HARVARD | BUSINESS | SCHOOL

Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA

Matthew S. Johnson
David I. Levine
Michael W. Toffel

Working Paper 20-019

Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA

Matthew S. Johnson
Duke University

David I. Levine
University of California

Michael W. Toffel
Harvard Business School

Working Paper 20-019

Funding for this research was provided in part by Harvard Business School, the Laura and John Arnold Foundation, and the Department of Labor DOL Scholars Program.

Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA

Matthew S. Johnson
Sanford School of Public Policy, Duke University

David I. Levine
Haas School of Business, University of California

Michael W. Toffel
Harvard Business School

February 26, 2020

We study how a regulator can best target inspections. Our case study is a US Occupational Safety and Health Administration (OSHA) program that randomly allocated some inspections. On average, each inspection averted 2.4 serious injuries (9%) over the next five years. We use new machine learning methods to estimate the effects of counterfactual targeting rules. OSHA could have averted over twice as many injuries by targeting the highest expected averted injuries and nearly as many by targeting the highest expected level of injuries. Either approach would have generated over $1 billion in social value over the decade we examine.

JEL Classifications: I18; L51; J38; J8

Acknowledgements: We are grateful for guidance from Dave Schmidt, Amee Bhatt, and Ricky Gonzalez on institutional details about the US Occupational Safety and Health Administration (OSHA), research assistance from Melissa Ouellet, research methods advice from Xiang Ao and Andrew Marder, and TMLE advice from Mark van der Laan and Cheng Ju. We received helpful comments from Dave Anderson, Jon Baron, Jon Davis, Ivan Fernandez-Val, Eric Frumin, Kevin Lang, Jim Rebitzer, Seth Sanders, and OSHA advisory board members Robin Baker and Lisa Brousseau. We benefited from comments by participants in the Harvard Labor Economics seminar, the Harvard Regulatory Policy Program seminar, RAND Santa Monica, the Duke Sanford School of Public Policy APPAM annual conference, and a presentation at OSHA. We gratefully acknowledge financial support from the Laura and John Arnold Foundation, the Harvard Business School Division of Research, and the Department of Labor DOL Scholars Program. Our pre-analysis plan is at https://osf.io/2snka/. Corresponding author email: matthew.johnson@duke.edu.

Government agencies spend billions each year inspecting establishments for worker safety, environmental protection, consumer protection, tax compliance, and other concerns (Shimshack 2014; US Food and Drug Administration 2016: 9; US OSHA 2017a). Most regulatory agencies' budgets only allow them to inspect a tiny share of the establishments they regulate. For example, workplace safety regulators in the United States inspected less than 1% of the 8 million workplaces they regulated in 2016 (US OSHA 2017a). The Internal Revenue Service, US Environmental Protection Agency, US Food and Drug Administration, and their global counterparts face similar budget constraints (Rubin 2017; US Department of Health and Human Services 2011). Agencies must therefore make difficult choices about how to target inspections, often relying on a combination of laws and heuristics. For example, worker safety regulators allocate many of their inspections to facilities that have recently experienced serious accidents, many injuries, or were the subject of employee complaints (US OSHA 2017b).

It is difficult to evaluate regulators' average historical performance because purposeful targeting makes it hard to find a valid comparison group for inspected workplaces. It is even more difficult to assess whether regulators could allocate their inspections *more* effectively, as the extent to which inspections would achieve their objectives (e.g., averting injuries, food poisoning, toxic emissions)—their *treatment effects*—is typically unobservable. Agencies are therefore vulnerable to critiques that they waste taxpayer dollars, target establishments to promote politicians' agendas (e.g., Weisman and Wald 2013), or serve the interests of those they regulate (Stigler 1971).

Fortunately, advances in machine learning methods to estimate heterogeneous treatment effects provide a new opportunity to assess how well inspections deliver on their objectives. While these methods can in principle enable agencies to identify how to optimally allocate their inspection resources, they can be challenging to implement in practice, as they require an experimental design to produce unbiased estimates of treatment effects and an enormous amount of data to produce reasonably precise estimates.

Regulators could also use machine learning to redirect inspection targeting by predicting more accurately where problems – whether injuries, emissions, violations, or so on – are most likely to occur (Glaeser et al. 2016; Hino, Benami, and Brooks 2018). This is a much simpler prediction problem than estimating heterogeneous treatment effects and does not require an

experimental design. However, it hinges on an implicit assumption that treatment effects of inspections (such as injuries averted due to inspections) are largest at workplaces that are predicted to have the most problems, which is not necessarily the case in a general setting (Athey 2017).

We develop an approach to assess the extent to which agencies are maximizing the effectiveness of their inspections. We combine randomization and machine learning to compare the effects of inspections as historically allocated to those of alternative targeting policies. We leverage randomization both to evaluate the regulator's historical performance and to estimate heterogeneous treatment effects via machine learning.

We apply this approach to OSHA, the US regulatory agency charged with assuring "safe and healthful working conditions."[1] For several reasons, OSHA is an important setting in which to examine regulatory effectiveness. First, US workplace injuries and illnesses impose a substantial burden on the economy, estimated at $250 billion in annual social costs (Leigh 2011). Second, OSHA has been a controversial agency ever since it was created in 1970. Supporters argue that it saves lives at little to no cost to employers (Feldman 2011), while critics charge that its regulations add costs but that some "don't add value to safety in the workforce"[2] or that its penalties are too low to affect behavior (Bartel and Thomas 1985). Resolving this debate on OSHA's ability to improve workplace safety therefore has important policy implications.

We focus on OSHA's Site-Specific Targeting (SST) program, which was the agency's largest inspection regime while the program operated (1999–2014). To allow the most "effective use of [OSHA's] enforcement resources," it targeted establishments with "serious safety and health problems" (US OSHA 2004: 25445-25446) by developing annual target lists of establishments that had high injury rates two years prior. When resources did not allow the agency to inspect all establishments on its target lists, it allocated inspections via random assignment.

We first evaluate the extent to which the subset of SST inspections that OSHA randomly assigned between 2001 and 2010 affected injuries. By focusing on inspections targeted using

[1] The Occupational Safety and Health Act of 1970 (OSH Act) 29 U.S.C. ch. 15 § 651 et seq, December 29, 1970.
[2] Spoken by Senator Heidi Heitkamp (D-ND) during a February 11, 2016 hearing with the Senate Homeland Security and Governmental Affairs Committee (Musick, Trotto, and Morrison 2016).

random assignment, our estimates are free of the selection bias that plagues most evaluations of workplace inspections.[3] Roughly 13,000 establishments, employing nearly 2 million workers, were at risk of being targeted for a randomized inspection over this 10-year period. Our primary outcome variable is serious injuries and illnesses—those leading to days away from work ("DAFW"). (For simplicity, we refer to both injuries and illnesses as "injuries.") We estimate effects of inspections over the five-year period comprising the year an establishment was placed on the SST target list (henceforth, the "directive year") and the four subsequent years.

Randomly assigned OSHA inspections reduced DAFW injuries at inspected establishments by an average of 9%, which equates to 2.4 fewer injuries over the five-year period we examined. This yields a social benefit of roughly $120,000 per inspection, roughly 35 times OSHA's cost of conducting an inspection. These inspections had no detectable impacts on business outcomes such as establishment survival, employment, or credit ratings.

Could OSHA have allocated its inspections to avert even more injuries? The ideal input into answering this question is an estimate of each establishment's conditional average treatment effect (CATE)—the difference between the number of injuries it would experience if it had been assigned to inspection and the number of injuries if it had not.[4] Many factors could influence how effective inspections are at reducing injuries. We use causal forest to estimate the CATE for each establishment on the historical SST target list. A causal forest is a flexible supervised machine-learning technique that predicts treatment effects based on high-dimensional nonlinear functions of observable characteristics (Wager and Athey 2018).

In theory, OSHA could maximize the number of injuries its inspections avert by targeting establishments with the largest estimated CATEs. In practice, estimating such heterogeneous treatment effects is difficult. We therefore also consider a second metric that OSHA could use to target inspections: the predicted number of serious injuries an establishment would experience absent an inspection. This metric is similar in spirit to the one OSHA actually used for SST;

---

[3] For example, because many OSHA inspections target establishments with recent accidents or complaints, those establishments likely have systematically different characteristics (both observable and unobservable) than non-inspected establishments. Furthermore, establishments experiencing high injury rates in one year (thus triggering an OSHA inspection) may experience fewer injuries the following year simply due to regression to the mean, in which case OSHA inspections correlate with lower injury rates without actually causing them. Similar endogeneity issues challenge the ability to evaluate the effects of inspections by other regulators, such as the US Environmental Protection Agency (Hanna and Oliva 2010).

[4] Our treatment is "assignment to SST inspection" rather than "inspection" because, while OSHA randomized assignments, such assignments imperfectly predicted actual inspections.

namely, an establishment's injury rate from two years prior. Because we use machine learning and more data, our metric plausibly more accurately predicts an establishment's underlying health and safety problems at the time an inspection would actually occur.

To derive consistent estimates of the number of injuries OSHA could avert under alternative targeting policies, we integrate our estimates into a method developed by Chernozhukov, Demirer, Duflo, and Fernandez-Val (2018)—henceforth "CDDF." The CDDF method incorporates estimates of heterogeneous treatment effects to yield consistent estimates of average treatment effects for specific subsets of the data.

We find that OSHA could avert many more injuries if it used machine learning to target inspections. For example, if OSHA had assigned the same number of its SST inspections each year to those establishments with the largest estimated CATEs, it would have averted 2.4 times as many injuries as it actually did. If OSHA had instead assigned this same number of inspections each year to the establishments predicted to have the most injuries, it would have averted 2.3 times as many injuries as it actually did. These estimates change very little if, rather than assigning the same number of inspections as the SST program, we instead assign the number of inspections that maintains OSHA's inspection budget under SST.

Targeting based on "largest estimated CATEs" or "most predicted injuries," however, eliminates randomization. This might reduce the threat of inspection for uninspected establishments (Cohen 2000; Shimshack and Ward 2005; Gray and Shadbegian 2007), and it would sacrifice opportunities for OSHA to continue learning where inspections are most effective. Thus, we also consider a policy in which OSHA allocates *all* inspections randomly: it creates larger target lists than its resources allow, as it did under SST, but populates these lists with establishments that have the largest estimated CATEs or most predicted injuries, rather than those with the highest injury rates from two years prior. It randomizes assignments within these target lists with a probability that mimics the historical SST policy. Compared to the actual SST policy, OSHA would avert 1.37 times as many injuries at assigned establishments when targeting on estimated CATE and 1.39 times as many when targeting on predicted injuries.

This paper contributes to a nascent literature on using machine learning to improve decisions. Our study is close in spirit to others that have examined how machine learning to *predict* outcomes can be used to understand and improve judges' decisions to release defendants before trial (Kleinberg et al. 2017), to help regulatory inspectors predict restaurants' health

violations (Glaeser et al. 2016) and facilities' water pollution violations (Hino, Benami, and Brooks 2018), to help electric utilities predict when unmaintained equipment will fail (Rudin et al. 2010), and to help firms select board members (Erel et al. 2018). We extend this literature by estimating variation in the *causal effects* of regulatory inspections, which Athey (2017) points out is the relevant criterion for an optimal resource allocation problem. Our approach to estimating heterogeneity in policy effectiveness is shared with Davis and Heller's (forthcoming) study of youth employment programs.

Our research also extends a large literature on the effects of regulatory inspections. Many studies have specifically examined the effects of OSHA inspections on injuries. Some found little to no effect (Smith 1979; Bartel and Thomas 1985; Viscusi 1986; Ruser and Smith 1991); others have found that federal or state OSHA inspections reduce injuries (Gray and Scholz 1993; Gray and Mendoloff 2005; Foley et al. 2012; Haviland et al. 2012), including two studies that leverage randomly assigned OSHA inspections (Levine, Toffel, and Johnson 2012; Lee and Taylor forthcoming). Our estimates of the effects of OSHA's SST inspections complement two recent studies of that program. Li and Singleton (2019), using a regression discontinuity design, find—as we did—that SST inspections reduced injuries at inspected establishments.[5] Peto et al. (2016), analyzing only SST inspections randomly assigned in 2011, find no statistically significant effects, perhaps because their small sample size left them underpowered.

Our findings also complement literatures on the effects of regulatory inspection in other domains, including food safety (Ibanez and Toffel forthcoming), environmental protection (Hanna and Oliva 2010; Telle 2013; see Shimshack 2014 for an overview), third-party social auditors (e.g., Short, Toffel, and Hugill 2016), and tax authority audits (e.g., Slemrod, Blumenthal, and Christian 2001; Kleven et al. 2011). By investigating how regulators can adjust their inspection strategies to best achieve their objectives, we complement Gonzalez-Lira and Mobarak (2019), who study how regulators should adjust inspection intensity to account for agents' adaptive behavior; Duflo et al. (2018), who illustrate the benefit of regulatory discretion

[5] One difference between Li and Singleton's (2019) design and ours is that their regression discontinuity design estimates the effect of SST inspections for the subset of establishments near the cutoff of eligibility for inclusion in the SST program, whereas our estimates correspond to the average effect among those inspected. A second difference is that we focus on the effects of inspections on especially serious injuries, whereas Li and Singleton (2019) consider a broader range of injuries. As we discuss below, less-serious injuries are more subject to measurement error (e.g., due to misreporting by workers to employers or by establishments to OSHA) than more serious injuries are; such error can confound how inspections affect actual injuries versus reporting of injuries.

in directing inspections; and Blundell et al. (2018), who estimate the gains to dynamic enforcement of environmental regulations for high-polluting plants.

We add to these literatures on regulation by (a) providing credible estimates based on a large set of randomly assigned inspections, (b) examining whether inspections have unintended effects on business outcomes such as employment and establishment survival, and (c) comparing the effectiveness of a regulatory agency's historical policy with counterfactual policies.

1 Setting and Data

OSHA's Site-Specific Targeting (SST) program is an excellent setting for our purposes. First, because OSHA allocated some inspections via random assignment, we can evaluate the extent to which its inspections led to fewer injuries on average and can use machine learning methods to estimate heterogeneous treatment effects. Second, SST was created to promote the most "effective use of [OSHA's] enforcement resources" (US OSHA 2004: 25446) and thus has a stated goal to compare alternative targeting policies. Finally, SST was an extremely large regulatory program, costing tens of millions of dollars and putting tens of thousands of establishments that employed millions of workers at risk of inspection.

1.1 OSHA Site-Specific Targeting (SST) Program

The SST program targeted high-injury workplaces in historically hazardous industries for inspection between 1999 and 2014. We focus on the 29 states that OSHA directly regulates.[6]

Each year from 1996 through 2011, OSHA conducted the OSHA Data Initiative (ODI), a survey of injury data from 60,000 to 80,000 establishments in the US in pre-defined high-risk industries that had at least 40 employees. Survey responses were based on the establishments' OSHA-required logs that documented every work-related injury and illness.[7] OSHA used ODI responses to create its SST target lists the following year: a "primary list" of the roughly 3,500 establishments with the highest injury rates (averaging roughly five times the national average) and a "secondary list" of the roughly 7,000 establishments with the next-highest injury rates

[6] Figure A.1 in Appendix A provides a map of those 29 states. The other 21 states operate state-run programs approved by OSHA. Appendix B gives a more complete description of SST in federal OSHA states.
[7] OSHA Form 300, which employers are required to complete, is available at https://www.osha.gov/recordkeeping/RKforms.html, accessed March 2019.

(averaging roughly three times the national average). Precise cutoffs for both lists varied by industry and year.

OSHA then sent each of its 81 Area Offices the list of establishments in its geographic boundary that were on the primary list. If an Area Office did not anticipate having sufficient resources to inspect its entire primary list, it told headquarters the number of inspections it anticipated being able to complete. OSHA's software then randomly assigned a subset of that many establishments from its primary list to inspect. If the Area Office completed these inspections before headquarters provided the following year's lists, the Office estimated how many more inspections it could conduct and the software generated a new random set of establishments from the remainder of its primary list. If an Area Office completed its entire primary list, the Area Office repeated this process with the secondary list. Thus, most Area Offices inspected a random subset of either their primary or their secondary list (but never both) each year.

Inspectors arrived unannounced to conduct SST inspections. As with other OSHA inspections, the inspector walked through the establishment to assess hazards and then met with managers and sometimes also with worker representatives. The inspector provided feedback on the workplace's safety program and explained any violations detected. OSHA typically assessed a fine for violations. Establishments could appeal fines and OSHA often reduced them if the violation was remediated immediately.

There are several reasons why OSHA inspections could improve workplace safety. Citing violations increases the incentives for managers to remediate those and other hazards. Even if the inspector did not identify any violations, inspections can heighten the salience to management of regulations and safety (Alm and Shimshack 2014). Inspections can also lead managers and workers to increase their perceived risk of being inspected—and potentially penalized—again (Kleven et al. 2011; Avis, Ferraz, and Finan 2016). Finally, inspectors sometimes share knowledge about safety practices with management (Choi and Almanza 2012).

Inspections might harm business outcomes if remediation raises costs or reduces productivity. If these effects are large enough, inspections might worsen an establishment's credit rating, reduce employment, or lead to plant closure. At the same time, if inspections increase knowledge about cost-effective safety practices, inspectors might not harm these business outcomes and might even improve them (Levine, Toffel, and Johnson 2012).

1.2 **Data**

We merged data from four sources: (a) OSHA's annual SST target lists (2001–2010);[8] (b) OSHA's annual ODI survey data on injuries and employment (1996–2011);[9] (c) OSHA inspection data from its Integrated Management Information Systems (IMIS) database (1990–2016);[10] and (d) annual Dun & Bradstreet data on employment, credit rating, and other business outcomes from the National Establishment Time Series (NETS) database (1990–2013).[11]

The annual SST primary and secondary target lists report the set of establishments at risk of SST inspection each year during 2001–2010, each establishment's associated Area Office, and whether or not OSHA assigned the establishment to receive an SST inspection.

The ODI dataset contains the annual survey results that establishments were required to report to OSHA from 1996 to 2011, including annual counts of DAFW injuries and injuries involving job transfers or restrictions, which together are called DART injuries (those resulting in "days away from work, restricted work, or a transfer"). ODI also contains an establishment's annual average employment and total labor hours worked and its DUNS number, a unique establishment-level identifier.[12] The ODI dataset is an unbalanced panel: it includes a different (but overlapping) set of establishments each year. The ODI sought to survey all establishments with at least 40 employees in hazardous industries every three years. Beginning around 2005, it resurveyed the following year those establishments reporting a DART rate of at least 7. Many establishments on the SST target lists report ODI data nearly every year.[13] This is because (a) OSHA's DART rate threshold for re-sampling establishments in the ODI survey was below its threshold for placing establishments on the SST primary target list (and at or above the threshold

[8] While the SST program operated from 1999 through 2014, OSHA's Office of Statistical Analysis could only locate and provide us with target lists for 2001 through 2010.
[9] We obtained SST target lists and ODI survey data from OSHA's Office of Statistical Analysis after signing a memorandum of understanding.
[10] We downloaded OSHA inspection records in January 2014 from the agency's publicly available IMIS database.
[11] NETS is a proprietary database distributed by Walls and Associates (Donald Walls, dwalls2@earthlink.net).
[12] The ODI dataset also includes the number of fatal injuries, the number of injuries and illnesses not meeting the DART criteria, and the number of illnesses in categories such as skin disorders, respiratory problems, poisoning, and hearing loss.
[13] Among the establishments that were ever on an SST target list, 25% reported injury data in at least 10 of the 16 years of the ODI program (1996–2011) and 50% reported injury data in at least 7 years of these 16 years.

used to place establishments on the secondary list) and (b) injury rates tended to be serially correlated.[14]

The IMIS database includes every inspection attempted by OSHA. Each record includes the name and address of the establishment, the inspection date, whether the inspector was unable to carry out the inspection (e.g., if the company had moved or gone out of business), what triggered the inspection (e.g., the SST program, a different program, a recent serious accident, or an employee complaint), and the number of violations and associated penalty values.

NETS is an annual panel dataset extracted from Dun & Bradstreet data. It seeks to include all establishments ever in operation since 1990. From NETS, we obtained each establishment's unique DUNS number, its first and last year in operation (to measure survival), characteristics such as whether it was part of a multi-unit firm, and business outcomes such as annual employment and credit rating.

To construct our sample, we begin with the list of all establishments on OSHA's 2001–2010 SST target lists. We linked an establishment's corresponding ODI and NETS records via its DUNS number. We found 100% and 97% of the SST target list establishments in ODI and NETS, respectively; we dropped the 3% that could not be matched to NETS. Because the IMIS database does not include DUNS numbers (or any other unique establishment identifier), we fuzzy-matched the remaining establishments to corresponding IMIS records using names, addresses, and industries. We implemented fuzzy matching by using *MatchIt* software, the Stata *reclink* command, and a manual process. We were thus able to link 82% of the establishments on the SST target list to an inspection record in IMIS.[15]

1.3 **Outcomes**

Our primary measure of workplace safety is an establishment's annual number of injuries that resulted in days away from work (DAFW injuries). We focus on these because they are (a) the most serious type reported to ODI, (b) more likely than other injuries to be reported to supervisors (Biddle and Roberts 2003), and (c) more likely to be recorded by employers on their

[14] For example, among all establishments ever on the SST list in successive years, the correlation between their ODI-reported DART rates in successive years is 0.55.

[15] We retain the 18% of establishments on the SST target list that did not link to IMIS records. While our matching algorithm might have failed to identify some of their corresponding IMIS records, some target list establishments were probably never inspected and thus do not have any inspection records in IMIS.

OSHA-mandated injury logs (Boden, Nestoriak, and Pierce 2010). DAFW injuries are thus likely the most accurate injury metric in ODI. (We discuss the validity of ODI injury data in Appendix C.)

DAFW injuries are enormously costly. There were 1.2 million in the US in 2005 (US Bureau of Labor Statistics 2007), the midpoint of our sample period. With an estimated cost of $52,000 per injury (in 2018 dollars; see Appendix D), these injuries cost the US over $60 billion. DAFW injuries therefore merit substantial policy interest. To reduce the effect of very large outliers of DAFW injury count (which exhibits positive skew), we top-code this variable at its 99th percentile, which is 54 per year.

We consider several business outcomes. We measure an establishment's (a) survival each year based on its continued presence in NETS, (b) annual employment based on data from NETS[16] and ODI, (c) total working hours from ODI, and (d) annual credit rating using Dun & Bradstreet's PAYDEX scores (ranging from 1 to 100, with higher scores reflecting more rapid payment of bills and thus greater creditworthiness).[17] For employment and total working hours, we reduce the effect of very small outliers by adding the variable's first percentile of non-zero values and we analyze log values to reduce skew.

2 Methods

In this section, we describe how we estimate the average treatment effects of randomized SST inspections, how we estimate heterogeneous treatment effects of inspections, and how we use these estimates to simulate the effects of alternative targeting policies.

[16] We change to "missing" the roughly 10% of observations in which NETS employment values were estimated by Dun & Bradstreet or Walls & Associates. Employment values from NETS and ODI reflect somewhat different definitions: NETS employment refers to the number of "jobs" in an establishment, whereas ODI refers to the number of employees working at an establishment. These definitions can lead to different counts. For example, at the time that NETS counts employment in a given year, if a worker had recently resigned but the establishment planned to fill the position in the future, employment reported by NETS would be 1 higher than that of ODI.

[17] For example, a PAYDEX score of 20 refers to establishments whose dollar-weighted value of bills are paid 120 days beyond terms, whereas a score of 80 denotes "prompt" payment. We analyze each establishment's lowest monthly PAYDEX score each year. Results are almost identical if we analyze the establishment's highest monthly score in a given year.

2.1 Estimating Average Treatment Effects of Randomized SST Inspections

We describe the sample and then the econometric models we use to estimate the average treatment effects of the randomized SST inspections.

2.1.1 Creating the Randomized Sample

During our 2001–2010 sample period, the SST target lists included 28,163 establishments that OSHA assigned to inspection and 63,663 not assigned.[18] To estimate the average treatment effects of SST inspections, we restrict our attention to the subset of establishments on the SST target lists that were eligible for a *randomized* inspection (henceforth, the "randomized sample").

We begin by mimicking the exclusions that OSHA applied to ensure that our sample includes only establishments that were at risk of an SST inspection. Specifically, we exclude from the target lists—just as OSHA Area Offices did—establishments that had already received a comprehensive safety inspection in the previous two years.[19] We also exclude the few establishments that were not in operation, according to NETS, two years prior to the directive year; OSHA assembled the target list based on injuries from two years prior and injury data for such establishments might therefore reflect measurement error. We also exclude establishments that were not in operation in the directive year and, therefore, could not have been inspected.

To focus on establishments at risk of a *randomized* inspection, we retained only those that were either randomly assigned for an SST inspection ("assigned to inspection")—our treatment group—or eligible for, but not assigned to, a randomized SST inspection ("not assigned to inspection")—our control group. Thus, we excluded establishments on an Area Office's primary list or secondary list in a given year if the Area Office assigned either none or all of the establishments on the list to inspection in that year.[20] Finally, we excluded

---

[18] Over this period, the target list also included 9,617 establishments that Area Offices explicitly marked as "deleted" from the SST target list because they were ineligible for SST inspection per OSHA's eligibility rules described above. Through conversations with Area Office directors, we learned that many Area Offices implemented their deletions, but did not input them into the SST target list database. This is one reason that we remove ineligible establishments manually.

[19] This criterion changed from two years to three years in 2009, which we mimic by deleting establishments on the 2009 and 2010 target lists that had a comprehensive inspection in the prior three years.

[20] In practice, we were more conservative and restricted our randomized sample to those primary and secondary lists in which between 5% and 95% of eligible establishments were assigned to inspection by an Area Office in a directive year, rather than including all lists where strictly between 0% and 100% of establishments were assigned to inspection. We do so to ensure that each Area Office directive in our randomized sample has a sufficient number of

establishments that OSHA had added to its target lists solely due to its concern that the exceptionally *low* injury rates reported to ODI might be inaccurate. (Table A.1 has more details on how we arrive at the randomized sample.) This results in a randomized sample with 6,977 establishments assigned to inspection and 9,164 not assigned. These 16,141 observations, which we refer to as establishment-directive dyads, correspond to 13,029 unique establishments because some establishments were included on SST target lists in multiple years. We use "directive year" to refer to the calendar year in which an establishment was placed on the target list and was thus eligible for assignment to inspection.[21] We construct a nine-year panel around these 16,141 establishment-directives: a "pre-period" of the four years prior to the directive year, a "post-period" of the directive year and the four subsequent years.[22]

We do not observe any ODI-reported outcomes in the post-period for 2,405 (15%) of the 16,141 establishment-directives in our randomized sample, largely due to (a) establishments on target lists in later directive years having fewer opportunities to appear again in the ODI because it ended in 2011 or (b) establishments shutting down or becoming ineligible for ODI by, for example, shrinking to fewer than 40 employees. These establishments are dropped from the sample for models in which the outcome variable is drawn from ODI (including injuries, hours, and employment); those models are thus estimated on 13,736 establishment-directives (11,083 unique establishments). Our attrition rate is much smaller for outcomes in the NETS database, which has employment values during the post-period for all but 390 establishments (2.4 %) in our randomized sample. We discuss sample attrition in Appendix E.

Table 1 reports the industry distribution of establishments in our sample. Table 2 reports summary statistics for our outcomes.[23]

---

establishments that were assigned to inspection and not assigned to inspection, because our regressions (described below) compare these two groups within the same Area Office directive.

[21] For example, all establishments placed on the target list issued on May 14, 2007 have a directive year of 2007.

[22] Our estimation samples contain fewer than 145,269 observations (16,141 × 9), as most establishments were not included in the ODI survey in all nine years. Establishments in our most recent directive year, 2010, have at most three subsequent years of NETS data (because our NETS dataset ends in 2013) and some ceased operation within five years of their directive year.

[23] We follow OSHA's rules and calculate injury rate as the number of injuries divided by (total working hours / 200,000). 200,000 is the number of hours 100 full-time employees (FTEs) would work in a year, so the denominator is effectively 100 FTEs.

Balancing tests to validate random assignment. To validate whether assignment to inspection was indeed random in our randomized sample, we regress a series of baseline variables on an *assigned to inspection* dummy and a set of fixed effects for each Area-Office–year dyad, which is the level at which the randomization took place.[24]

Establishments assigned to inspection are statistically indistinguishable from those not assigned to inspection in terms of all baseline characteristics except for total working hours and employment reported to ODI (Table 3). For both metrics, establishments assigned to inspection are roughly 2.5% larger—statistically significant at the 5% level. While it is not necessarily surprising to find a significant difference for two (highly correlated) variables given that we examined nine, we nonetheless estimated the evaluation models described below by also controlling for total working hours in *t-2* as a robustness test, with virtually identical results.

Addressing fuzziness in the treatment assignment. While our setting provides a clean experimental design due to the randomization of *assignment* to an SST inspection, there are several reasons why comparing those assigned to inspection to those not assigned does not yield the causal effect of *receiving* an SST inspection.

First, OSHA issued the annual SST directives between April and August, and Area Offices did not begin conducting those inspections until months later. As a result, only 18% of establishments randomly assigned to SST inspection received one by the end of the calendar year in which they had been placed on the target list (the directive year).

Second, *assignment* to an SST inspection does not perfectly correspond to *receiving* one. Sometimes the inspector could not find the establishment; sometimes an Area Office had successfully petitioned OSHA headquarters for permission to not inspect all of the establishments assigned to it. One year after their directive years, 73% of establishments assigned to inspection had received an SST inspection (Figure A.2).[25]

Third, some establishments eligible for assignment but not assigned to inspection in a given directive year were assigned to inspection in a subsequent year. Almost all establishments

[24] We use two years prior to the directive year because the target lists in a given year are constructed based on injury rates from two years prior. We obtain essentially identical results using four years.
[25] It is possible that OSHA did inspect some of these establishments, but our procedure to link SST with OSHA's information system (IMIS) failed to find their corresponding inspections in IMIS.

on the SST target list in one year also qualify to be ODI-surveyed in a subsequent year and many are placed on a future year's SST target list. Thus, 28% of our control establishments did receive an SST inspection within the next four calendar years.[26]

In short, some establishments assigned to SST inspection did not receive one and some that were not assigned in a given year *did* receive one (having been assigned in a later year). Thus, comparing injuries between establishments that OSHA did and did not randomly assign to SST inspection in a given year (the "intent-to-treat" estimate) will underestimate the effect of receiving an SST inspection on inspected establishments' injury rates.

Following standard practice for experiments with imperfect assignment, we instrument whether an establishment has been inspected with whether it was assigned to inspection in the directive year. This approach scales the intent-to-treat estimate by the extent to which assignment to inspection increases the probability of actually being inspected. The instrumental variable procedure estimates the local average treatment effect of SST inspections ("local" in the sense that it applies to randomized inspections).

2.1.2 Specification to Estimate Intent-to-treat Effects

To ensure that estimates are not vulnerable to large outliers or threats to identification, we conducted a series of tests to pre-specify our regression specification. We first blinded ourselves to each establishment's assignment status. In each of 500 simulation runs, we created an *assigned to placebo inspection* dummy that randomly assigned 0 or 1 to each establishment-directive in our randomized sample, with a probability corresponding to the proportion of establishments that OSHA historically assigned to inspection. We then estimated the effect of *assigned to placebo inspection* on various outcomes using several functional forms and

[26] Another potential source of fuzziness for our experimental design is that establishments might have had other types of inspections. For example, if establishments we classified as "not assigned to inspection" had differential rates of another type of OSHA inspection because they were, for example, targeted by a different OSHA emphasis program, that would mean we are not fully capturing differences in rates of OSHA inspections between establishments that were and were not assigned to SST inspection. However, we find essentially zero difference between the establishments assigned to inspection and not assigned to inspection in the likelihood of experiencing a non-SST inspection conducted by OSHA. Thus, we do not consider this potential source of bias to be an important factor in our context and do not discuss it further.

approaches to handle outliers.[27] Our objective was to identify which specifications most often yielded a precisely estimated zero coefficient on the *assigned to placebo inspection* dummies across all simulations. We describe this procedure in detail in the pre-analysis plan, which we summarize in Appendix F.[28]

Based on the results of these simulations, our preferred intent-to-treat specification is:

$$y_{ijt\tau}^{post} = F\left(\alpha_1 Assigned_{it} + \alpha_2 y_{it}^{pre} + \boldsymbol{\gamma X}_{it} + \mu_{jt} + \theta_\tau + \epsilon_{ijt\tau}\right), \quad (1)$$

where $y_{ijt\tau}^{post}$ is the annual outcome for establishment *i* located within the geographic boundary of Area Office *j*, on the SST target list in year *t*, realized τ years relative to SST directive year *t*. In this specification, τ ranges from 0 (the directive year) to 4 (four years following the directive year), so that we include up to 5 years (and at least 1 year) of data for each establishment-directive. When *y* refers to the establishment's injury count, we use a Poisson specification, modelling the right-hand side of Equation 1 as the conditional mean function of *y*. For all other outcome variables (survival, employment, hours, credit rating), we use OLS.

$Assigned_{it}$ is a dummy coded 1 if establishment *i* was randomly assigned to an SST inspection in directive year *t*, and 0 if it was eligible but was not assigned to an SST inspection. The coefficient on $Assigned_{it}$, α_1, represents the intent-to-treat (ITT) effect of assignment to inspection. To improve precision we control for establishment *i*'s average *y* value over the four years prior to the directive year, y_{it}^{pre}, in our OLS specifications, or the average of $log(y_{it}^{pre}+1)$ over this period in our Poisson specifications, sometimes called an analysis of covariance (ANCOVA) specification (McKenzie 2012).[29]

$\boldsymbol{X}_{it}$ refers to control variables for each establishment-directive. It includes the number of years of data on which the baseline mean y_{it}^{pre} is based (since we do not observe ODI- or NETS-reported data in all years for all establishments). In some specifications, we also include total working hours (or its log) in year *t-2*. μ_{jt} represents a series of fixed effects for each Area-

[27] For example, we considered OLS specifications in regressions with injury rate (both level and log) as the dependent variable.

[28] We posted our pre-analysis plan to the Open Science Framework on July 10, 2015 at https://osf.io/2snka/.

[29] In our Poisson specifications, we control for baseline $log(y+1)$ because a coefficient on y_{it}^{pre} estimates how much a 1-unit change in baseline *y* is correlated with a 1% change in post-period *y*, whereas a coefficient on $\log(y_{it}^{pre}+1)$ estimates how much a 1% change in baseline *y* is correlated with a 1% change in post-period *y*, which is a more appropriate metric to capture the correlation between baseline and post-period outcomes. We add 1 to y_{it}^{pre} before taking the log to account for zeroes.

Office×directive. θ_τ refers to a series of fixed effects for each τ year relative to the directive year. We cluster standard errors by establishment.

Equation 1 assumes that the effects of assignment to SST inspection are constant over the directive year and the four subsequent years, but the effects of inspection might vary substantially over time. Moreover, $\alpha_1$might yield a biased estimate of the effect of assignment to inspection if treatment and control establishments have differential pre-assignment trends. We address these concerns by estimating a distributed lag specification. Specifically, we estimate, using Equation 2 below, the annual difference in injury counts—in each of the four years prior to, the year of, and each of the four years following the directive year—between establishments that were and were not assigned to inspection:

$$\boldsymbol{y_{ijt\tau}} = \boldsymbol{F}(\textstyle\sum_{k\in[-4,4]} \boldsymbol{\beta_k D_{\tau=k}} * \boldsymbol{Assigned_{it}} + \boldsymbol{\mu_{jt}} * \mathbb{1}(\boldsymbol{\tau \geq 0}) + \boldsymbol{\lambda_{it}} + \boldsymbol{\theta_\tau} + \boldsymbol{\epsilon_{ijt\tau}}) \quad , \qquad (2)$$

where $\mathbb{1}(\tau \geq 0)$ equals 1 when $\tau \geq 0$, and 0 otherwise. This specification, unlike Equation 1, includes both pre- and post-period years for each establishment-directive ($\tau \in \{-4, 4\}$). $D_{\tau=k}$ is a dummy equal to 1 if $\tau = k$ for $k \in \{-4, 4\}$. β_k coefficients estimate the difference in outcome y between establishments assigned and not assigned to inspection for each of the four years prior to, the year of, and the four years following the directive year. λ_{it} is a fixed effect for each establishment-directive, which effectively replaces y_{it}^{pre} and X_{it} from Equation 1 to control for time-invariant differences across establishment-directives.[30] As in Equation 1, μ_{jt} is a fixed effect for each Area-Office×directive, but here we multiply it by $\mathbb{1}(\tau \geq 0)$, since it is invariant within establishment-directives and would otherwise be dropped from the regression.

2.1.3 Instrumental Variables Specification

In Equation 1, α_1 yields the difference in y^{post} between establishments assigned to and not assigned to inspection. However, as discussed above, the intent-to-treat specification will underestimate the average treatment effects of inspection. To estimate the average treatment effects of inspection on injuries and other outcomes, we instrument whether an establishment has been SST-inspected with whether it had been assigned to inspection in the directive year. Specifically, we estimate the following variant of Equation 1:

[30] Equation 2 includes a fixed effect for each establishment-directive because, unlike the ANCOVA regression, a distributed lag specification like this one does not allow for including baseline *y* as a control variable.

$$y_{ijt\tau}^{post} = F\left(\delta_1 \widehat{Inspected}_{it\tau} + \delta_2 y_{it}^{pre} + \boldsymbol{\kappa} \boldsymbol{X}_{it} + \mu_{jt} + \theta_\tau + \eta_{ijt\tau}\right). \quad (3)$$

Here, $\widehat{Inspected}_{it\tau}$ is the predicted value from the following first-stage equation, in which $Inspected_{it\tau}$ is a dummy coded 1 if establishment *i* was SST-inspected at any time between the directive year *t* and $t+\tau$ (where τ ranges from 0 to 4, as in Equation 1) and coded 0 otherwise:

$$Inspected_{it\tau} = \pi_1 Assigned_{it} + \pi_2 y_{it}^{pre} + \boldsymbol{\delta} \boldsymbol{X}_{it} + \mu_{jt} + \theta_\tau + \nu_{ijt\tau}. \quad (4)$$

For both Equations 3 and 4, $\boldsymbol{X}_{it}$, μ_{jt}, and θ_τ are the same as defined in Equation 1. $\eta_{ijt\tau}$ and $\nu_{ijt\tau}$ represent i.i.d. error terms.

Our instrumentation of *inspected* with *assigned* meets the two requirements for $\widehat{\delta_1}$ in Equation 3 to identify the effect of an SST inspection on outcome *y*. First, as we show below, the first-stage relationship modelled in Equation 4 is strong. Second, consider the exclusion restriction that *assigned* cannot directly affect outcome *y* except through its influence on *inspected.* There is no plausible reason that this exclusion restriction was violated because (a) establishments were never informed that they were assigned to SST inspection and (b) this assignment had no effect on inspectors' actions other than allocating SST inspections.

We use an IV-Poisson regression model to estimate the causal effect of being inspected on DAFW injury count (we consider other specifications in robustness checks). For our other outcome variables, we use linear IV.

2.2 Constructing Alternative Targeting Criteria

In this section, we analyze how OSHA could have used two machine-learning–based measures of the heterogeneous effects of inspections in order to target its inspections.

2.2.1 Estimating Heterogeneous Treatment Effects of Inspections

If the effect of inspections on safety is heterogeneous, OSHA could potentially have averted more injuries by targeting workplaces especially responsive to inspections. Here, we describe a machine learning approach to estimating heterogeneous treatment effects of assignment to inspection. We focus on heterogeneity in the effect of *assigning* an establishment to be inspected (the intention to treat), rather than the heterogeneity of whether an establishment was actually *inspected* (the treatment), because assignment is the lever at OSHA controls.

Assessing the extent to which OSHA could reallocate assignments to inspection to increase their effectiveness requires estimating each establishment's treatment effect given its

baseline characteristics; that is, its conditional average treatment effect (CATE). Following Rubin's (1974) potential outcomes framework, we define an establishment's CATE as:

$$s_0(Z) = E[\,Y(1)|Z] - E[Y(0)|\,Z],$$

where Z is a vector of baseline characteristics, $Y(1)$ is the establishment's outcomes if assigned to inspection, and $Y(0)$ is the establishment's outcomes if not assigned to inspection.

Constructing an estimate $S(Z)$ of the CATE ($s_0(Z)$) is challenging. For example, one could include dozens of candidate interaction terms in a regression model to test for heterogeneous effects, but including many interaction terms can lead to spurious over-fitted estimates that predict poorly out of sample. Further complicating matters, elements of Z could affect establishments' CATE in highly nonlinear ways. We therefore estimate each establishment's CATE using *causal forest* (Wager and Athey 2018), a supervised machine learning method that builds on Breiman's (2001) random forest algorithm. Random forest is a prediction algorithm that allows for flexible modeling of interactions in high-dimensional settings. A random forest first builds many regression trees. A regression tree is a form of nearest-neighbor matching in which the set of neighbors is determined by the data to maximize both similarity within a leaf and divergence across leaves. The random forest then averages predictions of the many small trees to reduce variance and improve predictive power.

Wager and Athey's (2018) causal forest modifies the random forest to estimate heterogeneity in causal effects. Causal forest searches for high-dimensional combinations of covariates that are associated with different treatment effects. To mitigate against overfitting, we create each tree with one subsample of the data and estimate the treatment effect at each leaf with a second subsample (which Wager and Athey refer to as the "honest" approach).

For causal forest to estimate unbiased CATEs, assignment to inspection must be independent of the potential outcomes, conditional on Z. This condition is satisfied among the establishments in our randomized sample because assignment to inspection was random conditional on Area-Office–year. In addition, there must be enough treatment and control observations in a given leaf, because small leaves can increase mean squared error (Athey and Imbens 2016). Thus, we include only leaves with at least 50 observations and for which the share of treatment or control observations is no less than 10 percent. Appendix G lists the covariates we include in Z. To simplify calculations in the causal forest, our outcome variable is

the average number of DAFW injuries over the directive year and the four subsequent (post-period) years, $\overline{y_{it}^{post}}$.

2.2.2 Predicting Establishments' Injuries

If causal forest could perfectly estimate establishments' CATEs of assignment to inspection, then using this metric would necessarily be the most effective way to allocate inspections to avert the most injuries. However, in practice this approach may not be optimal—or even feasible. First, estimating CATEs—with causal forest or any other method—is difficult in finite samples and estimates are subject to sampling variation. Second, targeting on predicted CATEs could be challenging for political economy reasons, as regulators might be leery of targeting based on a "black box" unobservable metric like expected injuries averted.

As a second metric for targeting, we consider the number of serious (DAFW) injuries an establishment would experience if not assigned to inspection: $b_0(Z) = E[Y(0)|\, Z]$. This metric is similar to the two-year lagged injury rate used for SST, but differs in three ways. First, $b_0(Z)$ predicts the injuries that would occur when inspectors would actually visit the establishment, as opposed to two years prior. Second, $b_0(Z)$ incorporates all the historical data we have, as opposed to being a single observation. Using multiple years of data and employer size is important because injury rates have substantial mean reversion, especially at smaller workplaces (Ruser 1995). Third, $b_0(Z)$ is a measure of injury *counts,* as opposed to *rates.*[31]

There are several reasons why targeting inspections to establishments with high estimated $b_0(Z)$ could improve safety. Establishments expected to have many injuries might reflect, in part, low levels of managerial and employee effort to implement safe work practices and limited knowledge on how to do so; inspections might increase their efforts and knowledge. A higher number of expected injuries might also indicate potential economies of scale in remediation, if remediating certain hazards would benefit many workers.

[31] A fourth distinction of our approach is that, whereas SST's targeting protocol relied on DART injuries (injuries resulting in days away from work, job restriction, or job transfer), we only consider the more serious subset: injuries resulting in days away from work (DAFW). As described in Section 2.3, DAFW injuries are likely to be reported more accurately than are job transfer or restriction injuries, which is one reason why we restrict attention to them.

Compared to $s_0(Z)$, $b_0(Z)$ has the advantage of being observed for establishments not assigned to inspection. This makes estimating $b_0(Z)$ a standard prediction problem. Using establishments not assigned to inspection, we use Super Learner, an ensemble machine learning procedure, to construct $B(Z)$, our estimate of $b_0(Z)$. Super Learner minimizes the mean squared error of out-of-sample predictions by using cross-validation to find the optimal weighted average of multiple machine learning methods (van der Laan, Polley, and Hubbard 2007).[32]

2.3 Evaluating Alternative Targeting Policies

Our machine learning methods are not designed to generate consistent estimates of $s_0(Z)$ and $b_0(Z)$ for individual observations (Chernozhukov et al. 2018). These raw estimates might therefore lead to misleading estimates of the effects of alternative policies that target inspections based on these metrics. To estimate how many injuries OSHA would avert under alternative policies, we follow the method developed in Chernozhukov et al. (2018) to produce consistent estimates of average CATEs for specific subsets of the data. We briefly describe the procedure here; see Chernozhukov et al. (2018) for details.

We first randomly partition the randomized sample into two equal subsets, which we refer to as the "auxiliary" and "holdout" samples.[33] Using the auxiliary sample, we use causal forest to construct $S(\cdot)$, the estimate of the function determining establishments' CATEs. Using those establishments in the auxiliary sample not assigned to inspection, we use Super Learner to construct $B(\cdot)$, the estimated function determining the number of injuries an establishment would experience if not assigned to inspection.

Using these functions estimated on the auxiliary sample, we compute the predicted CATE ($S(Z)$) and predicted baseline average ($B(Z)$) for establishments in the holdout sample. This sample-partitioning approach is a common machine learning method to avoid overfitting.

We then post-process $S(Z)$ and $B(Z)$ for all establishments in the holdout sample to estimate what Chernozhukov et al. (2018) call the "sorted group average treatment effects." For

[32] Our Super Learner library includes random forest (Breiman 2001), the Generalized Additive Model, and a linear interaction model. We initially ran Super Learner on the entire analysis sample with several additional learners in the library; these three learners are those to which Super Learner gave non-zero weight. We used the default parameters for each algorithm, except that we restricted the smallest leaves in the random forest to have at least 50 observations, because small leaves can increase mean squared error (Athey and Imbens 2016).

[33] Chernozhukov et al. (2018) refer to the holdout sample as the "main sample."

example, one policy that OSHA could follow is to assign the same number of inspections each year as it did historically, but allocate them to the establishments with that year's largest estimated CATEs—that is, the most negative $S(Z)$, or the most expected injuries averted. Define a group *G* such that (a) G_1 indicates that an establishment's $S(Z)$ is high enough to be assigned to inspection under this policy and (b) G_0 indicates otherwise. To estimate the number of injuries OSHA would avert under this policy, we need a consistent estimate of $E[s_0(Z) \mid G_1]$, which we obtain from the following weighted linear regression estimated on the holdout sample:

$$Y = \alpha_1 + \alpha_2 B(Z) + \sum_{k=0}^{1} \gamma_k \big(D - p(Z)\big) * \mathbb{1}(G_k) + \nu\,, \tag{5}$$

where *D* is an indicator for whether an establishment was assigned to treatment and $p(Z)$ is the probability an establishment would be assigned to treatment under the historical rule (the "propensity score"). Because OSHA assigned inspections to establishments in our randomized sample randomly conditional on Area-Office–year–list, an establishment's $p(Z)$ is just the proportion of establishments in its Area-Office–year–list eligible for inspection that were assigned to inspection. ν is an i.i.d. error term. Following CDDF, the regression is weighted by $\omega = \{p(Z) * \big(1 - p(Z)\big)\}^{-1}$.

The expected treatment effect among those in group G_k is:

$$\gamma_k = \; E[s_0(Z) \mid G_k] \text{ for } k \in \{0, 1\}.$$

Thus, $\widehat{\gamma_k}$ is a consistent estimate of the mean number of injuries averted per establishment among the establishments in group G_k. We then estimate the total number of injuries averted under a given targeting policy by computing $\sum_k(\widehat{\gamma_k} * \mathrm{N_k})$, where N_k is the number of establishments in group G_k that OSHA would assign to inspection under the policy.

However, relying on a single partition can be problematic if the holdout sample is, by chance, not representative of the entire randomized sample. We therefore conduct 250 iterations of the partitioning process, each time randomly partitioning the data into new auxiliary and holdout samples and saving the key coefficients and their associated standard errors. We then use the median point estimates and standard errors of $\widehat{\gamma_k}$ across these iterations as our estimates of the group average treatment effect associated with each targeting policy.

We make one final amendment to this procedure to obtain estimates of the overall effects of counterfactual targeting policies. While we seek to evaluate counterfactual targeting policies

that apply to the entire historical SST lists, we can apply the CDDF procedure only to our randomized sample, since it is impossible to know if the propensity score $p(Z)$ is correlated with potential outcomes among establishments in the nonrandomized sample.

We adapt the CDDF procedure to generate estimates that pertain to the entire historical SST target lists as follows. Each time we partition the randomized sample into auxiliary and holdout samples, we use the causal forest and Super Learner models from the auxiliary sample to predict $S(Z)$ and $B(Z)$ for both the holdout sample *and* a random 50% subset of the nonrandomized sample. We use the combined holdout sample and nonrandomized subsample to construct the *G* groupings that correspond to a particular targeting policy. We then estimate the coefficients ($\hat{\gamma}$s) that correspond to the mean number of injuries averted among the establishments in the group by running regressions of Equation 5 on the holdout sample. As described above, we estimate the number of injuries averted under a counterfactual policy by multiplying the $\hat{\gamma}$s by the number of establishments in group G_k (including those in both the randomized and nonrandomized samples) that OSHA assigns to inspection in the policy. Section 3.4.2 and Appendix J discuss potential threats to whether the estimates from this approach will apply to the entire target list (rather than just the randomized sample).

2.3.1 Using Historical Data to Predict the Effects of Counterfactual Targeting Policies

Alternative targeting rules change the threat of inspection that establishments face relative to the historical policy environment. Thus, managers might change their behavior under this new policy environment whether or not they are actually inspected (Lucas 1976). Indeed, in other enforcement regimes, the threat of higher fines has been shown to have a substantial effect on plants' EPA compliance (Blundell et al. 2018). By not accounting for such behavioral changes, our estimates of the effects of alternative targeting strategies could be misleading.

However, in practice we do not expect such behavioral changes to be a substantive concern in our setting. First, even though OSHA published its targeting rules in the Federal Register each year, anecdotal evidence indicates that most managers did not know OSHA's historical targeting rules.[34] In Appendix H, we provide empirical evidence that is consistent with

[34] For example, in March 2015, we spoke with a safety and health professional who had worked with thousands of establishments, many of which had recently experienced an OSHA inspection. He indicated that most of those establishments had no idea about the SST program, let alone its targeting criteria.

such lack of knowledge, including an absence of (a) bunching of reported DART rates just below the cutoffs for the primary or secondary lists and (b) an effect of the change in the threat of inspection on establishments' injury rates. This evidence implies that changes to OSHA's targeting rules are unlikely to meaningfully affect establishments' behavior; our estimates are thus likely to reflect the first-order effects of the counterfactual targeting policies we consider.

If we are wrong and the Lucas critique is important in this setting, it almost surely reinforces the benefits of targeting on $S(Z)$ or $B(Z)$. To the extent that establishments *could* have reacted to their own threat of inspection under SST, it would be more difficult under a targeting regime based on machine learning. Given the black-box nature of the machine learning algorithms underpinning our targeting regimes, establishments would be unable to tell if they were on the target lists; this uncertainty would limit the ability of those not on the target list to know with certainty that they faced zero chance of inspection.

Finally, if new targeting strategies *did* lead to behavioral changes, it would almost certainly result in even greater improvements in safety. Assume many employers *do* understand the targeting rules and that those at high risk of inspection preemptively remediate *prior* to inspection a share of the hazards they would have remediated after inspection. Then threatening employers with high $S(Z)$ will also maximize total injuries avoided by preemptive remediation at uninspected workplaces. Because we do not consider such benefits, our procedure, if anything, *underestimates* injuries averted under alternative targeting regimes (though, for reasons we argue in Appendix H, we do not expect this to be a large difference).

At the same time, such preemptive remediation by managers would mean that $S(Z)$ estimated in one year would have less predictive power for establishments' treatment effect of inspection in a later year. Given this dynamic response by employers, it would be important for the regulator to reestimate S(Z) as new data arrive. Though we do not undertake this dynamic estimation of S(Z), it would be straightforward to implement.

3 Results

3.1 Average Effects of OSHA's SST Inspections

3.1.1 Average Effects on Injuries

Table 4 reports estimates of the average effects of an SST inspection on the number of DAFW injuries. The first column displays Poisson regression results from the intent-to-treat specification from Equation 1. Establishments assigned to SST inspection experience 3.4% fewer injuries over the directive year and four following years (β = -0.035, SE = 0.017, p = 0.04) than those not assigned to inspection. This estimate is essentially unchanged when we control for baseline log total working hours from two years before the directive year (year *t*-2) (Column 2).

To investigate the extent to which the average effect on injuries of being assigned to inspection varies over time, Figure 1 shows the annual $ß_k$ coefficients and their 95% confidence intervals from a Poisson regression estimating the ITT distributed lag specification in Equation 2 (the omitted τ year is -2, two years prior to the directive year). For each of the four years prior to the directive year, the coefficient hovers around zero, consistent with random assignment. Beginning with the directive year (τ = 0), the coefficient becomes negative, hovering between -0.04 and -0.05 each year, and is statistically significant in years τ = 0 and $\tau = 1$.

To estimate the effect of having been inspected (treatment-on-the-treated effect), we must account for the fuzziness in inspection assignment described in Section 3.1.1. Column 3 reports an OLS estimate of the first-stage effect of assignment to SST inspection in the directive year on the probability of being SST-inspected, corresponding to Equation 4. Assignment to inspection increases the probability of actually being inspected over the directive year and four following years by 46 percentage points (p < 0.01) over the 17% inspection rate over this period among those not assigned to inspection in the directive year.[35] Column 4 reports the effect of receiving an SST inspection on DAFW injuries (treatment-on-the-treated effect), corresponding to the IV-Poisson specification in Equation 3. The average SST inspection leads to 8.7% fewer DAFW injuries per year (β = -0.091, SE = 0.042, p = 0.03). Because control establishments averaged

[35] This result is the regression-adjusted estimate of the average difference of the y-axis values between the two lines depicted in Figure A.2, which report the annual probability of having been inspected among those assigned to inspection in the directive year (solid line) and among those not assigned in the directive year (dashed line).

5.35 injuries per year over the directive year and the four following years, this estimate implies that the average SST inspection averted 2.3 DAFW injuries over the five-year period (5.35 * 5 * 8.7% = 2.3). Given our estimate (described above) that a DAFW injury cost $52,000 in our sample period, each randomized inspection averts $120,000 in injury costs over this five-year period,[36] roughly 35 times the cost of conducting an inspection.[37]

Our estimates are robust to a number of alternative specifications and other checks (presented in Appendix I). We ran our ANCOVA model in Equation 1, dropping establishments that had ever received a violation from OSHA for injury recordkeeping. We also estimated the effect of assignment to inspection using a difference-in-differences specification. Additionally, we averaged outcomes during the directive year and the following four years into a single observation, using OLS to estimate an ANCOVA model on this collapsed dataset. We also used two very different alternative models to estimate average intent-to-treat effects: targeted maximum likelihood estimation combined with Super Learner (van der Laan and Rose 2011) as well as the CDDF procedure. All results were economically similar to and statistically indistinguishable from the results reported in Table 4.

3.1.2 Average Effects on Business Outcomes

We next examine whether SST inspections had unintended effects on establishments' business outcomes. We estimate Equation 3, our instrumental variable approach, using linear models that predict establishment death, employment, total hours worked, and PAYDEX.[38] None

[36] Our estimate that SST inspections led to 8.7% fewer DAFW injuries is similar to a prior study's estimate that randomized inspections by California's Division of Occupational Safety and Health led to 9.9% fewer injuries that triggered workers' compensation claims (Levine, Toffel, and Johnson 2012). Whereas Levine, Toffel, and Johnson (2012) considered all injuries filed to workers' compensation, this study only considers DAFW injuries, a subset of workers-compensation–eligible injuries. This difference likely explains why our estimate that SST inspections have a $120,000 social benefit is much lower than Levine, Toffel, and Johnson's (2012) estimate that Cal-OSHA inspections had a $355,000 social benefit. We note that if SSST inspections also lead to a decline in non-DAFW injuries, then $120,000 substantially underestimates the social benefit of SST inspections.

[37] We estimate that it cost OSHA roughly $3,400 to conduct a typical inspection during our sample period. We derive this estimate by dividing OSHA's FY2009 federal enforcement budget of $194 million by the 37,700 inspections conducted by federal OSHA in FY2009 (US Department of Labor 2008). We assume that one-third of OSHA's enforcement budget is overhead and that SST inspections cost the same as other inspections.

[38] We originally planned to estimate the effect on establishment survival using a Cox proportional survival model. However, because it is not straightforward to estimate an instrumental variables Cox model and to be consistent with the rest of the table, we report here the linear specification. Using a Cox model to evaluate the effect on survival (not reported) yields a coefficient qualitatively similar to the intent-to-treat version of the coefficient we report.

of the coefficients on *SST inspected* are economically large or statistically significant (see Table 5).[39] Because these results provide no evidence that SST inspections harmed business outcomes, we do not analyze business outcomes when we evaluate counterfactual targeting policies.

3.2 How Heterogeneous Are Treatment Effects of Inspections—And Why?

Before assessing how many injuries OSHA could avert though alternative policies that target on $S(Z)$ or $B(Z)$, we first use our causal forest estimates to assess the distribution of CATEs among establishments on the historical SST target list, as well as the characteristics associated with high or low CATEs.

3.2.1 Heterogeneity in CATE

Figure 2 plots centiles of estimated CATEs for the establishments on the SST target lists. Each dot represents the median, across 250 sample splits, of the corresponding centile of the Best Linear Predictor of $s_0(Z)$.[40] These estimates are likely biased (Chernozhukov et al. 2018), but can give a sense of the potential heterogeneity in CATEs across the historical target lists. To ease interpretation, the figure plots the negative of the CATE estimates (that is, injuries averted).

The CATE levels increase dramatically beyond the 80th percentile: the estimated CATE for an establishment at the 90th percentile is 0.56 averted injuries per year, over three times the estimated CATE for an establishment at the 70th percentile (0.18). In contrast, the establishment at the 70th percentile has an estimated CATE only 1.8 times as large as an establishment at the median (0.10). Thus, prioritizing inspections to establishments with the largest CATEs could very likely substantially improve OSHA's effectiveness.

3.2.2 Sources of Heterogeneity in CATEs

Table 6 illustrates the association between establishments' estimated CATEs and their baseline characteristics *Z*. Following CDDF, we test whether the characteristics of

[39] The outcomes that yield point estimates with the largest negative magnitude are ODI hours and employment. While these point estimates could suggest that inspection leads to a reduction in employment among surviving establishments, the results in Column 1 suggest that being inspected slightly increases the likelihood of survival. When we estimated regression models with ODI-reported log working hours or employment as the outcome—but recoding values to 0 (rather than missing) in years when an establishment is not alive—the resulting coefficient *on SST inspected* shrinks in magnitude, essentially to 0 (results not shown).

[40] As detailed in Chernozhukov et al. (2018), the Best Linear Predictor of S(Z) is obtained from the regression coefficients in Equation G.1, as $\widehat{\beta_1} + \widehat{\beta_2} * (S - ES)$.

establishments with the highest and lowest CATEs differ. In each of our 250 iterations, we identify establishments with an $S(Z)$ in the top 20% or the bottom 20% of the combined holdout sample and nonrandomized sample. We then calculate the means of each group's characteristics. Table 6 reports the median of these top-20% and bottom-20% group means, the standard errors of the means, the difference between these means, and the p-value on this difference.

In the pre-period, establishments with the largest estimated CATEs have substantially higher DAFW injury counts and employment than those with the lowest estimated CATEs. Establishments with the largest estimated CATEs are also less likely to be nursing homes and more likely to be in the manufacturing sector.[41] In contrast, the DART rate from two years prior to the directive year—the metric OSHA used to construct SST inspection target lists—exhibits little variation between the largest- and smallest-CATE groups. Finally, establishments with the largest estimated CATEs have substantially higher predicted number of injuries in the post-period absent assignment to inspection ($B(Z)$) than do those with the lowest estimated CATEs.

3.3 **Effects of Alternative Targeting Policies**

We now estimate how different targeting rules affect the number of injuries OSHA could have averted. Each of our policy simulations maintains OSHA's rule that an establishment is ineligible for an inspection if it received one in the prior two years.[42]

3.3.1 Did OSHA's Targeting Avert as Many Injuries as Possible?

OSHA allocated its SST inspections by creating a target list of establishments with high DART injury rates two years prior and then prioritizing within these establishments by establishing a threshold that, roughly, placed the establishments with the top third of DART injury rates on the primary list and the rest on the secondary list. Among the establishments on

[41] The comparisons reported in Table 6 illustrate how a regulatory agency can use our approach not only to target inspections where they are more effective (e.g., manufacturing plants), but also to learn where they are relatively ineffective (e.g., nursing homes). OSHA's statutes provide a hint as to why its inspections would be less effective in nursing homes than in other industries. A large share of injuries in nursing homes are musculoskeletal disorders and ergonomics-related injuries, but OSHA does not have an ergonomics standard. Thus, OSHA inspectors may have less potential to facilitate improvement in this industry.

[42] We maintain this rule because our estimates of the effects of SST inspection are conditional on inspections of any particular establishment being conducted at least three years apart. Thus, because we cannot know if the treatment effect of inspections would differ if they were conducted within one or two years of each other, we do not allow for such instances in the policies we consider.

the 2001–2010 target lists that were eligible for SST inspection, OSHA assigned to inspection 43% of those on the primary lists and 10% of those on secondary lists.

OSHA could adjust two levers to develop alternative targeting policies. First, it could change the metric used to create its target lists. For example, rather than using establishments' DART rates two years prior, OSHA could focus on those establishments that it predicts will *respond* the most to inspections by reducing injuries (largest negative *S(Z)*s) or on those with the highest expected injuries in the absence of an inspection (largest *B(Z)*s). Second, OSHA could change the size and inspection probabilities of its primary and secondary lists. For example, it could create smaller primary lists of establishments that it would assign to inspection with certainty and larger secondary lists from which it could randomly assign some to inspection. OSHA could also abandon randomization altogether and instead create only primarily lists on which it assigns all listed establishments to inspection with certainty.

The first row of Table 7 reports, for comparison, the parameters and effects associated with OSHA's historical policy; specifically, the criterion to target inspections (Column 1, the DART rate from two years prior) and the number of establishments assigned to inspection from the primary and secondary lists under the 2001–2010 SST directives (Columns 2–3), which we generically refer to here as the "high-priority" and "low-priority" lists.[43] The table also reports the estimated mean treatment effect for the high- and low-priority lists (Columns 4–5), the mean number of annual injuries averted among establishments assigned under SST (Column 6),[44] and the estimated total number of injuries averted among assigned establishments over the five-year period comprising the directive year and four subsequent years (Column 7).

The next two rows consider our two benchmark policies, whereby OSHA targets with certainty a new high-priority list of the establishments with either the largest estimated CATEs ($S(Z)$), in Row 2, or the most predicted injuries ($B(Z)$), in Row 3. For both, OSHA assigns to inspection the same number of establishments each year as the SST policy did: 16,861 over the 10-year period. Targeting based on the largest estimated CATEs would have averted an average

[43] Note that the number of assignments to inspection in this row (16,861) differs from the number of establishments assigned to inspection on OSHA's 2001–2010 target lists (28,163) reported in Table A.1 in Appendix A. There are two reasons. First, for this analysis, we have excluded the 9,170 establishments on the 2001–2010 target lists without any post-period ODI data. Second, we restrict the analysis to establishments that were not SST-inspected in either of the prior two years, since they were ineligible for inspection under OSHA's rules.

[44] This estimate, also reported in Column 5 of Table A.2, is the median ($\widehat{\beta_1}$), from Equation G.1, across our 250 iterations of data partitioning.

of 0.433 (SE = 0.23) injuries per year among assigned establishments, or 36,625 injuries total—roughly 2.41 times as many injuries averted as the historical SST policy. Targeting based on the most predicted injuries would have averted 2.29 as many.

3.3.2 Variations on the Benchmark Targeting Policies

How might variations on targeting affect the expected number of injuries averted?

Maintaining OSHA's inspection budget. Establishments expected to have the greatest reduction in injuries following inspection tend have more employees than average (Table 6), as do those predicted to have the most injuries (results not shown). Because inspecting workplaces with more employees typically takes more inspector time, the agency might not have the resources to conduct our benchmark policy.

We therefore consider policies in which OSHA targets either the largest $S(Z)$ or the largest $B(Z)$, but maintains the estimated total *cost* (rather than *number*) of inspections under the historical SST policy. As a rough approximation, we model the cost of inspections as proportional to $\log_{10}$ of the establishment's full-time employees (FTEs); for example, if inspecting an establishment with 25 employees requires one day, we expect that one with 250 employees would require two days and one with 2,500 employees would require three days. The high-priority list each year includes those establishments with the largest $S(Z)$ or $B(Z)$, until the sum of log(FTEs) of these establishments equals that of those that were inspected that year under SST. Constraining total inspection costs to the historical policy's budget in this manner would reduce the number of assignments to inspection by 10% when targeting on $S(Z)$ and by and 12% when targeting on $B(Z)$. However, in both cases the estimated number of injuries averted remains essentially unchanged (Rows 4 and 5 of Table 7).[45]

Maintaining randomization. By assigning some inspections at random, the SST policy threatens a broad pool of establishments. This can motivate non-inspected establishments to

[45] The estimated number of injuries averted declines by a small amount when targeting on $S(Z)$ and actually *increases* slightly when targeting on $B(Z)$. This latter result arises, even though the number of inspections is lower in this policy than in the benchmark policy, due to the exclusion criteria we impose to mimic OSHA's rules that an establishment cannot be inspected if it was inspected in either of the prior two years. This restriction means that the set of establishments eligible each year for each policy is slightly different. If we omit this exclusion criterion, the gap reverses. In all cases, these differences are not statistically significant.

deter noncompliance and improve safety (Cohen 2000; Shimshack and Ward 2005; Gray and Shadbegian 2007).[46] Randomization also permits OSHA to continue to evaluate its effectiveness by comparing outcomes between establishments that it does and does not randomly assign to inspection. However, randomization implies that OSHA would not inspect some establishments where inspections would be expected to avert the most injuries.

To inform these tradeoffs, we consider policies whereby OSHA conducts *all* inspections randomly, maintaining the same sizes of the primary and secondary target lists as the historical SST policy (with the primary list consisting of the top 39% of eligible establishments and the secondary containing the rest), but placing establishments on these lists based on their $S(Z)$ or $B(Z)$ rather than on their DART rate from two years prior. We set the probability of assignment to inspection for the high-priority list to equal that of the historical primary list (43%) and that of the low-priority list to maintain the inspection budget of the agency's historical policy (resulting in a probability of 9%). This strategy would avert an average of 0.255 (SE = 0.133) and 0.261 (SE = 0.135) injuries per year per assigned establishment, or 1.37 and 1.39 times as many injuries as the historical SST policy, if OSHA targeted on $S(Z)$ or $B(Z)$, respectively.

These policies illustrate that OSHA could still avert more injuries aided by machine learning, even when randomizing all of its inspections and maintaining many of its historical procedures. Such a policy would ensure that OSHA could continue to learn where its inspections are most effective and could maintain general deterrence via randomization. But because randomization directs fewer inspections where they have the highest expected value, OSHA would avert substantially fewer injuries than it would with the benchmark policies.

3.4 Threats to the Validity of Our Estimates of Counterfactual Policies

Here, we address the robustness of our estimated effects of alternative targeting policies.

[46] It is worth noting that our benchmark policies that target inspections to establishments with the largest S(Z)s or B(Z)s also elicit general deterrence. Though OSHA assigns inspections deterministically under these policies, establishments would not know their estimated CATE or predicted injuries, due to the black-box nature of the machine learning algorithms. Thus, establishments would perceive some probability of being inspected, just as they would under a policy that randomizes inspections.

3.4.1 Assessing Stability of CATE Estimates over the Sample Period

Our analyses have used data from our entire 2001–2010 sample period to estimate CATEs. In reality, when OSHA targets inspections in a given year, it can access only data through the prior year. To assess whether our estimates would be materially different were they based only on data available to the agency when it was constructing its target lists, we estimate a causal forest and a Super Learner, using only the 2001–2006 randomized sample (the first half of our sample period), to estimate $S(\cdot)$ and $B(\cdot)$, respectively. We then use these models to generate predicted CATEs ($S(Z)$) and predicted baseline injuries ($B(Z)$) of establishments on the 2007–2010 randomized sample (the second half of our sample period). Encouragingly, there is a high correlation ($\rho = 0.8$) between the 2007–2010 sample's estimated CATEs when based on (a) the full sample or (b) just the earlier years (2001–2006).

To more formally assess this concern, we compare the estimated benefits of targeting inspections for the 2007–2010 target lists when we estimate $S(Z)$ and $B(Z)$ using all data (our main approach) to when we estimate $S(Z)$ and $B(Z)$ using only the earlier 2001–2006 data.

First, to estimate the average treatment effect (Table 7, Row 1) and the predicted benefits of our benchmark policies, we repeat the procedure used in our main analysis (Table 7, Rows 2 and 3), but use only the observations from the second half of our time period (2007–2010). Under these policies, the estimated number of averted injuries per year from assignment to inspection among establishments in the high-priority group is 0.744 (SE = 0.310) and 1.05 (SE = 0.321), respectively, as reported in Columns 2–3 of Table A.2 in Appendix A.

We then identify the establishments on the 2007–2010 target lists that would be in the high-priority group (those assigned to inspection with certainty) in the benchmark policies, based on the *S(Z)*s and *B(Z)*s estimated on the 2001–2006 sample. As in our main analysis, we run a regression corresponding to Equation 5 to estimate the average number of injuries averted. In contrast to our procedure in the main analysis, here we only estimate the regression once, since we are not randomly partitioning the data as we did in our main analysis. We therefore report $\widehat{\gamma_1}$ from this regression (rather than the median of $\widehat{\gamma_1}$ across 250 sample splits).

We report results in Columns 4–6 of Table A.2 in Appendix A. Column 4 reports that the estimated average treatment effect of assignment to inspection for the 2007–2010 randomized sample is -0.207 (SE = 0.109), essentially identical to what we get when $S(Z)$ and $B(Z)$ are

estimated on the whole sample (Column 1). Column 5 reports that the estimate of the average number of averted annual injuries among the establishments assigned to inspection in the benchmark policy when targeting on $S(Z)$ is -0.63 (SE = 0.35). Column 6 reports that the corresponding estimate when targeting on $B(Z)$ is -0. 87 (SE = 0.37). Each of these estimates is very similar to the corresponding estimates for the 2007–2010 target lists when $S(Z)$ and $B(Z)$ were estimated on the whole sample (Columns 2–3).

In short, these results suggest that the effects of alternative targeting policies, which we estimate using all years in our data (in Table 7), are similar to what OSHA could have produced with the data it had available each year.

3.4.2 Using the Randomized Sample to Estimate Gains to Re-targeting the Entire SST List

We use estimates of CATEs among establishments in the randomized sample to simulate the effects of counterfactual policies for the entire SST target lists. This approach implicitly assumes that our estimates of the effects of inspections on the randomized sample generalize to the nonrandomized sample:

$$E[s_0(Z) \mid G_k] = E[s_0(Z) \mid G_k, randomized = 1] = E[s_0(Z) \mid G_k, randomized = 0].$$

This assumption could fail to hold if establishments in the randomized and nonrandomized samples have different *Z*s (observable characteristics) or if the function that maps *Z* to treatment effects, $s_0(\cdot)$, differed between those two samples (if the two groups have different unobservables). In Appendix J, we provide evidence that the randomized and nonrandomized samples do not have meaningfully different observables associated with $S(Z)$ or $B(Z)$ and we do not find any evidence that our machine learning models have differential predictive power for the two samples. While we cannot rule out differences in the mapping between observable factors and treatment effects, we have no evidence that these issues affect our estimates.

4 Discussion

OSHA inspections of dangerous workplaces substantially improved workplace safety. Our estimates imply that the average inspection averted 2.4 DAFW injuries over five years, a 9% decline relative to what those establishments would have otherwise experienced. We do not find any large or statistically significant consequences of inspections on business outcomes such as survival and employment. We also find that the agency could have averted many more

injuries by targeting workplaces where the expected number of averted injuries is high—over twice as many if it had targeted establishments with the largest predicted treatment effects.

Targeting based on high expected injuries is consistent with OSHA's goal of targeting establishments with "serious safety and problems" (US OSHA 2004: 25445). Historically, OSHA measured "serious safety and health problems" based on a single year's injury rate, a measure with substantial mean reversion. Our measure of predicted injuries accounts for mean reversion and better identifies workplaces with persistent safety and health problems.

OSHA could have averted nearly as many, if not more, injuries by targeting on predicted injuries ($B(Z)$) than by targeting on estimated heterogeneous treatment effects (CATE, or $S(Z)$). This seems surprising because targeting where inspections are most valuable (that is, the true CATE) will (by construction) be the most effective. However, our estimated CATEs are noisy predictors of true CATEs.[47] In contrast, estimating the expected number of injuries absent an inspection is a much easier prediction problem, with very strong fit out of sample (see Appendix J). Table 6 showed that $S(Z)$ and $B(Z)$ are correlated. In short, in this setting, using a well-estimated proxy for CATE ($B(Z)$) will target as well or better than using a direct estimate of CATEs ($S(Z)$) that is not estimated as precisely.[48] Targeting baseline outcomes $B(Z)$ would be less useful in settings in which CATEs are easier to estimate precisely or where baseline outcomes are less correlated with the treatment effects. The extent to which these two conditions hold can illuminate which of these two criteria is more effective for targeting inspections.

Alternatively, if the agency created target lists that prioritized establishments with large predicted treatment effects but then *randomized* inspections among these lists in a way that mimicked its historical procedures, it would still avert more injuries than the historical policy did, but by a much smaller amount. This illustrates the tradeoff associated with randomizing inspections. Though randomizing ensures that the agency can both keep learning where it is

[47] One gauge of the extent to which our S(Z) are accurate estimates of establishments' underlying CATEs, $s_0(Z)$, is given by the estimated coefficient $\widehat{\beta_2}$ from Equation G.1. If S(Z) is a perfect proxy for $s_0(Z)$, then this coefficient is 1; if the estimates are complete noise, then the coefficient is 0 (Chernozhukov et al. 2018). Across our 250 sample splits, the median $\widehat{\beta_2}$ from this regression is 1.5—indicating that our approach yields an S(Z) that is a meaningful, slight under-estimate for underlying CATEs. However, the median SE of $\widehat{\beta_2}$ is 0.8, indicating that the estimate exhibits substantial sampling variation.

[48] We also ran our causal forest models adding B(Z) to the set of covariates in *Z*. This addition had essentially no effect on our estimates of the injuries averted under the policies targeting on S(Z).

effective and maintain general deterrence, it also leaves uninspected many establishments for which inspections would most effectively achieve the agency's goal. Using these policies as bounds is informative, as an agency could experiment with nonrandom targeting of establishments with the highest predicted benefits while randomizing the rest.

OSHA can also benefit from learning where its inspections are relatively *ineffective.* For example, we found that inspections of nursing homes—an industry with very high injury rates—avert fewer injuries than those in other sectors. OSHA could investigate why and attempt to improve. For example, OSHA has no standard for musculoskeletal diseases, which account for a large share of injuries in nursing homes. If that omission were responsible for OSHA's lack of effectiveness at nursing homes, this finding might help improve regulations.

Our study has several limitations. We do not consider effects of inspections beyond five years, we do not measure effects on workplace illnesses or on injuries that do not result in days away from work, and we cannot say anything about injuries sustained by temporary or contract workers, as their injuries are not recorded in ODI. Finally, it is unclear how well our results generalize to the 21 states that operate their own occupational safety programs.

With these limitations in mind, we show that combining randomization and machine learning could substantially improve regulatory agencies' performance. This approach could improve the effectiveness of many other organizations ranging from regulatory agencies such as the US Internal Revenue Service and the US Food and Drug Administration to accounting firms targeting audits and multinational firms assessing suppliers' production processes and product quality. Moreover, our study provides guidance to the nascent practice of regulatory agencies targeting inspections in part based on algorithms. For example, the US Food and Drug Administration targets inspections of foreign food manufacturers based on predicted risk of producing contaminated food (US Government Accountability Office 2016). In 2018, the US Bureau of Safety and Environmental Enforcement began targeting certain inspections of offshore oil and gas operations based on perceived risks of noncompliance and of safety incidents (US Bureau of Safety and Environmental Enforcement 2018). Chicago has begun using risk-based forecasting to help determine the order in which it inspects restaurants (Spector 2016). Our research reveals how agencies can estimate the relative benefits of alternative algorithms that vary in simplicity and transparency as well as in general deterrence to encourage compliance among non-inspected establishments.

5 References

Alm, James, and Jay Shimshack. 2014. "Environmental Enforcement and Compliance: Lessons from Pollution, Safety, and Tax Settings." *Foundations and Trends in Microeconomics* 10 (4): 209–274.

Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355: 483–485.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Avis, Eric, Claudio Ferraz, and Frederico Finan. 2016. "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians." NBER Working Paper No. w22443.

Azaroff, Lenore S., Charles Levenstein, and David H. Wegman. 2002. "Occupational Injury and Illness Surveillance: Conceptual Filters Explain Underreporting." *American Journal of Public Health* 92 (9): 1421–1429.

Bartel, Ann P., and Lacy Glenn Thomas. 1985. "Direct and Indirect Effects of Regulation: A New Look at OSHA's Impact." *Journal of Law and Economics* 28 (1): 1–25.

Biddle, Jeff, and Karen Roberts. 2003. "Claiming Behavior in Workers' Compensation." *Journal of Risk and Insurance* 70 (4): 759–780.

Blundell, W., Gautam Gowrisankaran, and Ashley Langer. 2018. "*Escalation of Scrutiny: The Gains from Dynamic Enforcement of Environmental Regulations* (No. w24810)." National Bureau of Economic Research.

Boden, Leslie I., Nicole Nestoriak, and Brooks Pierce. 2010. "Using Capture-recapture Analysis to Identify Factors Associated with Differential Reporting of Workplace Injuries and Illnesses." *2010 JSM Proceedings, Statistical Computing Section* (Alexandria, VA: American Statistical Association).

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2019. "Regression Discontinuity Designs Using Covariates." *Review of Economics and Statistics* 101 (3): 442–451.

Cattaneo, Matias D. Michael Jansson, and Xinwei Ma. 2018. "Manipulation Testing Based on Density Discontinuity." *Stata Journal* 18 (1): 234–261.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments." NBER Working Paper No. 24678.

Choi, Jinkyung, and Barbara Almanza. 2012. "Health Inspectors' Perceptions of the Words Used to Describe Violations." *Food Protection Trends* 32 (1): 26–33.

Cohen, Mark A. 2000. "Empirical Research on the Deterrence Effect of Environmental Monitoring and Enforcement." *Environmental Law Reporter* 30: 10245–10252.

Davis, Jonathan M. V., and Sara B. Heller. Forthcoming. "Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs." *Review of Economics and Statistics*.

Dong, Xiuwen S., Alissa Fujimoto, Knut Ringen, Erich Stafford, James W. Platner, Janie L. Gittleman, and Xuanwen Wang. 2011. "Injury Underreporting among Small Establishments in the Construction Industry." *American Journal of Industrial Medicine* 54 (5): 339–349.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2018. "The Value of Regulatory Discretion: Estimates from Environmental Inspections in India." *Econometrica* 86 (6): 2123–2160.

Erel, Isil, Lea Henny Stern, Chenhao Tan, and Michael S. Weisbach. 2018. "Selecting Directors Using Machine Learning." NBER Working Paper No. w24435.

ERG. 2013. "Analysis of OSHA's National Emphasis Program on Injury and Illness Recordkeeping (RK NEP)." Prepared for the US Occupational Safety and Health Administration's Office of Statistical Analysis [contract J-099-F-2-8441], November 1, 2013.

ERG and National Opinion Research Center. 2009. "OSHA Data Initiative Collection Quality Control: Analysis of Audits on CY 2006 Employer Injury and Illness Recordkeeping." Prepared for the U.S. Occupational Safety and Health Administration's Office of Statistical Analysis [contract no. J-099-F-2-8441], November 25, 2009, https://www.reginfo.gov/public/do/DownloadDocument?objectID=14894901

Feldman, Justin. 2011. *OSHA Inaction: Onerous Requirements Imposed on OSHA Prevent the Agency from Issuing Lifesaving Rules*. Washington, D.C.: Public Citizen's Congress Watch.

Foley, Michael, Z. Joyce Fan, Eddy Rauser, and Barbara Silverstein. 2012. "The Impact of Regulatory Enforcement and Consultation Visits on Workers' Compensation Claims Incidence Rates and Costs, 1999–2008." *American Journal of Industrial Medicine* 55: 976–990.

Glaeser, Edward L., Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy." *American Economic Review* 106 (5): 114–118.

Gonzalez-Lira, Andres, and Ahmed M. Mobarak. 2019. "Slippery Fish: Enforcing Regulation under Subversive Adaptation" Institute for the Study of Labor [IZA] no. 12179.

Gray, Wayne B., and John M. Mendeloff. 2005. "The Declining Effects of OSHA Inspections on Manufacturing Injuries, 1979–1998." *ILR Review* 58 (4): 571–587.

Gray, Wayne B., and John T. Scholz. 1993. "Does Regulatory Enforcement Work? A Panel Analysis of OSHA Enforcement." *Law and Society Review* 27 (1): 177–214.

Gray, Wayne B., and Ronald J. Shadbegian. 2007. "The Environmental Performance of Polluting Plants: A Spatial Analysis." *Journal of Regional Science* 47 (1): 63–84.

Hanna, Rema Nadeem, and Paulina Oliva. 2010. "The Impact of Inspections on Plant-level Air Emissions." *BE Journal of Economic Analysis & Policy* 10 (1): Article 19.

Haviland, Amelia M., Wayne B. Gray, John Mendeloff, Rachel M. Burns, and Teague Ruder. 2012. "A New Estimate of the Impact of OSHA Inspections on Manufacturing Injury Rates, 1998–2005." *American Journal of Industrial Medicine* 55 (11): 964–975.

Hino, M., Benami, E. and Brooks, N. 2018. "Machine Learning for Environmental Monitoring." *Nature Sustainability* 1 (10): 583–588.

Ibanez, Maria R., and Michael W. Toffel. Forthcoming. “How Scheduling Can Bias Quality Assessment: Evidence from Food Safety Inspections.” *Management Science*.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. “Human Decisions and Machine Predictions.” *Quarterly Journal of Economics* 133 (1): 237–293.

Kleven, Henrik Jacobsen, Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez, 2011. “Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark.” *Econometrica* 79 (3): 651–692.

Lee, Jonathan M., and Laura O. Taylor. Forthcoming. “Randomized Safety Inspections and Risk Exposure on the Job: Quasi-experimental Estimates of the Value of a Statistical Life.” *American Economic Journal: Economic Policy* 11 (4): 350–374.

Leigh, J. P. 2011. “Economic Burden of Occupational Injury and Illness in the United States.” *Milbank Quarterly* 89(4): 728–772.

Levine, David I., Michael W. Toffel, and Matthew S. Johnson. 2012. “Randomized Government Safety Inspections Reduce Worker Injuries with No Detectable Job Loss.” *Science* 336 (6083): 907–911.

Li, Ling, and Perry Singleton. 2019. “The Effect of Workplace Inspections on Worker Safety.” *ILR Review* 72 (3): 718–748.

Lucas Jr, Robert E. 1976. “Econometric Policy Evaluation: A Critique.” *Carnegie-Rochester Conference Series on Public Policy* 1: 19–46.

McKenzie, David. 2012. “Beyond Baseline and Follow-up: The Case for More T in Experiments.” *Journal of Development Economics* 99 (2): 210–221.

Messiou, Eleni, and Brian Zaidman. 2005. “Comparing Workers’ Compensation Claims and OSHA Data Initiative Cases.” St. Paul, MN: Minnesota Department of Labor and Industry.

Musick, Tom, Sarah Trotto, and Kyle Morrison. 2016. “Compliance Assistance—Not Fines—Should Be Priority, Senators Tell OSHA.” *Safety & Health* 193 (4): 10.

Nestoriak, Nicole, and Brooks Pierce. 2009. “Comparing Workers’ Compensation Claims with Establishments' Responses to the SOII.” *Monthly Labor Review* 132 (5): 57–64.

Oleinick, Arthur, Jeremy V. Gluck, and Kenneth E. Guire. 1995. “Establishment Size and Risk of Occupational Injury.” *American Journal of Industrial Medicine* 28 (1): 1–21.

Peto, Balint, Laura Hoesly, George Cave, David Kretch, and Ed Dieterle. 2016. “Evaluation of the Occupational Safety and Health Administration’s Site-Specific Targeting Program—Final Report.” Washington, D.C.: Summit Consulting LLC. https://www.dol.gov/asp/evaluation/completed-studies/SST_Evaluation_Final_Report.pdf.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66 (5): 688–701.

Rubin, Richard. 2017. “IRS Audits of Individuals Drop for Fifth Straight Year.” *Wall Street Journal*, Feb. 22.

Rudin, Cynthia, Rebecca J. Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac. 2010. “A Process for Predicting Manhole Events in Manhattan.” *Machine Learning* 80 (1): 1–31.

Ruser, John W. 1995. “Self-correction versus Persistence of Establishment Injury Rates.” *Journal of Risk and Insurance* 62 (1): 67–93.

Ruser, John W. 2008. “Examining Evidence on Whether BLS Undercounts Workplace Injuries and Illnesses.” *Monthly Labor Review* 131 (8): 20−32.

Ruser, John W., and Robert S. Smith. 1991. “Re-estimating OSHA's Effects: Have the Data Changed?” *The Journal of Human Resources* 26 (2): 212−235.

Shimshack, Jay P. 2014. “The Economics of Environmental Monitoring and Enforcement.” *Annual Review of Resource Economics* 6 (1): 339–360.

Shimshack, Jay P., and Michael B. Ward. 2005. “Regulator Reputation, Enforcement, and Environmental Compliance.” *Journal of Environmental Economics and Management* 50 (3): 519–540.

Short, Jodi L., Michael W. Toffel, and Andrea R. Hugill. 2016. “Monitoring Global Supply Chains.” *Strategic Management Journal* 37 (9): 1878–1897.

Slemrod, Joel, Marsha Blumenthal, and Charles Christian. 2001. “Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota.” *Journal of Public Economics* 79 (3): 455–483.

Smith, Rosen S. 1979. “The Impact of OSHA Inspections on Manufacturing Injury Rates.” *Journal of Human Resources* 14 (2): 145–170.

Spector, Julian. 2016. “Chicago Is Predicting Food Safety Violations. Why Aren’t Other Cities?” CityLab website, January 7, 2016, https://www.citylab.com/solutions/2016/01/chicago-is-predicting-food-safety-violations-why-arent-other-cities/422511/, accessed February 2020.

Stigler, George J. 1971. “The Theory of Economic Regulation.” *Bell Journal of Economics and Management Science* 2 (1): 3–21.

Telle, Kjetil. 2013. “Monitoring and Enforcement of Environmental Regulations: Lessons from a Natural Field Experiment in Norway.” *Journal of Public Economics* 99: 24–34.

U.S. Bureau of Labor Statistics. 2007. “Nonfatal Occupational Injuries and Illnesses Requiring Days Away from Work, 2005.” https://www.bls.gov/iif/oshwc/osh/case/osnr0027.pdf.

U.S. Bureau of Labor Statistics. 2015. “Industry Injury and Illness Data.” http://www.bls.gov/iif/oshsum.htm#10Supplemental_News_Release_Tables.

U.S. Bureau of Safety and Environmental Enforcement. 2018. “BSEE Launches Risk-Based Inspection Program.” March 12. https://www.bsee.gov/newsroom/latest-news/statements-and-releases/press-releases/bsee-launches-risk-based-inspection.

U.S. Department of Health and Human Services. 2011. “Fiscal Year 2012 Food and Drug Administration, Justification of Estimates for Appropriations Committees.” https://oig.hhs.gov/publications/docs/budget/FY2012_HHSOIG_Congressional_Justification.pdf.

U.S. Department of Labor. 2008. "Congressional Budget Justification: Occupational Safety and Health Administration, FY 2009." https://www.dol.gov/dol/budget/2009/PDF/CBJ-2009-V2-08.pdf.

U.S. Food and Drug Administration. 2016."2016 Annual Report on Inspections of Establishments in FY 2015." https://www.fda.gov/downloads/RegulatoryInformation/LawsEnforcedbyFDA/SignificantAmendmentstotheFDCAct/FDASIA/UCM483994.pdf.

U.S. Government Accountability Office. 2016. "FDA's Targeting Tool Has Enhanced Screening, But Further Improvements Are Possible." http://www.gao.gov/assets/680/677538.pdf.

U.S. OSHA. 2004. "Nationwide Site-Specific Targeting (SST) Inspection Program Request for Comments." *Federal Register* 69 (88): 25445–25446.

U.S. OSHA. 2008. "Site-Specific Targeting 2008 Directive 08-03 (CPL 02), Effective Date May 19, 2008." https://www.osha.gov/OshDoc/Directive_pdf/CPL_02_08-03.pdf, accessed March 2019.

U.S. OSHA. 2016. "OSHA Field Operations Manual, Directive CPL-02-00-160, Effective Date August 2, 2016." https://www.osha.gov/OshDoc/Directive_pdf/CPL_02-00-160.pdf, accessed March 2018.

U.S. OSHA. 2017a. "Commonly Used Statistics." , https://www.osha.gov/oshstats/commonstats.html, accessed February 2017.

U.S. OSHA. 2017b. "OSHA Fact Sheet: OSHA Inspections." https://www.osha.gov/OshDoc/data_General_Facts/factsheet-inspections.pdf, accessed March 2018.

van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1): Article 25.

van der Laan, Mark J., and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer Science & Business Media.

Viscusi, W. Kip. 1986. "The Impact of Occupational Safety and Health Regulation, 1973–1983." *Bell Journal of Economics* 17 (4): 567–580.

Waehrer, Geetha M., Xiuwen S. Dong, Ted Miller, Elizabeth Haile, and Yurong Men. 2007. "Costs of Occupational Injuries in Construction in the United States." *Accident Analysis & Prevention* 39 (6): 1258–1266.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

Weisman, Jonathan, and Matthew L. Wald. 2013. "I.R.S. Focus on Conservatives Gives G.O.P. an Issue to Seize On." *New York Times*, May 12. https://www.nytimes.com/2013/05/13/us/politics/republicans-call-for-irs-inquiry-after-disclosure.html.

Table 1: Industry Tabulation for SST Target Lists and Randomized Sample Subset

	(1)	(2)	(3)	(4)
	All SST target lists		Randomized sample	
	Count	% of total	Count	% of total
Agriculture, forestry, fishing	764	0.8%	150	0.9%
Mining	59	$< .01\%$	4	$< .01\%$
Construction	38	$< .01\%$	4	$< .01\%$
Manufacturing	50,349	54.8%	9,148	56.7%
Wholesale trade	10,436	11.4%	1,770	11.0%
Retail trade	4,827	5.3%	863	5.3%
Transportation, warehousing	14,894	16.2%	2,291	14.2%
Other services	1,173	1.3%	221	1.4%
Nursing homes	9,291	10.1%	1,690	10.5%
Number of establishment-directives	91,831		16,141	

An establishment-directive corresponds to a unique instance of an establishment being included on an annual SST target list from 2001 to 2010 (some establishments are included in multiple years' target lists). The sample in Columns 1 and 2 includes the entire 2001–2010 SST target lists in states under federal OSHA jurisdiction. The subsample in Columns 3 and 4 includes the subset of establishment-directives on the SST target lists that are included in our randomized sample, as described in Table A.1.

Table 2: Summary Statistics for the Randomized Sample, +/- 4 Years from Directive Year

	n	mean	sd	median	min	max
Number of times on prior SST target list	143,757	1.2	1.6	1.0	0.0	9.0
Injuries with Days Away, Restricted or Transferred / 100 FTE[a]	90,343	7.6	5.0	6.7	0.0	23.6
Injuries with Days Away from Work / 100 FTE[a]	90,343	3.9	3.6	3.1	0.0	16.6
Number of Days Away from Work (DAFW) injuries	90,343	6.5	8.9	4.0	0.0	54.0
Total hours worked, 000s [ODI]	90,346	284.8	331.2	183.0	0.0	2369.8
Average number of employees [ODI]	90,346	149.5	176.8	96.0	1.0	1257.0
Number of employees [NETS]	137,566	135.1	153.1	89.0	1.0	1000.0
Minimum PAYDEX score [NETS][b]	125,794	67.7	10.7	70.0	2.0	96.0
Number of OSHA inspections in calendar year	143,757	0.2	0.5	0.0	0.0	3.0
Number of SST inspections in calendar year	143,757	0.1	0.3	0.0	0.0	2.0

The sample consists of the 16,141 establishment-directives on the 2001–2010 annual SST target lists included in our randomized sample. Establishment-directive refers to a specific instance of an establishment being on an annual SST target list. The criteria for the randomized sample are summarized in Table A.1.
The table includes data from a 9-year window, consisting of the 4 years prior to the directive year (the year the establishment was placed on the target list), the directive year, and the four years following the directive year. Variables from ODI are only observed in years in which an establishment was included in the ODI survey. NETS variables are observed for all years that an establishment reports to Dun & Bradstreet being in operation.
All unbounded variables are top-coded at their 99th percentiles.
[a] FTE (full-time employees) is calculated as the total number of hours worked divided by 2,000 (the number of hours a full-time employee would work in a year).
[b] PAYDEX is a monthly score ranging 0–100 assigned to an establishment by Dun & Bradstreet to reflect the speed with which an establishment pays its creditors, with higher scores reflecting faster payment. Min PAYDEX is the establishment's minimum score over all monthly reports in a year. This variable is missing when Dun & Bradstreet lacks sufficient payment information to create a score.

Table 3: Balance Tests on Baseline Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			SST and ODI variables (in year t-2)					NETS variables (in year t-2)	
	# OSHA inspections t-1 though t-4[a]	# times previously on SST target list[b]	DART[c] injuries per 100 FTE[e]	DAFW[d] injuries per 100 FTE[e]	# of DAFW injuries	log (total hours worked)	log (# employees)	log (# employees)	Minimum PAYDEX score[f]
Assigned to inspection	0.019	-0.0021	-0.036	0.019	0.24	0.025	0.023	0.018	-0.27
	(0.016)	(0.025)	(0.056)	(0.059)	(0.17)	(0.012)*	(0.012)*	(0.015)	(0.20)
# observations	16,141	16,141	16,141	16,141	16,141	16,141	16,141	15,021	14,265
# assigned to inspection	6,977	6,977	6,977	6,977	6,977	6,977	6,977	6,506	6,158
# not assigned to inspection	9,164	9,164	9,164	9,164	9,164	9,164	9,164	8,515	8,107
# Area-Office–directives	383	383	383	383	383	383	383	383	383
Mean dep var, estabs not assigned	.522	1.22	10.3	5.4	8.16	12.4	4.89	4.6	67.6

This table reports results of OLS regression models that regress the dependent variable indicated in the header row on an *Assigned to inspection* dummy and Area-Office–directve fixed effects, within the randomized sample. The unit of analysis is the establishment-directive. The reported coefficient on *Assigned to inspection* is an estimate of the level change in the dependent variable associated with being assigned to SST inspection within the randomized sample. Standard errors clustered by establishment are reported in parentheses. **p<.01, *p<.05, +p<.1.

The ODI and NETS variables are from two years prior to the directive year. *# times previously on SST target list* is evaluated as of the directive year.

The sample includes all establishments eligible for randomized SST inspections in Area-Office–years that randomized their target lists, as described in Table A.1.

[a] OSHA inspections include those triggered by an incident (i.e., a serious accident, complaint, or referral) or pre-planned via one of OSHA's programs (including SST). The variable used is the number of inspections in the 4 years prior to the directive year.

[b] *# times previously on SST target list* is the number of years the establishment has appeared on the SST target list prior to the directive year.

[c] DART injuries refer to injuries that result in days away from work, job restriction, or job transfer.

[d] DAFW injuries refer to injuries that result in days away from work.

[e] Injury rate variables are calculated by multiplying the number of injuries in a calendar year by 20,000, and then dividing that by that year's hours worked. To mitigate the influence of outliers, the numerator and denominator are each top-coded at the 99th percentile and the first percentile of hours worked is added to the denominator.

[f] PAYDEX is a monthly score, ranging 0–100, assigned to an establishment by Dun & Bradstreet to reflect the speed with which an establishment pays back its creditors, with higher scores reflecting faster payment. Min PAYDEX is the minimum score over all monthly reports in a year. This variable is missing when Dun & Bradstreet lacks sufficient payment information to create a score.

Table 4: Effects of SST Inspection on Injuries Resulting in Days Away from Work (DAFW)

	(1)	(2)	(3)	(4)
	# of DAFW injuries		SST inspected	# of DAFW injuries
	Intent-to-treat		(First stage)	Treatment-on treated
Assigned to inspection	-0.035	-0.037	0.458	
	(0.017)*	(0.017)*	(0.006)**	
SST-inspected				-0.091
				(0.042)*
log(hours) in t-2		0.258		0.255
		(0.019)**		(0.019)**
# observations	40,993	40,993	40,993	40,993
# establishment-directives	13,736	13,736	13,736	13,736
# establishments	11,083	11,083	11,083	11,083
# Area-Office–directives	383	383	383	383
Mean dep var, estabs not assigned	5.35	5.35	0.17	5.35
Specification	Poisson	Poisson	OLS	IV-Poisson

All regressions include Area-Office–directive and tau-year (number of years since the directive year) fixed effects. Each regression also controls for the mean of the establishment's dependent variable (or log(1+ dependent variable) in Poisson regressions) over the 4 years prior to the directive year and for the number of years over which this baseline mean is calculated. SEs, in parentheses, are clustered by establishment. +p<.1, *p<.05, **p<.01.

Regressions restricted to randomized sample, described in Table A.1, and a 5-year window of the directive year and 4 years following.

Columns 1–2 report Poisson regression estimates of the effect of being assigned to SST inspection on an establishment's annual number of DAFW injuries, which are intent-to-treat estimates. Column 3 reports an OLS estimate of the increased probability (in percentage points) that establishments assigned to SST inspection in the directive year actually received an SST inspection. Column 4 reports the IV-Poisson regression estimate of the effect of receiving an SST inspection on an establishment's annual number of DAFW injuries.

Table 5: Effects of SST Inspections on Business Outcomes: IV Results

	(1)	(2)	(3)	(4)	(5)
		ODI variables[a]		NETS variables[b]	
	Establishment dies [NETS][c]	Log(# employees)	Log(total hours worked)	Log(# employees)	Min PAYDEX score[d]
SST-inspected	-0.006	-0.015	-0.014	-0.005	-0.027
	(0.006)	(0.011)	(0.012)	(0.016)	(0.281)
# observations	79,193	40,993	40,993	70,257	67,639
# establishment-directives	16,141	13,736	13,736	15,751	14,973
# establishments	13,029	11,083	11,083	12,754	12,111
# Area-Office–directives	383	383	383	383	383
Mean dep var, estabs not assigned	0.05	4.88	12.42	4.61	67.84
Specification	OLS	OLS	OLS	OLS	OLS

The table shows the results of instrumental-variable linear regressions in which *SST-inspected* is instrumented with *Assigned to inspection*. All regressions include Area-Office-directive and tau-year (number of years since the directive year) fixed effects. Each regression also controls for the mean of the establishment's dependent variable over the 4 years prior to the directive year and the number of years over which this baseline mean is calculated. SEs, in parentheses, clustered by establishment. +p<.1, *p<.05, **p<.01.

Regressions restricted to analysis sample, described in Table A.1, and to a window of the directive year and 4 years following.

SST inspected is a dummy equal to 1 in years beginning with the directive year for establishments that have received an SST inspection in or prior to the corresponding calendar year, and 0 otherwise.

[a] Variables representing measures reported by establishments to OSHA via the ODI Survey.

[b] Variables representing measures in the NETS database based on Dun & Bradstreet data.

[c] A dummy equal to 1 if, according to the NETS database, the establishment has ceased being in operation during or prior to the current year.

[d] PAYDEX is a monthly score, ranging 0–100, assigned to an establishment by Dun & Bradstreet to reflect the speed with which an establishment pays back its creditors, with higher scores reflecting faster payment. Min PAYDEX is the minimum score over all monthly reports in a year.

Table 6: Differences in Characteristics among Establishments with Estimated CATEs in the Top and Bottom 20% of the Distribution

	(1) Establishments with estimated CATE in: Top 20%	(2) Bottom 20%	(3) Absolute difference	(4) Percent difference*
DART rate t-2	12.547	11.532	1.115	8.8%
	(0.048)	(0.042)	(0.063)	
			[0.000]	
DAFW count averaged t-1 to t-4	19.496	4.778	14.422	308.0%
	(0.110)	(0.025)	(0.113)	
			[0.000]	
# employees [NETS]	375.569	103.096	279.63	264.3%
	(6.515)	(0.350)	(6.525)	
			[0.000]	
State-year leave-one-out-mean DAFW injury rate t-2	3.009	2.896	0.12	3.9%
	(0.008)	(0.007)	(0.010)	
			[0.000]	
Nursing homes	0.08	0.144	-0.056	-44.4%
	(0.002)	(0.003)	(0.004)	
			[0.000]	
Manufacturing	0.551	0.473	0.081	16.5%
	(0.004)	(0.004)	(0.006)	
			[0.000]	
Injuries with other recordable cases / 100 FTE in t-2	5.284	3.821	1.503	38.3%
	(0.040)	(0.034)	(0.053)	
			[0.000]	
ln(Total days away from work) in t-2	5.812	3.928	1.902	48.0%
	(0.012)	(0.012)	(0.017)	
			[0.000]	
Any fatal injuries, t-2	0.014	0.006	0.008	133.3%
	(0.001)	(0.001)	(0.001)	
			[0.000]	
Standalone firm t-1	0.253	0.39	-0.137	-35.1%
	(0.003)	(0.004)	(0.005)	
			[0.000]	
Establishment age t-1	29.135	27.143	1.886	7.3%
	(0.238)	(0.207)	(0.314)	
			[0.000]	
Minimum PAYDEX score [NETS] in t-2	67.431	67.763	-0.505	-0.5%
	(0.074)	(0.075)	(0.106)	
			[0.000]	
Establishment has ever been inspected prior to this year	0.687	0.505	0.181	36.0%
	(0.004)	(0.004)	(0.006)	
			[0.000]	
Establishment had a complaint inspection in t-1 through t-3	0.201	0.071	0.131	183.1%
	(0.003)	(0.002)	(0.004)	
			[0.000]	
Has ODI data in t-1	0.909	0.896	0.008	1.5%
	(0.002)	(0.002)	(0.003)	
			[0.000]	
Has ODI data in t-3	0.808	0.713	0.094	13.3%
	(0.003)	(0.004)	(0.005)	
			[0.000]	
Number of times previously on SST target list	1.872	1.281	0.6	46.1%
	(0.017)	(0.013)	(0.021)	
			[0.000]	
B(Z)	12.779	2.955	9.664	332.5%
	(0.073)	(0.014)	(0.074)	
			[0.000]	

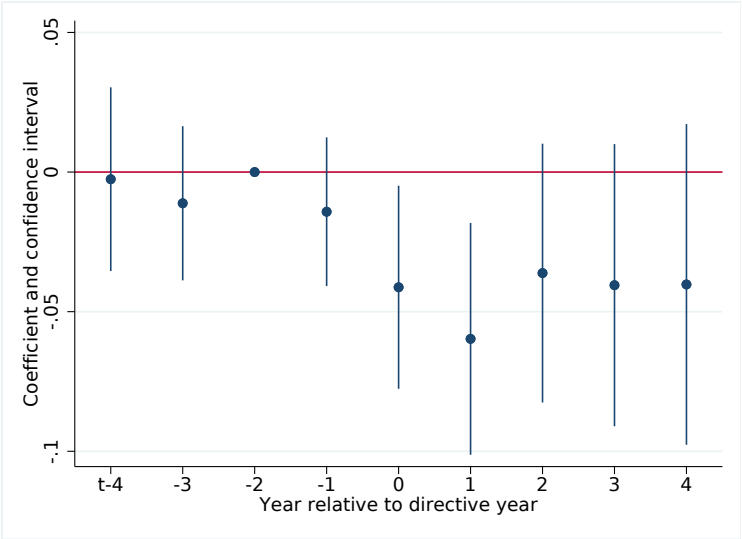
We conduct 250 random even splits of the randomized sample. In each iteration we train a causal forest on the auxiliary sample to predict CATE for estabishments in the holdout and nonrandomized samples. Among establishments in the holdout and nonrandoized samples, we identify those with the top 20% and bottom 20% of CATEs and calculate the means of the characteristic in each row for each of those two groups. Column 1 reports the medians of these 250 means for the top-20% groups with standard errors in parentheses. Column 2 reports these for the 250 bottom-20% groups. We also calculate the difference of these two means in each iteration. Column 3 reports the median of these 250 differences, with standard errors in parentheses and the p-values on a two-tailed t-test in brackets. See Section 3.3 for further information on how these sample splits and CATE estimates are obtained.

Table 7: Number of Injuries OSHA Would Avert under Alternative Targeting Policies

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Targeting criterion	Number of establishments assigned to inspection on the...		Estimated Group Average Treatment Effect for establishments on the...		Average number of annual injuries averted per establishment among assigned	Total number of injuries averted over five years
		...High-priority list	...Low-priority list	...High-priority list	...Low-priority list		
Historical SST policy	DART rate t-2	12,458	4,403	-0.143 (0.156)	-0.254 (0.188)	-0.180 (0.120)	15,170 (10,102)
Inspect highest S(Z) or B(Z), preserving the historical number of inspections	S(Z)	16,861	0	-0.433 (0.230)	-0.105 (0.130)	-0.433 (0.230)	36,520 (19,431)
	B(Z)	16,861	0	-0.411 (0.232)	-0.089 (0.127)	-0.411 (0.232)	34,665 (19,572)
Inspect highest S(Z) or B(Z), preserving the historical cost of inspections	S(Z)	15,132	0	-0.450 (0.245)	-0.113 (0.127)	-0.450 (0.245)	34,026 (18,516)
	B(Z)	14,915	0	-0.486 (0.248)	-0.072 (0.123)	-0.486 (0.248)	36,223 (18,488)
Preserve size and Pr(inspection) of lists from historical policy, preserving the historical cost of inspections	S(Z)	10,084	6,195	-0.340 (0.195)	-0.096 (0.155)	-0.255 (0.133)	20,738 (10,859)
	B(Z)	9,968	6,169	-0.367 (0.200)	-0.079 (0.150)	-0.261 (0.136)	21,063 (10,960)

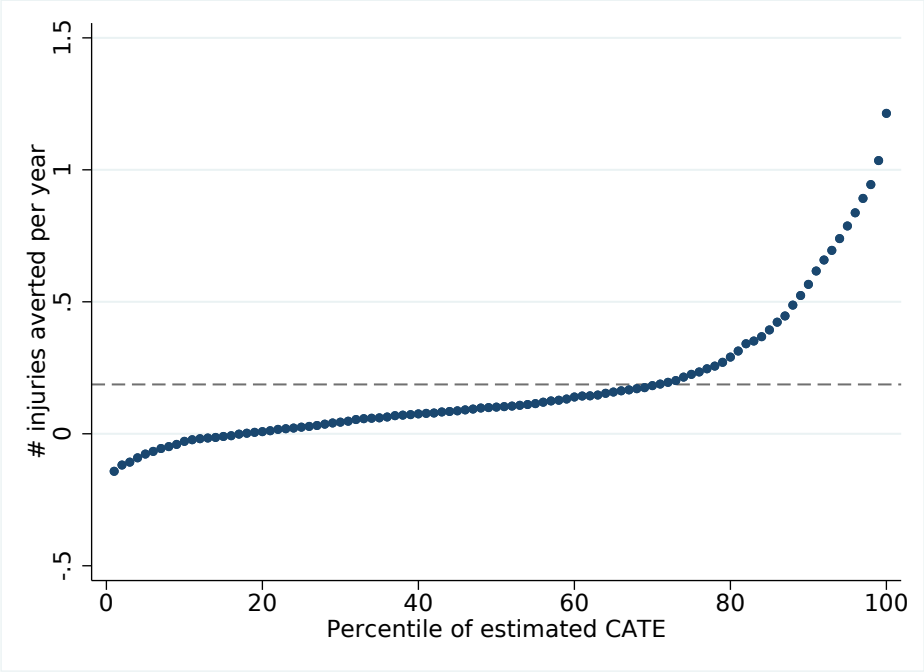
DART rate t-2 = DART rate from 2 years prior to the directive year.
S(Z) = estimate of difference of establishment's number of annual DAFW injuries if or if not assigned to inspection (i.e., its CATE).
B(Z) = estimate of establishment's number of annual injuries if not assigned to inspection.
The estimates in Columns 4 and 5 correspond to the gamma coefficients, specified in Equation 5 in the text, to estimate Group Average Treatment Effects. Each reported estimate (and standard error in parentheses below) is the median coefficient across 250 random splits of the randomized sample. See Section 3.3 for details.
The estimate in Column 7 is the number of establishments assigned to inspection (the sum of Columns 2 and 3), multiplied by the average treatment effect among assigned establishments (Column 6), multipled by 5 (the window of years over which we estimate the effects of assignment to inspection).

Figure 1: Temporal Effects of Assignment to Inspection on Injuries, by Year Relative to Directive Year



Results from a distributed lag intent-to-treat regression specification (corresponding to Equation 2 in the text) with the dependent variable equal to the number of DAFW injuries (those resulting in Days Away from Work) an establishment experiences in a year. Each dot is a coefficient on *Assigned to inspection* interacted with a dummy for each corresponding tau year, with a 95% confidence interval. The omitted year is t-2 (two years prior to the directive year).

Figure 2: Percentiles of Estimated Number of Injuries Averted per Year If Assigned to SST Inspection



Estimates of the percentiles of Conditional Average Treatment Effect (CATE), or S(Z), of assignment to SST inspection among the set of establishments on OSHA's SST target lists from 2001–2010. Each dot represents the median, across 250 sample splits, of the corresponding centile of the Best Linear Predictor of S(Z), based on the method in Chernozhukov et al. (2017). See Section 4.2.1 for details. The dashed horizontal line is the mean estimated CATE in the sample (0.18).

A Appendix Tables and Figures

Table A.1: Pipeline from Full SST Target Lists to the Randomized Sample

	Primary list				Secondary list				Target list (Pri & Sec)			
	Assigned to inspection		Not assigned		Assigned to inspection		Not assigned		Assigned to inspection		Not assigned	
	#	% of total	#	% of total	#	% of total	#	% of total	#	% of total	#	% of total
Number of establishment-directives[a] on 2010 SST target lists and...												
In states under federal OSHA jurisdiction[b]	20,708	100.0	13,314	100.0	7,455	100.0	50,349	100.0	28,163	100.0	63,663	100.0
Restrict to Area-Office–directives that randomized lists												
On Primary list that was started but not exhausted [c]	8,084	39.0	10,063	75.6	.	.	.	.				
Primary exhausted, and on Secondary started but not exhausted [c]	.	.	.	.	5,150	69.1	7,472	14.8				
Located in an Area-Office–directive that randomized (overall)	.	.	.	.	.	.	.	.	13,234	47.0	17,535	27.5
Restrict to establishments eligible for SST inspection												
Not subject to deletion criteria [d]	6,589	31.8	7,747	58.2	4,145	55.6	5,234	10.4	10,734	38.1	12,981	20.4
Drop establishments targeted for concerns with ODI reporting quality												
Has non-missing ODI data in directive year [e]	6,363	30.7	7,474	56.1	4,032	54.1	5,094	10.1	10,395	36.9	12,568	19.7
DART/DAFW meets selection criteria for corresponding list [e]	5,935	28.7	7,099	53.3	3,923	52.6	5,003	9.9	9,858	35.0	12,102	19.0
Cross-checks that establishment exists												
Found in NETS	5,813	28.1	6,941	52.1	3,850	51.6	4,913	9.8	9,663	34.3	11,854	18.6
Alive in year t-2 [NETS] [f]	5,496	26.5	6,556	49.2	3,684	49.4	4,721	9.4	9,180	32.6	11,277	17.7
Alive in year t [NETS] [g]	5,316	25.7	6,341	47.6	3,593	48.2	4,585	9.1	8,909	31.6	10,926	17.2
Final steps for analysis sample												
Not a nursing home in 2002 directive [h]	5,151	24.9	5,980	44.9	3,398	45.6	4,079	8.1	8,549	30.4	10,059	15.8
SST cycle is opened [i]	4,806	23.2	5,980	44.9	3,308	44.4	4,079	8.1	8,114	28.8	10,059	15.8
Focal DART, emp, hours in common support	4,805	23.2	5,968	44.8	3,305	44.3	4,073	8.1	8,110	28.8	10,041	15.8
Area-Office–directive has ≥ 1 assigned and not-assigned meeting restrictions	4,279	20.7	5,253	39.5	2,698	36.2	3,911	7.8	6,977	24.8	9,164	14.4

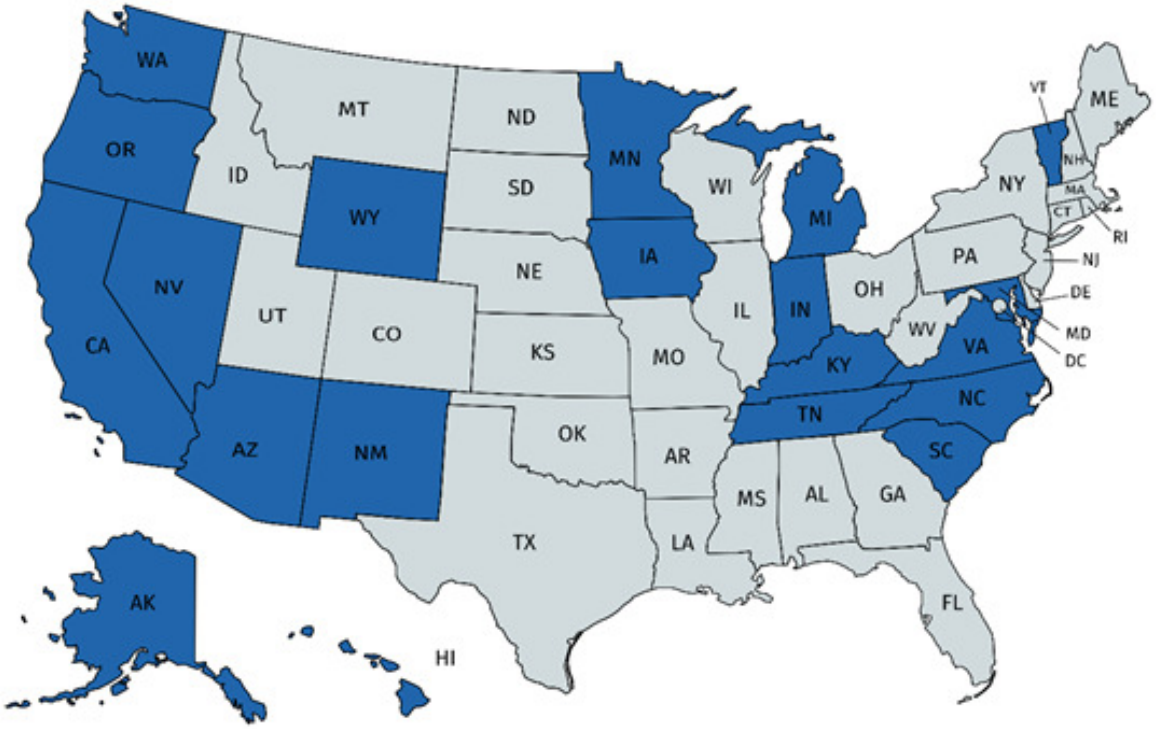
[a] An establishment-directive refers to a specific instance of an establishment being placed on a particular year's SST target list.
[b] Restricts to the 29 states under federal OSHA jurisdiction. While a few of the 21 states with state-run OSHA offices participated in SST, they were not subject to oversight from the federal office. See Figure A.1 for details.
[c] Restricts to establishments on (a) the primary list and with the percent of the primary list in its Area-Office–directive assigned to inspection strictly between 5 and 95, or (b) the secondary list and with the percent of the corresponding primary list assigned to inspection equal to 1 and the percent of the secondary list assigned to inspection strictly between 5 and 95. This is the subset of the target lists that was randomized.
[d] An establishment is subject to the deletion criteria if, within 2 years of the directive start date—or 3 years, beginning with the 2009 SST directive start date—it had an inspection in IMIS coded as a comprehensive safety inspection or as a records-only inspection or if it is a nursing home and had a focused inspection.
[e] A random sample of establishments that either do not respond to the ODI survey or report very low injury rates are placed on the target list each year to assess the reliability of their reported data. Because these establishments are targeted precisely because of concerns over the accuracy of their data, we remove them from our sample.
[f] Drops establishments that were not in operation two calendar years before the directive year, according to NETS.
[g] Drops establishments not alive at the start of the directive year, as such establishments were ineligible for SST inspection.
[h] The 2002 SST directive said nursing homes were to be excluded from the 2002 target list, due to OSHA's concurrent National Emphases Program on nursing homes. We therefore drop such establishments from the sample.
[i] An Area Office could assign subsets of its target list to inspection in cycles. Once it inspected each establishment in a cycle, it could create another one. In some cases, if an Area Office created a cycle, but did not actually open it (i.e., begin inspecting it), it was allowed to move on. We therefore drop such unopened cycles from our sample. We identify unopened SST cycles as those in which less than 5% of eligible establishments in the cycle show up in IMIS with an SST inspection in the directive year.

Table A.2: Estimating the Benefit of Our Benchmark Targeting Policy If Predicted CATEs Are Based Only on Data OSHA Would Have Available

Sample =	2007–2010 target lists					
S(Z) and B(Z) estimated on:	2001–2010 sample			2001–2006 sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Average number of DAFW injuries averted among establishments that...						
Were on 2007–2010 randomized sample	-0.212			-0.208		
	(0.154)			(0.109)+		
Would be assigned in benchmark policy, targeting on $S(Z)$		-0.744			-0.628	
		(0.310)**			(0.352)+	
Would be assigned in benchmark policy, targeting on $B(Z)$			-1.05			-0.874
			(0.321)***			(0.369)*

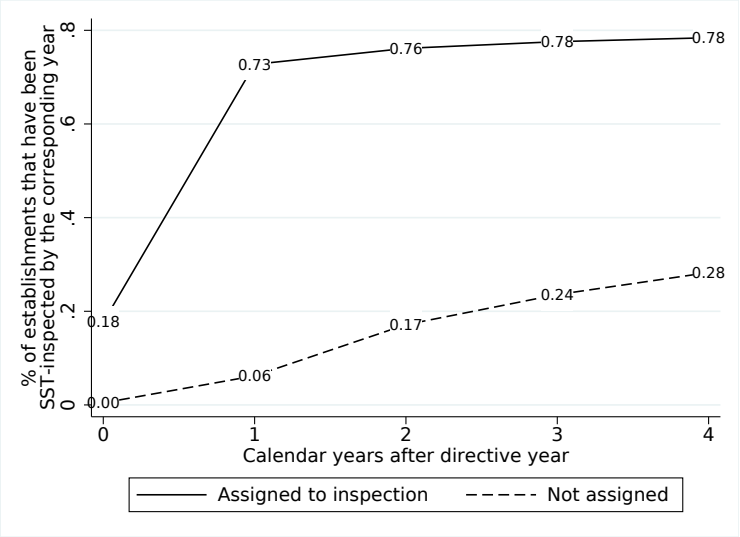
The dependent variable in each column is equal to the average number of injuries an establishment experienced over the 5-year period comprising the directive year and 4 subsequent years. The sample is establishment-directives in the randomized sample on the 2007–2010 SST target lists. *Would be assigned in benchmark policy* is a dummy equal to 1 if an establishment's S(Z) (predicted CATE, or difference in the number of annual injuries if and if not assigned to SST inspection over a 5-year period), or B(Z) (predicted number of injuries absent assignment to inspection), is high enough to be assigned to inspection in this policy. See Section 4.4.1 for details. In Columns 1–3, S(Z) (B(Z)) is estimated from a causal forest (Super Learner) run on the 2001–2010 target lists, using the CDDF method described in Section 3.3. The policy estimates are evaluated for the 2007–2010 target lists only (rather than 2001–2010), and the estimates correspond to the median $\hat{\gamma_1}$ from Equation 5 across 250 sample splits. In Columns 4–6, the models underlying (Z)and(Z) are estimated using the 2001–2006 samples, then applied out of sample to the 2007–2010 samples, and the estimates correspond to $\hat{\gamma_1}$ from Equation 5 estimated once on the 2007–2010 sample.
Robust standard errors in parentheses. +p<.1, *p<.05, **p<.01.

Figure A.1: States under OSHA's Jurisdiction



Private-sector establishments in the 29 states in white are are under federal OSHA jurisdiction. Source: `www.osha.gov/dcsp/osp/`. Map created with `https://mapchart.net/usa.html`.

Figure A.2: SST Inspection Rates by Year Relative to Directive Year



The figure shows the percent of establishments in our randomized sample with at least one completed SST inspection by the end of each calendar year relative to the directive year, separately for those assigned and not assigned to inspection.

Appendix B **Further Details about the SST Program**

OSHA first identified a set of hazardous industries based on Bureau of Labor Statistics (BLS) data.[49] Each year, the OSHA Data Initiative (ODI) surveyed between 60,000 and 80,000 workplaces—with at least 40 employees—in these industries.[50] OSHA sent the ODI survey mid-year and establishments reported summary data on their injuries, illnesses, and employment from the previous calendar year. Establishments based their survey responses on recorded logs that OSHA required them to keep of every work-related injury and illness.[51]

Beginning in 1999, each year—between April and August[52]—OSHA created a primary and a secondary SST target list based on the prior year's ODI survey. The primary list consisted of the roughly 3,500 establishments that had reported the highest rates of Days Away From Work, Restricted Work, or a Transfer (collectively referred to as DART) injuries or Days Away from Work (DAFW) injuries in the prior year's ODI survey. Both rates were measured as the number of such injuries reported over a year per 100 full-time employees working 40-hour weeks that year. The secondary list contained the roughly 7,000 establishments with the next-highest rates. For example, in 2008, the primary list included establishments reporting DART rates of at least 11 or DAFW rates of at least 9 and the secondary list included establishments with DART rates between 7 and 11 or DAFW rates between 5 and 9. The specific cutoffs for the primary and secondary lists changed each year and, beginning in 2009, varied by industry. We restrict our analysis to the 2001–2010 target lists because those are the only years for which we were able to obtain primary and secondary lists from OSHA. Establishments on the primary and secondary lists in those years reported average DART rates of 12.8 and 7.0, respectively. These rates are several times the average DART rate of 2.3 for all private-sector establishments over this period (US Bureau of Labor Statistics 2015).

---

[49] Specifically, the Bureau of Labor Statistics (BLS) Survey of Occupational Injuries and Illnesses gathered data each year from a sample of approximately 200,000 establishments drawn from all private-sector industry establishments. OSHA selected for the SST program a subset of industries that BLS classified as "high hazard industries." OSHA used BLS annual "high hazard industries" lists until 2003, when BLS stopped updating them, and from then on used the 2003 edition.

[50] OSHA intended to survey each establishment meeting these criteria (with at least 40 employees and in the specified hazardous industries) at least once every three years.

[51] OSHA Form 300, which OSHA standard 29 CFR 1904 requires employers to complete, is available at https://www.osha.gov/recordkeeping/RKform300pkg-fillable-enabled.pdf.

[52] For example, the 2007 SST directive was issued on May 14, 2007 and was in effect until the 2008 directive was issued on May 19, 2008.

OSHA then sent each of its 81 Area Offices the list of all establishments on the primary list that were located in that Area Office's region. If an Area Office did not anticipate having sufficient resources to inspect its entire primary list, a "cycle" ensued whereby the Area Office entered the number it anticipated being able to inspect into OSHA software. The software then randomly assigned the subset of establishments from the primary list that the Area Office was to inspect. If the Area Office inspected all of these establishments before OSHA headquarters issued the next year's list, another cycle ensued whereby the Area Office estimated how many additional inspections it could conduct and the software generated a new random set of establishments from the remainder of its primary list. When an Area Office had attempted inspections at all of its primary list, it repeated this process with the secondary list (for details, see US OSHA 2008). Thus, most Area Offices inspected a random subset of either their primary or secondary list (but never both).

When an OSHA inspector arrived to conduct an SST inspection, he or she explained that the establishment was being inspected because it had a relatively high injury rate.[53] Subsequent actions were similar to other types of OSHA inspection: the inspector walked through the establishment to assess hazards that could lead to injuries or illnesses, then conducted a closing conference with representatives from management and (sometimes) the employees. The inspector typically discussed any violations and also "the strengths and weaknesses of the employer's occupational safety and health system and any other applicable programs, and advises the employer of the benefits of an effective program and provides information, such as OSHA's website, describing program elements" (US OSHA 2016: 3–20). If the inspector discovered violations of OSHA regulations, a few weeks later OSHA would issue a citation and typically assess a fine. Establishments could appeal fines and OSHA often reduced them if the violation was remediated immediately.

[53] OSHA did not inform establishments of whether they were on the SST target list until an inspector showed up unannounced for an inspection.

Appendix C **Validity of ODI-reported Injury Rates**

Because our analysis relies on injury data that establishments self-report to OSHA as part of the OSHA Data Initiative (ODI), data accuracy could be a concern.

First, we note that measurement error in ODI-reported injuries—at least for those injuries employees report to their employer—might not be a significant concern in practice. Messiou and Zaidman (2005) compared establishment-level workers' compensation data to ODI-reported data in 2003 and—while they found some differences—they found no systematic underreporting of injuries to ODI. Moreover, OSHA routinely audits a random sample of ODI respondents to verify the accuracy of their ODI responses by comparing them to the establishment's OSHA log forms, assessing large fines if the ODI response is found to be inaccurate. The threat of such audits provides employers incentives to report accurately to ODI and OSHA's prior audits have found low rates of misreporting (ERG and National Opinion Research Center 2009; ERG 2013).

Still, one may be concerned about measurement error affecting our estimates. There is very likely classical measurement error (that is, pure noise) in injuries reported to ODI. In addition, there is evidence in other contexts that injuries reported to government surveys are often an undercount. Many factors could explain this divergence, but two primary ones are that (a) some employees might not report some injuries to their employers and (b) some employers might not report some injuries to OSHA (for a thorough discussion of these factors, see Azaroff et al. 2002). Younger employees are less likely to report their injuries, as are employees who suffer less serious injuries and those who work in states that offer less generous workers' compensation benefits (Biddle and Roberts 2003). Smaller employers are less likely to report injuries to OSHA (Oleinick, Gluck, and Guire 1995; Dong et al. 2011) and all employers are less likely to report less-serious injuries to OSHA (Boden et al. 2010).[54]

As long as these sources of measurement error in injury reporting are unaffected by OSHA inspections, they will increase the standard errors of our estimates but not bias the coefficients. More worrisome is the potential for inspections to affect the accuracy of self-

[54] A few studies have compared the Bureau of Labor Statistics' Survey on Occupational Injuries and Illnesses (SOII) to workers' compensation data to estimate the reliability of data collected by the SOII. While the SOII is distinct from ODI, its format is quite similar and both rely on employers' logs of OSHA-recordable injuries; thus, lessons from these studies probably apply to ODI. A consistent finding is that injuries which are more acute and easier to diagnose (such as amputations) are reported quite accurately in the BLS survey, whereas chronic injuries (such as carpal tunnel syndrome), injuries that are more difficult to diagnose (such as those that result in hearing loss), and occupational illnesses are more likely to be mis-reported (Ruser 2008; Nestoriak and Pierce 2009).

reported injuries. On the one hand, inspections could *increase* reported injuries; for example, if OSHA issues recordkeeping violations that motivate employers to keep more complete injury records. In this case, even if inspections truly lead to lower injuries, this effect would bias regression estimates towards inspections *increasing* injuries reported to the ODI. On the other hand, inspections could *decrease* the reporting of subsequent injuries by leading establishments to perceive that this would reduce the likelihood of future inspections. In this case, inspections could lead to fewer *reported* injuries, even if they have no effect on their actual occurrence.

While the extent of such bias is unobservable, we nonetheless address this concern in several ways. First, our primary outcome is DAFW injuries, which are the most serious class of injuries and for which the scope for measurement error is smaller than for total injuries (Boden et al. 2010). Second, given evidence that underreporting injuries is more common among smaller establishments, its extent should be mitigated by the fact that ODI is restricted to relatively large employers (the minimum was 40 employees until 2009 and 20 thereafter). Third, we conducted a robustness check which excluded from our analysis any establishment that OSHA had ever cited for a recordkeeping violation, which in our sample constitutes roughly 7% of establishments assigned to inspection and 4% of establishments not assigned to inspection.

Appendix D **Estimating the Social Cost of DAFW Injuries during Our Sample Period**

Waehrer et al. (2007) estimate that the social cost of a DAFW injury—the combined costs to employers, workers, and the rest of society—in 2002 was $37,016. Our goal is to estimate the cost in 2005—the median year of our sample—but in 2018 dollars.

According to Leigh (2011), medical costs make up roughly a quarter of the social costs of injuries, with the remaining three-quarters made up of indirect costs such as foregone wages and loss to home production. Because medical care spending rose 24% from 2002 to 2005,[55] we scale up 25% of the $37,016 figure ($9,254) by 24% to estimate that the medical cost portion was $11,475 in 2005. We assume that indirect costs grew at the rate of inflation, noting that, according to the Bureau of Labor Statistics, the consumer price index (CPI) rose by 5.4% from 2002 to 2005. We therefore scale up the indirect portion (75% of 37,016) by 5.4% to estimate it as $29,261 in 2005. Thus, we estimate a DAFW injury in 2005 to cost $40,736 ($11,475 + $29,261) in 2005 dollars.

To convert this to 2018 dollars, we note that the Bureau of Labor Statistics reports that the CPI rose by 28.6% from 2005 to 2018. Thus, the social cost of a 2005 DAFW injury in 2018 dollars is $40,717 * 1.286 = $52,362.

[55] Peterson-Kaiser Health Systems Tracker, "U.S. Health Expenditures 1960-2015," http://www.healthsystemtracker.org/interactive/?display=U.S.%2520%2524%2520Billions&service=Hospitals%252CPhysicians%2520%2526%2520Clinics%252CPrescription%2520Drug, accessed July 2017.

Appendix E **Sample Attrition**

As discussed in Section 3.1, we do not observe ODI-reported injury data in any year of the post-period (the directive year and four following calendar years) for 15% of our randomized sample (though most have post-period data for other outcomes we examine). Such attrition might be a concern if, for example, it is correlated with assignment to inspection. Here we discuss the sources of our sample attrition and assess its relationship with assignment to SST inspection.

Table E.1 illustrates factors leading to sample attrition. While the overall attrition rate is 14.9%, this falls to 10.9% among those establishments on the 2001–2007 SST target lists. Because the ODI survey ended in 2011, establishments on the 2008–2010 lists had fewer opportunities to be surveyed. When we further restrict the sample to those establishments that never change industry and whose employment never drops below 40 in the post-period, the attrition rate drops slightly from 10.9% to 9.8%. This suggests that factors that would render establishments ineligible for the ODI survey are not a major source of sample attrition.

The final row of Table E.1 assesses the role of establishment survival: one clear way to exit the sample is to shut down. Indeed, further restricting the sample to those establishments alive during the entire post-period reduces the attrition rate to 5.9%. Thus, over 60% of the sample attrition can be explained by straightforward observable characteristics.

Table E.1: Sources of ODI Attrition

	(1) Number of establishment-directives in the randomized sample that… …lack ODI data in the post-period	(2) …have ODI data in the post-period	(3) % with no ODI data: (1)/[(1)+(2)]
Analysis sample	2,405	13,736	14.9%
…and in 2001–2007 target lists	1,269	10,353	10.9%
…and employment [NETS] remains above 40	1,064	9,674	9.9%
…and never change industry	916	8,457	9.8%
…and remained alive during sample period	491	7,864	5.9%

More pressing than why sample attrition occurs is whether it is correlated with assignment to SST inspection. In Table E.2, we report the coefficients from a series of regressions that predict an indicator variable equal to 1 if an establishment has ODI-reported data in any of the post-period years, with the key explanatory variable being *assigned to SST inspection* and controlling for directive-year fixed effects. The columns report estimates of this model on each of the sample restrictions in Table E.1.

Reassuringly, in all columns, the coefficient on *assigned to SST inspection* is tiny and statistically indistinguishable from zero, implying that the attrition in ODI-reported data is unlikely to bias our estimates of the effects of SST inspections on ODI-reported outcomes.

Table E.2. Does Assignment to Inspection Predict ODI Attrition?

Sample =	(1) Randomized sample	(2) 2001–2007 target lists	(3) …and employment [NETS] remains above 40	(4) …and never changed industry	(5) …and remains alive
Assigned to SST inspection	0.0022	0.0077	0.0086	0.0087	0.0062
	(0.0055)	(0.0058)	(0.0058)	(0.0062)	(0.0052)
Directive-year fixed effects	Y	Y	Y	Y	Y
Observations	16,141	11,622	10,738	9,373	8,355
R-squared	0.041	0.003	0.003	0.003	0.003
Dependent variable sample mean	0.851	0.891	0.901	0.902	0.941

Each column reports estimates from a separate OLS regression, in which the dependent variable is an indicator of whether an establishment has ODI-reported data in any of the five years made up of the directive year and the four following years. Robust standard errors in parentheses. +p<.1, *p<.05, **p<.01.

Appendix F **Pre-specification**

We pre-specified our design and posted our subsequent pre-analysis plan to the Open Science Framework at https://osf.io/2snka/.

The first version of our pre-analysis plan, posted in July 2015, provided the basic outline of our study and described our primary outcome variables and our planned empirical specifications to estimate the baseline overall effects of inspections. We also uploaded the Stata code we would use to estimate our regressions.

After posting this plan, we found several minor glitches in our pre-specified design, which we therefore updated over the next months. For example, because we initially believed a large share of establishments assigned to control in one year would become assigned to treatment (that is, assigned to inspection) in later years, we originally planned to estimate the effects of inspections using outcomes within a window of three years before and after the focal year. However, while creating our analysis sample, we learned that this "crossover" of controls was not as large as we thought and that our power would increase if we estimated outcomes using a window of four years before and after the focal year. As another example, we pre-specified that one specification would control for "employment" but we had intended "ln(employment)."

Additionally, after specifying our randomized sample in the original pre-analysis plan, we learned of some unique features of the SST program in 2002 and 2003 that we deemed important to incorporate into our analysis. We also made some improvements to our fuzzy linking between the SST target lists and IMIS, which slightly changed our analysis sample.

We incorporated these changes in an updated version of our pre-analysis plan, which we uploaded to the Open Science Framework in January 2016.

Our initial and updated pre-analysis plans included two analyses that we subsequently decided were not suitable for our paper. First, we had initially planned to use establishments' sales, gathered from NETS, as an outcome. However, we discovered that NETS often reports estimated sales—rather than actual sales—for standalone establishments and always reported estimates for branch establishments of multi-unit firms (based on either firm-wide sales or an establishment's size and industry). We concluded that sales values from NETS would be an uninformative outcome and therefore omitted it from our analysis.

Second, we initially planned to use Dun & Bradstreet's Composite Credit Appraisal as an additional measure (besides PAYDEX) of establishments' creditworthiness. This is an annual

measure of Dun & Bradstreet's overall assessment of risk of default and slow payments and is rated on an ordinal scale of limited, fair, good, and high. We discovered that this measure was missing for roughly half the establishments in our sample and that we obtained very similar estimates whether using this or PAYDEX for the overlapping sample that had both measures, so we decided to omit this measure from our analysis.

Appendix G **Predictors in the Machine Learning Analyses**

This appendix lists the variables we included in the two machine learning exercises:

1) Using causal forest to estimate establishments' CATE, $s_0(Z)$
2) Using Super Learner to estimate establishments' baseline conditional average, $b_0(Z)$

When any establishment was missing a variable, we replaced it with that variable's sample mean.

Location and year variables

- Dummies for 10 OSHA regions
- Dummy if the establishment is located in a large metro area
- Number of days after a work-related injury until the injured worker can receive workers' compensation, as determined by the establishment's state
- State leave-one-out mean[56] annual DAFW rate, lagged 2 years
- Dummies for directive year

Industry and size variables

- Establishment's annual total working hours, lagged 1 and 2 years
- Establishment's annual log employment reported in NETS, lagged 1 year
- 4-digit SIC leave-one-out mean annual DAFW rate, lagged 2 years
- 3-digit SIC leave-one-out mean annual penalties assessed at OSHA inspections, lagged 1 year
- Dummy for manufacturing sector
- Dummy for nursing home sector

Compliance-related variables

- Dummy if establishment had any OSHA inspection prior to directive year
- Dummy if establishment had any OSHA complaint inspection from t-1 to t-3

Other establishment characteristics

- Establishment age reported in NETS, lagged 2 years
- Dummy for standalone firm, lagged 1 year

[56] A leave-one-out mean is the mean of the variable, excluding the focal establishment.

- Establishment's minimum monthly PAYDEX score, lagged 2 years

Variables related to injuries and to ODI

- Number of years establishment was previously on an SST target list
- Establishment’s average annual number of DAFW injuries, t-1 to t-4
- Establishment’s annual DAFW injury rate, lagged 1, 2, and 3 years
- Establishment’s annual transfer/restriction injury rate, lagged 2 years
- Establishment’s annual other recordable injury rate, lagged 2 years
- Establishment’s annual DAFW injury rate, squared, lagged 2 years
- Establishment’s total annual number of days away from work (DAFW), lagged 2 years
- Dummy for "has ODI data in t-1"
- Dummy for "has ODI data in t-3"
- Establishment’s annual DAFW rate from t-2, interacted with dummies for 4 employment quartiles (from NETS) in t-2.

The causal forest to estimate CATEs also included the percentage of establishments selected for inspection in the establishment’s Area-Office–directive–year primary or secondary list.

Appendix H Assessing the Lucas Critique: Can We Estimate Effects of New Targeting Rules Using Historical Data?

Our approach to estimating the effects of alternative targeting strategies does not consider the potential behavioral effects on *uninspected* establishments. Because alternative targeting strategies change the *threat* of inspection, such behavioral changes could be important, rendering our estimates misleading. Here, we provide evidence that such effects are, in fact, unlikely to be important in this setting.

In particular, we find no evidence that establishments were responsive to marginal changes in the threat of inspection in the historical SST program. Such responsiveness could have manifested in two ways. First, establishments might have strategically misreported injury rates to reduce their risk of inspection under SST targeting rules. Second, establishments facing a higher threat of SST inspection might have reduced their injury rates relative to those facing a lower threat. We find no evidence of either effect and provide details below. These non-results provide confidence that our main approach, which ignores their potential effects on *uninspected* establishments, yields accurate estimates of the effects of alternative targeting regimes.

Do Establishments Manipulate Reported Injuries to Reduce Threat of Inspection?

Establishments with injury rates near the cutoff for primary or secondary list face discontinuously higher threat (risk) of inspection if they report injuries above the cutoff than just below. If the threat effect were important and managers knew OSHA's (publicly available) targeting rule, establishments would have an incentive to underreport their injuries to keep them below these cutoff values. We assess whether there is bunching of reported injury rates just below the limits for either the primary or secondary lists.

Such bunching is unlikely. First, it is difficult for establishments to misreport their injuries to ODI (Appendix C). Second, the cutoffs for the SST directive in a given year were based on injury rates from two years prior, making them difficult to predict, although there were some instances in which the cutoffs remained the same for a few consecutive years.

To test for the presence of bunching, we reviewed OSHA's archived SST directives for each year 2001–2010 to identify the relevant injury rate cutoffs for that year's directive. In the early years of this range, these cutoffs were based exclusively on DART rates (for example, the 2002 primary list included those establishments whose DART rates exceeded 14 in the year 2000). In

the later years, the cutoffs were a function of DART *and* DAFW rates (for example, the 2006 primary list consisted of establishments with *either* a 2004 DART rate above 12 *or* a 2004 DAFW rate above 9). For each directive year *t*, we calculate the difference between an establishment *i*'s relevant injury rate from *t-2* and the relevant cutoff, which we designate as *c*. For example, $DART_{i,t-2} - c_{pri,dart,t}$ is this quantity for the DART rate relative to the primary list cutoff. The "running variable" determining whether an establishment is eligible for the primary list in year *t* is:

$$\boldsymbol{Rate_{it}^{pri} = \max\{(DART_{i,t-2} - c_{pri,dart,t}), (DAFW_{i,t-2} - c_{pri,dafw,t})\}.} \quad (\mathbf{H.1})$$

An establishment is eligible for the primary list in year *t* if $Rate_{it}^{pri} > 0$. We define $Rate_{it}^{sec}$ for the secondary list analogously.

Our sample restrictions for this analysis are similar to those in our main analysis: we drop establishments (a) whose recent inspection histories make them ineligible for SST inspection, (b) whose ODI data quality were flagged as questionable, (c) which were not alive in year t-2 according to NETS, or (d) which reported fewer than 40 employees to ODI in year t-2. See Table A.1 for details about these criteria.[57]

The top row of Figure H.1 demonstrates that this running variable corresponds a nearly sharp discontinuity in placement on the primary list (left column) and secondary list (right column). The second row shows a discontinuous increase in the probability that an establishment actually receives an SST inspection in the calendar year of or following the directive year (that is, in either *t* or *t+1*). Establishments that just barely make it onto the primary list have a 30-percentage-point–higher likelihood of receiving an SST inspection than those that are just below; for the secondary cutoff, that difference is roughly 9 percentage points.

Finally, Figure H.2 reports the density of this running variable. If establishments are strategic and if they are able to misreport their injuries, then we expect to see bunching of reported injuries just below the cutoffs for the primary and secondary lists. However, the density

[57] We also drop nursing homes for all years except 2009–2010 for this analysis. Nursing homes were excluded from the secondary list in all years prior to 2009. While nursing homes were included on the primary list in earlier years, their inclusion was not based on the same cutoff rule as other industries.

appears smooth in both cases and a formal manipulation test based on Cattaneo, Jansson, and Ma (2018) fails to reject that the density is continuous.[58] Thus, there is no evidence that establishments strategically misreport injuries to reduce their threat of inspection.

Does a Higher **Threat** *of Inspection Reduce Establishments' Injuries?*

The analysis above might be a weak test of whether changes in the threat of inspection elicit behavioral changes: because OSHA determined its annual SST cutoffs based on injury rates from two years prior and because establishments cannot easily misreport their injuries, the scope for strategic behavior is limited. Therefore, in this section we conduct a complementary test: whether a change in the *threat* of inspection leads to a *subsequent* change in injuries.

Establishments just above the primary or secondary cutoffs faced a higher threat of inspection than those just below. If establishments adjust their safety hazards based on their threat of inspection (because, for example, reducing hazards reduces the expected costs of future inspections), then those just above the cutoff will decrease their injury rates more than those just below, even if they are not actually inspected.

We can evaluate whether establishments are responsive to higher threats by investigating whether those just above the primary or secondary list cutoffs subsequently experience fewer injuries than those just below. However, without further medication, this exercise combines the effect of both the threat of inspection and actually being inspected on injuries. Consider the following regression discontinuity (RD) design:

$$\boldsymbol{y_{it\tau} = \alpha + \gamma * \mathbb{1}(Rate_{it}^{pri} > 0) + f(Rate_{it}^{pri}) + X_{it}\beta + \epsilon_{it\tau}.} \qquad (\mathbf{H.2})$$

Here, the outcome variable *y* is a function of injuries experienced by establishment *i* that is being considered for the SST list in year *t*, evaluated τ years after *t*. We define *y* as log(DAFW injuries + 1) and use $\tau = 1$ (one year following *t*). $Rate_{it}^{pri}$, the "running variable," is defined as above, and $\mathbb{1}(\cdot)$ is an indicator function. The coefficient γ thus estimates the effect of just barely being eligible for the primary list. To improve precision, we include the following controls in X_{it}: year fixed effects, OSHA region fixed effects, a manufacturing industry dummy, and

[58] This evidence corroborates Li and Singleton (2019), who also find no evidence of bunching around the SST cutoffs.

$\log(hours)_{t-2}$. We use an analogous specification—but with $Rate_{it}^{sec}$—for the secondary list cutoff. To establish the sample for these regressions, we use the approach developed by Calonico et al. (2019) to select the bandwidth around the cutoffs—that is, the range of $Rate_{it}^{pri}$ and $Rate_{it}^{sec}$ above and below zero included in our regressions—that minimizes mean squared error (MSE) and which we refer to as the MSE-optimal bandwidth.

With no further modifications, the coefficient γ estimates the combination of two effects: (a) the specific deterrence effect imposed on establishments that *were inspected* and (b) the general deterrence effect on *all* establishments above the cutoff due to greater *threat* of inspection.

However, we aim to isolate the first of these effects: whether establishments become safer due solely to greater *threat* of inspection as injury rates cross the primary or secondary list cutoff. To do so, we leverage variation across OSHA's Area Offices in what portion of their two target lists they actually inspected. In any given year, some Area Offices began inspecting establishments on their primary list but did not begin their secondary list, whereas others completed their primary list and began inspecting establishments on their secondary list.

Using this variation, we estimate Equation H.2 for the following subsamples of the data for which γ should reflect purely an estimate of an increased threat of inspection. First, we focus on establishments that (a) were in the jurisdiction of an Area Office that began its primary list but not its secondary list, (b) had injury rates just above or just below the primary list cutoff (defining "just above" and "just below" using the MSE-optimal bandwidth approach described above), and (c) were *not* selected for SST inspection. Analogously, we focus on establishments that (a) were in the jurisdiction of an Area Office that completed its primary list and began its secondary list, (b) had injury rates just above or just below the *secondary* list cutoff, and (c) were *not* selected for SST inspection. Establishments in these two subsamples faced a discontinuous threat of inspection at the cutoff but, due to OSHA's resource constraints, were not actually inspected. Thus, γ represents the effect of a change strictly in the *threat* of inspection on subsequent injuries.

Table H.1 reports our decomposition of the specific and general deterrence effects of SST inspections and reports $\hat{\gamma}$ from Equation H.2 for various subsets of the data. Panel A reports estimates using the sample of all establishments with injury rates around the primary and

secondary cutoffs, whether or not selected for inspection. In Columns 1 and 3, the dependent variable is an indicator coded 1 when an establishment was SST-inspected in year t or t+1, which means that $\hat{\gamma}$ estimates the effect of crossing the primary or secondary cutoff on the likelihood of subsequently being inspected (each cell reports the MSE-optimal bandwidth around the cutoff used in that regression in the table footer). The results indicate that just barely making it above the primary cutoff yields a 31-percentage-point increase in the probability of receiving an SST inspection in that or the following year (Column 1); we find a similar 30-percentage-point increase for those just barely above the secondary cutoff (Column 3).[59] Thus, just crossing the primary or secondary list cutoffs substantially increased the risk (or *threat*) of being inspected.

Columns 2 and 4 report estimates from the same specification but with the dependent variable being log(DAFW injuries) in *t+1.* As described above, given this sample, the estimate in Columns 2 and 4 of Panel A represents a mixture of the effects on injuries of a heightened threat of inspection and of an actual inspection. Point estimates indicate that establishments just above the cutoff subsequently experience 7% (p=0.07) and 7% (p=0.13) fewer DAFW injuries for the primary and secondary cutoffs, respectively, though the estimate is not quite statistically significant for the secondary cutoff. Thus, the combined effect of the higher *threat* of inspection and higher proportion of realized inspections yielded fewer injuries for establishments just above the cutoff than for those below.[60]

In Panel B, we aim to isolate the general deterrence effect of a heightened *threat* of inspection on establishments' injuries. We reestimate the regressions from Panel A but restrict the sample to establishments *not* selected for inspection. Even though Panel A's results in Columns 1 and 3 indicated that the *threat* of inspection changed substantially at the cutoffs, Panel B's estimates in these columns confirm that the *realized* inspection rates barely change across the cutoff for this subsample. Columns 2 and 4 of Panel B indicate no difference in

[59] Note that this increase in the probability of inspection at the secondary cutoff (30 percentage points) is higher than the increase in the probability at the secondary cutoff indicated in the second row, right column of Figure H.1 (roughly 9 percentage points). This difference arises because in Table H.1 we have restricted to establishments in the jurisdiction of an Area Office that began inspecting its secondary list (which many Area Offices did not do).
[60] This evidence corroborates Li and Singleton (2019), who estimate the effects of SST inspections using a similar research design and find similar magnitudes (though they only consider the primary cutoff).

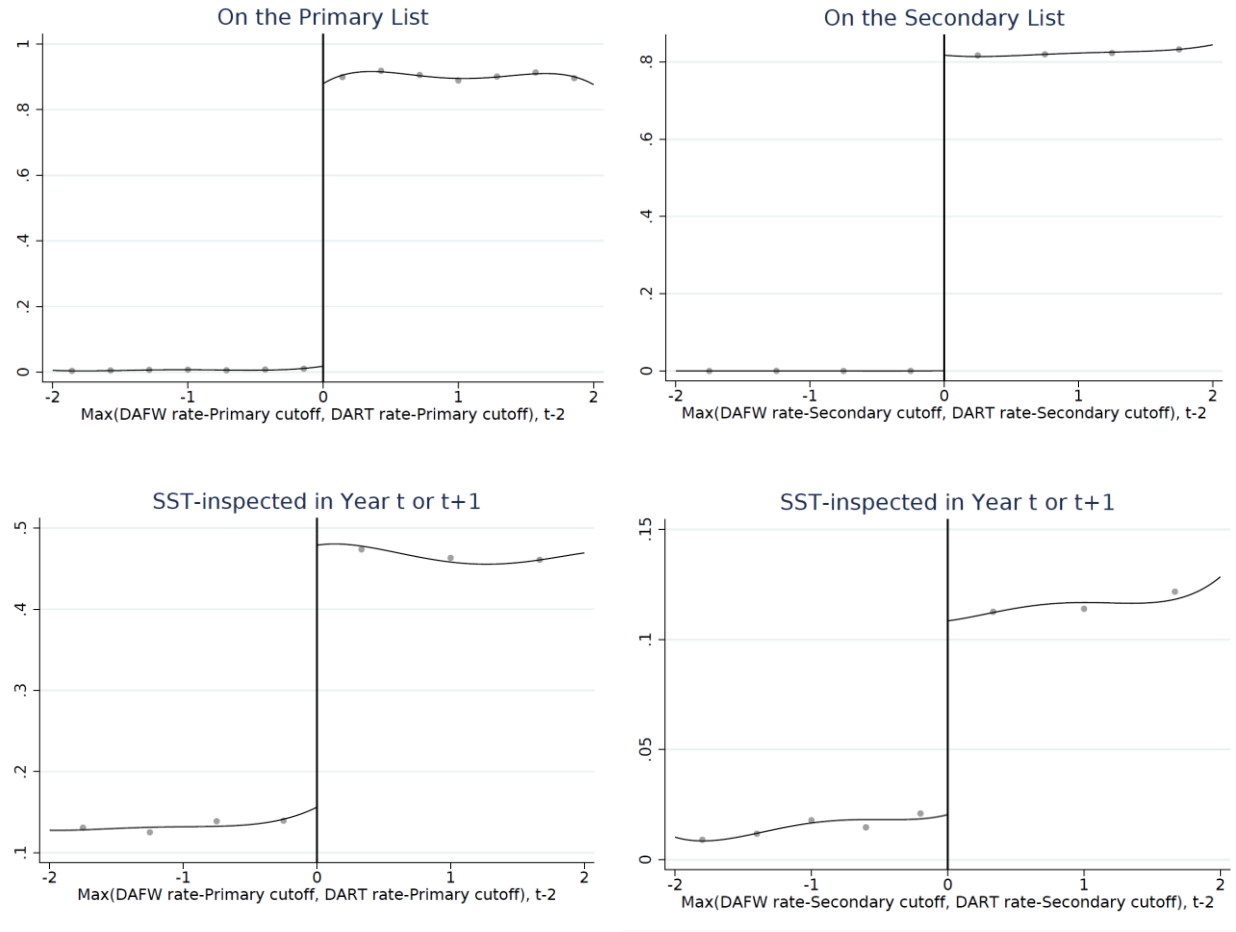
subsequent injuries between those just above and just below the primary cutoff ($\beta = 0.00$, SE = 0.06) and similarly for the secondary cutoff ($\beta = -0.02$, SE = 0.05). Both estimates are small in magnitude and statistically indistinguishable from zero. These estimates provide no evidence of a general deterrence effect or that establishments improve safety due purely to a change in the *threat* of inspection.

We conduct one final test to assess general deterrence effects. In Panel C, we focus on all establishments that (a) were in the jurisdiction of an Area Office that did not begin its secondary list and (b) had injury rates just above or below the secondary cutoff (again, defining "just above" and "just below" using the MSE-optimal bandwidth). Column 3 shows that the change in the *realized* probability of inspection at the cutoff for this sample is essentially zero. That is, although establishments just above the secondary cutoff faced a higher overall risk of inspection, essentially none in this sample were ultimately inspected, since these particular Area Offices were unable to start on their secondary lists. Corroborating the estimates in Panel B, Panel C's Column 4 shows precisely no difference in subsequent injuries at the secondary cutoff for this subsample ($\beta=0.01$; SE=0.03).

Overall, the estimates in this section provide no evidence that establishments' safety changed in response to merely a heightened *threat* of inspection. First, reported injuries were not strategically bunched just below the primary or secondary list cutoffs, which would have suggested strategic misreporting to reduce the threat of inspection. Second, the evidence in Table H.2—and especially the results reported in Panel B's Columns 2 and 4 and Panel C's Column 4—indicates that injuries were unaffected by a heightened threat of inspection.

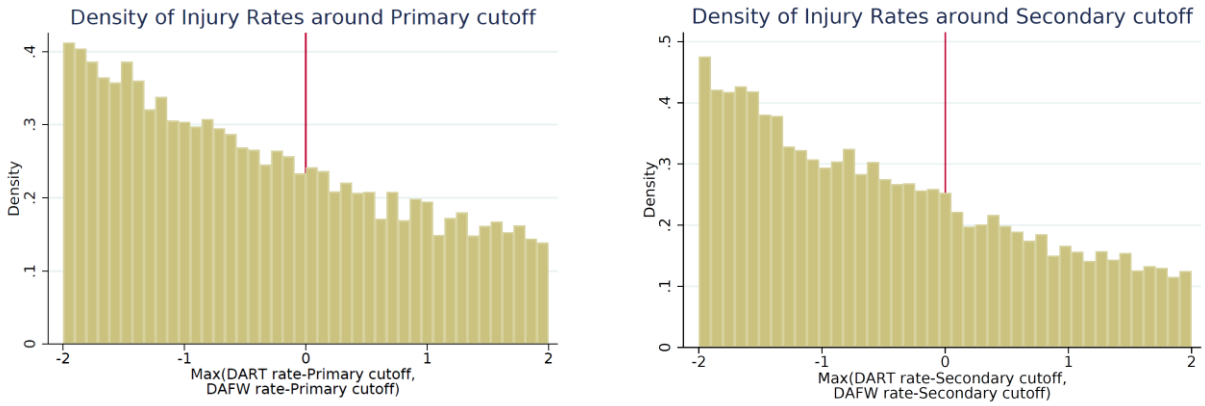
This lack of general deterrence effects implies that such potential behavioral changes by establishments in response to a change in targeting strategy would be second-order relative to the specific deterrence effects of *realized* inspections. Indeed, the evidence in Panel A's Columns 2 and 4 bolsters the evidence in the main text that realized inspections improve safety, even as the evidence in Columns 2 and 4 in Panels B and C indicates that changes in threats do not. Thus, these results give us confidence that our main approach estimates the effects of alternative targeting strategies on injuries, and that effects on uninspected establishments (general deterrence effects) are not a serious confounding factor in our context.

Figure H.1: Did Establishments' Risk of OSHA Inspection Increase at the Primary and Secondary List Cutoffs?



Note: Each figure displays a binned scatterplot, in which the x-variable is either $Rate_{it}^{pri}$ (left column) or $Rate_{it}^{sec}$ (right column). $Rate_{it}^{pri}$ and $Rate_{it}^{sec}$—defined in Equation H.1—are the "running variables" that determine whether an establishment's injury rate from year *t-2* is high enough to make it eligible for the SST primary or secondary list, respectively, in year *t.* In each row, the y-variable is the indicator variable provide in the figure's title. The sample includes those establishments that were considered for the SST Target Lists between 2001-2010 (except nursing homes for all years prior to 2009), that had not received an inspection in the prior two years, and that reported injury rates that resulted in the absolute value of their $Rate_{it}^{pri}$ or $Rate_{it}^{sec}$ being less than 2.0.

Figure H.2: Density of Establishments' Injury Rates Relative to the Primary and Secondary List Cutoffs



Note: These figures display the density of $Rate_{it}^{pri}$ and $Rate_{it}^{sec}$—defined in Equation H.1—in the left and right columns, respectively. $Rate_{it}^{pri}$ and $Rate_{it}^{sec}$ are the "running variables" that determine whether an establishment's injury rate from year *t-2* is high enough to make it eligible for the SST primary or secondary list, respectively, in year *t.* The sample includes those establishments that were considered for the SST Target Lists between 2001-2010 (except nursing homes for all years prior to 2009), that had not received an inspection in the prior two years, and that reported injury rates that resulted in the absolute value of their $Rate_{it}^{pri}$ or $Rate_{it}^{sec}$ being less than 2.0.

Table H.1: Decomposing the Specific and General Deterrence Effects of SST Inspections

	(1)	(2)	(3)	(4)
	Dep Var =			
	SST-inspected in t or t+1	log DAFW, t+1	SST-inspected in t or t+1	log DAFW, t+1
Panel A: Specific + General Deterrence				
Sample= establishments in boundary of Area Office that...				
	began Primary (c = Primary cutoff)		began Secondary (c = Secondary cutoff)	
$Rate_{it}^{pri,sec} > 0$	0.31	-0.07	0.30	-0.07
	(0.01)**	(0.04)+	(0.01)**	(0.04)
Robust p-value	0.000	0.069	0.000	0.128
# observations	22,395	7,638	8,733	4,788
Left bandwidth around c	2.28	0.99	1.13	1.55
Right bandwidth around c	1.61	1.87	1.84	1.44
Panel B: General Deterrence only				
Sample = establishments not selected for inspection and in boundary of Area Office that...				
	began Primary (c = Primary cutoff)		began Secondary (c = Secondary cutoff)	
$Rate_{it}^{pri,sec} > 0$	0.02	0.00	0.02	-0.02
	(0.01)	(0.06)	(0.01)+	(0.05)
Robust p-value	0.140	0.984	0.066	0.707
# observations	14,426	4,791	8,038	4,316
Left bandwidth around c	1.87	1.20	1.20	1.34
Right bandwidth around c	1.79	1.01	1.19	1.26
Panel C: General Deterrence only				
Sample = establishments in boundary of Area Office that...				
			did not begin Secondary (c = Secondary cutoff)	
$Rate_{it}^{pri,sec} > 0$			-0.00	0.01
			(0.00)	(0.03)
Robust p-value			0.893	0.759
# observations			24,073	11,683
Left bandwidth around c			1.74	1.51
Right bandwidth around c			0.88	1.42

The table reports regression discontinuity estimates using the approach from Cattaneo et al. (2016) to estimate the MSE-optimal bandwidth around the cutoff. The running variable in Columns 1 and 2, defined in Equation H.1, determines whether an establishment's injury rates from two years prior render it eligible for the SST primary list. The running variable in Columns 3 and 4 is defined analogously, but for eligibility for the secondary list. In each panel, *# observations* refers to the number of establishments with injury rates in year *t-2* within the left and right bandwidths around the cutoff (also reported in each panel). The MSE-optimal bandwidths are selected, based on the method in Cattaneo et al. (2016), separately for each regression. All regressions include controls for region fixed effects and year fixed effects, a manufacturing dummy, and the log of the establishment's working hours from *t-2*.

Appendix I Robustness Checks on Estimate of ITT Effect on Injuries

We consider several other specifications as robustness checks for our ANCOVA model (Equation 1).

In a pre-specified robustness check, we estimated the effect of assignment to inspection using a difference-in-differences specification.[61]

We also sought to minimize the chances that our estimates were contaminated by inspections leading to more complete reporting of injuries. Therefore, we reestimated the ITT specification corresponding to Column 1 of Table 4 but excluded any establishment that OSHA had cited for violating recordkeeping regulations at any point during our sample period (7% and 4% of establishments assigned and not assigned to inspection, respectively).[62]

We also averaged the outcomes during the directive year and four following years into one observation per establishment-directive to construct a new dependent variable, $\overline{y_{it}^{post}}$. We reestimate the ANCOVA model (omitting the τ-year fixed effects, θ_τ) on this outcome, using OLS rather than Poisson because the outcome includes non-integer values.

We also estimated the ITT effect using targeted maximum likelihood estimation (TMLE) combined with Super Learner (van der Laan and Rose 2011), again using $\overline{y_{it}^{post}}$ as our outcome. TMLE is a double-robust approach to estimating treatment effects in the presence of potential misspecifications of the treatment assignment process.[63]

As a final check, we report the average ITT estimate from the CDDF procedure. CDDF also lets us estimate the overall average effect of assignment to inspection (that is, the average intent-to-treat effect), $E[s_0(Z)]$). The CDDF estimate of the ITT is β_1 from the following regression model estimated on the holdout sample:

---

[61] We estimate the following difference-in-differences regression model for this robustness check:

$$y_{ijt\tau} = \alpha_1 Assigned_{it} * \mathbb{1}(\tau \geq 0) + \alpha_2 \mathbb{1}(\tau \geq 0) + \mu_{jt} * \mathbb{1}(\tau \geq 0) + \lambda_{it} + \theta_\tau + \epsilon_{jit\tau} .$$

All variables here are defined as in Equations 1 and 2. We control for P_τ separately to account for changes in injuries (and other outcomes) following the directive year that would have occurred even without an SST inspection.

[62] As one other robustness check, we drop observations in which, according to the NETS database, an establishment is no longer in operation. We lose only a few hundred observations and obtain essentially identical estimates (results not shown).

[63] In our pre-analysis plan posted to the Open Science Framework, we pre-specified that we would estimate CATEs using targeted maximum likelihood estimation (TMLE) (van der Laan and Rose 2011)). We later discovered that TMLE was not well suited to simulate counterfactual policies, so we do not report these estimates. However, TMLE consistently and efficiently estimates the average treatment effect. We used the same library of learners in Super Learner as we did when estimating $B(\cdot)$ in the CDDF procedure (Section 4.3).

$$Y = \alpha_1 + \alpha_2 B(Z) + \beta_1\big(D - p(Z)\big) + \beta_2\big(D - p(Z)\big) * (S - ES) + \epsilon\,. \qquad (G.1)$$

We estimate $\widehat{\beta_1}$ from the holdout sample. As noted above, we estimate on 250 iterations of the partition process and use the median point estimate and standard error as $\widehat{\beta_1}$.

We report results from these specification checks in Table H.1. The difference-in-differences estimate (Column 1: β = -0.033, SE = 0.017) is nearly identical to the estimate from the ANCOVA specification (β = -0.035, SE = 0.017). Excluding establishments that had ever had a recordkeeping violation during our sample period (Column 2) yields a coefficient (β = - 0.040, SE=0.018) slightly larger in magnitude than that from our baseline specification, which is consistent with the idea that establishments cited with recordkeeping violations are subsequently less likely to underreport injuries. Our OLS estimate on the collapsed outcome variable (Column 3) yields a point estimate of β = - 0.178 (SE = 0.081), which, as a percent of the control mean (-0.178/ 4.62 = 3.8%), is essentially identical to the estimate from the Poisson model. Finally, the point estimates of TMLE (β = -0.208, SE = 0.087) and CDDF (β = -0.180, SE = 0.118) are slightly larger in magnitude than our OLS estimates, but the differences are not statistically significant or economically meaningful. In short, our results are robust to these several specification checks.

Table I.1: Intent-to-treat Effects of SST Inspection on DAFW Injuries: Robustness and Alternative Specification

	(1)	(2)	(3)	(4)	(5)
	Dep var = # of DAFW injuries				
	Fixed effects	Drop record-keeping violators	Collapse to mean over the post-period [OLS]	[TMLE]	[CDDF]
Assigned to inspection	-0.033	-0.040	-0.178	-0.208	-0.180
	(0.017)*	(0.018)*	(0.081)*	(0.087)+	(0.12)
# observations	89,509	38,171	13,736	13,736	13,736
# establishment-directive years	15,715	12,909	13,736	13,736	13,736
# establishments	12,630	10,509	11,083	11,083	11,083
# Area-Office–directives	383	383	383	383	383
Specification	Poisson	Poisson	OLS	OLS	OLS
Mean dep var, estabs not assigned, post-period	5.37	5.26	4.62	4.62	4.62

The table shows the results of Poisson or OLS regressions with coefficient on an *Assigned to inspection* dummy and SEs in parentheses. *Assigned to inspection* is equal to 1 in years beginning with the directive year, for establishments selected for SST inspection in the directive year. OLS and Poisson coefficients are estimates of the level change and percent change, respectively, in the dependent variable associated with *Assigned to inspection*.

Poisson drops establishments with only one observation or constant values across observations.

Columns 2 and 3 report results from an ANCOVA regression. Column 1 reports results from a specification with establishment fixed effects. If an establishment appears more than once, a separate establishment fixed effect is included for each directive year. Standard errors in Columns 1–3 clustered by establishment. **P<.01., *P<.05, +P<.1.

Column 4 reports the estimate of the ITT using Targeted Maximum Likelihood; Column 5 reports the estimate of the average ITT from the procedure in Chernozhukov et al. (2018), described in the text.

Each regression is restricted to the randomized sample, described in Table A.1, and is restricted to a window of 4 years before and after the directive year.

Appendix J **Extrapolating Estimates from Randomized to Nonrandomized Sample**

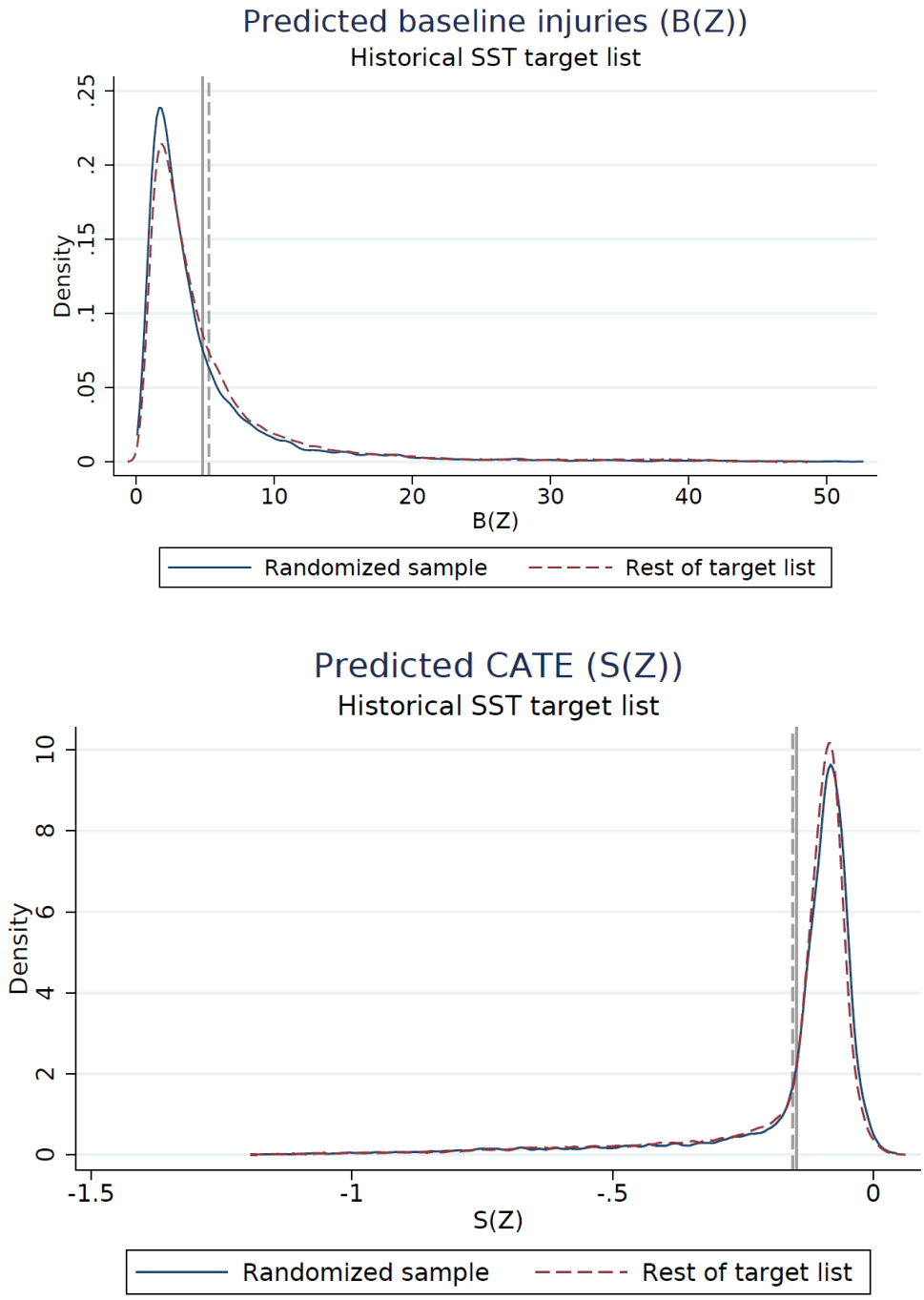
Our estimation of the number of injuries OSHA could avert under alternative targeting policies relies on using predictions of establishments' baseline conditional average ($B(Z)$) and of their CATEs if assigned to inspection ($S(Z)$)—predictions trained and validated on the randomized sample—to estimate the effects of counterfactual policies that target establishments from the overall historical SST target lists, including establishments beyond our randomized sample. Those excluded from our randomized sample include (a) those on a target list of an Area Office in a directive year in which fewer than 5% or more than 95% of the establishments listed were assigned to inspection, (b) those that had been inspected under SST in the prior two years, and (c) those that met other criteria described in Table A.1.

As described in Section 4.4.2, there are two reasons that the estimates from our procedure might not generalize to the entire historical SST target lists and thus might result in misleading estimates of targeting policy counterfactuals. First, if establishments in the randomized sample have different observable characteristics (different *Z*s), then $E[s_0(Z) \mid G_k, randomized = 1]$ could differ from $E[s_0(Z) \mid G_k, randomized = 0]$. Second, establishments in the two groups could have different *unobservable* characteristics, in which case a causal forest model estimated on the randomized sample would poorly predict CATEs for the nonrandomized sample, even if the distribution of Zs did not differ between the two samples. We address these concerns below.

Do establishments in the two samples have different observable characteristics?

If the distribution of $s_0(Z)$ or $b_0(Z)$ in the randomized sample differs substantially from the distribution for establishments in the nonrandomized sample, then our estimates of the impacts of counterfactual targeting policies—estimated on the randomized sample—would not consistently estimate policy effects for the entire historical target list. However, these distributions appear to be very similar. Panels (a) and (b) of Figure I.1 plot the distribution of, respectively, $B(Z)$ and $S(Z)$. Here, we estimate $B(Z)$ using a Super Learner on the set of all establishments not assigned to inspection in the randomized sample and estimate $S(Z)$ using a causal forest on the set of all establishments in the randomized sample (including those assigned to inspection). Distributions of both measures are very similar for the two groups. It is therefore unlikely that any differences in *observable* characteristics between the randomized and nonrandomized samples render our estimates of counterfactual policies inconsistent.

Figure J.1. Distribution of $S(Z)$ and $B(Z)$ for Establishments on the SST Historical Target List in the Randomized versus Nonrandomized Sample



The sample for these kernel density plots is the establishments on the 2001–2010 historical SST target lists that were eligible for an SST inspection (i.e., that had not received one in the prior two calendar years). *B(Z)* is estimated using a Super Learner to predict $\overline{y_{it}^{post}}$ among all establishments not assigned to inspection in the randomized sample and extrapolating the estimates to the rest of the target list. *S(Z)* is estimated using a causal forest to predict CATEs for all establishments in the randomized sample and extrapolating the estimates to the rest of the target list.

Does the predictive power of the estimates from our machine learning models differ for establishments in the randomized and nonrandomized samples?

Our machine learning estimates of $S(\cdot)$ and $B(\cdot)$—the estimates of the models underlying an establishment's CATE if assigned to inspection and baseline number of injuries if not assigned to inspection—for the randomized sample could have poor out-of-sample predictive power for the nonrandomized sample if establishments in the two samples had different *unobservable* characteristics.

Unfortunately, we cannot test the out-of-sample predictive power of $S(\cdot)$ because $s_0(Z)$ is unobservable. Fortunately, we *can* test whether the predictive power of $B(\cdot)$—the model behind the baseline number of injuries—fits worse in the nonrandomized sample than in the randomized sample. We can do the same for what we will call $T(\cdot)$, the number of injuries an establishment would experience if assigned to inspection, $T(Z)$.

Recall that in the CDDF algorithm, we estimate $B(Z)$ using a Super Learner to predict $\overline{y_{it}^{post}}$ among the establishments assigned to control in the auxiliary sample. We then use the results to predict $\overline{y_{it}^{post}}$ for those in the holdout sample and the nonrandomized sample. We repeat this process 250 times.

To test whether the predictions of $B(Z)$ are more accurate or less accurate for the holdout and nonrandomized samples, we conduct the following exercise. For each of our 250 iterations, among establishments not assigned to inspection we regress realized $\overline{y_{it}^{post}}$ on $B(Z)$ separately for those in the holdout sample and those in the nonrandomized sample. Among those establishments that OSHA did not assign to inspection, $\overline{y_{it}^{post}}$ equates to $b_0(Z)$. Thus, we are regressing establishments' realized baseline number of injuries ($b_0(Z)$) on its predicted value, $B(Z)$. We save the median coefficient and standard error on $B(Z)$, as well as the median $R^{2,}$ across the 250 iterations.

For establishments not assigned to SST inspection, both in the holdout sample and in the nonrandomized sample, the median coefficients on $B(Z)$ are close to 1 and the estimates are statistically indistinguishable from each other (Table I.1, Columns 1 and 2). The R^2s are high and nearly identical. In other words, $B(\cdot)$ estimated on half of the randomized sample has equal predictive power for realized injuries ($\overline{y_{it}^{post}}$) for the other half of the randomized sample and for the nonrandomized sample.

We can go one step further and perform a similar exercise to asses predictions of the number of injuries an establishment would experience if assigned to inspection; that is, the flip side of $B(Z)$. We will call this potential outcome $t_0(Z)$ and its estimate $T(Z)$. In each of our 250 iterations, we construct $T(\cdot)$ using a Super Learner to predict injuries among those in the auxiliary sample that are selected for inspection. We then use this model to construct $T(Z)$ for both the main and nonrandomized samples. Analogous to what we did before, in each iteration we regress realized $\overline{y_{it}^{post}}$ on its predicted value $T(Z)$ separately among establishments assigned to inspection in the main sample and in the nonrandomized sample.

As revealed in Columns 3 and 4 of Table J.1, $T(Z)$ has essentially equal predictive power for the main sample and the nonrandomized sample.

Overall, this exercise bolsters the case that our estimates from the CDDF procedure on the randomized sample generate unbiased estimates of the number of injuries OSHA could avert by re-targeting the *entire* SST target list.

Table J.1. Comparing the Predictive Power of Establishments' Estimated Injuries If Not Treated, $B(Z)$, for the Randomized Controls and for the Nonrandomized Samples

Dep var =	Average annual # of DAFW injuries over post-period			
Sample =	Establishments *not assigned* to inspection in…		Establishments *assigned* to inspection in…	
	Holdout sample	Nonrandomized sample	Holdout sample	Nonrandomized sample
B(Z)	1.028 (0.032)	1.061 (0.008)		
T(Z)			1.026 (0.036)	1.087 (0.013)
R^2	0.686	0.702	0.681	0.709

The table assesses whether our machine-learning–based estimates of the number of annual DAFW injuries an establishment would experience if not assigned to inspection (*B(Z))* and the number it would experience if it were assigned to inspection (*T(Z))* have differential predictive power for establishments in the randomized sample vs. the nonrandomized sample. For each of the 250 splits of the data used in the CDDF algorithm, we train the Super Learner algorithm on establishments not assigned to inspection in the auxiliary sample (a random half of the randomized sample) to construct $B(\cdot)$ and apply the model to estimate $B(Z)$ in the holdout sample (the other half of the randomized sample) and the nonrandomized sample. In each split, focusing on establishments that were not historically assigned to inspection, we regress establishments' realized mean annual DAFW injury count on their $B(Z)$, separately for establishments in the holdout sample (Column 1) and the nonrandomized sample (Column 2). We use an analogous procedure in Columns 3 and 4 to assess the predictive power of *T(Z),* restricting to establishments that were assigned to inspection in the historical policy. The table reports the median coefficient and SE on $B(Z)$ and the median R^2, across the 250 splits.