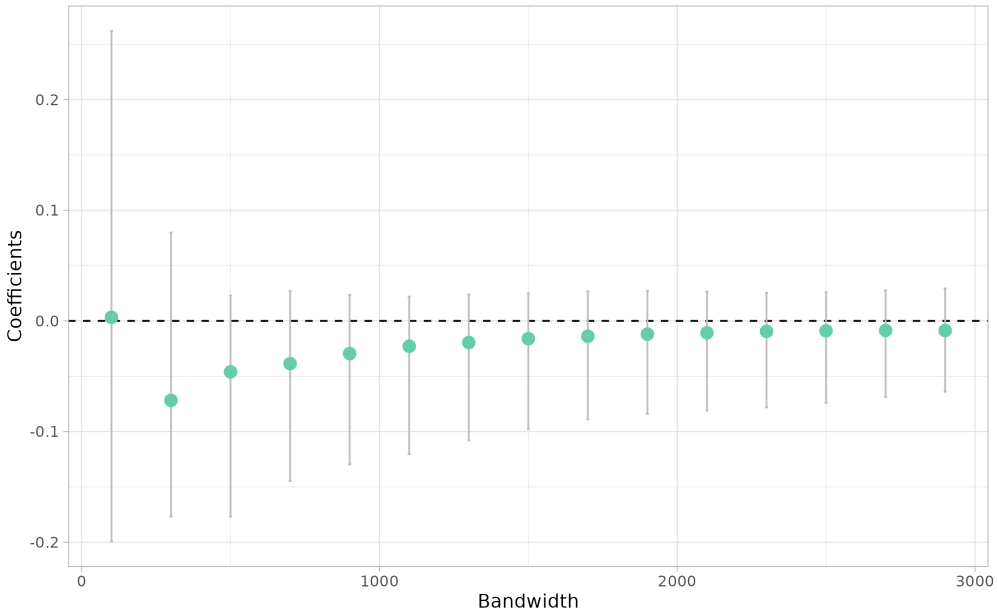
Notes: p-value = 0.983. Data includes only observations within two kilometers of the district boundary.

Figure A.8: Pre-Treatment Placebo Test, by Bandwidth (100 - 2900m, in Increments of 200)



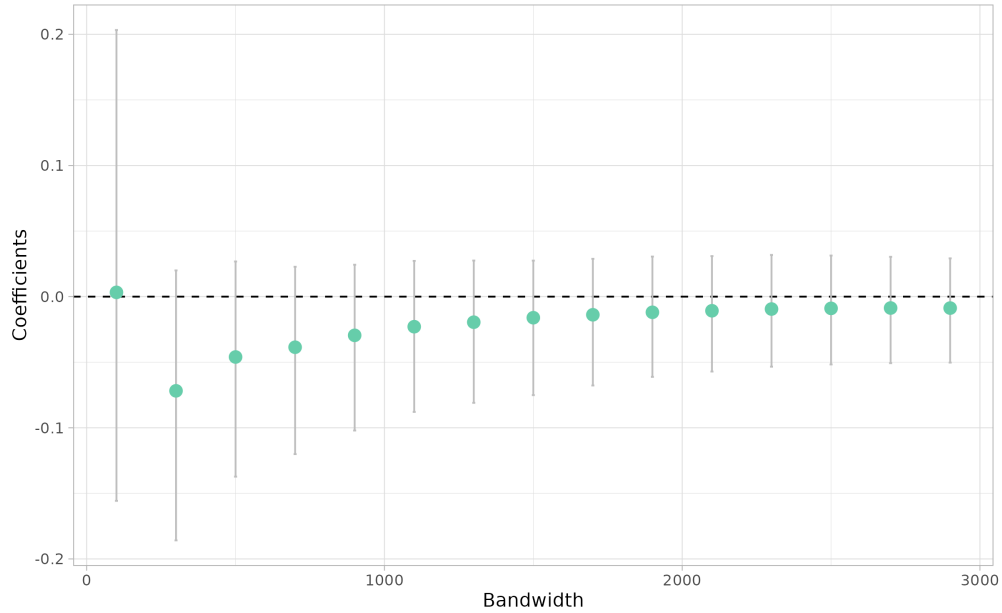
Notes: 95% robust confidence intervals in gray. Estimates are from the pre-ShotSpotter period and vary the bandwidth from 100 to 2,900 meters in increments of 200 meters. The ratio of the h to b bandwidths is held fixed to make estimates comparable to the baseline optimal-bandwidth results.

Figure A.9: Sharp RD Estimates, by 1:1 h/b Ratio Bandwidths (100 - 2900m, in Increments of 200)



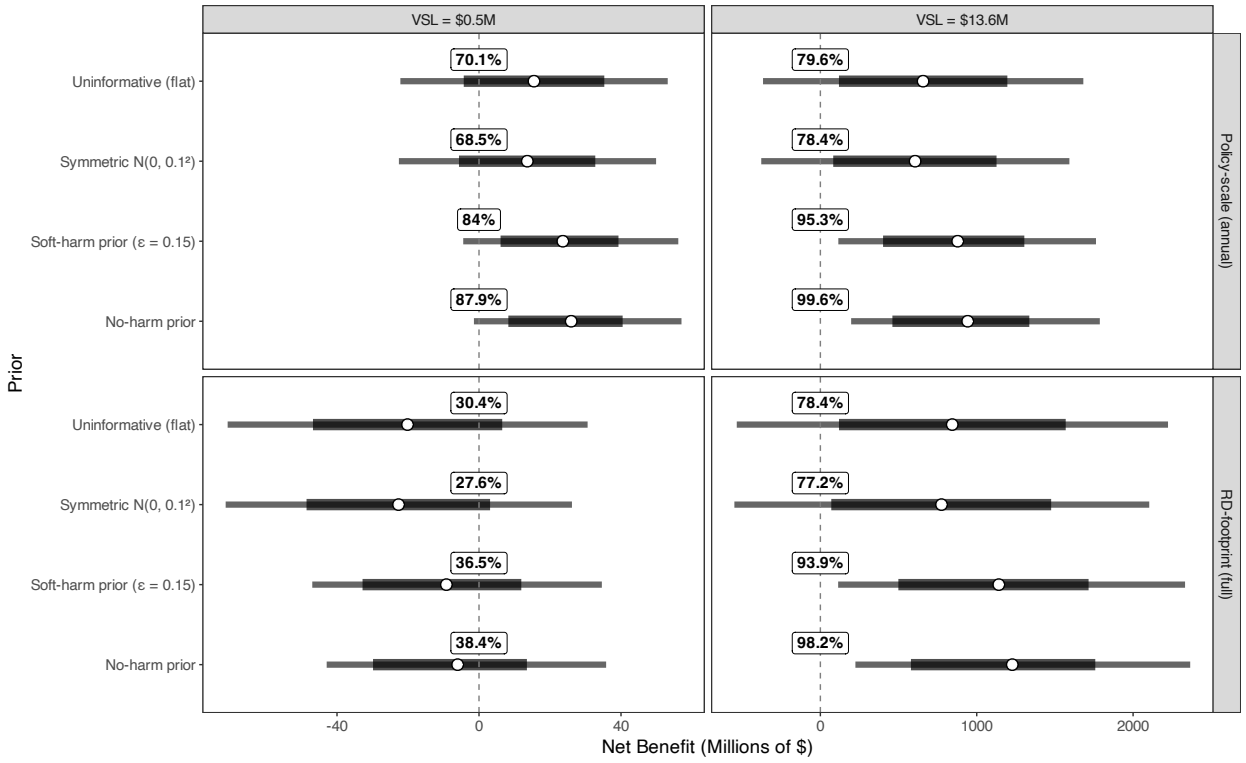
Notes: 95% robust confidence intervals in gray. Estimates vary the bandwidth from 100 to 2,900 meters in increments of 200 meters, setting the h and b bandwidths equal at each value. Appendix Figure A.10 presents equivalent estimates holding fixed the baseline ratio of the h to b bandwidths.

Figure A.10: Sharp RD Estimates, by Bandwidth (100 - 2900m, in Increments of 200)



Notes: 95% robust confidence intervals in gray. Estimates hold fixed a ratio of the h to b bandwidths to make estimates comparable to the baseline optimal bandwidth results. Appendix Figure A.9 presents equivalent estimates without this fixed ratio and qualitatively identical statistical significance.

Figure A.11: Posterior Net Benefits Across Priors, VSL Assumptions, and Exposure Definitions



Notes: Boxed labels report the posterior probability that program benefits exceed costs, $P(B > C)$. White circles report posterior mean annual net benefits. Lighter horizontal segments span the 10th–90th percentiles and darker segments span the interquartile range. Columns vary the value of a statistical life. Rows compare the policy-scale annual exposure definition with the full RD-footprint exposure definition. Results are reported for the uninformative flat, symmetric Normal, soft-harm, and no-harm priors.

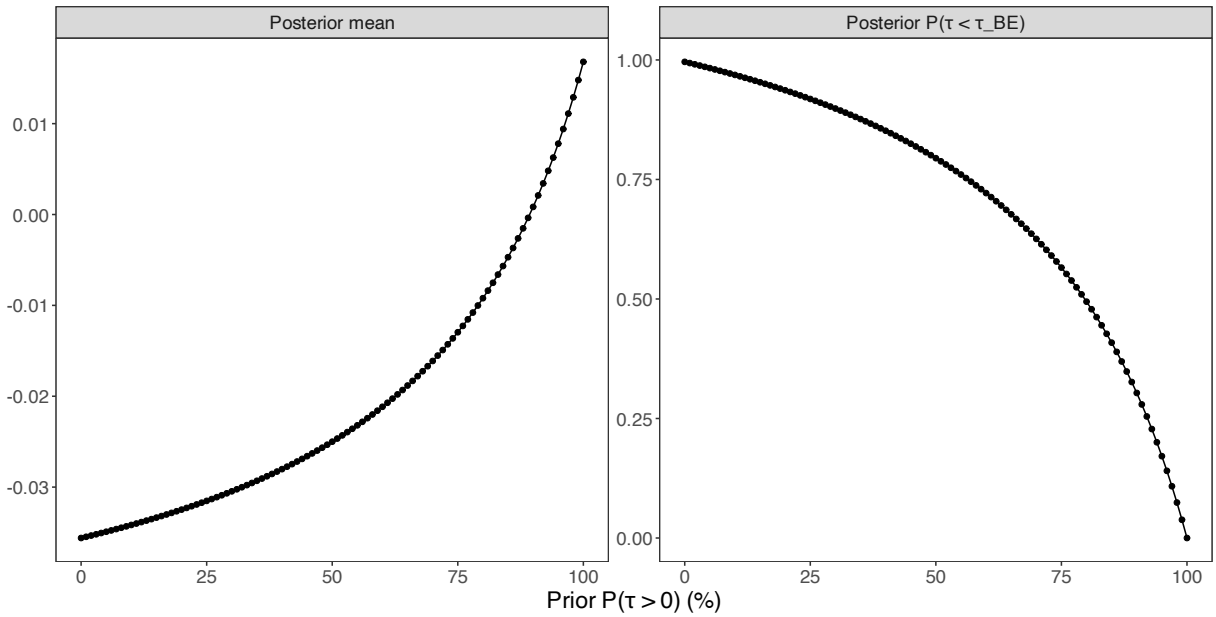


Figure A.12: Sensitivity to Prior Mass on Harm (ε-priors)

Notes: The left panel plots the posterior mean effect of ShotSpotter on the shooting fatality rate as the assumed prior probability of harm (ε) increases from 0 to 1. The right panel shows the corresponding posterior probability that the program passes the benefit-cost (break-even) threshold.

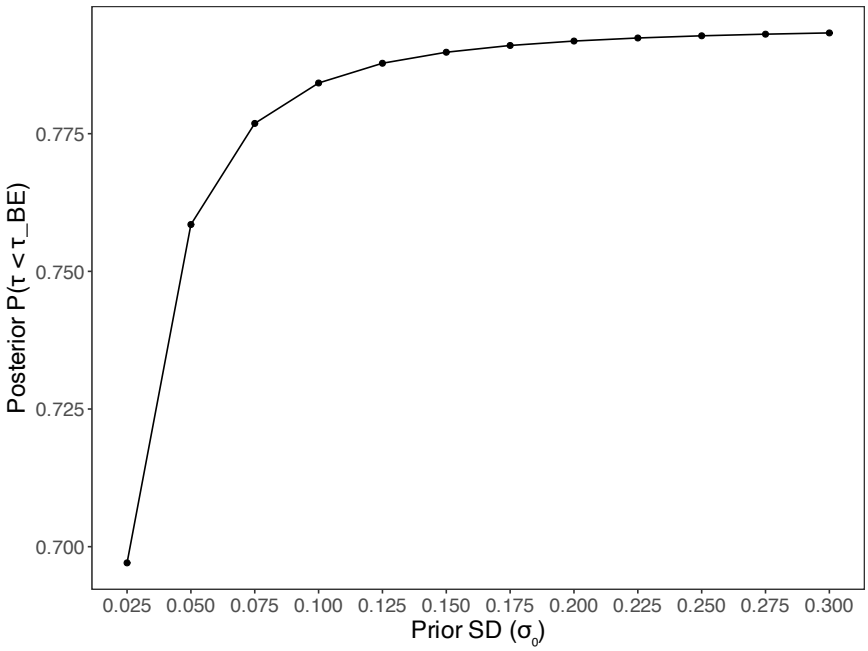


Figure A.13: Sensitivity to Symmetric Prior Dispersion (σ_0), $N_{\text{exposed}} = 1966$

Notes: The figure plots the posterior probability that ShotSpotter passes the benefit-cost (break-even) threshold as the dispersion of the symmetric prior (σ_0) varies, holding $N_{\text{exposed}} = 1966$. Tighter priors (smaller σ_0) shrink the posterior toward zero, reducing the probability of cost-effectiveness, while highly diffuse priors approach the likelihood-only case, converging to the uninformative prior benchmark.

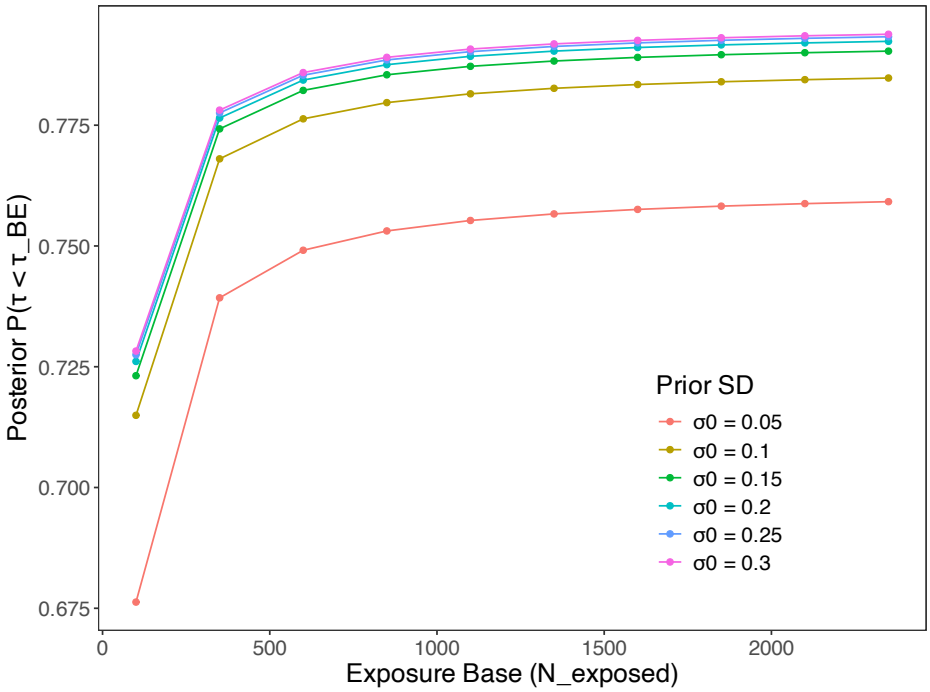


Figure A.14: Posterior Pass Probability vs Exposure Base under Symmetric Prior

Notes: The figure plots the posterior probability that ShotSpotter passes the benefit-cost (break-even) threshold by exposure base under the symmetric prior, shown for several prior dispersions (σ_0). Larger exposure bases lower the break-even effect and increase the posterior probability of cost-effectiveness.

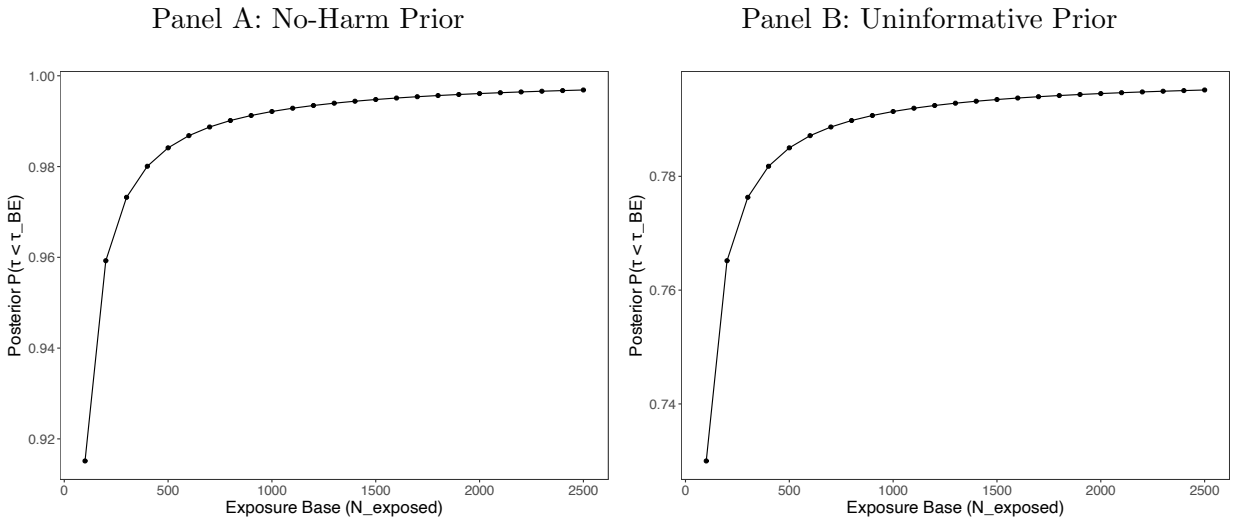


Figure A.15: Probability of Passing Break-Even Threshold by Exposure Base

Notes: The left panel plots the posterior probability that ShotSpotter passes the benefit-cost (break-even) threshold by exposure base under the no-harm prior. The right panel shows the same relationship under an uninformative (flat) prior, where the posterior coincides with the likelihood implied by the RD estimate.

B Appendix Tables

Table A1: Bayesian Methods and Decision Analysis in Recent AEJ Articles

	Number	% of Bayesian indexed hits	% of all articles
Panel A. Search and manual review			
All articles, 2020–2025	760	–	100.0%
Articles identified by "Bayesian" search	43	100.0%	5.7%
Articles with substantive Bayesian use after manual review	31	72.1%	4.1%
Panel B. Comparison keyword searches			
Articles identified by "null" search	316	–	41.6%
Articles identified by "hypothesis" search	450	–	59.2%
Articles identified by "significant" search	662	–	87.1%
Panel C. Primary classification of substantive Bayesian uses			
Bayesian learning, beliefs, or persuasion	13	30.2%	1.7%
Bayesian equilibrium or theory	4	9.3%	0.5%
Empirical Bayes or shrinkage	6	14.0%	0.8%
Bayesian statistical or structural estimation	4	9.3%	0.5%
Bayesian meta-analysis or predictive synthesis	2	4.7%	0.3%
Mixed empirical Bayes and Bayesian decision model	1	2.3%	0.1%
Bayesian policy optimization under uncertainty	1	2.3%	0.1%
Panel D. Closest decision-relevant uses			
Policy-relevant predictive distribution over effects	1	2.3%	0.1%
Bayesian structural estimation followed by policy counterfactuals	1	2.3%	0.1%
Explicit policy optimization under Bayesian uncertainty	1	2.3%	0.1%
Posterior program-impact uncertainty mapped into explicit adoption criterion	0	0.0%	0.0%

Notes: The sample consists of all 760 articles published in the *American Economic Journal: Applied Economics* and *American Economic Journal: Economic Policy* from 2020 through 2025. Google Scholar identified 43 unique articles containing the term "Bayesian." Manual review identified 31 articles with substantive Bayesian use. The remaining 12 articles were excluded because the term appeared only in references (6), only as the Bayesian information criterion (3), or only in an incidental footnote (1), or did not appear in the version reviewed (2). Panel B reports unvalidated full-document keyword searches and is included only to provide descriptive context; occurrences of "null," "hypothesis," and "significant" were not manually reviewed. Panel C assigns each of the 31 articles with substantive Bayesian use to a mutually exclusive category based on its primary use of Bayesian methods or reasoning. Panel D identifies selected decision-relevant uses and is not mutually exclusive with Panel C. An explicit adoption criterion requires uncertainty about an estimated program impact to be mapped into a policy payoff or threshold, such as expected net benefits or the probability that benefits exceed costs.

Table A2: Naive Comparison of ShotSpotter on Fatality Rates, All Districts

	Control Mean	Treatment-Control Diff.	p-value	N
Fatality Rate	0.160	0.017	0.021	19,198

Notes: Robust standard errors. Each observation is a shooting victimization.

Table A3: Continuity Tests of Victim/Incident Covariates, Pre-Treatment

Covariate	Control Mean	Treatment-Control Diff.	p-value	Observations
Age 0-19	0.259	-0.030	0.532	3511
Age 20-29	0.485	-0.009	0.799	3511
Age 30-39	0.176	0.047	0.171	3511
Age 40-49	0.057	-0.013	0.395	3511
Age 50-59	0.015	-0.002	0.735	3511
Age 60-69	0.007	0.006	0.215	3511
Age 70-79	0.001	0.002	0.184	3511
Age 80+	0.000	0.000	0.198	3511
Age (Missing)	0.022	-0.020	0.027	3599
API/Other NH	0.004	0.003	0.319	3484
Black NH	0.584	0.086	0.207	3484
Hispanic/Latino	0.374	-0.071	0.326	3484
White NH	0.037	-0.018	0.287	3484
Race (Missing)	0.029	-0.008	0.351	3599
Male	0.901	-0.000	0.884	3505
Sex (Missing)	0.023	-0.014	0.133	3599
Jan	0.059	-0.040	0.014	3599
Feb	0.031	-0.005	0.715	3599
Mar	0.076	-0.045	0.008	3599
Apr	0.082	-0.007	0.707	3599
May	0.088	0.085	0.000	3599
Jun	0.107	0.005	0.596	3599
Jul	0.113	-0.019	0.192	3599
Aug	0.114	0.047	0.002	3599
Sep	0.093	0.015	0.332	3599
Oct	0.093	0.001	0.775	3599
Nov	0.066	-0.016	0.216	3599
Dec	0.079	-0.020	0.411	3599
Monday - Friday	0.566	-0.025	0.220	3599
CPD Watch #1 11:00pm-7:00am	0.395	0.051	0.202	3599
CPD Watch #2 7:00am-3:00pm	0.142	0.013	0.845	3599
CPD Watch #3 3:00pm-11:00pm	0.463	-0.064	0.168	3599

Notes: Each test carried out using the baseline specification outlined in equation (1), using a local linear regression, optimal bandwidths from the main RD, bandwidth selection robust inference, and a triangular kernel. Observation totals are comprised of treatment and control, within bandwidth.

Table A4: Continuity Tests of Victim/Incident Covariates, During-Treatment

Covariate	Control Mean	Treatment-Control Diff.	p-value	Observations
Age 0-19	0.217	-0.045	0.131	3957
Age 20-29	0.402	0.048	0.160	3957
Age 30-39	0.221	0.020	0.774	3957
Age 40-49	0.110	-0.019	0.461	3957
Age 50-59	0.034	0.010	0.405	3957
Age 60-69	0.016	-0.016	0.047	3957
Age 70-79	0.000	-0.000	0.250	3957
Age 80+	0.000	0.002	0.345	3957
Age (Missing)	0.015	0.010	0.126	4021
API/Other NH	0.003	0.008	0.068	3938
Black NH	0.659	0.082	0.019	3938
Hispanic/Latino	0.299	-0.083	0.013	3938
White NH	0.039	-0.007	0.542	3938
Race (Missing)	0.017	0.013	0.105	4021
Male	0.881	0.019	0.507	3969
Sex (Missing)	0.012	0.006	0.302	4021
Jan	0.063	-0.017	0.441	4021
Feb	0.046	-0.018	0.149	4021
Mar	0.041	0.023	0.143	4021
Apr	0.060	0.010	0.507	4021
May	0.108	0.012	0.334	4021
Jun	0.118	0.054	0.034	4021
Jul	0.126	-0.014	0.496	4021
Aug	0.107	0.035	0.183	4021
Sep	0.117	-0.007	0.915	4021
Oct	0.063	0.011	0.637	4021
Nov	0.074	-0.054	0.001	4021
Dec	0.076	-0.035	0.033	4021
Monday - Friday	0.598	-0.058	0.042	4021
CPD Watch #1	0.413	0.036	0.301	4021
CPD Watch #2	0.151	-0.029	0.127	4021
CPD Watch #3	0.437	-0.008	0.825	4021

Note: Each test carried out using the baseline specification outlined in equation (1), using a local linear regression, optimal bandwidths from the main RD, bandwidth selection robust inference, and a triangular kernel. Observation totals are comprised of treatment and control, within bandwidth.

Table A5: Sharp Regression Discontinuity Estimates, by Polynomial and Kernel

	Outcome: Fatal Shooting		
In ShotSpotter Area	-0.018 [-0.083, 0.03] (0.353)	-0.036 [-0.109, 0.028] (0.246)	-0.034 [-0.106, 0.025] (0.228)
Optimal Bandwidth	883.73	1521.323	1369.125
Polynomial Order	1	2	2
Kernel	Uniform	Triangular	Uniform
Observations	3551	5568	5024

Notes: Bandwidth selection robust confidence interval in brackets, p-value in parentheses.

Table A6: Conceptual Frameworks and Decision Rules for Policy Evaluation

Approach	**Conceptual Framework**	**Decision Rule**	**Implication for ShotSpotter**
Two-sided hypothesis test	There exists a true effect τ to be discovered	Reject $H_0 : \tau = 0$ at fixed α	Fails to reject; treated as "no effect," regardless of welfare implications
One-sided hypothesis test	True effect compared to a threshold τ_{BE}	Reject $H_0 : \tau \geq \tau_{BE}$ at fixed α	Would require unrealistically precise estimates; changing α does not incorporate magnitude of benefits
Bayesian decision analysis	Uncertainty represented by a posterior distribution	Adopt if expected net benefit is positive (or $P(B > C)$ exceeds a policy-relevant threshold)	Supports adoption under a wide range of priors by directly evaluating welfare outcomes
Robust control (Hansen-Sargent)	Posterior treated as an approximating model subject to misspecification	Evaluate adoption under worst-case local perturbations of the posterior	Identifies when Bayesian conclusions are sensitive to model credibility and when they are robust

Table A7: Break-even robustness thresholds (θ^*) under alternative priors and VSL assumptions

Prior	θ^* (VSL = \$13.6M)	θ^* (VSL = \$0.5M)
No-harm prior	\$1.2 million	\$2.7 million
Soft-harm prior ($\varepsilon = 0.15$)	\$312 million	\$14.1 million
Uninformative (flat) prior	\$485 million	\$27.9 million
Symmetric Normal prior	\$492 million	\$29.5 million

Notes: θ^* denotes the robustness level at which robust net benefits equal zero. Larger values indicate greater sensitivity of the policy conclusion to pessimistic reinterpretations of the posterior distribution. Robust net benefits are computed using the closed-form expression found in Section B, applied to posterior draws of τ. The break-even threshold θ^* is obtained numerically as the unique solution to $NB_{\text{rob}}(\theta^*) = 0$.

Table A8: Interpretation of break-even robustness thresholds: tail-weight concentration

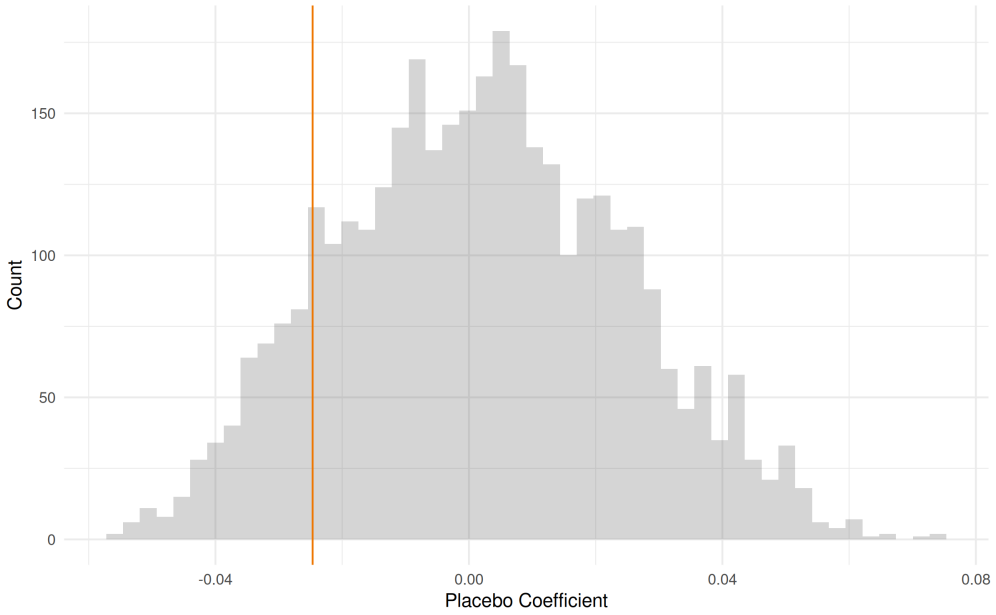
Prior	VSL	θ^* (millions)	Weight on worst 1%	Weight on worst 5%
No-harm prior	\$13.6M	1.21	1.000	1.000
Soft-harm prior ($\varepsilon = 0.15$)	\$13.6M	312.28	0.733	0.843
Uninformative (flat) prior	\$13.6M	485.05	0.245	0.501
Symmetric Normal prior	\$13.6M	492.23	0.228	0.471
No-harm prior	\$0.5M	2.67	0.250	0.750
Soft-harm prior ($\varepsilon = 0.15$)	\$0.5M	14.11	0.478	0.640
Uninformative (flat) prior	\$0.5M	27.89	0.102	0.278
Symmetric Normal prior	\$0.5M	29.54	0.086	0.247

Notes: For each prior and VSL, we report the break-even robustness threshold θ^* along with the implied concentration of decision weight on the lower tail of the posterior distribution. For a given θ, robust control assigns normalized weights to posterior draws proportional to $\exp(-NB/\theta)$. The reported shares indicate how much of the total decision weight is placed on the worst 1% and 5% of net-benefit realizations at θ^*, where "worst" is defined by ranking draws from lowest to highest NB.

C Appendix: Geographic Placebo Test

To provide an additional robustness check on the statistical inference of our baseline geographic RD estimate, we carry out a placebo test in geography (Figure A.16) by estimating regression discontinuity models that follow our main specification at "pseudo-cutoffs" where no effect is expected. We find generally similar p-values for this test, assuaging concerns that our findings thus far are due to finite samples or other incorrect inferential assumptions. As the main identification strategy relies on ShotSpotter being discontinuously available on one side of a police district boundary versus another, we create placebo boundaries by shifting the existing boundary by a large number of randomly drawn constant values.[14] Under the placebo null distribution, our baseline estimate results in a p-value of 0.3, similar to that found in our baseline regression.

Figure A.16: Placebo in Geography



Note: Baseline post-treatment RD estimate in orange.

[14] We constrain the placebo cutoff generation process to only keep draws that would allow enough of a sample size to estimate an effect (e.g., if a placebo cutoff would yield an in-bandwidth sample that has very few placebo control or placebo treatment observations, we do not count it towards the placebo null distribution). We generate 10,000 placebo cutoffs, then limit our placebo sample to those that would have at least 2,000 observations within a bandwidth equivalent to that of our baseline RD estimate, resulting in 3,558 valid placebo draws. Given that the choice of 2,000 is arbitrary, we assess the change in p-values to cutoffs other than 2,000, and find that between 1,000 and 5,000 observations, we would obtain p-values between 0.265 and 0.385.

D Appendix: Shooting to ShotSpotter (SST) Alert Matching

Due to the absence of unique identifiers between shooting victimization records and SST alert data, we perform spatial-temporal linking by identifying the SST alerts closest to shooting victimizations within defined time and distance parameters.

Spatial matches were conducted using within-distance thresholds between 200 and 450 meters to account for known data adjustments and SST sensor error. Thus, if an alert signaled that the gunfire originated within that radius of a shooting victimization, it would qualify as a matched pair. To determine these thresholds, we considered known data adjustment and sensor errors. According to SoundThinking, the owners of SST, the distance between the crime scene location and the location of where SST estimates the gunfire came from is typically within 25 meters (~82 feet) (SoundThinking, 2024). In addition, geospatial coordinates are deliberately jittered for both shooting and SST alert data by roughly the size of an average city block, with a new location randomly selected within the circumference of this circle (City of Chicago, 2024b,a). As a result, the locations of shootings and alerts may deviate by as much as two city blocks total (one block for each set of coordinates jittered), plus additional error from the alert data. Given that the average city block in Chicago is about 660 feet (~200 meters), we adjusted the matching distance parameters to range from within 200 meters to within 450 meters to account for this jittering and alert error. This theoretically captures the majority of relevant matches.

According to SoundThinking documentation (SoundThinking, 2024), the average time between a shooting and an alert being dispatched is approximately one minute. To balance capturing relevant matches while minimizing the risk of linking alerts to shootings at significantly different times, we adopted a less restrictive approach by exploring a time window of ±5 to 20 minutes. These buffers assume the SST sensor-generated timestamps are more precise while accommodating potential inaccuracies in shooting timestamps. In practice, a narrow time difference within a reasonable geographic buffer is likely to provide more meaningful information than spatial distance alone, particularly given the previously noted coordinate jittering, and the fact that shootings in Chicago are highly geographically concentrated (Lopez and Ballard, 2024).

Although shooting incidents can generate multiple SST alerts, identifying victimizations that took up treatment only requires one reasonably matched alert. While the take up indicator would not change, we can select the alert that has the highest ranking using two methods: Time-Space Normalized, which scales time and distance values between 0 and 1, and Time-Space Normalized & Weighted, which does the aforementioned scaling exercise and assigns a specified weight to time and distance (i.e. 70% weight to the time difference and the remaining 30% to the spatial distance). The normalized ranking helps identify the single best match for each shooting incident and a deduplication step ensures no unique SST alert matches more than one shooting.

To evaluate how match rates vary with different constraints, we tested multiple combinations of these parameters using the following criteria.

- The observation period is dynamic to begin on the SST district-level implementation dates as defined by Piza et al. (2024) and the cutoff for all districts is the day after SST stopped logging alerts following the Mayor's decision to end the contract. These dates are outlined in

Appendix Table A11.

- Although it is possible for an SST alert to estimate the origin of gunfire in a location of a non-SST district, the match rates only review alerts that matched to shooting cases that occurred in SST districts.
- Similarly, though it is possible for an alert to be matched to an indoor shooting incident, those cases are removed from the match rates analysis.

The overall match rates at the parameter-level (distance & time) are enumerated in the table below. The rates range from 51.3% to 56.9%, with an average rate of 54.6%. These results suggest that adjusting the match function parameters does not lead to a substantial variation in match quality across the treated districts. This outcome further implies that for every two outdoor or exterior shootings in an SST district, only one has a matched alert.

The average match rates across all parameters at the district-level are outlined in Table A10 below. These results indicate that there is variability on matching at the district-level, ranging from 47.6% in District 15 to 61% in District 5, a difference of 13.4 percentage points. An inspection of the average district-level match rates broken out by distance and time indicate that match rates within districts are also not largely impacted by toggling the distance and time window parameters as shown in the previous table.

An analysis using the parameters within 450 meter distance with a $\pm$15 minute time window provides descriptive evidence suggesting matching rate trends evolved over time, with varying levels and year-over-year directions observed at the district level (see Figure A.17 below). Generally, most districts have experienced a gradual increase in match rates since their activation, indicating an improvement in the likelihood of linking an alert to a shooting incident within a specified time and spatial range. However, total outdoor shooting incidents within districts do not consistently align with these match rate trends. For example, in District 11, outdoor shooting incidents decreased beginning in 2021, even as match rates showed a steady increase during the same period. This divergence highlights potential differences in trends between shooting incidents and SST alert accuracy or detection rates.

Table A9: Match Rates by Distance and Time Thresholds

Shooting to Alert Distance (Meters)	Shooting to Alert Time (Mins)	Matched Shooting Cases	Total Shooting Cases	Match Rate
200	5	7246	14130	0.513
200	10	7455	14130	0.528
200	15	7554	14130	0.535
200	20	7602	14130	0.538
250	5	7401	14130	0.524
250	10	7614	14130	0.539
250	15	7717	14130	0.546
250	20	7770	14130	0.550
300	5	7488	14130	0.530
300	10	7705	14130	0.545
300	15	7816	14130	0.553
300	20	7869	14130	0.557
350	5	7545	14130	0.534
350	10	7766	14130	0.550
350	15	7882	14130	0.558
350	20	7938	14130	0.562
400	5	7584	14130	0.537
400	10	7808	14130	0.553
400	15	7925	14130	0.561
400	20	7984	14130	0.565
450	5	7629	14130	0.540
450	10	7855	14130	0.556
450	15	7977	14130	0.565
450	20	8042	14130	0.569

Table A10: Rollout Schedule and Match Rates by District

Rollout Date	District	Average Match Rate
2017-02-06	7	0.558
2017-02-06	11	0.489
2017-04-24	15	0.476
2017-07-14	9	0.550
2017-09-26	6	0.580
2017-10-30	10	0.576
2017-12-31	3	0.552
2018-01-31	4	0.600
2018-03-07	5	0.610
2018-04-11	8	0.501
2018-04-11	25	0.565
2018-05-16	2	0.552

Figure A.17: Annual Shooting Cases & SST-Alerts-to-Shooting-cases Match Rates



Note: Numerator and denominator records are filtered to SST Districts & Outdoor/Exterior cases only.

E Appendix: ShotSpotter Implementation Dates

There are three main sources of information on ShotSpotter installation dates: Piza et al. (2024), Topper and Ferrazares (2024), and the ShotSpotter alert data from the City of Chicago data portal (City of Chicago, 2024a).

According to Piza et al. (2024, p. 25), there was a pilot phase of ShotSpotter in micro geographies within District 11 and 7 between 9/1/2012 and 2/5/2017. During this pilot phase, the system did not retain alert data. The remaining implementation across districts went through May 2018. It is unknown where this source of installation dates originated.

In Topper and Ferrazares (2024, p. 44), the rollout dates used span from 2017 - 2018. The authors credit a combination of the Chicago Police Department as well as Max Kapustin and Ashna Arora for assisting with details regarding rollout dates in Chicago.

The ShotSpotter alert data from the City of Chicago comes with alert dates and District labels. As such, it was possible to discern the earliest alert per district and to detect when alerts start to exceed an average weekly threshold of alerts. The latter analysis begins by filtering alert data to include only alerts occurring within ShotSpotter districts. To minimize driving up the average with potential false positives, alerts from New Year's weekends across all years are excluded, as alert counts typically spike significantly around this holiday. Weekly alert counts are then calculated at the district level for each week from the start of 2017 through present, with any null observations filled with a count of zero. Then, for each district, the average weekly counts are computed on an annual basis which serves as the average alert threshold. The data are then structured to indicate whether a week's alert counts meet or exceed the average alert threshold for each district-year. For districts that started using ShotSpotter in 2017, based on the Piza et al. (2024) dates, the 2018 average alert count serves as the threshold; for those implemented in 2018, the 2019 average is used. Finally, the earliest date at which the alert count meets or surpasses the threshold in each district is designated as the "rollout date" supplied in Table A11.

Table A11: Comparison of ShotSpotter Implementation Dates Across Sources

District	Piza et al. (2024) [1]	Topper & Ferrazares (2024) [2]	First Alert (City of Chicago) [3]	Weekly Alerts Exceed Average (City of Chicago) [4]	Days Between [1] - [2]	Days Between [1] - [3]	Days Between [1] - [4]
7	2017-02-06	2017-01-13	2017-01-13	2017-03-26	24	24	-48
11	2017-02-06	2017-03-01	2017-01-19	2017-04-16	-23	18	-69
15	2017-04-24	2017-05-13	2017-03-16	2017-05-14	-19	39	-20
9	2017-07-14	2017-06-01	2017-02-14	2017-07-09	43	150	5
6	2017-09-26	2017-09-24	2017-08-07	2018-02-11	2	50	-138
10	2017-10-30	2017-10-16	2017-03-19	2017-11-26	14	225	-27
3	2017-12-31	2018-01-04	2017-06-04	2018-02-04	-4	210	-35
4	2018-01-31	2018-02-01	2018-01-29	2018-01-28	-1	2	3
5	2018-03-07	2018-03-07	2018-03-13	2018-04-29	0	-6	-53
8	2018-04-11	2018-04-01	2017-01-21	2018-05-13	10	445	-32
25	2018-04-11	2018-04-24	2017-01-25	2018-04-29	-13	441	-18
2	2018-05-16	2018-05-16	2018-02-03	2018-05-20	0	102	-4

To identify the set of rollout dates to use for this analysis, we first analyzed the average absolute differences between available sources. The average discrepancy between Piza et al. and Topper & Ferrazares was 12.8 days, suggesting a high degree of consistency between these sources. In contrast, the alert data from the City of Chicago showed a larger average difference. For instance, comparing dates from Piza et al. with the initial alert date per district from the city data revealed an average difference of 142.7 days (over 4.5 months)—in all but one district, the first alert date was always earlier than the corresponding Piza et al. date. To assess the implications of this discrepancy, we calculated the number of alerts observed during this period, finding an additional 574 alerts (ranging from 1 to 196 in each district). Given that in 2019 there was an average of 1,586.25 citywide alerts per month (over 7,000 in 4.5 months), this relatively low alert volume recovered within the 4.5-month timeframe suggests ShotSpotter was not yet in full operation. Therefore, we concluded that the first alert date is not representative of a fully functional ShotSpotter deployment and excluded it as an implementation date.

Additionally, we compared the Piza et al. dates with the "alerts-exceed-average" method, which identified rollout dates based on when weekly alerts began to exceed the district's average. The average difference between these two approaches was 37.7 days (or 28.5 days when excluding an outlier in District 6, where the difference was 138 days). In this comparison, the alerts-exceed-average dates consistently followed the Piza et al. dates in all but two districts, suggesting we would be using a later set of rollout dates in general. Across these discrepancies, the alerts-exceed-average dates captured 827 alerts within roughly a month of delay—still lower than the monthly average of 1,586.25 alerts in 2019 but closer to an operational volume.

Ultimately, given the findings discussed, we opted to use the Topper & Ferrazares dates for the present study.