IRLE
Institute for Research on Labor and Employment

IRLE WORKING PAPER
#107-19
August 2019

Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA

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

Cite as: Matthew S. Johnson, David I. Levine, and Michael W. Toffel. (2019). "Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA". IRLE Working Paper No. 107-19.
http://irle.berkeley.edu/files/2019/09/Improving-Regulatory-Effectiveness-through-Better-Targeting.pdf

http://irle.berkeley.edu/working-papers

Berkeley
UNIVERSITY OF CALIFORNIA

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

August 23, 2019

We study how a regulator can best allocate its limited inspection resources. We direct our analysis to a US Occupational Safety and Health Administration (OSHA) inspection program that targeted dangerous establishments and allocated some inspections via random assignment. We find that inspections reduced serious injuries by an average of 9% over the following five years. We use new machine learning methods to estimate the effects of counterfactual targeting rules OSHA could have deployed. OSHA could have averted over twice as many injuries if its inspections had targeted the establishments where we predict inspections would avert the most injuries. The agency could have averted nearly as many additional injuries by targeting the establishments predicted to have the most injuries. Both of these targeting regimes would have generated over $1 billion in social value over the decade we examine. Our results demonstrate the promise, and limitations, of using machine learning to improve resource allocation. *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

1 Introduction

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 Occupational Safety and Health Administration 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 Occupational Safety and Health Administration 2017a). Similarly, the Internal Revenue Service audited less than 1% of the tax returns it processed (Rubin 2017). The US Environmental Protection Agency, the US Food and Drug Administration, and their counterparts in most other nations face similar budget constraints (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 recently experienced serious accidents, many injuries, or were the subject of employee complaints (US Occupational Safety and Health Administration 2017b). Such rules could in theory ensure that regulators are allocating inspections effectively, especially if they have private information that enables them to target enforcement where problems are most severe (Duflo et al. 2018).

However, assessing whether agencies' choices are leading them to direct inspections where they most effectively deliver on their regulatory mission is fraught with empirical challenges. It is difficult enough to evaluate regulators' performance because purposeful targeting makes it hard to find a valid comparison group for the inspected workplaces. Moreover, assessing whether regulators could allocate their inspections *more* effectively is even more challenging, as it requires estimates of the effects of alternative policies that were never implemented (Athey 2017). In theory, the ideal benchmark policy would allocate inspections where they most effectively deliver on the regulator's objectives. Unfortunately, the extent to which inspections achieve these objectives (e.g., averting injuries, emissions, toxic spills, or food poisonings inspections)—their *treatment effects*—is typically unobservable. As a result, regulatory agencies are left 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 an unparalleled opportunity to estimate how well inspections deliver on their objectives, and whether they could be targeted more effectively. Regulators can also use machine learning to improve inspection targeting in other ways, such as predicting more accurately where the problems are—that is, where injuries, emissions, spills, and poisonings are likely occur.

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 against those of alternative targeting policies regulators could enact. We leverage randomization both to evaluate the regulator's historical performance and to estimate heterogeneous treatment effects via machine learning. Using these estimates, we use machine learning to simulate the effects of alternative policies the regulator could have used to target inspections.

We apply this approach to OSHA, the US regulatory agency charged with assuring "safe and healthful working conditions."[1] For several reasons, this is an ideal setting in which to examine regulatory effectiveness. First, workplace injuries and illnesses impose a massive burden on the US economy, with an estimated annual social cost of $250 billion (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 surrounding OSHA's ability to improve workplace safety—and whether it could do better—has enormous social implications.

We focus on OSHA's Site-Specific Targeting (SST) program, enacted precisely to improve the agency's own effectiveness. It ran from 1999 to 2014, during which time it was OSHA's largest inspection regime, accounting for nearly 10% of the agency's enforcement budget. Specifically, in an effort to allow "the most effective use of OSHA's limited resources" (ERG and National Opinions Research Center 2009), OSHA designed the SST program to prioritize inspections to establishments with "serious health and safety problems" (US Occupational Safety and Health Administration 2004) by developing annual "target lists" of establishments that had high injury rates two years prior. When the agency's resources did not allow it to inspect all establishments on its target lists, it allocated inspections via random assignment.

---

[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 February 11, 2016 hearing with Senate Homeland Security and Governmental Affairs Committee (*Safety+Health Magazine* 2016).

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 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 will refer to 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 serious injuries and illnesses at inspected establishments by an average of 9%, which equates to 2.4 fewer DAFW injuries over the five-year period we examined. This equates to a social benefit of roughly $120,000 per inspection, which is 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 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

[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).

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—a flexible supervised machine-learning technique that predicts treatment effects based on high-dimensional nonlinear functions of observable characteristics (Wager and Athey 2018)—to estimate the CATE for each establishment on the historical SST target list.

While 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—also constructed with machine learning—that OSHA could use to target inspections: the predicted number of serious injuries an establishment would experience absent an inspection (that is, Y_0 in Rubin's (1974) potential outcomes framework). This metric is similar in spirit to the one OSHA actually used in the SST program—establishments' injury rate from two years prior—but our metric is likely to more accurately measures 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 CATE and Y_0 estimates into a method developed by Chernozhukov, Demirer, Duflo, and Fernandez-Val (2018)—henceforth "CDDF"—to estimate the effects of alternative targeting policies. The CDDF method incorporates estimates of heterogeneous treatment effects to yield consistent estimates of average treatment effects for specific subsets of the data that are robust to sampling variation; we apply this method to the subsets that OSHA would target for inspection under our alternative policies.

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

We find that OSHA could avert many more injuries if it used these machine learning techniques 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, the agency would have averted 2.41 times as many injuries as it actually did. If OSHA had instead assigned this same number of inspections each year to those establishments predicted to have the most injuries, it would have averted 2.29 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 the SST program.

Targeting based on "largest estimated CATEs" or "most predicted injuries," however, might not be optimal or even feasible for the agency. In particular, eliminating randomization might reduce the threat of inspections for uninspected establishments (Cohen 2000; Shimshack and Ward 2005; Gray and Shadbegian 2007) and would sacrifice opportunities for OSHA to continue learning where its 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, just as it did under the SST program, 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. These two alternative policies avert substantially fewer injuries at assigned establishments than the policies above that forego randomization. However, these alternative policies—because they use randomization—

ensure that the agency can continue learning where its efforts are most effective, while still averting more injuries than the historical policy did.

Our results highlight the promise and limitations of using machine learning to estimate heterogeneous treatment effects as a way to improve resource allocation. OSHA could substantially improve the effectiveness of its inspections by targeting establishments with high predicted CATEs, but it could do about as well using the simpler metric of the predicted number of injuries an establishment would have if were not assigned to inspection. Targeting on predicted number of injuries is especially likely to be as effective (if not more) as targeting on estimated CATEs under two conditions, both of which are plausible in our setting: (a) when estimates of an observation's underlying CATE are noisier than estimates of an observation's outcome if not treated and (b) when one expects that these underlying measures are highly correlated. The extent to which these two conditions hold can illuminate which of these two criteria is more effective for targeting inspections.

This paper contributes to a nascent literature that examines how machine learning can improve decisions. Our approach is close in spirit to studies that have examined how machine learning can be used to understand and improve judges' decisions to release defendants before trial (Kleinberg et al. 2017), to help food safety inspectors predict restaurants' health and safety violations (Glaeser et al. 2016), 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).[5] We extend this literature on machine learning, which has focused on predicting outcomes, to instead

[5] Our results add to research exploring how data-driven techniques can predict which workplaces are likely to have the most OSHA violations (by Cary Coglianese and Adam Finkel) or workplace injuries (by Alison Morantz and co-authors, presented by Alison Morantz at MIT Sloan on February 25, 2014).

estimate variation in the *causal effects* of regulatory inspections, which Athey (2017) points out is the relevant criterion to judge an optimal resource allocation problem. Our approach to estimating heterogeneity in policy effectiveness is shared with a recent study in the context of youth employment programs (Davis and Heller 2017). Relatedly, by illustrating how machine learning can help identify if agents' behavior is best achieving their objectives, our paper is related to work diagnosing deviations from "optimal" behavior, such as pricing in the car rental market (Cho and Rust 2010) and by retail chains (DellaVigna and Gentzkow 2019).

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 directly 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.[6] Peto et al. (2016), analyzing only SST inspections randomly assigned in 2011, find no significant effects, perhaps because their small sample size left them underpowered.

[6] 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 and such measurement error can confound how inspections affect actual injuries versus reporting of injuries.

Our findings also complement literatures examining the effects of regulatory inspections 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, and Blundell et al. (2018), who estimate the gains to dynamic enforcement of environmental regulations against 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.

2 Setting and Data

OSHA's Site-Specific Targeting (SST) program provides a compelling setting for our purposes. First, for over a decade, OSHA allocated a subset of these inspections via random assignment. Thus, we can evaluate the extent to which its inspections led to fewer injuries on average and—using the rich data we have on establishments' characteristics and historical performance—we can apply machine learning methods to estimate heterogeneous treatment effects. Second, the SST program was created to promote "the most effective use of OSHA's limited resources" (ERG and National Opinions Research Center 2009) and thus provides an ideal benchmark for OSHA's targeting priorities against which to compare alternative targeting policies. Finally, SST was an extremely large regulatory program; understanding its effects is therefore important in assessing the deployment of public resources. Tens of thousands of establishments that employed millions of workers were at risk of an SST inspection and OSHA spent tens of millions of dollars on the program.

2.1 OSHA Site-Specific Targeting (SST) Program

The SST program targeted high-injury workplaces in historically hazardous industries for inspection between 1999 and 2014. We discuss here only those details about the program's implementation that are essential for our analysis; see Appendix B for a more complete description. We focus our analysis on the 29 states that OSHA directly regulates.[7]

Each year from 1996 through 2011, the OSHA Data Initiative (ODI) collected injury data from 60,000 to 80,000 establishments. These represented the universe of establishments with at least 40 employees in a set of pre-defined high-risk industries. Establishments based their ODI survey responses on records ("OSHA logs") that OSHA required them to maintain in order to document every work-related injury and illness.[8]

OSHA used these 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 (averaging roughly three times the national average). The precise cutoffs for both lists varied by industry and year.

OSHA then sent each of its 81 Area Offices, located across the 29 states, the list of establishments in its geographic boundary that were on the primary list. If an Area Office did

---

[7] Figure A.1 in Appendix A provides a map of the 29 states under OSHA's jurisdiction. The other 21 states operate state-run programs that are approved by OSHA.

[8] OSHA Form 300, which OSHA standard 29 CFR 1904 requires employers to complete, is available at https://www.osha.gov/recordkeeping/RKforms.html, accessed March 2019.

not anticipate having sufficient resources to inspect its entire primary list, it estimated 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 before the following year's lists were announced, it 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. 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; worker representatives were also sometimes present. The inspector provided feedback on the workplace's safety program and explained any violations detected. OSHA typically assessed a fine when violations were found. 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. Inspections that cite violations increase the incentives for managers to take corrective actions and 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 at inspected establishments 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 during the meeting with management (e.g., 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 have no net impact on these business outcomes or might even improve them (Levine, Toffel, and Johnson 2012).

2.2 Data

We acquired and merged data from four sources: (a) OSHA's annual SST target lists (2001–2010);[9] (b) OSHA's annual ODI survey data on injuries and employment (1996–2011);[10] (c) OSHA inspection data from its Integrated Management Information Systems (IMIS) database (1990–2016);[11] and (d) annual Dun & Bradstreet data on employment, credit rating, and other business outcomes from the National Establishment Time Series (NETS) database (1990–2013).[12]

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 each 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 comprise DART injuries (those resulting in

---

[9] 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.
[10] We obtained SST target lists and ODI survey data from OSHA's Office of Statistical Analysis after signing a memorandum of understanding.
[11] We downloaded OSHA inspection records in January 2014 from the agency's publicly available IMIS database.
[12] NETS is a proprietary database distributed by Walls and Associates (Donald Walls, dwalls2@earthlink.net).

"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.[13] The ODI dataset is an unbalanced panel because it includes a different (but overlapping) set of establishments each year. The ODI sought to survey all establishments with at least 40 employees in a set of hazardous industries every three years. Beginning around 2005, it resurveyed the following year those establishments that reported a DART rate of at least 7. Many establishments on the SST target lists report ODI data nearly every year.[14] This is because (a) OSHA's DART rate threshold for re-sampling establishments in the ODI survey was below OSHA's threshold for placing establishments on the SST primary target list (and at or above the threshold used to place establishments on the secondary list) and (b) injury rates tended to be serially correlated.[15]

The IMIS database includes records for every inspection attempted by OSHA. Each record includes the name and address of the inspected 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, an employee complaint), and the number of violations and associated penalties.

NETS is an annual panel dataset extracted from Dun & Bradstreet data. It seeks to include all establishments in operation at any point since 1990. From NETS, we obtained each

[13] 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.

[14] 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.

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

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 establishment 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 establishments on the SST target list to an inspection record in IMIS.[16]

2.3 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 DAFW injuries because they are (a) the most serious type reported to ODI, (b) more likely than other injuries to be reported by employees to supervisors (Biddle and Roberts 2003), and (c) more likely to be recorded by employers on their 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.)

[16] 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.

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 U.S. 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[17] 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[18]). We use each establishment's lowest monthly PAYDEX score each year.[19] 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.

3 Methods

In this section, we first describe our methods to estimate the average treatment effects of randomized SST inspections in order to establish a baseline for comparison. We then describe

[17] 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 ODI.

[18] 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.

[19] We obtain essentially identical results if we instead use the establishment's highest monthly PAYDEX score in a given year.

our methods to estimate heterogeneous treatment effects of inspections and how this enables us to estimate the effects of alternative targeting policies that OSHA could conduct.

3.1 Estimating Average Treatment Effects of Randomized SST Inspections

In this subsection, we describe the sample and the econometric models we use to estimate the average treatment effects of the SST inspections OSHA allocated using random assignment.

3.1.1 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.[20] 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—those establishments that had already received a comprehensive safety inspection in the previous two years.[21] 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.

[20] 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.

[21] 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.

To focus on those establishments at risk of a *randomized* inspection, we retained only those establishments that were either (a) randomly assigned for an SST inspection ("assigned to inspection"), which collectively serve as our treatment group, or (b) eligible for but not assigned to a randomized SST inspection ("not assigned to inspection"), which collectively serve as our control group. To implement this, 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, as those establishments were not subject to random assignment.[22] Finally, we excluded establishments that OSHA had added to its target lists solely due to the agency's concern that their exceptionally *low* injury rates reported to ODI might be inaccurate.

Implementing these steps yielded a randomized sample consisting of 6,977 establishments assigned to inspection and 9,164 that were not assigned. (See Table A.1 in Appendix A for more details on how we trimmed the full SST target lists to arrive at this randomized sample.) 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 SST target list and was thus eligible for assignment to inspection.[23] We construct a nine-year panel dataset around these 16,141 establishment-

[22] 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 were assigned to inspection and not assigned to inspection, because our regressions (described below) compare these two groups within the same Area Office directive.
[23] For example, all establishments placed on the target list issued on May 14, 2007 have a directive year of 2007.

directives: the four years prior to the directive year (the "pre-period") and the five years comprising the directive year and the four subsequent years (the "post-period").[24]

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. This primarily affects establishments on target lists in later directive years, as they had fewer opportunities to appear again in the ODI because it ended in 2011. Some of these instances are due to 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 our 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 (made up of 11,083 unique establishments). Our attrition rate is much smaller for outcomes in the NETS database: we observe employment values in NETS during the post-period for all but 390 (2.4 %) of the establishments in our randomized sample. We discuss sources of this sample attrition in Appendix E.

Table 1 reports the industry distribution of establishments in our sample. Table 2 reports summary statistics for our outcomes.[25]

Balancing tests to validate random assignment. To validate whether assignment to inspection was truly 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. In separate regressions, we predict the

[24] 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 establishments ceased operation within five years of their directive year.

[25] We follow OSHA's rules and calculate injury rates 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.

number of years the establishment had previously appeared on the SST target list and the number of OSHA inspections the establishment had experienced in the four years prior to the directive year. We also predict the establishment's DAFW injury rate, DART injury rate, employment, total hours worked, and credit rating (minimum PAYDEX score), using values from the two years before the directive year.[26]

These regressions indicate that 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 of these metrics, establishments assigned to inspection are roughly 2.5% larger, a difference that is 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, which produced 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, a comparison of those assigned to inspection to those not assigned does not yield the causal effect of *receiving* an SST inspection for three reasons.

First, OSHA issued the annual SST directives between April and August each year and Area Offices did not begin conducting those SST inspections until several months later. As a result, only 18% of establishments that OSHA randomly assigned to inspection received an SST

[26] 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 if we instead use values four years prior to the directive year.

inspection by the end of the calendar year in which it was placed on the target list (recall that we refer to this as the directive year).

Second, *assignment* to an SST inspection does not perfectly correspond to eventually *receiving* an inspection. OSHA did not inspect some assigned establishments because the inspector could not physically locate the establishment or because an Area Office successfully petitioned OSHA headquarters for permission not to inspect all of the establishments that had been assigned for SST inspection. One year after the directive year, 73% of establishments assigned to inspection had received an SST inspection (see Figure A.2 in Appendix A for how this proportion changed over time).[27]

Third, some establishments eligible but not assigned to inspection in a given directive year (in which they thus serve as controls) were assigned to inspection in a subsequent year. Almost all establishments 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 subsequently received an SST inspection within the next four calendar years.[28]

In short, some establishments assigned to an SST inspection did not receive one and some that were not assigned in a given year *did* subsequently receive one (having been assigned to an SST inspection 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"

---

[27] It is possible that OSHA inspected some of these establishments, but our procedure to link SST with OSHA's information system (IMIS) failed to find their corresponding inspections in IMIS.

[28] Another potential source of fuzziness for our experimental design is that establishments might have experienced other types of inspections. For example, if establishments we classified as "not assigned to inspection" experienced 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.

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 average treatment effect of randomized SST inspections.

3.1.2 Specification to Estimate Intent-to-treat Effects

To ensure that our estimates are not vulnerable to large outliers or to threats to identification, we conducted a series of tests to pre-specify our regression specification. To implement this procedure, 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 actually were assigned to inspection in the sample. We then estimated the effect of *assigned to placebo inspection* on various outcomes using several functional forms and approaches in order to handle outliers.[29] 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.[30]

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

---

[29] For example, we considered OLS specifications in regressions with injury rate (both level and log) as the dependent variable.
[30] We posted our pre-analysis plan to the Open Science Framework on July 10, 2015 at https://osf.io/2snka/.

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

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. We use an approach sometimes called analysis of covariance (ANCOVA) to improve precision (McKenzie 2012) by controlling 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.[31]

$\boldsymbol{X}_{it}$ refers to additional control variables for each establishment-directive, meant to improve precision. 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

[31] In our Poisson specifications, we control for baseline *log(y+1)* instead of baseline unlogged values of *y* for the following reason. In a Poisson regression, 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 more appropriate to reflecting the correlation between baseline and post-period outcomes. We add 1 to y_{it}^{pre} before taking the log to account for zeroes.

a series of fixed effects for each Area-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, α_1 might yield a biased estimate of the effect of assignment to inspection if treatment and control establishments have differential pre-trends prior to assignment. We address these concerns by estimating a distributed lag specification. Specifically, we estimate the annual difference in injury counts—in each of the four years prior to, the year of, and the four years following the directive year—between establishments that were assigned to inspection and those that were not assigned to inspection using Equation 2:

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

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 $k \in \{-4, 4\}$. The set of β_k coefficients estimate the difference in outcome y between establishments assigned to inspection and those not assigned, 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 in order to control for time-invariant differences across establishment-directives.[32] 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 μ_{jt} is invariant within establishment-directives and would otherwise be dropped from the regression.

---

[32] 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.

3.1.3 Instrumental Variables Specification

In Equation 1, α_1 yields the difference in y^{post} between establishments assigned to inspection and those not assigned. However, as discussed in the section above on fuzziness in the treatment assignment, that 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:

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

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:

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

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, specifying Equation 3 as the conditional mean function for *y*, assuming *y* conditional on our set of explanatory variables is Poisson-distributed. For our other outcome variables, we use linear IV. As before, we cluster standard errors by establishment.

3.2 Constructing Alternative Targeting Criteria

The approach in Section 3.1 yields unbiased estimates of the average effects on injuries of assignment to SST inspection and of actually receiving an SST inspection—among the subset of inspections allocated randomly. That analysis serves as an evaluation of OSHA's targeting of inspections to establishments with high DART injury rates two years prior. In this section, we describe two machine-learning–based measures OSHA could have used to target its inspections and which we use here to assess whether OSHA's effectiveness could thus have been improved.

3.2.1 Estimating Heterogeneous Treatment Effects of Inspections

If the effect of inspections on safety is heterogeneous, OSHA could potentially have averted more injuries had it targeted those workplaces that were especially responsive to inspections. In this section, 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's disposal to allocate differently.

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 an establishment's 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 (or $s_0(Z)$) is challenging. For example, while one could include numerous interaction terms in a regression model to test for heterogeneous effects, there are dozens of candidate interactions and including many interaction terms can lead to spurious overfitted 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*, a supervised machine learning method developed by Wager and Athey (2018) 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.

With causal forest, Wager and Athey (2018) adapt the random forest to estimate the heterogeneity in causal effects. Causal forest searches for high-dimensional combinations of covariates that are associated with different treatment effects. To mitigate against biased estimates due to 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 requires that assignment to inspection is 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 that have at least 50 observations and for which the share of treatment or control observations in the leaf is no less than 10 percent. For more description of the causal forest approach, see Wager and Athey (2017). 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, $\overline{y_{it}^{post}}$, instead of each of these annual outcomes.

3.2.2 Predicting Establishments' Injuries

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

As a second metric that OSHA could use for targeting, we consider the number of 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 that OSHA used in its SST program, but differs in two 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, $b_0(Z)$ is a measure of injury *counts*, as opposed to *rates*; this is important because injury rates can be subject to substantial mean reversion, especially for smaller establishments (Ruser 1995). For these two reasons, predicted injury count is likely a better measure of an establishment's underlying health and safety problems at the time of inspection than is its (realized) injury rate two years earlier.[33]

There are several reasons why targeting inspections to establishments with high estimated $b_0(Z)$ could be an effective approach to 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. Thus, 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.

Compared to $s_0(Z)$, $b_0(Z)$ has the advantage of being observed for establishments not assigned to SST inspection. This makes estimating $b_0(Z)$ a standard machine learning prediction problem. Using the establishments not assigned to inspection, we use an ensemble machine learning procedure called Super Learner to construct $B(Z)$, our estimate of $b_0(Z)$. Super Learner minimize the mean squared error of out-of-sample predictions by using cross-validation

[33] A third distinction of our approach is that, whereas the SST program's targeting protocol relied on DART injuries (that is, 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.

to find the optimal weighted average among various machine learning methods (van der Laan, Polley, and Hubbard 2007).[34]

3.3 Evaluating Alternative Targeting Policies

Causal forest (Super Learner) might not generate consistent estimates of $s_0(Z)$ ($b_0(Z)$) for individual observations (Chernozhukov et al. 2018). These "raw" estimates might therefore lead to misleading estimates of the effects of alternative inspection 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.[35] Using the auxiliary sample, we use causal forest to construct $S(Z)$ — the estimate of the function determining establishments' CATEs. Using those establishments in the auxiliary sample not assigned to inspection, we use Super Learner to estimate $B(Z)$, the number of injuries an establishment would experience if not assigned to inspection.

Using these functions estimated on the auxiliary sample, we then compute the predicted CATE ($S(Z)$) and predicted baseline average ($B(Z)$) for each establishment in the holdout

[34] Our Super Learner library includes random forest (Breiman 2001), the Generalized Additive Model, and a linear interaction model. We selected these three learners by initially running 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 leafs in the random forest to have at least 50 observations, because small leafs can increase mean squared error (Athey and Imbens 2016).

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

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 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 $S(Z)$. 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:

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

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.

While we seek to evaluate counterfactual targeting policies that apply to the entire historical SST lists, we can only apply the CDDF procedure to our randomized sample because the propensity score $p(Z)$ is an input into the regression models described in Equation 5 and we do not know if it is uncorrelated with potential outcomes for 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 per establishment 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 4.4.2 discusses potential threats to whether the estimates from this approach will apply to the entire target list (rather than just the randomized sample) and how we address them.

Our estimates are based on the assumption that employers would not change their behavior in response to our alternative inspection targeting regimes. We believe this assumption is reasonable in our setting. Under OSHA's historical SST policy, establishments could presumably know whether they were at risk of SST inspection if they knew the DART rate thresholds OSHA used to create its target lists (though our discussions with OSHA officials and safety officials suggest that most establishments were unaware of OSHA's precise targeting strategy). Given the black-box nature of the machine learning algorithms that underpin our targeting regimes, it would be far more difficult for establishments to know whether they were placed on the target lists, greatly thwarting their ability to identify their likelihood of inspection.

4 Results

4.1 Average Effects of OSHA's SST Inspections

4.1.1 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 corresponding to Equation 1. Establishments assigned to SST inspection experience 3.4% fewer injuries over the directive year and four following years ($\beta = -0.035$, 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 from being assigned to inspection varies over time, Figure 1 displays the set of annual $ß_k$ coefficients and their 95% confidence intervals from a Poisson regression that estimates 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 ($\tau = 0$), the coefficient becomes negative, hovering between -0.04 and -0.05 each year, and is statistically significant in years $\tau = 0$ and $\tau = 1$.

To estimate the effect of having been inspected (the treatment-on-the-treated effect), we must account for some establishments assigned for inspection not being inspected and some not assigned being inspected in later years (see Section 3.1.1). Column 3 reports an OLS estimate of the first-stage effect of being assigned to an 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 and above the 17% inspection rate over this period among those not assigned to inspection in the directive year.[36] Column 4 reports the effect of receiving an SST inspection on DAFW injuries (the 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 ($\beta = -0.091$, SE = 0.042, $p = 0.03$). Because control

[36] 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).

establishments averaged 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, this translates to each randomized inspection averting $120,000 in injury costs over this five-year period,[37] which is roughly 35 times OSHA's cost of conducting an inspection.[38]

Our estimates are robust to a number of alternative specifications and other checks (presented in Appendix H). We ran our ANCOVA model, dropping establishments that had ever received a violation from OSHA for injury record-keeping. 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. First, we used targeted maximum likelihood estimation (TMLE) combined with Super Learner (van der Laan and Rose 2011). Second, we estimated with mean intent-to-treat effect with the CDDF procedure. All results were economically similar to and statistically indistinguishable from the

---

[37] 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 in the types of injuries studies likely explains why our estimate that SST inspections have a $120,000 social benefit is much lower than Levine, Toffel, and Johnson (2012)'s 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.

[38] 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.

results reported in Table 4.

4.1.2 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.[39] None of the coefficients on *SST inspected* are economically large or statistically significant (see Table 5).[40] Because these results provide no evidence that SST inspections harmed business outcomes, we do not analyze business outcomes when we evaluate counterfactual targeting policies.

4.2 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.

---

[39] 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 that is qualitatively similar to the intent-to-treat version of the coefficient we report.

[40] We obtain essentially identical estimates using a fixed-effects specification (results not shown).
The outcomes in Columns 3–8 that yield point estimates with the largest magnitude (most negative) are ODI hours and employment. While these point estimates could suggest that assignment to inspection leads to a reduction in employment among surviving establishments, the results in Column 1 suggest that being assigned to inspection slightly increases the likelihood that an establishment survives. When we estimated regression models with ODI-reported log total working hours or employment as the outcome—but where these outcomes take the value of 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).

4.2.1 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 $S(Z)$.[41] These estimates are unlikely to be unbiased (Chernozhukov et al. 2018), but they can give us a sense of the potential heterogeneity in conditional average treatment effects 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 that is 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.

4.2.2 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 establishments with the largest CATEs differ from those of establishments with the lowest CATEs. 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

---

[41] 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 $\widehat{\beta_1} + \widehat{\beta_2} * (S - ES)$.

bottom-20% group means, the standard errors of the means, the difference between these means, and the p-value on this difference.

During the pre-period, establishments with the largest estimated CATEs have substantially higher DAFW injury counts and employment than establishments 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.[42] 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. That DART rates are poor predictors of CATEs implies that other targeting criteria might better identify establishments with high CATEs. 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.

4.3 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 an inspection in the prior two years.[43]

[42] 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.

[43] 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.

4.3.1 *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 placed roughly 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 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 instead focus on those establishments that it predicts will *respond* to inspections by reducing injuries most aggressively—that is, those with the most negative $S(Z)$s—or those that it predicts will experience the most injuries in the absence of an inspection—that is, those with most positive $B(Z)$s. Second, OSHA could change the size and inspection probabilities of its primary and secondary lists. For example, it could create smaller primary lists made up of establishments that the agency 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 purposes, the parameters and effects associated with OSHA's historical policy. Specifically, it reports the criterion OSHA used to target its inspections (Column 1, the DART rate from two years prior) and the number of establishments assigned to inspection from its 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.[44] 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),[45] 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).

In the next two rows, we 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 of these benchmark policies, we assume that OSHA assigns to inspection the same number of establishments each year as the SST policy did: 16,861 over the 10-year period. If OSHA had targeted based on the largest estimated CATEs, it would have averted an average of 0.433 (SE = 0.23) injuries per year among assigned establishments, or 36,625 injuries total, which is roughly 2.41 times as many injuries averted as the historical SST policy. Alternatively, OSHA would have averted 2.29 as many injuries had it targeted based on the most predicted injuries.

4.3.2 Variations on the Benchmark Targeting Policies

These benchmark policies might not be feasible or even preferable. We therefore consider next how variations on the targeting strategies in these benchmarks would affect the number of injuries averted.

---

[44] 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. This difference arises for 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.

[45] 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.

Maintaining OSHA's inspection budget. Establishments expected to have the greatest reduction in injuries following inspection tend have more employees than average (Table 6). This is also the case for 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 next 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).[46]

[46] 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.

Maintaining the threat of inspection and other benefits of randomization. By assigning some inspections at random, the SST policy imposed the threat of being inspected on a broad pool of establishments. This threat motivates non-inspected establishments to deter noncompliance and improve safety (Cohen 2000; Shimshack and Ward 2005; Gray and Shadbegian 2007).[47] 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 comes with the opportunity cost that OSHA would not inspect some of the establishments where inspections are predicted to avert the most injuries.

To inform these tradeoffs, we consider policies whereby OSHA conducts *all* of its 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 (which results in a probability of 9%). We estimate that 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.

[47] 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.

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 it also highlights the magnitude of the opportunity cost of randomization relative to the benchmark policies. Because randomization results in many of the establishments for which inspections most effectively achieve the agency's objectives going uninspected, OSHA would avert substantially fewer injuries at assigned establishments than it would with the benchmark policies.

4.3.3 Discussion

In each of the policies we consider, OSHA averts nearly as many, if not more, injuries by targeting on $B(Z)$ than by targeting on $S(Z)$. At first blush, this is a surprising result. In fact, if our estimates of establishments' CATEs ($S(Z)$) were perfect proxies for their true underlying CATEs ($s_0(Z)$), this result would be impossible. However, estimating heterogeneous treatment effects is a challenging problem (Chernozhukov et al. 2018). Thus, our estimated CATEs are noisy predictors of true CATEs, likely subject to much sampling variation. [48] In contrast, estimating the expected number of injuries absent an inspection ($b_0(Z)$) is a much more standard prediction problem. $B(Z)$ is a very strong predictor of $b_0(Z)$, with very little sampling variation (see Appendix I). Above, we showed that $S(Z)$ and $B(Z)$ are correlated (Table 6). In short, in

[48] 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 (SE = 0.8), which indicates that our SIZ) is a meaningful—but somewhat noisy—proxy for underlying CATEs.

this setting, using a well-estimated proxy for $s_0(Z)$—that is, $B(Z)$—targets as well or better than using a direct estimate of $s_0(Z)$—namely, $S(Z)$— that is not estimated as precisely.[49] In other settings—for example, when $s_0(Z)$ and $b_0(Z)$ are unlikely to be highly correlated—this would not necessarily be the case.

4.4 **Threats to the Validity of Our Estimates of Counterfactual Policies**

In this section, we address the robustness of our estimates of the effects of alternative targeting policies.

4.4.1 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 $B(\cdot)$, respecitvely. We then use these models to generate predicted CATEs ($S(Z)$) and predicted baseline number of 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 the (a) 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

[49] 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).

main approach) to the benefits when we estimate $S(Z)$ and $B(Z)$ using only the earlier 2001–2006 data.

First, we use an identical procedure to estimate the average treatment effect (Table 7, Row 1) and the predicted benefits of our benchmark policies, as we did in our main analysis (Table 7, Rows 2 and 3), but using only the observations from the second half of our time period (2007–2010). The estimated number of averted injuries per year from assignment to inspection among establishments in the high-priority group in each of these policies 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 separately the *B(Z)*s—estimated on the 2001–2006 sample. To estimate the average number of injuries averted among establishments in the high-priority groups under these policies, we run a regression corresponding to Equation 5. 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 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.

4.4.2 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 very different observable characteristics (i.e., different *Z*s) or if the function that maps *Z* to treatment effects, $s_0(\cdot)$, differed between the randomized and nonrandomized samples (i.e., if the two groups have different unobservables). In Appendix I, we provide evidence that the randomized and nonrandomized samples do not have meaningfully different observables associated with $S(Z)$ or $B(Z)$. We also do not find any evidence that our machine learning model to estimate $B(Z)$ has differential predictive power for the two samples. While we cannot rule out differences on the mapping between observable factors and treatment effects, we have no evidence that these issues affect our estimates.

5 Discussion

OSHA's 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 establishment survival and employment.

At the same time, the agency could have averted many more injuries by targeting workplaces where the expected number of averted injuries is high. We first used machine learning methods to estimate heterogeneous treatment effects. We then simulated alternative targeting policies that OSHA could have conducted with the aid of machine learning. The agency could have averted over twice as many injuries if it had targeted its inspections to the establishments with the largest predicted treatment effects.

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, we estimate that it would still avert more injuries than the historical policy, but by a substantially smaller amount. This illustrates the magnitude of the tradeoff associated with randomizing inspections in our setting. Though randomizing ensures that the agency can both keep learning where it is effective and maintain general deterrence for uninspected establishments, it also means that many establishments for which inspections would most effectively achieve the agency's goal will go uninspected. Using these policies as bounds is informative, as an agency could experiment with a mixture of nonrandom targeting of establishments with the highest predicted benefits while randomizing the rest.

Regardless of its use of randomization, OSHA could avert just as many—if not *more*—injuries by targeting establishments predicted to have the most injuries. This result highlights that when heterogeneous treatment effects are difficult to predict, it can more effective to target on a simpler outcome that is easier to predict and that is likely to be correlated with underlying treatment effects. In fact, targeting on predicted injuries is consistent with OSHA's goal of targeting inspections to establishments with "serious health and safety problems" (US Occupational Safety and Health Administration 2004). OSHA measured "serious health and safety problems" based on a single year's injury rate; a measure with substantial mean reversion. In contrast, our measure of predicted injuries accounts for mean reversion and better identifies workplaces with persistent safety and health problems.

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 inspections in other sectors. OSHA could investigate why and attempt to improve its effectiveness. For example, it does not have a 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. Alternatively, OSHA might find that lower effectiveness in some sectors is due to poor training; improved training for those sectors would be called for.

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

It is possible that employers will change their safety behaviors if they perceive a change in the probability of inspection. Our estimates would not reflect such behavioral changes under a new targeting regime, but we expect such reactions to be minimal. Firms had little information about OSHA’s targeting strategies under SST, so it is unlikely they would adjust to alternative targeting policies. Furthermore, the opacity of the machine learning procedure makes it even more difficult to game the system by predicting one’s likelihood of inspection.

With these limitations in mind, we show that combining randomization and machine learning could substantially improve regulatory agencies’ performance. We also expect 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 to multinational firms that assess their 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 their predicted risk of producing contaminated food (US Government Accountability Office 2016). In 2018, the US Bureau of Safety and Environmental Enforcement began targeting some of its 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). And while Chicago, like many American cities, conducts food safety inspection of all of its restaurants every year, it has begun using risk-based forecasting to help determine the inspection sequence (Spector 2016). Our research reveals how agencies can estimate the relative benefits of alternative algorithms that feature varying levels of simplicity and transparency as well as general deterrence to encourage compliance among non-inspected establishments.

DUKE UNIVERSITY

UNIVERSITY OF CALIFORNIA, BERKELEY

HARVARD BUSINESS SCHOOL

6 References

Alm, James, and Jay Shimshack, "Environmental Enforcement and Compliance: Lessons from Pollution, Safety, and Tax Settings," *Foundations and Trends in Microeconomics,* 10, no. 4 (2014), 209–274.

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

Athey, Susan, and Guido Imbens, "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113, no. 27 (2016), 7353–7360.

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

Azaroff, Lenore S., Charles Levenstein, and David H. Wegman, "Occupational Injury and Illness Surveillance: Conceptual Filters Explain Underreporting," *American Journal of Public Health*, 92, no. 9 (2002), 1421–1429.

Bartel, Ann P. and Lacy Glenn Thomas, "Direct and Indirect Effects of Regulation: A New Look at OSHA's Impact," *Journal of Law and Economics*, 28, no. 1 (1985), 1–25.

Biddle, Jeff, and Karen Roberts, "Claiming Behavior in Workers' Compensation," *Journal of Risk and Insurance,* 70, no. 4 (2003), 759–780.

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

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

Breiman, Leo, “Random Forests,” *Machine Learning*, 45, no. 1 (2001), 5–32.

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

Choi, Jinkyung, and Barbara Almanza, “Health Inspectors’ Perceptions of the Words Used to Describe Violations,” *Food Protection Trends*, 32, no. 1 (2012), 26–33.

Cho, Sungjin, and John Rust, “The Flat Rental Puzzle,” *The Review of Economic Studies*, 77, no. 2 (2010), 560–594.

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

Davis, Jonathan M. V., and Sara B. Heller, “Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs,” NBER Working Paper No. 23443, 2017.

DellaVigna, Stefano and Matthew Gentzkow, “Uniform Pricing in US Retail Chains,” Working Paper, 2019.

Dong, Xiuwen S., Alissa Fujimoto, Knut Ringen, Erich Stafford, James W. Platner, Janie L. Gittleman, and Xuanwen Wang, “Injury Underreporting among Small Establishments in the Construction Industry,” *American Journal of Industrial Medicine*, 54, no. 5 (2011), 339–349.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, “The Value of Regulatory Discretion: Estimates from Environmental Inspections in India,” *Econometrica*, 86, no. 6 (2018), 2123–2160.

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

ERG, *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 Opinions Research Center, *OSHA Data Initiative Collection Quality Control: Analysis of Audits on CY 2006 Employer Injury and Illness Recordkeeping*, prepared for the US Occupational Safety and Health Administration’s Office of Statistical Analysis [contract no. J-099-F-2-8441], November 25, 2009.

Feldman, Justin, *OSHA Inaction: Onerous Requirements Imposed on OSHA Prevent the Agency from Issuing Lifesaving Rules* (Washington, DC: Public Citizen’s Congress Watch, 2011).

Foley, Michael, Z. Joyce Fan, Eddy Rauser, and Barbara Silverstei. 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 (2012), 976–990.

Glaeser, Edward L., Andrew Hillis, Scott Duke Kominers, and Michael Luca, "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy," *American Economic Review*, 106, no. 5 (2016), 114–118.

Gonzalez-Lira, Andres, and Ahmed M. Mobarak, *Slippery Fish: Enforcing Regulation under Subversive Adaptation* (No. 12179). Bonn, Germany: Institute for the Study of Labor (IZA), 2019.

Gray, Wayne B., and John M. Mendeloff, "The Declining Effects of OSHA Inspections on Manufacturing Injuries, 1979–1998," *ILR Review*, 58, no. 4 (2005), 571–587.

Gray, Wayne B., and John T. Scholz, "Does Regulatory Enforcement Work? A Panel Analysis of OSHA Enforcement," *Law and Society Review*, 27, no. 1 (1993), 177–214.

Gray, Wayne B., and Ronald J. Shadbegian, "The Environmental Performance of Polluting Plants: A Spatial Analysis," *Journal of Regional Science*, 47, no. 1 (2007), 63–84.

Hanna, Rema Nadeem, and Paulina Oliva, "The Impact of Inspections on Plant-level Air Emissions," *BE Journal of Economic Analysis & Policy*, 10, no. 1 (2010), Article 19.

Haviland, Amelia M., Wayne B. Gray, John Mendeloff, Rachel M. Burns, and Teague Ruder, "A New Estimate of the Impact of OSHA Inspections on Manufacturing Injury Rates, 1998–2005," *American Journal of Industrial Medicine*, 55, no. 11 (2012), 964–975.

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

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133, no. 1 (2017), 237–293.

Kleven, Henrik Jacobsen, Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez, "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark," *Econometrica*, 79, no. 3 (2011), 651–692.

Lee, Jonathan M., and Laura O. Taylor, "Randomized Safety Inspections and Risk Exposure on the Job: Quasi-experimental Estimates of the Value of a Statistical Life," *American Economic Journal: Economic Policy*, forthcoming.

Leigh, J. P., "Economic Burden of Occupational Injury and Illness in the United States," *The Milbank Quarterly*, 89, no. 4 (2011), 728–772.

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

Li, Ling, and Perry Singleton, "The Effect of Workplace Inspections on Worker Safety," *Industrial and Labor Relations Review*, 72, no. 3, (2019): 718–748.

Messiou, Eleni, and Brian Zaidman, "Comparing Workers' Compensation Claims and OSHA Data Initiative Cases" (St. Paul, MN: Minnesota Department of Labor and Industry, 2005).

Nestoriak, Nicole, and Brooks Pierce, "Comparing Workers' Compensation Claims with Establishments' Responses to the SOII," *Monthly Labor Review*, 132, no. 5 (2009), 57–64.

Oleinick, Arthur, Jeremy V. Gluck, and Kenneth E. Guire, "Establishment Size and Risk of Occupational Injury," *American Journal of Industrial Medicine*, 28, no. 1, (1995), 1–21.

Peto, Balint, Laura Hoesly, George Cave, David Kretch, and Ed Dieterle, *Evaluation of the Occupational Safety and Health Administration's Site-Specific Targeting Program—Final Report* (Washington, DC: Summit Consulting LLC, 2016), available at: https://www.dol.gov/asp/evaluation/completed-studies/SST_Evaluation_Final_Report.pdf.

Rubin, Donald B., "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, no. 5 (1974): 688–701.

Rubin, Richard, "IRS Audits of Individuals Drop for Fifth Straight Year," *Wall Street Journal*, Feb. 22, 2017, https://www.wsj.com/articles/irs-audits-of-individuals-drop-for-5th-straight-year-1487794717?shareToken=stf0c27866341641198c8ae23fe037012b.

Rudin, Cynthia, Rebecca J. Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac, "A Process for Predicting Manhole Events in Manhattan," *Machine Learning*, 80, no. 1 (2010), 1–31.

Ruser, John W., "Workers' Compensation and Occupational Injuries and Illnesses," *Journal of Labor Economics*, 9, no. 4, (1991), 325–350.

Ruser, John W., "Self-correction versus Persistence of Establishment Injury Rates," *Journal of Risk and Insurance*, 62, no. 1 (1995), 67–93.

Ruser, John W., "Examining Evidence on Whether BLS Undercounts Workplace Injuries and Illnesses," *Monthly Labor Review*, 131, no. 8 (2008), 20−32.

Ruser, John W. and Robert S. Smith, "Re-estimating OSHA's effects: Have the data changed?" *The Journal of Human Resources*, 26, no. 2 (1991), 212−235.

Safety+Health Magazine, "Compliance Assistance—Not Fines—Should Be Priority, Senators Tell OSHA," Feb. 12, 2016, https://www.safetyandhealthmagazine.com/articles/13662-compliance-assistance-not-fines-should-be-priority-senators-tell-osha, accessed March 2018.

Shimshack, Jay P., "The Economics of Environmental Monitoring and Enforcement," *Annual Review of Resource Economics*, 6, no. 1 (2014), 339–360.

Shimshack, Jay P., and Michael B. Ward, "Regulator Reputation, Enforcement, and Environmental Compliance," *Journal of Environmental Economics and Management*, 50, no. 3 (2005), 519–540.

Short, Jodi L., Michael W. Toffel, and Andrea R. Hugill. “Monitoring Global Supply Chains.” *Strategic Management Journal* 37, no. 9 (2016): 1878–1897.

Slemrod, Joel, Marsha Blumenthal, and Charles Christian, “Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota,” *Journal of Public Economics*, 79, no. 3 (2001), 455–483.

Smith, Rosen S., “The Impact of OSHA Inspections on Manufacturing Injury Rates,” *Journal of Human Resources*, 14, no. 2 (1979), 145–170.

Spector, Julian, “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/.

Stigler, George J., “The Theory of Economic Regulation,” *Bell Journal of Economics and Management Science*, 2, no. 1 (1971), 3–21.

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

US Bureau of Labor Statistics, *Nonfatal Occupational Injuries and Illnesses Requiring Days Away from Work, 2005* (2007), https://www.bls.gov/iif/oshwc/osh/case/osnr0027.pdf.

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

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

US Department of Health and Human Services, *Fiscal Year 2012 Food and Drug Administration, Justification of Estimates for Appropriations Committees* (2011).

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

US Food and Drug Administration, *2016 Annual Report on Inspections of Establishments in FY 2015* (2016), https://www.fda.gov/downloads/RegulatoryInformation/LawsEnforcedbyFDA/SignificantAmendmentstotheFDCAct/FDASIA/UCM483994.pdf.

US Government Accountability Office, *FDA’s Targeting Tool Has Enhanced Screening, But Further Improvements Are Possible* (May 2016), http://www.gao.gov/assets/680/677538.pdf.

US Occupational Safety and Health Administration, “Nationwide Site-Specific Targeting (SST) Inspection Program Request for Comments,” *Federal Register* 69 (May 6, 2004), 25445–25446.

US Occupational Safety and Health Administration, *Site-Specific Targeting 2008 Directive 08-03 (CPL 02), Effective Date May 19, 2008* (2008), https://www.osha.gov/OshDoc/Directive_pdf/CPL_02_08-03.pdf, accessed March 2019.

US Occupational Safety and Health Administration, *OSHA Field Operations Manual*, *Directive CPL-02-00-160, Effective Date August 2, 2016* (2016), https://www.osha.gov/OshDoc/Directive_pdf/CPL_02-00-160.pdf, accessed March 2018.

US Occupational Safety and Health Administration, "Commonly Used Statistics" (2017a), https://www.osha.gov/oshstats/commonstats.html, accessed February 2017.

US Occupational Safety and Health Administration, *OSHA Fact Sheet: OSHA Inspections* (2017b), 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, "Super Learner," *Statistical Applications in Genetics and Molecular Biology,* 6, no. 1 (2007), Article 25.

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

Viscusi, W. Kip, "The Impact of Occupational Safety and Health Regulation, 1973–1983," *Bell Journal of Economics*, 17, no. 4 (1986), 567–580.

Wager, Stefan, and Susan Athey. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, no. 523 (2018), 1228–1242.

Waehrer, Geetha M., Xiuwen S. Dong, Ted Miller, Elizabeth Haile, and Yurong Men, "Costs of Occupational Injuries in Construction in the United States," *Accident Analysis & Prevention*, 39, no. 6 (2007), 1258–1266.

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

Table I: 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 II: 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. The 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 III: 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 insp- ections 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 (# emp- loyees)	log (# emp- loyees)	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. The 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 IV: Effects of SST Inspection on Injuries Resulting in Days Away From Work (DAFW)

	(1)	(2)	(3)	(4)
	# of DAFW injuries		SST insp-ected	# 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 (i.e., 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 V: 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 (i.e., 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. The Min PAYDEX is the minimum score over all monthly reports in a year.

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

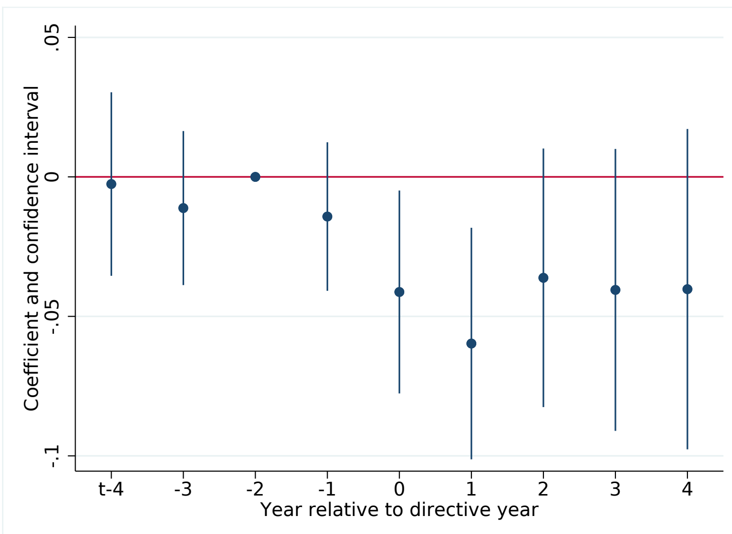
We conduct 250 random even splits of the randomized sample. In each iteration we train a causal forest on the auxiliary sample to predict CATE for estabishments in the holdout and nonrandomized samples. Among establishments in the holdout and nonrandoized samples, we identify those with the top 20% and 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 VII: 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)

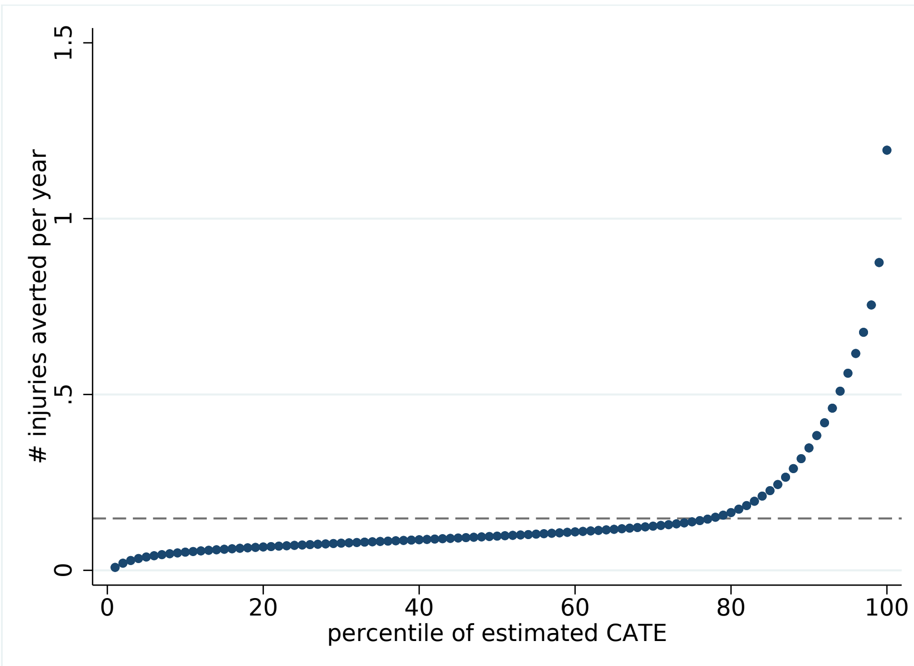
DART rate t-2 = DART rate from 2 years prior to the directive year.
S(Z) = estimate of difference of establishment's number of annual DAFW injuries if or if not assigned to inspection (i.e., its CATE).
B(Z) = estimate of establishment's number of annual injuries if not assigned to inspection.
The estimates 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 I: Temporal Effects of Assignment to Inspection on Injuries, by Year Relative to Directive Year



The figure displays 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 II: Percentiles of Estimated Number of Injuries Averted per Year If Assigned to SST Inspection



The graph presents estimates of the percentiles of Conditional Average Treatment Effect (CATE), or
(Z), $ofassignmenttoSSTinspectionamongthesetofestablishmentsonOSHA'sSSTtargetlistsfrom2001-$
$-2010.Eachdotrepresentsthemedian, across250samplesplits, ofthecorrespondingcentileoftheBestLinearPredictorof(Z),$
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 establishments exist												
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

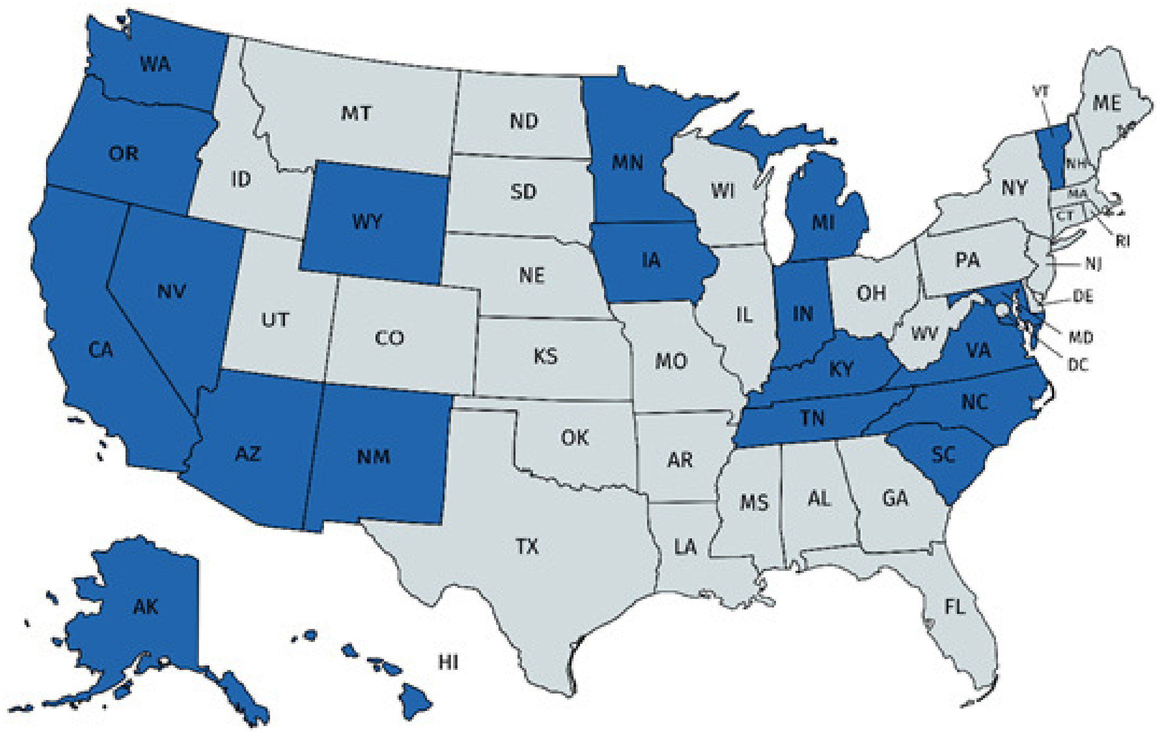
[a] An establishment-directive refers to a specific instance of a unique establishment being placed on a particular year's SST target list.
[b] Federal states restrict to the 29 states under federal OSHA jurisdiction. While a few of the 21 states with state-run OSHA offices participated in SST, they 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 listand with the percent of the corresponding primary list assigned to inspection equal to 1 and the percent of the seconday 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 prior to the directive start date—or 3 years prior, 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 prior to 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] Area Offices could assign subsets of their target lists 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)*

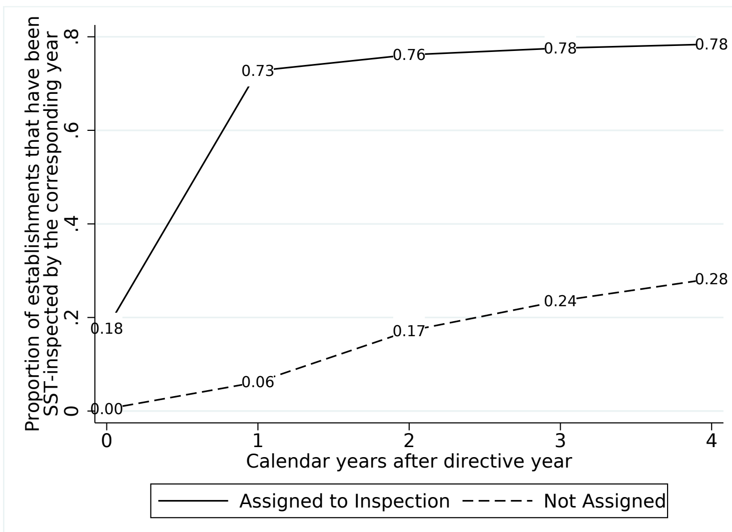
The dependent variable in each column is equal to the average number of injuries an establishment experienced over the 5-year period comprising the directive year and 4 subsequent years. The sample is establishment-directives in the randomized sample on the 2007–2010 SST target lists. *Would Be assigned in benchmark policy* is a dummy equal to 1 if an establishment's (Z)$(predictedCATE, ordifferenceinthenumberofannualinjuriesifandifnotassignedtoSSTinspectionovera5-yearperiod), or(Z)$ (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, (Z)$((Z))$ 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 map 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.[50] Each year, the OSHA Data Initiative (ODI) surveyed between 60,000 and 80,000 workplaces—with at least 40 employees—in these industries.[51] 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.[52]

Beginning in 1999, each year—between April and August[53]—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

---

[50] Specifically, the 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.

[51] 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.

[52] 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.

[53] 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.

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).

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 Occupational Safety and Health Administration 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.[54] Subsequent actions were similar to other types of inspections that OSHA conducts: the inspector walked through the establishment to assess hazards that could lead to injuries or illnesses. After

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

completing an inspection, the inspector conducted a closing conference with representatives from management and the employees. Inspectors typically discuss 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 Occupational Safety and Health Administration 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.

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.

Frist, 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. 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 2013).

Still, one may be concerned about measurement error affecting our estimates. There is very likely classical measurement error (i.e., 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 of the true number of injuries. 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 (see Azaroff et al. 2002 for a thorough discussion of these factors). 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 benefit levels (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).[55]

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-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, the extent of underreporting should be mitigated by the fact that ODI is restricted to relatively large employers (the minimum was 40 employees until 2009, and then it became 20). Third, we conducted a robustness check which excluded from our analysis any

---

[55] 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. The BLS surveys under-report 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) (Ruser 2008; Nestoriak and Pierce 2009).

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,[56] 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 the CPI rose by 5.4% from 2002 to 2005 according to the Bureau of Labor Statistics. 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 2016. Thus, the social cost of a 2005 DAFW injury in 2016 dollars is $40,717*1.286 = $52,362.

[56] 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. In this Appendix, we discuss the sources of our sample attrition and assess its relationship with assignment to SST inspection.

Table E.1 Illustrates some of the 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 assess 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)	(2)	(3)
	Number of establishment-directives in the randomized sample that…		
	…lack ODI data in the post-period	…have ODI data in the post-period	% 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 establishments' 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 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 three 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 large metro area
- The 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[57] 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

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

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

Related to injuries and ODI variables

- 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 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.[58]

We also sought to minimize the chances that our estimates were contaminated by inspections leading to more complete reporting of injuries. Therefore, we re-estimated 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).[59]

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 re-estimate 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.

[58] 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.

[59] 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).

TMLE is a double-robust approach to estimating treatment effects in the presence of potential misspecifications of the treatment assignment process.[60]

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:

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

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 recordkeeping violations lead to less under-reporting of 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

---

[60] 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 to 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).

OLS estimates, but the differences are not statistically significant or economically meaningful. In short, our results are robust to these several specification checks.

Table H.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 I **Extrapolating Machine Learning Estimates from the Randomized Sample to the Nonrandomized Sample**

Our estimation of the number of injuries that OSHA could avert under various alternative targeting policies relies on using the randomized sample to construct a prediction of establishments' baseline conditional average ($B(Z)$) and their CATEs if assigned to inspection ($s_0(Z)$) and then using these predictions to estimate the effects of counterfactual policies that target establishments from the overall historical SST target lists, which includes 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.

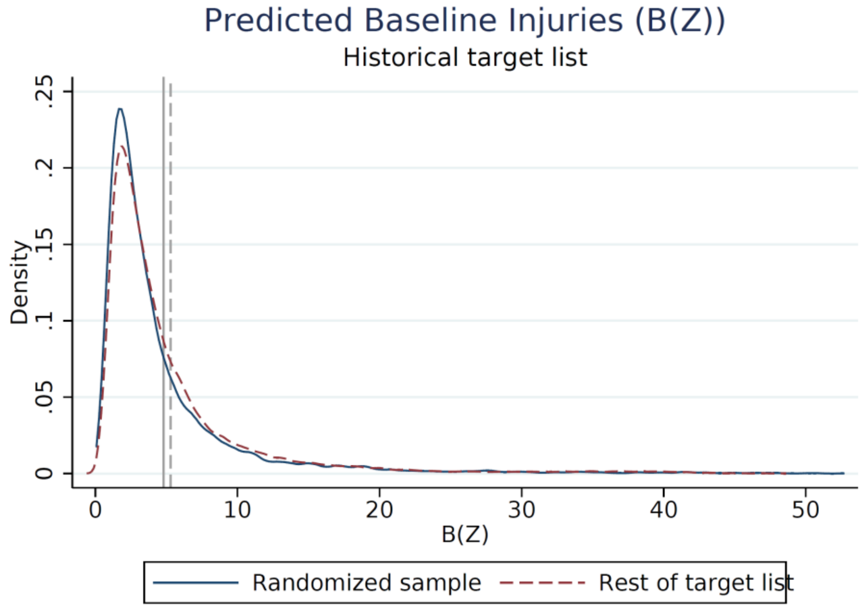
As described in Section 4.4.2, there are two reasons that the estimates from the CDDF procedure might not generalize to the entire historical SST target lists and thus might result in misleading estimates of targeting policy counterfactuals. First, if establishments in the randomized sample have different observable characteristics (i.e., different Zs), 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.

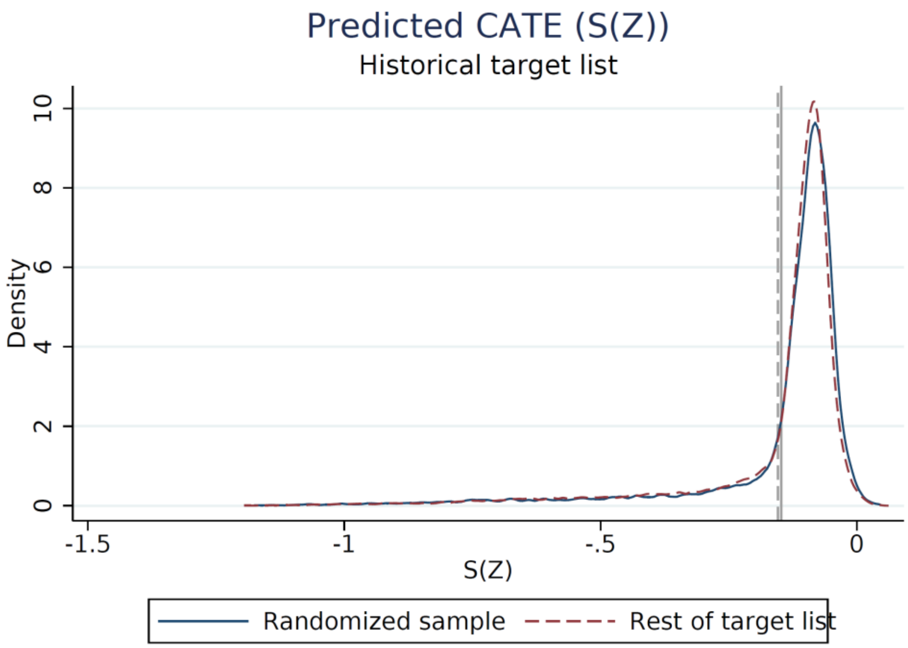
1) **Do establishments in the randomized and nonrandomized samples have different observable characteristics?**

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—which are estimated on the randomized sample—would not consistently estimate effects of policies as they pertain to the entire historical target list. However, these distributions appear to be very similar. Panels (a) and (b) of Figure H.1 plot the distribution of, respectively, $B(Z)$—an estimate of establishments' underlying baseline number of injuries ($b_0(Z)$)—and $S(Z)$—an estimate of establishments' underlying CATE ($s_0(Z)$). 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). The distributions of both measures are very similar for the two groups of establishments.

Given the similarity of these distributions, it is unlikely that any differences in *observable* characteristics between the randomized and nonrandomized samples render our estimates of counterfactual policies inconsistent.

Figure I.1. The distribution of $S(Z)$ and $B(Z)$ for establishments on the SST historical target list that are in the randomized sample vs. those in the 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 an SST inspection 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.

2) Does the predictive power of our 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 baseline number of injuries—fits worse in the nonrandomized sample than in the randomized sample.

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, $B(Z)$.

For establishments not assigned to SST inspection, both in the holdout sample and in the nonrandomized sample, the coefficients on $B(Z)$ are close to 1 and the estimates are statistically indistinguishable from each other (Table I.1). The R^2s are both high and nearly identical.

It remains possible that the model estimating the *causal* effect ($S(Z)$ has less predictive power for the nonrandomized sample than for the randomized sample. Nevertheless, these results are consistent with machine learning estimates from the randomized sample generalizing to the nonrandomized sample.

Table I.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	
	Holdout sample of controls	Nonrandomized sample
B(Z)	1.028	1.061
	(0.032)	(0.008)
R^2	0.686	0.702

The table assesses whether our machine-learning–based estimates of establishments' baseline conditional average—the number of annual DAFW injuries they would experience if not assigned to inspection—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) and apply the model to 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 1). The table reports the median coefficient and SE on $B(Z)$ and the median R^2, across the 250 splits.