American Economic Journal: Applied Economics 2023, 15(4): 30–67
https://doi.org/10.1257/app.20200659

Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA[†]

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

We study how a regulator can best target inspections. Our case study is a US Occupational Safety and Health Administration (OSHA) program that randomly allocated some inspections. On average, each inspection led to 2.4 (9 percent) fewer serious injuries over the next 5 years. Using new machine learning methods, we find that OSHA could have averted as much as twice as many injuries by targeting inspections to workplaces with the highest expected averted injuries and nearly as many by targeting the highest expected level of injuries. Either approach would have generated up to $850 million in social value over the decade we examine. (*JEL* C63, J28, J81, K32, L51)

While government agencies spend billions each year inspecting establishments for worker safety, environmental protection, tax compliance, and other concerns (Shimshack 2014; US Food and Drug Administration 2016; US Occupational Safety and Health Administration 2017a), most agencies' budgets enable them to inspect only a tiny share of the establishments they regulate (US Department of Health and Human Services 2011; Rubin 2017). For example, workplace safety regulators in the United States inspected less than 1 percent of the 8 million workplaces they regulated in 2016 (US Occupational Safety and Health Administration 2017a). Therefore, a crucial question for regulators is how to target their scarce inspections.

Many regulators direct inspections to establishments that have had the most severe problems, such as injuries, accidents, or emissions. For example, US occupational

*Johnson: Sanford School of Public Policy, Duke University (email: matthew.johnson@duke.edu); Levine: Haas School of Business, University of California (email: levine@berkeley.edu); Toffel: Harvard Business School, Harvard University (email: mtoffel@hbs.edu). Benjamin Olken was coeditor for this article. 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 Cheng Ju and Mark van der Laan. We received helpful comments from Dave Anderson, Jon Baron, Jon Davis, Ivan Fernandez-Val, Kris Ferreira, Eric Frumin, Daniel Jacob, 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. The authors declare no relevant or material financial interests that relate to the research described in this paper.

[†]Go to https://doi.org/10.1257/app.20200659 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

safety regulators use establishments' historical accident records to target many of their inspections (US Occupational Safety and Health Administration 2017b). A concern with this approach is that prior problems are a noisy signal of where problems continue to be severe. Machine learning approaches could improve predictions of future problems, as Glaeser et al. (2016) found with restaurant inspections and Hino, Benami, and Brooks (2018) found with environmental inspections.

Targeting inspections based on past problems might be poorly suited to promote regulators' objective of *improving* the safety, health, and environmental impact of establishments ranging from offices to restaurants to factories. Inspectors could instead seek to maximize the extent to which inspections *reduce* injuries, food-borne illnesses, emissions, and so on—that is, *treatment effects* (also taking into account any change in the threat effects to uninspected establishments). Targeting based on treatment effects could lead a regulator to target a different set of establishments. For example, workplaces with the most severe problems might have production processes that make them the least *responsive* to inspections (Athey 2017).

New machine learning methods that have enabled the estimation of inspections' heterogeneous treatment effects could be deployed to target inspections where inspections are estimated to avert the most injuries, illnesses, emissions, and so on. However, generating estimates of establishments' treatment effects is challenging because it requires both a large amount of data and inspections that were assigned randomly (or quasi-randomly); even then, estimates might still be quite imprecise. Thus, it is unclear whether the idea of targeting on estimated treatment effects can yield targeting lists that will, in practice, outperform regulators' conventional approaches.

We develop an approach to compare the effectiveness of a wide array of inspection targeting regimes. We first combine randomization and machine learning to evaluate the effects of inspections as historically allocated. We then evaluate the relative effectiveness of three alternative targeting policies based on (a) a sharper estimate of past problems, (b) a machine-learning–based prediction of future problems, and (c) a machine-learning-based prediction of treatment effects of inspection. We evaluate these alternatives using methods that avoid overfitting and enable valid inference.

We apply this approach to the US Occupational Safety and Health Administration (OSHA), the regulatory agency charged with assuring "safe and healthful working conditions" (US Occupational Safety and Health Administration 2021) and an important setting in which to examine regulatory effectiveness. First, workplace injuries and illnesses impose a substantial burden on the US economy, estimated at $250 billion in annual social costs (Leigh 2011). Second, OSHA has been controversial ever since its founding in 1970. Supporters argue that it saves lives at little to no cost to employers (Feldman 2011), whereas critics charge that its penalties are too low to affect behavior (Bartel and Thomas 1985) or that its regulations "don't add value to safety in the workforce."[1] Understanding whether OSHA inspections

[1] Senator Heidi Heitkamp (D-ND) during a February 11, 2016 hearing with the Senate Homeland Security and Governmental Affairs Committee (Musick, Trotto, and Morrison 2016). In other sectors, concerns about regulators'

are effective in improving safety and whether alternative targeting regimes could make them more effective therefore has important policy implications.

We focus on the Site-Specific Targeting (SST) program, which, during its 16 years of operation (1999–2014), was OSHA's largest inspection regime. To promote the most "effective use of [OSHA's] enforcement resources," it targeted establishments with "serious safety and health problems" by developing annual target lists of establishments that had high injury rates two years prior (US Occupational Safety and Health Administration 2004). When resource constraints prevented OSHA from inspecting all establishments on its target lists, the agency allocated inspections via random assignment.

As a basis of comparison for the alternative targeting regimes we will examine, we first evaluate the extent to which injuries were reduced by those SST inspections that OSHA randomly assigned between 2001 and 2010. By comparing establishments that were randomly assigned to inspections to establishments that were also eligible but not assigned, our estimates are free of the selection bias that plagues most evaluations of workplace inspections.[2] Roughly 13,000 establishments, employing nearly two million workers, were at risk of being targeted for a randomized inspection over this 10-year period. Our primary outcome variable is an establishment's annual number of injuries and illnesses leading to days away from work—henceforth "serious injuries." (For simplicity, we will refer to both injuries and illnesses as "injuries.") We estimate effects of inspections on serious injuries 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.

We find that randomly assigned OSHA inspections reduced serious injuries at inspected establishments by an average of 9 percent, which equates to 2.4 fewer injuries, over the five-year post-period. Each inspection thus yields a social benefit of roughly $125,000, which is roughly 35 times OSHA's cost of conducting an inspection.

Could OSHA have allocated its inspections to avert more injuries? Intuition suggests so. For one thing, the metric OSHA used to target SST inspections—injury rates from two years prior—is a noisy signal of which facilities face persistent serious health and safety problems; a more precise measure of persistent problems might avert more injuries. Furthermore, if establishments with prior safety problems or establishments predicted to have persistent safety problems are not those most *responsive* to inspections, the regulator could perhaps avert more injuries by targeting inspections on predicted *treatment effects*.[3]

targeting include political motivations, regulatory capture (Stigler 1971; Weisman and Wald 2013), and the desire to collect funds from violation penalties.

[2] 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 uninspected 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).

[3] We consider heterogeneity in the effect of *assignment* to SST inspection (the "intention to treat") rather than of inspection (the "treatment"), because, while OSHA randomized assignments to inspection, such assignments imperfectly predicted actual inspections.

We thus consider three alternative targeting criteria. First, we use the establishment's average annual number of serious injuries over the prior four years (*historical injuries*). Second, we use many observable characteristics and an ensemble of machine learning methods to predict the number of serious injuries an establishment would experience over the next five years if it were not assigned to inspection (*predicted injuries*). Third, we use a causal forest to estimate each establishment's treatment effect of assignment to inspection (*estimated treatment effects*). The causal forest is a flexible machine learning technique that predicts treatment effects based on high-dimensional nonlinear functions of observable characteristics (Wager and Athey 2018).

To estimate the number of injuries OSHA could avert under alternative targeting policies, we apply a method from Chernozhukov et al. (2020) that incorporates estimates of heterogeneous treatment effects with repeated sample splitting to yield consistent estimates of average treatment effects for subsets of the sample.

We find that OSHA could have averted many more injuries had it targeted inspections using any of these alternative criteria. If OSHA had assigned to those establishments with the highest *historical injuries* the same number of inspections that it assigned in the SST program, it would have averted 1.9 times as many injuries as the SST program actually did. If OSHA had instead assigned the same number of inspections to those establishments with the highest *predicted injuries* or to those with the highest *estimated treatment effects*, it would have averted 2.1 or 2.2 times as many injuries as the SST program, respectively. Based on our estimated social cost of injuries, these three alternative regimes would have generated social value of $643 million, $844 million, and $876 million, respectively. These estimates change very little if, rather than assigning the same number of inspections as the SST program did, we assign the number of inspections that would maintain SST's inspection budget.

This alternative policy that allocates all inspections nonrandomly, however, has two drawbacks: (a) it forgoes opportunities for the regulator to continue learning where inspections are most effective and (b) it potentially reduces the "threat effect" that motivates uninspected establishments to preemptively improve safety (often referred to as "general deterrence"). We therefore also consider a targeting policy that maintains the SST program's threat level, but ranks establishments using our three alternative targeting criteria rather than the SST program's criteria (injury rate from two years prior). Here, we create a high-priority list with the same number of establishments that were on the SST's primary list each year and place the remaining eligible establishments on a low-priority list. This policy randomizes assignment-to-inspection within each of these two target lists using the same probabilities that the SST policy used, constraining the total the number of inspections to maintain SST's inspection budget. This policy yields intermediate benefits, averting fewer injuries than our nonrandom policies, but more than the SST policy. Compared to the SST policy, OSHA would avert 1.36, 1.36, and 1.17 times as many injuries at assigned establishments when targeting on *historical injuries* (the average number of annual serious injuries in the prior four years), *predicted injuries*, and *estimated treatment effects*, respectively.

We address several considerations that could threaten the validity of our estimated effects of these alternative policies. The most important is that a change in the

targeting policy could elicit behavioral changes at uninspected establishments that our estimates do not capture (Lucas 1976). In particular, we consider whether our alternative targeting policies might temper uninspected establishments' motivations to foster safe workplaces, eroding the general deterrence effects of inspections. We provide several pieces of evidence indicating that such behavioral changes would be minimal in our context.

Our results provide insight into the question of how and when machine learning can improve the allocation of a scarce treatment. Let $Y(1)_i$ and $Y(0)_i$ be the number of injuries establishment i experiences if it is or is not assigned to inspection. By definition, targeting inspections based on treatment effects $(Y(1)_i - Y(0)_i)$ is aligned with the regulator's objective: to improve workplace safety and health. However, it is difficult to estimate treatment effects precisely, even with the very large randomized sample that SST provides, because one only observes $Y(1)_i$ or $Y(0)_i$ but never both (Rubin 1974). Indeed, we find that OSHA averted just as many, if not more, injuries by targeting on a different metric (namely, a prediction of $Y(0)_i$) than it did by targeting on an estimate of treatment effects. Intuitively, while predicted injuries is further from the regulator's objective, we show that in our setting it is both easier to estimate and is correlated with estimated treatment effects. In other settings in which treatment effects are easier to estimate, or when they are less correlated with predicted outcomes, estimated treatment effects would be a more effective targeting criterion.

Our work contributes to the large literature on the effects of OSHA inspections. Some studies found little to no effect (Smith 1979; Bartel and Thomas 1985; Viscusi 1986; Ruser and Smith 1991), while others have found that inspections do reduce injuries (Gray and Scholz 1993; Gray and Mendoloff 2005; Foley et al. 2012; Haviland et al. 2012). Our paper is closest to three studies whose research designs rely on inspections that were randomly or quasi-randomly assigned. These studies found that randomized OSHA inspections conducted during ther period of 1987–1997 reduced fatal injuries by as much as 50 percent (Lee and Taylor 2019), that randomized inspections conducted by California's state occupational safety and health agency during 1996–2006 reduced injuries triggering workers' compensation claims by an average of 9 percent over the subsequent four years (Levine, Toffel, and Johnson 2012), and that OSHA SST inspections conducted during 1996–2011 reduced injuries resulting in days away from work, job restrictions, or job transfers by 20 percent in the year following inspection, based on a regression discontinuity design (Li and Singleton 2019). We extend Lee and Taylor (2019) by considering a larger set of injuries than the very rare fatal injuries they studied, extend Levine, Toffel, and Johnson (2012) by investigating a much larger inspection program covering many states, and extend Li and Singleton (2019) by estimating the average effect of inspections, rather than a local average treatment effect.[4] More importantly, our study goes beyond all of these by estimating heterogeneous effects of inspections and by comparing the effects of alternative targeting approaches.

[4] We go beyond Li and Singleton (2019) in two other ways. We focus on the effects of inspections on especially serious injuries (those causing days away from work), whereas they focus on a broader range of injuries (including those causing job restrictions and job transfers) which, as we discuss, are more subject to measurement error. Finally, Li and Singleton (2019) only estimate the effect of inspections on injuries one year later, whereas we find evidence that inspections lead to a decrease in injuries lasting several years.

Our research also contributes to a broader literature that analyzes the effects of inspections in other domains, including food safety (Ibanez and Toffel 2020), environmental protection (e.g., Hanna and Oliva 2010; Telle 2013; see Shimshack 2014 for an overview), tax authorities (e.g., Slemrod, Blumenthal, and Christian 2001; Kleven et al. 2011), and working conditions (e.g., Short, Toffel, and Hugill 2016).

Our study also extends a literature on optimal inspection strategies (e.g., Gonzalez-Lira and Mobarak 2019; Blundell et al. 2020). Particularly close to our work is Duflo et al. (2018), which assesses the benefits of providing regulators discretion in allocating inspections in the context of pollution enforcement in India. Duflo et al. (2018) assumes that an inspection's effectiveness—the extent to which it averts pollution—is proportional to levels of pollution. Our study tests that assumption in our setting and finds, consistent with their assumption, that heterogeneous treatment effects are highly correlated with predicted injury counts.

Finally, our paper contributes to a rapidly growing literature that uses machine learning to improve decisions. Studies have examined how using machine learning to predict outcomes can improve judges' decisions to release defendants before trial (Kleinberg et al. 2017), help regulatory inspectors more accurately predict which establishments are violating standards governing hygiene (Glaeser et al. 2016) and water pollution (Hino, Benami, and Brooks 2018), and help electric utilities predict when unmaintained equipment will fail (Rudin et al. 2010). We extend this literature by estimating variation in the *causal effects* of inspections, which Athey (2017) points out is the relevant criterion for optimal resource allocation.[5]

I. Setting and Data

OSHA's Site-Specific Targeting (SST) program is an ideal setting for our purposes. First, because OSHA allocated some SST inspections via random assignment, we can both evaluate their average causal effect on injuries and use machine learning methods to estimate heterogeneous treatment effects. Second, SST was created to promote the most "effective use of [OSHA's] enforcement resources" to ensure "safe working conditions for employees" (US Occupational Safety and Health Administration 2008), a goal that implies the need to compare alternative targeting policies in order to learn what is most effective. Finally, SST was a large regulatory program: it cost tens of millions of dollars and put at risk of inspection tens of thousands of establishments that collectively employed millions of workers.

A. *OSHA Site-Specific Targeting (SST) Program*

The SST program, which operated between 1999 and 2014, targeted inspections at high-injury workplaces within historically hazardous industries. We focus on the 29 states that OSHA directly regulates and not the 21 states that operate their own safety regulatory agencies.[6]

[5] Davis and Heller (2020) also estimate heterogeneity in the effectiveness of youth employment programs.

[6] Online Appendix A gives a more complete description of SST in federal OSHA states. Figure B.1 in online Appendix B provides a map of those 29 states. The other 21 states operate state-run programs approved by OSHA.

The SST program relied on data from an annual survey OSHA conducted between 1996 and 2011, the OSHA Data Initiative (ODI), that gathered injury data from 60,000 to 80,000 establishments per year (all of which had at least 40 employees and were within a predefined set of high-risk industries). Establishments' responses were based on logs that OSHA required them to maintain to document every work-related injury and illness.[7] OSHA used ODI responses from the prior year to create annual SST directives that specified two target lists: 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). Precise cutoffs for both lists varied by year.

OSHA then sent each of its 81 area offices (distributed across the 29 states) the names and addresses of establishments on the primary list that were within the area office's geographic territory. Area offices were told to ignore any establishments on the list that had received a comprehensive inspection during the two previous years (which in 2009 increased to three previous years), as they were ineligible for SST inspection. If an area office did not anticipate having sufficient resources to inspect its entire eligible primary list, it told headquarters the number of inspections it anticipated being able to complete. OSHA's software then randomly assigned a subset of that many establishments from its primary list to inspect. If the area office completed these inspections before headquarters provided the next year's lists, the office estimated how many more inspections it could conduct and the software generated a new random set of establishments from the remainder of its primary list. If an area office completed its entire primary list, the office repeated this process with the secondary list. Thus, OSHA randomly assigned most area offices a subset of either their primary or their secondary list to inspect each year.[8] Online Appendix A provides more details about the SST program.

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

There are several ways that OSHA inspections could lead establishments to improve workplace safety. Penalties for detected violations provide incentives for managers to remediate those and other hazards. Even when an inspection does not result in any violations, it can heighten manager's awareness of regulations and safety (Alm and Shimshack 2014) and increase managers' and workers' perceived risk of being inspected again (Kleven et al. 2011; Avis, Ferraz, and Finan 2018).

[7] OSHA Form 300, which employers are required to complete, is available at https://www.osha.gov/recordkeeping/RKforms.html, accessed March 2019.

[8] Area offices did not inspect every establishment that they assigned for inspection, for reasons we discuss below. Our estimation approach, described in Section IIA, accounts for this fact.

Moreover, inspectors sometimes share knowledge about safety practices with management (Choi and Almanza 2012).

B. *Data*

We combined data from (a) OSHA's annual SST target lists (2001–2010, the years for which OSHA's Office of Statistical Analysis could locate and provide us with target lists); (b) OSHA's annual ODI survey data on injuries and employment (1996–2011);[9] (c) OSHA inspection data from its Integrated Management Information Systems (IMIS) database (1990–2014),[10] and (d) annual Dun and Bradstreet data on employment, credit rating, and other business outcomes from the National Establishment Time Series (NETS) database (1990–2013).[11]

OSHA's annual SST primary and secondary target lists (US Occupational Safety and Health Administration 2011) report the name, address, and corresponding area office of each establishment for which injury rates in the prior year's ODI survey exceeded the SST directive's injury rate thresholds. These target lists also indicate which of these establishments were assigned to inspection.

The ODI dataset (US Occupational Safety and Health Administration 2013) contains the annual survey results that establishments reported to OSHA from 1996 to 2011, including annual counts of injuries involving days away from work (serious injuries) and of injuries involving job transfers or restrictions—which together are called DART (days away from work, restricted work, or a transfer) injuries. ODI also reports injury rates, which are annual injury counts per 100 full-time workers, and contains each establishment's annual average employment and total labor hours worked and its DUNS number, a unique establishment-level identifier. The ODI dataset is an unbalanced panel: it includes a different (but somewhat overlapping) set of establishments each year. The ODI sought to survey all establishments with at least 40 employees in hazardous industries every three years. Beginning around 2005, ODI resurveyed the following year those establishments reporting at least seven DART injuries per 100 full-time workers. Many establishments on the SST target lists report ODI data nearly every year[12] because (a) OSHA's DART rate threshold for resampling establishments in the ODI survey was below (more lax than) its 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 tend to be serially correlated.[13]

The IMIS database (US Occupational Safety and Health Administration 2014) includes every inspection attempted by OSHA. Each record includes the establishment's name and address, the inspection date, what triggered the inspection (e.g., the SST program, a recent serious accident, an employee complaint), whether the

[9] We obtained SST target lists and ODI survey data from OSHA after signing a memorandum of understanding.

[10] We downloaded OSHA inspection records in December 2014 from the agency's publicly available database, available at: https://enforcedata.dol.gov/views/data_summary.php.

[11] NETS is a proprietary database distributed by Walls and Associates (Donald Walls, dwalls2@earthlink.net).

[12] Among the establishments that were ever on an SST target list, 25 percent reported injury data in at least 10 of the 16 years of the ODI program (1996–2011) and 50 percent reported injury data in at least 7 of these 16 years.

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

inspector was unable to carry out the inspection (e.g., if the company had moved or gone out of business), and the number of violations and value of penalties.

NETS (NETS 2016) is an annual panel dataset extracted from Dun and Bradstreet data that seeks to include all establishments in the United States. We extracted each establishment's unique DUNS number, its first and last year in operation, whether it was part of a multiunit firm, and annual employment and credit rating.

To construct our sample, we used the DUNS numbers to link establishments on any of OSHA's SST 2001–2010 target lists to their corresponding records in ODI and NETS. We found 100 percent of the SST target list establishments in ODI and 97 percent in NETS; we dropped the 3 percent that we could not match to NETS. Because the IMIS database does not include DUNS numbers, we linked the establishments on the SST target list to IMIS records by fuzzy-matching names, addresses, and industries using *MatchIt* software, the Stata *reclink* command, and a manual process. We linked 82 percent of the establishments on the SST target lists to at least one inspection record in IMIS.[14]

During our 2001–2010 sample period, annual SST target lists ranged from 8,404 to 12,153 establishments, totaling 101,463 across all these years. We refer to the calendar year of an annual SST target list as a "directive year"[15] and we refer to each time OSHA placed an establishment on an SST target list (in a given directive year) as an "establishment-directive." These 101,463 establishment-directive observations comprised 40,946 unique establishments, because some establishments were on target lists in multiple directive years.

C. *Creating the Randomized Sample*

For most of our analyses, we narrowed our dataset to the subset of establishments on the SST target lists that were at risk of a randomized inspection (henceforth, the "randomized sample"). To do so, we first excluded the establishments on area office target lists in a given directive year when the area office did not assign inspections using randomization: when they assigned either none or all of the establishments on their target list to inspection. Second, we excluded establishments that OSHA told its area offices to inspect with certainty: those OSHA manually added to its SST target lists solely due to its concern that they had reported exceptionally *low* (and thus potentially inaccurate) injury rates to ODI or because they did not respond to the survey at all. Third, we omit those establishments that, according to the IMIS database, had already received a comprehensive safety inspection within the previous two years (which OSHA extended to three years, as of 2009) and were therefore ineligible for SST inspection.[16] Finally, we exclude the few establishments that, according to NETS, were not in operation two years prior to the directive year (the

[14] We retain the 18 percent 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.

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

[16] We also omitted the 9,617 cases in which area offices explicitly marked such ineligible establishments as "deleted" in the target list database. Through conversations with area office directors, we learned that many area offices implemented their deletions, but did not input them into the SST target list database. This is one reason that we remove ineligible establishments manually.

baseline injury-rate year SST relied on) or were not in operation during the directive year and thus could not have been inspected.

The remaining establishments were either randomly assigned for an SST inspection ("assigned to inspection")—our treatment group—or were eligible but not assigned ("not assigned to inspection")—our control group. Table B.1 in online Appendix B has more details on how we constructed the randomized sample.

The randomized sample includes 16,141 establishment-directives at risk of being randomly assigned to inspection, with 6,977 assigned to inspection and 9,164 not assigned. These 16,141 establishment-directives correspond to 13,029 unique establishments. We construct a panel around these 16,141 establishment-directives: a "pre-period" of the four years prior to the directive year and a "post-period" of the directive year and the four subsequent years.[17]

We do not observe any ODI-reported outcomes in the post-period for 2,405 (15 percent) of the 16,141 establishment-directives in our randomized sample, largely due to (a) establishments on target lists in later directive years having fewer opportunities to appear again in the ODI because the ODI survey ended in 2011 and (b) establishments shutting down or becoming ineligible for ODI by, for example, shrinking to fewer than 40 employees. In online Appendix C, we further discuss sources of sample attrition and—importantly—show that attrition of those assigned to inspection was statistically indistinguishable from attrition of those not assigned.

Table 1 reports summary statistics[18] and Table B.2 in online Appendix B reports the industry distribution of our sample.

Randomization implies that those assigned and those not assigned to inspection should be balanced on baseline characteristics. To assess whether this was the case in our randomized sample, we regress an *assigned to inspection* dummy on a series of baseline characteristics (using lagged values pertaining to the years before the directive year) and a set of fixed effects for area-office $\times$ directive dyads, the level at which the randomization took place. We report results in Table 2[19]; a Wald test failed to reject the null that the coefficients on the nine variables jointly differed from zero ($p = 0.19$), bolstering confidence in the randomization. [20]

[17] Our estimation samples contain fewer than 145,269 observations (16,141 $\times$ 9) because (a) most establishments were not included in the ODI survey in all nine years, (b) some ceased operation within five years of their directive year, and (c) the ODI survey ended in 2011 (one year after our final directive year, 2010).

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

[19] The two variables in this model from NETS (employment and PAYDEX score) had missing values for a small number of observations. We replaced missing values with the variable means and included dummy variables denoting these recodings.

[20] Establishments assigned to inspection are statistically indistinguishable from those not assigned to inspection in terms of all baseline characteristics except for DART injury rate and total working hours, both from two years prior. The difference in DART rates ($p = 0.043$) is miniscule in magnitude—only 0.02 percent of the variable mean of 10.34. Establishments assigned to inspection have 3.8 percent more total working hours ($p = 0.088$). While it is not surprising to find a statistically significant difference for two variables, given that we examined nine, we report robustness tests that estimated the evaluation models described below when controlling for total working hours in $t - 2$, with virtually identical results.

Table 1—Summary Statistics for the Randomized Sample, +/− 4 Years from Directive Year

	N	mean	sd	median	min	max
Number of times on prior years' SST target lists [SST]	143,757	1.2	1.6	1.0	0.0	9.0
Number of serious injuries [ODI]	90,343	6.5	8.9	4.0	0.0	54.0
Serious injuries[a]/100 FTE[a] [ODI]	90,343	3.9	3.6	3.1	0.0	16.6
Injuries with days away, restricted or transfer/100 FTE (DART rate)[b] [ODI]	90,343	7.6	5.0	6.7	0.0	23.6
Total hours worked, 000s [ODI]	90,346	284.8	331.2	183.0	0.0	2,369.8
Average number of employees [ODI]	90,346	149.5	176.8	96.0	1.0	1,257.0
Number of employees [NETS]	137,566	135.1	153.1	89.0	1.0	1,000.0
Minimum PAYDEX score [NETS][c]	125,794	67.7	10.7	70.0	2.0	96.0
Number of OSHA inspections in calendar year[d] [IMIS]	143,757	0.2	0.5	0.0	0.0	3.0
Number of SST inspections in calendar year [IMIS]	143,757	0.1	0.3	0.0	0.0	2.0

Notes: 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 online Appendix Table A.1. The table includes data from a nine-year window, consisting of the four 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 data source from which the variable is drawn is indicated in brackets. 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 and Bradstreet that it is in operation. For other variables, the number of observations is less than 145,269 ($= 16{,}141 \times 9$) due to right-censoring (our final directive year is 2010 and our data end in 2013). Unbounded variables (all except Number of times on prior years' SST target lists and Minimum PAYDEX score) are top-coded at their ninety-ninth percentiles.

[a]Serious injuries are those that cause days away from work.

[b]Injury rate variables are reported as the number of injuries per 100 full-time employees (FTE). They are calculated by dividing the number of injuries in a calendar year by the number of FTE (which is calculated as the total number of hours worked divided by 2,000), and multiplying this ratio by 100.

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

[d]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).

D. *Serious Injuries*

Our primary measure of workplace safety is an establishment's annual number of serious injuries (those resulting in at least one day away from work). We focus on these injuries because they are the most serious type reported to ODI and are less prone to measurement error due to misreporting (Biddle and Roberts 2003; Boden, Nestoriak, and Pierce 2010); see online Appendix D for a discussion of the validity of ODI injury data. Moreover, serious injuries are enormously costly: in 2005, the midpoint of our sample period, the 1.2 million injuries causing days away from work in the United States (US Bureau of Labor Statistics 2007) triggered more than $60 billion in costs, based on an estimated cost of $53,000 per injury (in 2018 dollars; see online Appendix E). To reduce the effect of large outliers of the number of serious injuries per year, we top-code this variable at its ninety-ninth percentile in our sample, which is 54.

Table 2—Regression Results Indicate That Baseline Characteristics Are Balanced between Those Assigned and Not Assigned to SST Inspection

Total OSHA inspections $t-1$ through $t-4$	0.003
	(0.004)
Number of times on prior years' SST target lists	−0.002
	(0.003)
Number of serious injuries, $t-2$	−0.000
	(0.001)
Serious injuries / 100 FTE, $t-2$	0.001
	(0.002)
Injuries with days away, restricted or transfer / 100 FTE, $t-2$	−0.002
	(0.001)
ln(Average number of working hours), $t-2$	0.038
	(0.022)
ln(Average number of employees [ODI]), $t-2$	−0.017
	(0.022)
ln(Employment [NETS]), $t-2$	−0.003
	(0.007)
Minimum PAYDEX score [NETS], $t-2$	−0.001
	(0.000)
Area-office–year FE	Y
Observations	16,141
R^2	0.238
p-value on joint significance	0.190

Notes: This table reports results of an OLS regression model in which the dependent variable is an indicator variable equal to 1 if an establishment is assigned to SST inspection and the explanatory variables are the variables reported in the table and area-office–directive fixed effects. The sample includes all establishments eligible for randomized SST inspections in area-office–years that randomized their target lists, as described in online Appendix Table A.1. The unit of analysis is the establishment-directive. Standard errors clustered by establishment are reported in parentheses. Variables indicated with "$t-2$" are from two years prior to the directive year. NETS data on employment and PAYDEX are missing in a small share of observations. We recode missing values to the variable mean and control for (but do not report) indicator variables that denote these observations.

II. Methods

We first describe how we estimate the average effect of randomized inspections. We then explain how we estimate heterogeneous treatment effects and use those estimates to evaluate alternative targeting policies.

A. *Estimating Average Treatment Effects of Randomized SST Inspections*

We used the following methods to estimate the intention-to-treat effect of assignment to a randomized SST inspection and then use an instrumental variable specification to estimate the treatment effect of receiving an inspection.

Estimating Intention-to-treat Effects.—We use the following specification (prespecified in a pre-analysis plan; see online Appendix F) to estimate the intention-to-treat effect of OSHA *assigning* an establishment to inspection:

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

$y_{ijt\tau}^{post}$ is the annual count of serious injuries for establishment i—within the geographic boundary of area office j and on the SST target list in year t—realized τ years relative to directive year t. In this specification, τ ranges from 0 (the directive year) to 4 (four years after the directive year), meaning we include up to five years of data for each establishment-directive. In our main specification, we use a Poisson specification, modelling the right-hand side of equation (1) as the conditional mean function of y.

$Assigned_{it}$ 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 not assigned. The coefficient on $Assigned_{it}$, α_1, represents the intention-to-treat effect of assignment to inspection. To improve precision, we control for y_{it}^{pre}: establishment i's injury count averaged over the four years prior to the directive year, sometimes called an analysis of covariance (ANCOVA) specification (McKenzie 2012).[21] (In robustness checks, we instead include the pre-period observations in the sample—rather than including their average as a control variable—and use a difference-in-differences design, which yields essentially identical results.)

$\mathbf{X}_{it}$ refers to control variables, which includes the number of years of data on which the historical mean y_{it}^{pre} is based (to account for varying precision in y_{it}^{pre}, since we do not observe ODI-reported data in all years for all establishments) as well as a dummy equal to one for nursing homes on the 2003 SST directive.[22] In some specifications, $\mathbf{X}_{it}$ also includes total working hours (or its log) in year $t-2$. μ_{jt} represents fixed effects for area-office–directive-year dyads. θ_τ represents fixed effects for each τ year relative to the directive-year. We cluster standard errors by establishment.

Equation (1) estimates the *average* annual effect of assignment to SST inspection over the directive year and four subsequent years, but this average might mask effects that vary over time. We thus also estimate equation (2), a distributed lag model, to estimate the annual difference in injuries—in each of the four years prior to, the year of, and each of the four years following the directive year—between establishments that were and were not assigned to inspection:

$$(2)\qquad y_{ijt\tau} = F\Bigg(\sum_{k\in[-4,4]} \beta_k D_{\tau=k} \times Assigned_{it} + \mu_{jt} \times \mathbf{1}\{\tau \geq 0\} + \mathbf{X}_{it} \times \mathbf{1}\{\tau \geq 0\} + \lambda_{it} + \theta_\tau + \epsilon_{ijt\tau}\Bigg),$$

[21] When we estimate Equation 1 with a Poisson regression, this control variable y_{it}^{pre} is the average of historical $\log(y+1)$, rather than the average of historical y. In a Poisson regression, a coefficient on the latter estimates how much a one-unit change in historical y is correlated with a 1 percent change in post-period y, whereas a coefficient on the former estimates how much a 1 percent change in historical y is correlated with a 1 percent change in post-period y, which is a more appropriate metric to capture the correlation between historical and post-period outcomes. We add 1 to y_{it}^{pre} before taking the log to account for zeroes. When we instead estimate equation (1) with OLS (as we do in robustness checks), we change y_{it}^{pre} to equal the average of historical y in levels.

[22] We include this dummy because OSHA did not originally include nursing homes on its SST Target List for the 2003 SST directive, which was initiated in June 2003, due to another concurrent national program on nursing homes. However, in September 2003 OSHA revised the 2003 directive and added nursing homes to the primary list of the 2003 Target List. Because of this delay, relatively few nursing homes were assigned to inspection in 2003. This dummy variable accounts for this differential treatment.

where $\mathbf{1}\{\tau \geq 0\}$ equals 1 when $\tau \geq 0$, and 0 otherwise. This specification 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 $\tau = k$. β_k coefficients estimate the difference in outcome y between establishments assigned and not assigned to inspection if $\tau = k$, for $k \in \{-4,4\}$. λ_{it} is a fixed effect for each establishment-directive, which effectively replaces y_{it}^{pre} from equation (1) to control for time-invariant differences across establishment-directives. As in equation (1), μ_{jt} is a fixed effect for each area-office–directive-year, and $\mathbf{X}_{it}$ is a vector of establishment-directive controls (here, log hours in the year $t-2$ and the dummy for nursing homes on the 2003 directive); we multiply these by $\mathbf{1}\{\tau \geq 0\}$ because otherwise μ_{jt} and $\mathbf{X}_{it}$ would be absorbed by λ_{it}. We cluster standard errors by establishment.

Instrumental Variables Specification.—While the randomization of *assignment* to an SST inspection provides a clean experimental design, assignment imperfectly predicts receiving an inspection for several reasons.

First, not all establishments assigned to an SST inspection were actually inspected. In some instances, inspectors could not find establishments assigned to inspection. Some area offices successfully petitioned OSHA headquarters for permission to not inspect all of the establishments it had assigned them. Also, OSHA issued the annual SST directives and target lists assigning establishments to inspection between April and August, but area offices did not begin conducting those inspections until months later. Together, these factors resulted in only 18 percent of establishments randomly assigned to SST inspection actually being inspected by the end of the calendar year in which OSHA placed them on the target list (the directive year), and only 73 percent by the end of the following calendar year (online Appendix Figure A.2).[23]

Second, some eligible establishments that were not assigned to inspection in a given directive year 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 OSHA placed many on a later SST target list. Thus, 28 percent of our control establishments (those eligible but not assigned to inspection in a given directive year) received an SST inspection within the next four calendar years.[24]

Given this imperfect adherence, comparing injuries between establishments that were randomly assigned to SST inspection in a given year to those that were eligible but not assigned (the "intention-to-treat" estimate) underestimates the effect of receiving an SST inspection.

To estimate the average treatment effects of inspection on injuries, we instrument whether an establishment has been SST-inspected with whether OSHA assigned it to inspection in the directive year. This approach scales the intention-to-treat esti-

[23] It is possible that OSHA did inspect some of the seemingly uninspected establishments, but our procedure to link SST with OSHA's information system (IMIS) failed to find their corresponding inspections in IMIS.

[24] Establishments might receive other types of inspections, but we find essentially zero difference between the establishments assigned to inspection and not assigned to inspection with respect to the likelihood of experiencing a non-SST OSHA inspection. Thus, we do not consider this potential source of bias to be an important factor in our context.

mate by the extent to which assignment to inspection increases the probability of being inspected. Specifically, we estimate the following variant of equation (1):

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

$\widehat{Inspected}_{it\tau}$ 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:

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

For equations (3) and (4), $y_{ijt\tau}^{post}$, y_{it}^{pre}, $\mathbf{X}_{it}$, μ_{jt}, and θ_τ are the same as in equation (1). $\eta_{ijt\tau}$ and $\nu_{ijt\tau}$ are error terms assumed to be independent and identically distributed (i.i.d.).

Instrumenting *inspected* with *assigned* in equation (3) meets the two requirements for the estimate of δ_1 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, the exclusion restriction is satisfied: *assigned* only affects outcome y through its influence on *inspected* because (a) an establishment only learned it had been assigned to SST inspection when an inspector arrived and (b) assignments only affected an inspector's actions by allocating SST inspections.

We use an IV-Poisson regression model to estimate the causal effect of being inspected on serious injury count. We consider other specifications in robustness checks described below.

B. *Alternative Targeting Policies*

OSHA might have averted more injuries by using different approaches to targeting. The potential gains from alternative targeting regimes depend on the degree of heterogeneity in treatment effects across establishments and on OSHA's ability to identify establishments where those treatment effects are largest. Historically, OSHA sought to target establishments with "serious health and safety problems" and, to identify them, focused on those with high injury rates two years prior. We consider three alternative targeting metrics OSHA could have used to assign inspections. The first two metrics are better measures of "serious health and safety problems" (one based on historical data, the other on machine learning) and the third is a machine-learning–based estimate of the treatment effects of inspections.

Historical Counts of Serious Injuries.—Our first alternative targeting metric is *historical injuries* (Inj_{it}^{pre}): the average annual number of serious injuries at an establishment over the four years prior to the directive year. This metric is similar to the two-year lagged injury rate that OSHA used for SST, but differs in three important ways. First, *historical injuries* incorporates four years of historical data, as opposed to being a single observation from two years prior. Using multiple years of data can overcome mean reversion because a single year of high injuries could

reflect bad luck (Ruser 1995). Second, *historical injuries* is a measure of injury *counts*, as opposed to *rates*. Counts are a more useful targeting criterion if there are economies of scale in inspecting and remediating hazards. Third, whereas SST's targeting regime relied on injuries resulting in days away from work, job restriction, or job transfer ("DART"), we only consider the more serious subset: injuries resulting in days away from work. As described in Section IIC, such injuries are less prone to misreporting than are less-consequential injuries that result in job transfers or restrictions.

Predicting Serious Injuries with Machine Learning.—Historical injuries might not be the best predictor of which establishments *currently* have the most serious health and safety problems. Thus, as a second alternative targeting metric, we consider *expected injuries*, which is the number of serious injuries (those leading to days away from work) an establishment would experience if not assigned to inspection, given its baseline characteristics Z, averaged across the assignment year and four subsequent years ($E\left[Y(0) \mid Z\right]$, or simply $Y(0)$, in the Rubin (1974) causal framework).

This metric has two potential advantages over the average historical injury count. First, *expected injuries* refers to the number of injuries that would actually occur (absent an inspection) in the year when inspectors are considering visiting the establishment, as opposed to historical injuries that have already occurred. Second, *expected injuries* recognizes that injuries might be a function not only of historical injury counts but also of many other variables—such as employer size, regional characteristics, and business conditions—and that these factors could affect injury counts in nonlinear ways.

We predict *expected injuries* focusing on establishments not assigned to inspection (those for which *expected injuries* is actually observed). We use the ensemble machine-learning procedure Super Learner, which minimizes the mean squared error of out-of-sample predictions by using cross-validation, to find the optimal weighted average of multiple machine-learning methods (van der Laan, Polley, and Hubbard 2007). Our Super Learner library includes random forest (Breiman 2001), the Generalized Additive Model, and a linear interaction model.[25] Vector Z (baseline characteristics) includes a host of factors including the establishment's size, age, injury record, inspection record, safety compliance record, geographic setting, industry characteristics, and temporal factors. Online Appendix G lists the full set of variables included in Z. We refer to the resulting prediction of *expected injuries* as simply *predicted injuries*.

Estimating Heterogeneous Treatment Effects of Inspections.—To serve its goal of making the most "effective use of its enforcement resources," OSHA would in principle want to maximize the number of injuries inspections averted—that

[25] These three learners are those that received non-zero weight when we initially ran Super Learner on the entire randomized sample with several additional learners in the library (generalized boosting regression, neural network, generalized linear model, and Bayesian generalized linear model). We used the default parameters for each algorithm, except that we restricted the smallest leaves in the random forest to have at least 50 observations because small leaves can increase mean squared error (Athey and Imbens 2016).

is, to direct inspections where their *treatment effects* are largest. These treatment effects might or might not be highly correlated with historical or predicted injuries. For example, if inspectors are better able to identify fixable safety issues at establishments with many serious injuries, then establishments with high historical injury counts will also have large treatment effects. But if such establishments face high costs of remediating hazards, they might instead be relatively unresponsive to inspections (Athey 2017).

We refer to an establishment's conditional average treatment effect—its treatment effect given its baseline characteristics (Z)—as its "treatment effect." Following Rubin's (1974) potential outcomes framework, we define an establishment's expected treatment effect as

$$\left(E\left[Y(1)\,|\,Z\right]\right) - \left(E\left[Y(0)\,|\,Z\right]\right);$$

that is, the difference between the establishment's outcome if it *were* assigned to inspection given a set of characteristics Z $\left(E\left[Y(1)\,|\,Z\right]\right)$ and if were *not* assigned to inspection $\left(E\left[Y(0)\,|\,Z\right]\right)$, the latter being *expected injuries* in our scenario.

A challenge to estimating treatment effects is that, unlike injuries, they require an unobservable counterfactual. We estimate each establishment's treatment effect using a causal forest (Wager and Athey 2018), a method that builds on Breiman's (2001) flexible random forest. A random forest first builds many regression trees, each of which is a form of nearest-neighbor matching in which the set of neighbors maximizes both similarity within a leaf and divergence across leaves. The random forest then averages predictions of the many trees to reduce variance and improve predictive power.

Wager and Athey's (2018) causal forest modifies the random forest to estimate heterogeneous treatment effects by searching for high-dimensional combinations of covariates associated with different treatment effects. To avoid overfitting, we create each regression 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 a causal forest to estimate unbiased treatment effects, two considerations must hold. First, conditional on covariates Z, assignment to inspection must be independent of the potential outcomes. This consideration is satisfied among the establishments in our randomized sample because assignment to inspection was random conditional on area-office–directive-year. Second, there must be enough treatment and control observations in a given leaf (Athey and Imbens 2016), so we include only leaves with at least 50 observations and for which the share of treatment or control observations is no less than 10 percent. We include the same covariates Z as used to predict injuries (see online Appendix G).

C. *Evaluating Alternative Targeting Policies*

Evaluating the number of injuries that OSHA could have averted had it used any of these alternative targeting policies is challenging for a few reasons. Most importantly, relevant to our two alternative policies that rely on machine learning,

predicted values of *expected injuries* and of *treatment effects* might reflect noise or sampling variation, as opposed to true heterogeneity. Additionally, multiple considerations could invalidate our ability to estimate the number of injures the regulator could avert under different targeting regimes. We discuss these concerns in the following two sections.

Estimating the Effects of Inspections for Different Groups.—Our machine learning methods might not generate consistent estimates of *expected injuries* or the *treatment effect* for individual observations and might therefore lead to misleading estimates of the effects of alternative policies that target inspections based on these metrics (Chernozhukov et al. 2020). To estimate how many injuries OSHA would avert under each alternative policy, we follow the method developed in Chernozhukov et al. (2020) to produce consistent estimates of average treatment effects for specific subsamples.

To apply their method, we first randomly partition the establishments in the randomized sample into two halves, which we refer to as the training sample and the holdout sample.[26] We use the training sample to estimate model parameters, and then apply these parameters to the holdout sample to predict *expected injuries* and *treatment effects.* First, focusing on the subset of establishments in the training sample that were not assigned to inspection, we use Super Learner to construct a function that predicts *expected injuries*—the number of injuries an establishment would experience when not assigned to inspection. Second, we use causal forest on the entire training sample to construct a function that estimates establishments' *treatment effects*. Third, we apply these two functions to the holdout sample to compute each of those establishments' *predicted injuries* and *estimated treatment effect*.

Because a given individual establishment's *estimated treatment effect* might be biased, we follow Chernozhukov et al. (2020) to estimate average treatment effects for specific subsets of the data. For example, one policy that OSHA could follow is to assign the same number of inspections each year, but allocate them to those establishments at which it expects inspections to avert the most injuries (i.e., establishments with the largest *estimated treatment effects*). Define a group G such that G_1 indicates that an establishment's estimated treatment effect is high enough to be assigned to inspection under this policy and G_0 indicates otherwise. To estimate the number of injuries OSHA would avert under this policy, we estimate the following weighted linear regression on the holdout sample:

$$(5) \qquad y_{it}^{post} = \alpha_1 + \alpha_2 Y\hat{(0)}_{it} + \sum_{k=0}^{1} \gamma_k \left[D_{it} - p(Z_{it})\right] \times \mathbf{1}\left\{G_{it}^{k}\right\} + \nu_{it},$$

where y_{it}^{post} is an establishment's annual number of serious injuries averaged over the five "post-period" years comprising the directive year and four subsequent years. D is an indicator for whether an establishment was assigned to treatment, $p(Z)$ is the proportion of establishments in its area-office directive list eligible for inspection

[26] Chernozhukov et al. (2020) refer to the training sample as the "auxiliary sample" and the holdout sample as the "main sample."

that were assigned to inspection that directive year (the "propensity score"), $\hat{Y(0)}$ is our prediction of an establishment's *expected injuries*,[27] and ν is an i.i.d. error term. Following Chernozhukov et al. (2020), the regression is weighted by $\omega = \left\{p(Z) \times \left[1 - p(Z)\right]\right\}^{-1}$.

Here, $\hat{\gamma}_k$ is a consistent estimate of the mean number of injuries averted per establishment among the establishments in group G_k. We estimate the total number of injuries averted under a given targeting policy, $\sum_k(\hat{\gamma}_k \times N_k)$, where N_k is the number of establishments in group G_k that OSHA would assign to inspection under that policy.

Due to sampling variation, the holdout sample might not be representative of the entire randomized sample. We follow Chernozhukov et al. (2020) to obtain point estimates and confidence intervals of the average treatment effects among establishments assigned to inspection in each policy. We conduct 250 iterations of the above process, each time randomly splitting the data into new training and holdout samples and saving the key coefficients ($\hat{\gamma}_k$). For each of these 250 iterations, we also reverse the roles of the training and holdout samples and repeat the same procedure outlined above, obtaining another set of coefficients for each iteration pertaining to the other half of the sample split. For each iteration, we then take the average of these two coefficients, akin to twofold cross-validation.[28] Examining the distribution of these 250 average coefficients, we use the median (fiftieth percentile) as our point estimate of the group average treatment effect for the policy and, pursuant to Chernozhukov et al. (2020), use the 2.5th and 97.5th percentiles as the range of our 90-percent confidence interval.[29]

We seek to evaluate the effect of alternative targeting policies that apply to the entire historical SST target list, which includes both our randomized sample and the subset for which inspections were assigned nonrandomly (the nonrandomized sample). However, we can apply Chernozhukov et al.'s (2020) procedure only to our randomized sample because the propensity score $p(Z)$ might be correlated with potential outcomes of establishments in the nonrandomized sample, which could bias our estimates of treatment effects. In online Appendix H, we explain how we adapt the above procedure to estimate the effects of policies that retarget the entire historical SST target list, and we show that our approach is robust to various potentially confounding factors.

[27] Chernozhukov et al. (2020) refer to $\hat{Y(0)}$ as $B(Z)$.

[28] This process to obtain point estimates and confidence intervals for the group average treatment effects deviates slightly from Chernozhukov et al. (2020). That technique does not take the step of reversing the roles of the training and holdout samples for each sample split; rather, it generates estimates only from the holdout sample and then uses the median coefficient values (and standard errors) across these iterations as its estimates of the group average treatment effect (and its standard error). Our approach—to instead average the coefficients estimated on the holdout sample and training sample for each split, and then use the median of these averages across the 250 iterations as the estimated group average treatment effect—is inspired by Chernozhukov et al. (2018) and Jacob (2020). Our version yields essentially identical point estimates, but smaller confidence intervals, than Chernozhukov et al. (2020). We thank Daniel Jacob for guidance in implementing this approach.

[29] The 90 percent confidence interval on these estimates corresponds to the 2.5th and 97.5th percentiles of the distribution, rather than the fifth and ninetieth, because Chernozhukov et al. (2020) note that the repeated sample splitting results in additional uncertainty.

Table 3—Regression-Estimated Effects of SST Inspection on Serious Injuries

	# of serious injuries		SST inspected	# of serious injuries
	Intention-to-treat		(First stage)	Treatment-on treated
	(1)	(2)	(3)	(4)
Assigned to SST 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)
Average marginal effect	−0.19	−0.20		−0.48
Specification	Poisson	Poisson	OLS	IV-Poisson
# 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

Notes: Serious injuries refers to those resulting in days away from work. All regressions include area-office–directive and τ-year (number of years since the directive year) fixed effects. Each regression also controls for the mean of the establishment's dependent variable (or log(1 + dependent variable) in Poisson regressions) over the 4 years prior to the directive year and for the number of years over which this baseline mean is calculated. SEs, in parentheses, are clustered by establishment. 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 serious 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 serious injuries.

Potential Threats to Estimating the Effects of Alternative Policies.—Two considerations might challenge our ability to estimate the effects of alternative targeting policies, even with unbiased estimates of average treatment effects for subsets of the data. First, a new targeting scheme might change behaviors—and thus outcomes—of inspected as well as uninspected establishments. Second, our analysis uses data from our entire 2001–2010 sample period to construct establishments' *predicted injuries* and *estimated treatment effects*, but each year the regulator would only have data through the prior year to construct these measures. Section IIID discusses these concerns in more detail, explains how we address them, and shows that none meaningfully invalidates our estimates of alternative targeting policies.

III. Results

A. *Average Effects of OSHA's SST Inspections*

Columns 1 and 2 of Table 3 report estimates of the average effects of assignment to an SST inspection on an establishment's number of serious injuries. The first column displays Poisson regression results from the intention-to-treat specification

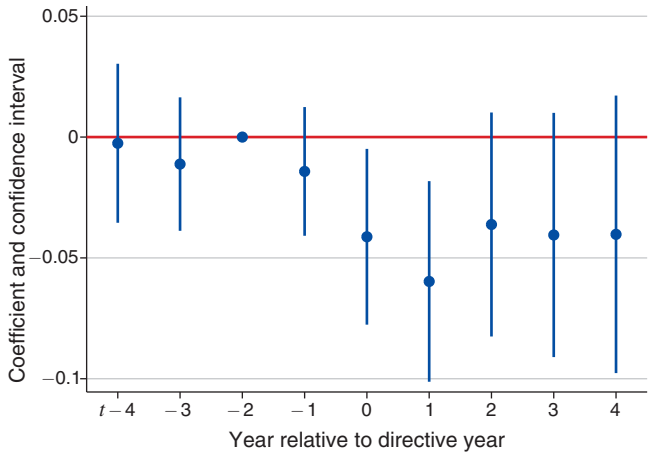

FIGURE 1. TEMPORAL EFFECTS OF ASSIGNMENT TO INSPECTION ON SERIOUS INJURIES, BY YEAR RELATIVE TO DIRECTIVE YEAR

Notes: 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 serious 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 τ year, with a 95 percent confidence interval. The omitted year is $t - 2$ (two years prior to the directive year).

from equation (1). Establishments assigned to SST inspection experience 3.4 percent fewer injuries over the directive year and four following years ($\beta = -0.035$, SE $= 0.017$, $p = 0.04$) than do those not assigned. The average marginal effect indicates that being assigned to inspection leads, on average over this five-year period, to a decline of 0.19 in the annual number of serious injuries from the sample annual average of 5.35. This estimate is essentially unchanged when we control for baseline log total working hours from two years before the directive year (year $t - 2$) (column 2).

To investigate the extent to which the average effect on injuries of being assigned to inspection varies over time, Figure 1 shows the annual *assigned to inspection* coefficients and their 95 percent confidence intervals from a Poisson regression estimating the intention-to-treat distributed lag specification in equation (2) (the omitted year is $t - 2$, two years prior to the directive year). During the four years prior to the directive year, the coefficients hover around zero, consistent with random assignment. Beginning with the directive year ($t = 0$), the coefficients become consistently negative, hovering between -0.04 and -0.05, and are statistically significant in years $t = 0$ and $t = 1$.

To estimate the effect of inspection (treatment-on-the-treated effect), we need to account for the fuzziness in inspection assignment described in Section IIIA. We use OLS to estimate the first-stage effect of assignment to SST inspection in the directive year on the probability of actually being SST-inspected, corresponding to equation (4). Being assigned to inspection increases the probability of actually being inspected over the directive year and the four following years by 46 percentage points ($p < 0.01$) above the 17 percent inspection rate over this period among

those not assigned in the directive year (column 3). Column 4 reports the effect of an SST inspection on serious injuries, an IV-Poisson estimate of equation (3). On average, SST inspections lead to 8.7 percent fewer serious injuries per year ($\beta = -0.091, \text{SE} = 0.042, p = 0.03$). The average marginal effect indicates that, over this five-year period, SST inspections lead to an average decline of 0.48 in the annual number of serious injuries, implying that the average SST inspection averted ($5 \times 0.48 =$) 2.4 serious injuries over the five-year period. Given our estimate that a serious injury costs $53,000 during our sample period (see online Appendix E), the average randomized inspection averted just over $125,000 in injury costs over this five-year period[30]—roughly 35 times the cost of conducting an inspection.[31]

Our estimates are robust to a number of alternative specifications and other checks that we summarize here and describe in detail in online Appendix I. We estimated equation (1) using a difference-in-differences design, and we also estimated it dropping establishments that had ever received a violation from OSHA for injury recordkeeping. We also estimated equation (1) with an ANCOVA model in which we collapsed all of an establishment's post-period annual observations into a single (averaged) observation. We also estimated average intention-to-treat effects with targeted maximum likelihood estimation combined with Super Learner (van der Laan and Rose 2011) and with the method outlined in Chernozhukov et al. (2020). All of these yielded results, reported in columns 1–5 of Table I.1 in online Appendix I, that were economically similar to and statistically indistinguishable from the results reported in Table 3.

B. *How Heterogeneous Are Treatment Effects of Inspections—and Why?*

As described at the start of Section IIB, the potential gains from alternative targeting depend on the degree of heterogeneity in treatment effects, as well as on OSHA's ability to identify the establishments where inspections will avert the most injuries. Before assessing how many injuries OSHA could avert though alternative targeting policies, we first present visual evidence of the degree of estimated treatment effect heterogeneity and examine the establishment characteristics associated with high or low estimated treatment effects.

Heterogeneity in Treatment Effects.—To assess the extent to which the estimated treatment effects exhibit variation, in each of our 250 sample splits (described in Section IIC) we retain the Best Linear Predictor of the treatment effect for each

[30] Our estimate that SST inspections led to 8.7 percent fewer DAFW injuries is similar to Levine, Toffel, and Johnson's (2012) estimate that inspections by California's Division of Occupational Safety and Health led to 9.9 percent fewer injuries that triggered workers' compensation claims. This paper only considers DAFW injuries, a subset of injuries eligible for workers compensation considered by Levine, Toffel, and Johnson (2012). This difference likely explains why our estimate that SST inspections have a $125,000 social benefit is lower than Levine, Toffel, and Johnson's (2012) estimate that Cal-OSHA inspections had a $355,000 social benefit.

[31] 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.

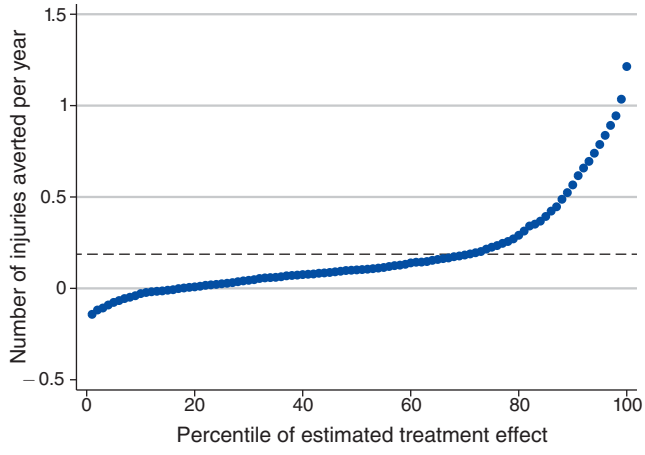

FIGURE 2. PERCENTILES OF ESTIMATED NUMBER OF SERIOUS INJURIES AVERTED PER YEAR IF ASSIGNED TO SST INSPECTION

Notes: Estimates of the percentiles of the treatment effect of assignment to SST inspection among the set of establishments on OSHA's SST target lists from 2001–2010. Each dot represents the median, across 250 sample splits, of the corresponding centile of the Best Linear Predictor of establishments' treatment effect. See Section IIB for details. The dashed horizontal line is the mean estimated treatment effect in the sample (0.18).

establishment in the holdout or nonrandomized sample (Chernozhukov et al. 2020).[32] We categorize the distribution of these treatment effects into centiles, calculate the average treatment effect among the establishments in each centile, and save these 100 averages. We repeat this process for each of the 250 sample splits. Next, we calculate the median across the 250 splits of the average estimated treatment effect values of the first centile and plot this as the first (leftmost) point on Figure 2. We then repeat this for each of the other centiles to plot the 100 points in Figure 2. While these estimates are likely biased (Chernozhukov et al. 2020), they nonetheless provide a sense of the heterogeneity in treatment effects. Figure 2 indicates that the estimated treatment effects of assignment to inspection exhibit substantial heterogeneity, particularly at the very top. The integral under the curve indicates that if OSHA assigned every establishment in this sample to inspection, 68 percent of the total injuries it would avert would occur among the establishments in the top 20 percent of treatment effects. Treatment effects being concentrated in a relatively small subset of establishments suggests that there might be large gains to changes in targeting OSHA's limited inspections.

Sources of Heterogeneity in Treatment Effects.—We next describe the differences in baseline characteristics between the group of establishments with the highest versus the lowest estimated treatment effects. Following Chernozhukov et al. (2020), within each of our 250 sample splits, we identify the establishments with

[32] The Best Linear Predictor of an establishment's treatment effect is obtained from the regression coefficients in online Appendix equation (H.1), as $\hat{\beta}_1 + \hat{\beta}_2 \times (S - ES)$. See Chernozhukov et al. (2020) for details on notation.

Table 4—Differences in Characteristics among Establishments with Estimated Treatment Effects in the Top and Bottom 20 Percent of the Distribution

	Mean and SE of variable for establishments with *estimated: treatment effect in:*		Difference in means	Percent difference
	Top 20% (1)	Bottom 20% (2)	(3)	(4)
Injuries with days away, restricted or transferred/100 FTE (DART rate), t − 2	12.510 (0.047)	11.541 (0.041)	1.101 (0.063) [0.000]	8.4%
Historical injuries[a]	19.823 (0.110)	4.699 (0.023)	15.183 (0.113) [0.000]	321.9%
Predicted injuries[b]	12.969 (0.072)	3.050 (0.013)	9.946 −0.075 [0.000]	325.2%
Nursing homes	0.078 (0.002)	0.162 (0.003)	(0.089) −0.004 [0.000]	−51.9%
Manufacturing	0.550 (0.004)	0.453 (0.004)	0.094 −0.006 [0.000]	21.4%
# employees [NETS], t − 2	389.004 (6.524)	105.546 (0.338)	281.809 (6.534) [0.000]	268.6%
Injuries with other recordable cases/100 FTE, t − 2	5.219 (0.04)	4.100 (0.036)	1.159 −0.054 [0.000]	27.3%
ln(*Total days away from work resulting from injuries*), *t − 2*	5.834 (0.012)	3.941 (0.012)	1.967 (0.017) [0.000]	48.0%
Any fatal injuries, t − 2	0.015 (0.001)	0.006 (0.001)	0.009 (0.001) [0.000]	150.0%
Standalone firm, t − 1	0.244 (0.003)	0.381 (0.004)	−0.142 (0.005) [0.000]	−36.0%
Establishment age, t − 1	29.219 (0.239)	26.612 (0.204)	2.527 (0.313) [0.000]	9.8%
Minimum PAYDEX score [NETS], t − 2	67.434 (0.074)	67.790 (0.075)	−0.394 (0.106) [0.000]	−0.5%
Establishment has ever been inspected prior to this year	0.602 (0.004)	0.402 (0.004)	0.199 (0.006) [0.000]	49.8%
Establishment had a complaint inspection in t − 1 through t − 3	0.205 (0.003)	0.066 (0.002)	0.138 (0.004) [0.000]	210.6%
Number of times on prior years' SST target lists	1.878 (0.017)	1.230 (0.013)	0.667 (0.021) [0.000]	52.7%
State-year leave-one-out-mean serious injury rate t − 2	3.013 (0.008)	2.947 (0.007)	0.068 (0.010) [0.000]	2.2%

Notes: We conduct 250 random even splits of the randomized sample. In each iteration, we train a causal forest on the training sample to predict treatment effects for establishments in the holdout and nonrandomized samples. Among establishments in each iteration's sample, we identify those with the top 20 percent and bottom 20 percent of predicted treatment effects 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 percent groups, with the median associated standard error in parentheses. Column 2 reports these for the 250 bottom-20 percent 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 IIIC for further information.

[a]Historical injuries is the annual number of serious injuries, averaged $t-1$ to $t-4$

[b]Predicted injuries is the predicted annual number of serious injuries if not assigned to inspection, averaged t to $t+4$

an estimated treatment effect in the top 20 percent of the combined holdout sample and nonrandomized sample and the establishments in the bottom 20 percent. We conduct *t*-tests to assess whether these two groups' average baseline characteristics differ. Across these 250 splits, we calculate and save (a) the medians of the relevant variable means, (b) the medians of those variables' within-split standard deviations, and (c) the median *t*-test *p*-values.

The DART injury rate two years prior to the directive year—the metric OSHA used to construct the SST target lists—is only slightly higher among establishments with the top 20 percent treatment effects versus the bottom 20 percent (12.5 versus 11.5; $p < 0.01$, Table 4, row 1). In contrast, establishments with the top 20 percent estimated treatment effects have substantially more *historical injuries* (19.5 versus 4.8; $p < 0.01$, row 2) and more *predicted injuries* (12.8 versus 3.0; $p < 0.01$, row 3). Establishments with high estimated treatment effects also have more employees, are more likely to be in the manufacturing sector, are less likely to be nursing homes, experienced more fatal injuries two years prior to randomization, and had more prior inspections triggered by a worker complaint.

C. *Effects of Alternative Targeting Policies*

We now estimate how different targeting rules affect the number of injuries OSHA could have averted. Each of our policies maintains OSHA's rule under the historical SST program that an establishment is ineligible for an inspection if it received one in the prior two years.[33]

How Many Injuries Could OSHA Have Averted through Alternative Targeting?—OSHA allocated its SST inspections by creating a target list of establishments with high DART injury rates two years before and then prioritized within them by establishing a threshold that, roughly, placed the establishments with the top third of DART injury rates on the primary list and the rest on the secondary list. Among the establishments on the 2001–2010 target lists that were eligible for SST inspection, OSHA assigned to inspection 43 percent of those on the primary lists and 10 percent of those on secondary lists. Using methods from Chernozhukov et al. (2020) (and described further in online Appendix I), we estimate that the historical SST program averted an average of 0.168 (90 percent confidence interval of −0.193 to −0.140) serious injuries per year per establishment assigned to inspection (Table 5, row 1). As expected, this result is quite similar to the average marginal effect obtained from our Poisson regression (column 1 of Table 3). This new estimate of the historical policy's average effect serves as the benchmark against which we evaluate the benefits of alternative targeting policies.

We now estimate how many injuries OSHA would have averted had it assigned *all* inspections nonrandomly each year to establishments with the highest value of

[33] 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.

Table 5—Number of Serious Injuries OSHA Averts under Alternative Targeting Policies

Policy (1)	Targeting metric (2)	Average number of annual serious injuries averted per establishment, among those assigned (3)	Number of establishments assigned to inspection on the … high-priority list (4)	… low-priority list (5)	Total number of serious injuries averted over 5 years (6)	Additional injuries averted compared to historical policy (7)
OSHA's policy (high-priority = OSHA's "primary list"; low-priority = OSHA's secondary list)	DART rate, $t-2$[a]	−0.168 [−0.193; −0.140]	12,458	4,403	14,163	
Inspect those with highest metric, preserving the historical total number of inspections	Historical injuries	−0.312 [−0.393; −0.222]	16,861	0	26,303	12,140
	Predicted injuries	−0.357 [−0.491; −0.227]	16,861	0	30,097	15,934
	Estimated treatment effect	−0.364 [−0.620; 0.072]	16,861	0	30,687	16,524
Inspect those with highest metric, preserving the historical total cost of inspections	Historical injuries	−0.332 [−0.414; −0.233]	15,126	0	25,109	10,945
	Predicted injuries	−0.419 [−0.567; −0.274]	14,944	0	31,308	17,144
	Estimated treatment effect	−0.393 [−0.681; 0.050]	15,223	0	29,914	15,751
Maintain size and Pr(inspection) of high- and low-priority lists from historical policy, preserving the historical total cost of inspections	Historical injuries	−0.240 [−0.363; −0.118]	11,788	4,276	19,277	5,114
	Predicted injuries	−0.241 [−0.387; −0.083]	11,798	4,188	19,262	5,099
	Estimated treatment effect	−0.205 [−0.365; −0.046]	11,890	4,307	16,602	2,439

Notes: The estimates in column 3 correspond to the γ coefficients, specified in equation (5) in the text, to estimate Group Average Treatment Effects. Each reported estimate in this column is the median of the average of two coefficient across 250 random splits of the sample into training and holdout samples: one coefficient where the machine learning models are trained on the training sample and the regression is estimated on the holdout sample, the other where the roles of the training and holdout samples are reversed. The 90 percent confidence intervals, reported below the coefficients in brackets, are estimated analogously. See Section IIC for details. The estimate in column 6 is the number of establishments assigned to inspection (the sum of columns 4 and 5), multiplied by the (negative of the) average treatment effect among assigned establishments (column 3), multiplied by 5 (the window of years over which we estimate the effects of assignment to inspection).

[a]DART rate, $t-2$ = Injuries with days away, restricted or transferred/100 FTE, from two years prior to the directive year.

each of our three targeting metrics: *historical injuries*, *predicted injuries*, and *estimated treatment effects*. In each scenario, we have OSHA assigning to inspection the same number of establishments each year as were assigned under the actual SST policy: 16,861 over the 10-year period.[34]

Targeting establishments with the most *historical injuries* would have averted an average of 0.312 (90 percent CI of −0.393 to −0.222, Row 2) serious injuries per year among assigned establishments, nearly twice as many as the 0.168 averted

[34]This number (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 online Appendix A. There are two reasons. First, for this analysis, we have excluded the 9,170 establishments on the 2001–2010 target lists without any post-period ODI data. Second, we restrict the analysis to establishments that were not SST-inspected in either of the prior two years, since they were ineligible for inspection under OSHA's rules.

under the historical SST policy (row 1). Targeting establishments using the highest *predicted injuries* would have averted an average of 0.357 (90 percent CI of −0.491 to −0.227, row 3) serious injuries per year, averting 2.1 times as many as averted under the historical SST policy. And targeting based on the highest *estimated treatment effects* would have averted an average of 0.364 (90 percent CI of −0.620 to 0.072, a wider CI that includes zero, row 4) serious injuries per year, 2.2 times as many as averted under the historical SST policy.

Given the number of establishments assigned to inspection, we can also estimate the total number of injuries that OSHA would avert under each policy. Considering injuries over a five-year period, targeting on *historical injuries*, *predicted injuries*, and *estimated treatment effects* would avert 12,140, 15,934, and 16,524 more injuries, respectively, than OSHA's historical SST policy (based on numbers in column 6). Based on our estimate that an injury causing days away from work imposes a social cost of $53,000, these additional averted injuries would have generated social value of $643 million, $844 million, and $876 million, respectively.

Variations on the Benchmark Targeting Policies.—We now estimate the number of injuries OSHA could avert, relaxing some assumptions embedded in the benchmark targeting policies.

Maintaining OSHA's Inspection Budget: Establishments with more employees tend to have larger estimated treatment effects (Table 4) as well as more *historical injuries* and *predicted injuries* (results not shown). But inspecting larger workplaces likely takes more inspector time.

We therefore consider policies in which OSHA targets the establishments with the largest values of each of our targeting metrics, but maintains the 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 employees. That is, if inspecting an establishment with 25 employees requires one day, we assume that inspecting an establishment with 250 employees requires two days. To create targeting groups based on *historical injuries*, we begin with the establishment with the most historical injuries and successively add those with the next highest *historical injuries* until the sum of log(FTEs) of the establishments in this group equals that of those that were assigned to inspection that year under SST. We refer to this group as the "high-priority group" and to the remaining establishments as the "low-priority group." We repeat this process for our other two targeting metrics.

Constraining total inspection costs to the historical policy's budget reduces the number of assignments to inspection by roughly 10 percent for each policy. Despite slightly fewer establishments assigned to inspection, the estimated number of injuries averted (rows 5–7 of Table 5) remains essentially unchanged from the benchmark policies, regardless of the targeting metric.[35]

[35] The estimated number of injuries averted declines by a small amount when targeting on historical injuries and treatment effects, and actually *increases* slightly when targeting on predicted injuries. This latter result arises, even though the number of inspections is lower in this policy than in the benchmark policy, due to the exclusion

Maintaining Randomization: The policies described above, being deterministic, prevent the regulator from using randomization to rigorously estimate the effectiveness of inspections. Deterministic policies might also reduce the threat effect associated with the possibility of being inspected that can motivate uninspected establishments to improve safety, an effect known as general deterrence (Cohen 2000; Shimshack and Ward 2005; Gray and Shadbegian 2007).

Thus, we also consider a targeting policy that incorporates randomization that preserves the SST policy's probability of assignment-to-inspection and thus the threat of inspection. In this approach, OSHA maintains the same proportion of establishments on the primary and secondary target lists as it did with the SST policy (with the primary list consisting of the top 39 percent of eligible establishments and the secondary containing the rest). With our policy, OSHA ranks establishments based on our three targeting criteria instead of the SST program's criterion (the DART rate from two years prior). OSHA randomly assigns all inspections, setting the probability of assignment-to-inspection within the high-priority list to equal that of the SST primary list (43 percent) and within the low-priority list to maintain the inspection budget of the SST program (resulting in a probability of 9 percent).

Our results indicate that OSHA could avert an average of 0.24 injuries per year per assigned establishment if this targeting policy were based on *historical injuries* (90 percent CI of −0.363 to −0.118; see Table 5 row 8), 0.241 injuries per year per assigned establishment if based on *predicted injuries* (90 percent CI of −0.387 to −0.083, row 9), or 0.205 injuries per year per assigned establishment if based on *estimated treatment effects* (90 percent CI of −0.365 to −0.046, row 10). These correspond to 136 percent, 136 percent, and 117 percent, respectively, of the number of injuries averted under the SST policy. These treatments effects are all statistically significantly different from zero, but not statistically significantly different from each other or from the estimate based on the historical policy.

D. *Threats to the Validity of Our Estimates of Counterfactual Policies*

Our estimates are robust to changes in threat effects on uninspected establishments (the Lucas critique) and to using data from the future to estimate treatment effects.

Assessing Potential Bias from a Threat Effect of Inspections.—An inspection regime can reduce injuries via direct effects on inspected establishments—the focus of our analysis—and via threat effects on uninspected establishments (Cohen 2000; Shimshack and Ward 2005; Gray and Shadbegian 2007; Blundell, Gowrisankaran, and Langer 2020). Thus, managers at uninspected establishments might change their behavior under a new targeting rule (Lucas 1976), which could make our estimated treatment effects misleading.

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 restriction, the gap reverses. In all cases, these differences are small and not statistically significant.

Alternative targeting policies will change the threat effect only if managers adjust their behavior in response to changes in their perceived risk of being inspected. If managers are perfectly informed of OSHA's targeting rules, those anticipating a high probability of being inspected might increase their effort to improve compliance, while those anticipating a low probability might reduce their effort.

It is unlikely that a weaker threat effect could offset the benefits of targeting that we report. Consider the alternative policy in which OSHA reassigns all inspections to those establishments with the highest *estimated treatment effects* each year (Table 5, row 4). For diminished threat effects to offset the benefits yielded by our analysis, we estimate that the remaining 80 percent of uninspected establishments would need to *increase* their injury rates by 68 percent of the amount that their injuries would *decrease* if they were assigned to inspection.[36]

We consider such a magnitude of the threat effect implausible. Even though OSHA published its targeting rules in the Federal Register each year, they almost never appeared in industry publications. Our conversations with health and safety professionals indicated that most managers did not know OSHA's targeting rules.[37]

We can also directly test for the importance of threat effects by seeing if workplaces that faced a higher risk of inspection decreased injuries more than similar workplaces facing lower risk of inspection. Specifically, we estimate a regression discontinuity model using OSHA's threshold for determining its SST primary list. We look for evidence of whether uninspected establishments that had injury rates just above the cutoff—and thus faced a much higher risk of inspection—exhibit evidence of greater safety effort than uninspected establishments just below the cutoff.

To calculate how far an establishment is from the cutoff, we subtract the injury cutoff that OSHA used for the primary list in directive year t, which is based on injury rates in year $t-2$, from the establishment's own injury rate in year $t-2$. (See online Appendix J for details.) We provide graphical evidence to illustrate the discontinuous change in the risk of inspection at the primary list threshold. For Figure 3, panels A and B, the x-axis depicts how far an establishment is from the cutoff (the primary list's injury rate threshold). Figure 3, panel A shows that those with a lagged injury rate above this threshold faced a nearly sharp discontinuity in the probability of being on the primary list, from zero to roughly 0.9. Figure 3, panel B shows that those just above this threshold were also roughly 30 percentage points more likely to receive an SST inspection in the calendar year of or following the directive year, compared to those just below the threshold.

[36] We obtain the 68 percent estimate as follows. We estimated that OSHA averts 0.365 annual injuries at establishments assigned to inspection in this alternative policy, which is 0.20 more than was averted than under the historical SST policy (Table 5, column 3). For threat effects to undermine these benefits, the remaining 80 percent of uninspected establishments would thus need incur on average 0.05 *additional* annual injuries. To put this 0.05 value in perspective, the data underlying Figure 2 revealed that the establishments in the bottom 80 percent of estimated treatment effect had an average of 0.073 injuries averted if assigned to inspection. That is, the change in the threat of inspection would need to increase injuries by 68 percent of the amount that injuries would fall if these establishments were assigned to inspection ($0.05/0.073 = 68$ percent).

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

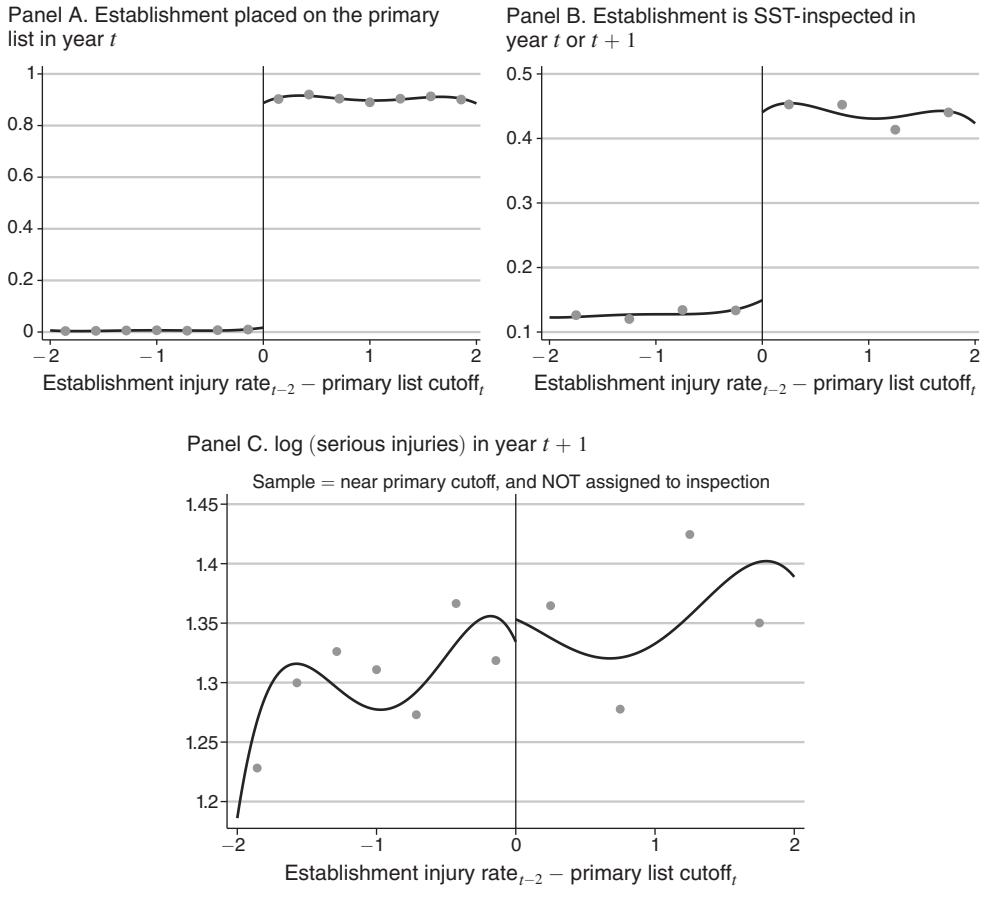

FIGURE 3. REGRESSION DISCONTINUITY RESULTS PROVIDE NO EVIDENCE THAT A CHANGE IN ESTABLISHMENTS' RISK OF INSPECTION AFFECTS INJURY RATES

Notes: Each panel displays a binned scatterplot in which the *y*-variable is provided in the panel title and the *x*-variable, defined in the text, determines whether an establishment's injury rate from year $t-2$ is high enough to make it eligible for the SST primary list in year t. In panels A and B, the sample includes those establishments considered for the 2001–2010 SST target lists that had not received an inspection in the prior two years. In panel C, the sample is further restricted to establishments that were on either the secondary or primary lists but were not assigned to SST inspection. See text for further details.

If establishments adjust their safety hazards based on their risk of being inspected, then those that were just above the primary cutoff (thus facing a higher risk of inspection) will, *even if they are not actually inspected*, engage in more safety efforts that decrease injury rates than those just below the threshold. To isolate the effect of the higher threat of inspection, we restrict attention to the subsample of establishments that (a) had injury rates in $t-2$ that put them just above or just below the primary list cutoff, (b) were eligible for an SST inspection, and (c) were *not* assigned one.

Figure 3, panel C illustrates the results of our test of the threat effect. It depicts a binned scatterplot of residuals from regressing the log number of serious injuries during the calendar year following the directive year (year $t+1$) on OSHA-region fixed effects, year fixed effects, a manufacturing industry dummy, and $\log(hours)_{t-2}$.

The solid lines depict best-fit lines with separate slopes for each side of the cutoff. There is no change in the number of injuries in year $t + 1$ between uninspected establishments that just barely did and did not make it onto the primary list. Online Appendix J provides the corresponding regression estimates and other details of this analysis.[38]

If establishments adjusted their safety efforts due to changes in the risk of inspection, we would expect to observe those that just barely made it onto the primary list experiencing fewer injuries than those that just barely did not, even among those not actually assigned to inspection. Figure 3, panel C and online Appendix J provide no such evidence.

This analysis suggests that establishments are unlikely to adjust their injuries in response to the changes to OSHA's targeting rules that we consider—changes that would be subtler than the discontinuity at the primary list cutoffs examined here. These results do not imply that threat effects are irrelevant, but rather that the differing threat of inspections would not substantially change the behavior of uninspected establishments under the targeting regimes we analyze.

If—contrary to the evidence just presented—threat effects *were* important in our setting, then most of our results would likely *underestimate* the benefits of our targeting policies. Consider our policies that exclusively target inspections nonrandomly. While our policy of rank-ordering establishments on *historical injuries* might eliminate any threat effect (because establishments could in theory know if they rank highly enough to be targeted), policies that rely on machine learning might *enhance* the threat effect. Assume that (a) establishments have a noisy signal of their ranking based on *predicted injuries* or *estimated treatment effect* and (b) the threat effect increases attention to safety in the same way that inspections do (for example, by triggering additional management hours dedicated to promoting safety). Then, our policies that target based on criteria generated by machine learning algorithms increase the threat of inspection among establishments (those with high *predicted injuries* or *estimated treatment effects*) for which we expect greater injury reductions than we expect for the establishments that faced a higher threat of inspection under the SST program (those with one year of high injury rates). Furthermore, we found that OSHA would still avert more injuries than it did under the SST policy if it used our policy that maintains the level of threat that held under SST but prioritized inspections using any of our three alternative metrics. In short, if threat effects are important, we have little reason to believe that they attenuate the value of the proposed alternative targeting approaches.

Assessing Potential Bias from Using Data from the Future.—Our analyses have used data from our entire 2001–2010 sample period to construct establishments' *predicted injuries* and *estimated treatment effects.* In reality, when OSHA constructs its target lists in a given year, only data through the prior year are available. To assess whether our estimates of the number of injuries averted under alternative targeting policies might be materially different if based only on data available to the agency

[38] We can tweak this regression discontinuity design to estimate the *local* average treatment effect of being SST-inspected on subsequent injuries, in the spirit of Li and Singleton (2019). We do so in online Appendix K, where we apply this approach to a similar sample, but include the assigned-to-inspection establishments.

when it was constructing its target lists, we reestimate our models using only data from the 2001–2006 randomized sample (the first half of our sample period) and then use those results to generate *predicted injuries* and *estimated treatment effects* for establishments in the 2007–2010 randomized sample (the second half of our sample period). We compare these estimated benefits of targeting inspections during 2007–2010, when we use these two metrics generated based only on the 2001–2006 data, with the results of our main approach, which generated these metrics based on all of the data (2001–2010).

Specifically, we use *predicted injuries* or *estimated treatment effects* trained on the 2001–2006 sample to identify which establishments to place in the high- and low-priority groups during 2007–2010. As in our main analysis, we run a regression corresponding to Equation 5 to estimate the average number of injuries averted. In contrast to our main analysis, however, we only estimate the regression once, since we are not randomly partitioning the data as we did in our main analysis.[39] We therefore report the estimated average treatment effect $\hat{\gamma}_1$ from this regression (rather than the median of $\hat{\gamma}_1$ across 250 sample splits). This yields an estimated number of averted injuries per year from assignment to inspection among establishments in the high-priority group of 0.792 (SE $=$ 0.367) when targeting on *predicted injuries* and 0.589 (SE $=$ 0.323) when targeting on *estimated treatment effects* (columns 1 and 2 of online Appendix Table A.3). Both estimates are only slightly smaller than (and not statistically significantly different from) the corresponding estimates for inspections assigned during the second half of our time period (2007–2010) based on parameters derived from the entire sample period (i.e., our main approach described in Section IIC); namely, 0.967 (SE $=$ 0.292) when targeting on *predicted injuries* and 0.664 (SE $=$ 0.286) when targeting on *estimated treatment effects* (columns 3 and 4 of online Appendix Table A.3).[40]

In short, these results suggest that our use of all years of data to construct our targeting metrics does not materially overestimate the benefits of alternative targeting policies.

IV. Why Did Targeting on Predicted Injuries Perform Better Than Targeting on Treatment Effects?

OSHA wants to allocate its scarce inspections where they avert the most injuries; that is, where the treatment effect is largest. However, in most of the alternative targeting approaches we considered, we estimated that OSHA would avert the most injuries by targeting inspections not on *estimated treatment effects*, but rather on *predicted injuries.* Furthermore, our estimates for targeting based on *estimated treatment effects* were much less precise. To understand why, it is useful to think

[39] We do not conduct the sample-splitting procedure here because we do not have to worry about overfitting leading to bias in our machine learning estimates. That is, we use sample splitting in our main analysis to estimate the models on one half of the sample, then apply these models to the second half of the sample, repeating this process many times because one random partition of the sample might not be representative. Here, because we estimate the models on the first half of the sample (2001–2006) and are applying it to the second half of the sample (2007–2010), we do not have to perform this repeated sample splitting.

[40] For simplicity, in this exercise we omit the exclusion criteria that establishments are ineligible for inspection if they were inspected in either of the prior two years.

of the regulator's problem of how to target inspections as a kind of *bias-variance tradeoff*. In particular, the regulator might better achieve its objective (of averting problems) by choosing to target inspections based on a criterion (such as the number of problems) that is a biased measure of that objective but can be estimated more precisely.

While inspecting establishments with high *treatment effects* clearly serves the regulator's objective to avert the most injuries, it is not obvious that inspecting establishments with high *expected injuries* serves this goal as well. In other words, *expected injuries* is a biased measure of the regulator's objective. However, in our case, this bias is likely small. Conceptually, it is plausible that OSHA inspectors could avert more injuries at workplaces with more injuries—as would occur if there are economies of scale in remediating hazards. Empirically, we found high correlation among establishments' *estimated treatment effects* and *predicted injuries*; across our 250 sample splits, the median correlation coefficient (r) was 0.78.

Furthermore, because treatment effects are fundamentally difficult to estimate, they will inherently be estimated with high variability. Indeed, even with a very large sample size like ours, treatment effect estimates were quite unstable. This imprecision means that ranking establishments based on their *estimated* treatment effects does an imperfect job of ranking them by their *true* treatment effects. As a result, targeting those with the highest *estimated treatment effects* will result in some establishments with low actual treatment effects being inspected (a Type I error), and some establishments with high actual treatment effects being uninspected (a Type II error). The regulator can predict injuries with much more precision and, as a result, more accurately target those with high *expected injuries.* Indeed, *predicted injuries* was much more stable across sample spits than *estimated treatment effects*. We estimate that the instability in establishments' *predicted injuries* across sample splits was nearly 10 times smaller than the instability in their *estimated treatment effects*.[41]

Given that the regulator could more accurately identify establishments with the highest *expected injuries* than those with the highest *estimated treatment effects* and given that establishments with many injuries tend to be those with the largest treatment effects, it is perhaps not surprising that we found that OSHA would avert just as many (if not more) injuries targeting on *predicted injuries* than on *estimated treatment effects.* We expect targeting on estimated treatment effects will be more effective than targeting on predicted problems in settings in which treatment effects (a) can be estimated with more precision, or (b) are not as strongly correlated with the level of problems.

[41] We measure this instability by considering the standard deviation in establishments' *estimated treatment effects* and *predicted injuries* across our 250 sample splits. Across all establishments in our sample, the mean *estimated treatment effect* was −0.169, with a standard deviation of 0.15. For the average establishment in our dataset, the standard deviation of *estimated treatment effects* across the 250 sample splits was 0.12, which is 80 percent of the 0.15 standard deviation in the *overall* distribution of treatment effects. In comparison, across all establishments in our sample, the mean *predicted injuries* was 5.1 with a standard deviation of 5.8. For the average establishment in our dataset, the standard deviation in its *predicted injuries* across the 250 sample splits was 0.49, a mere 8.4 percent of the 5.8 standard deviation in the *overall* distribution of predicted injuries. The 8.4 percent instability in establishments' *predicted injuries* across sample splits is nearly 10 times smaller than the 80 percent instability in establishments' *estimated treatment effects* across sample splits.

V. Conclusion

OSHA inspections of dangerous workplaces substantially improved workplace safety. Our estimates imply that the average inspection caused a 9 percent decline in serious injuries (those causing days away from work), averting an average of 2.4 serious injuries over five years at each inspected establishment. We estimate that each inspection yielded a social benefit of roughly $125,000, roughly 35 times OSHA's cost of conducting an inspection.

However, we also found that the agency could have averted substantially more injuries had it used any of our alternative targeting policies. OSHA could have averted over twice as many serious injuries by targeting its inspections to establishments with the highest estimated treatment effects and nearly as many—and in some cases more—by targeting based on average historical serious injury counts or on predicted serious injury counts. It was surprising to find that, in some scenarios, more injuries could be averted from a targeting regime based on predicted injuries than from a regime based on estimated treatment effects. However, estimating establishments' expected number of injuries absent an inspection is a much easier prediction problem than estimating establishments' treatment effects. Moreover, in our setting, the two metrics turned out to be highly correlated. Targeting historical outcomes or predicted outcomes would be less likely to outperform targeting based on treatment effects in settings in which treatment effects can be estimated precisely or where predicted problems and treatment effects are less correlated.

Our approach also enables regulators to learn where inspections are relatively *ineffective.* For example, OSHA could investigate why nursing homes have high injury rates but avert fewer-than-average injuries after being inspected. Perhaps this is due to OSHA lacking standards for musculoskeletal diseases, which account for a large share of injuries in nursing homes; if so, this finding might inform OSHA about the possible need to create such standards.

Our study has several limitations. One is that we estimated the effects of alternative targeting policies only on establishments that had been on OSHA's historical SST target lists. In reality, OSHA could choose its annual inspection targets from a much larger set of establishments. Our not considering this broader population means that our results are weakly conservative, as the restriction might have caused our results to *underestimate* the gains of our alternative targeting policies. Other limitations of our study include our not considering effects of inspections beyond five years and our focusing only on the 29 states where OSHA is the primary regulator. We also did not measure the effects of inspections on illnesses or on less consequential injuries not resulting in days away from work. Finally, data limitations prevented us from considering injuries sustained by temporary or contract workers.

With these limitations in mind, we show that combining randomization and machine learning can substantially improve regulatory agencies' performance. This approach could improve the effectiveness of many other organizations that target inspections, from tax and food regulatory agencies to multinational firms assessing suppliers' process quality and working conditions. 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), the US Bureau of Safety and Environmental Enforcement has begun targeting inspections of offshore oil and gas operations based on their perceived risk of safety incidents (US Bureau of Safety and Environmental Enforcement 2018), and the city of Chicago has begun using risk-based forecasting to help determine the order in which it inspects restaurants (Spector 2016). Our research reveals how agencies can estimate the relative benefits of algorithms that vary in simplicity and transparency as well as in threat effects to encourage compliance among uninspected establishments.

REFERENCES

Alm, James, and Jay Shimshack. 2014. "Environmental Enforcement and Compliance: Lessons from Pollution, Safety, and Tax Settings." *Foundations and Trends in Microeconomics* 10 (4): 209–74.

Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355 (6324): 483–85.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *PNAS* 113 (27): 7353–60.

Avis, Eric, Claudio Ferraz, and Frederico Finan. 2018. "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians." *Journal of Political Economy* 126 (5): 1912–64.

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

Biddle, Jeff, and Karen Roberts. 2003. "Claiming Behavior in Workers' Compensation." *Journal of Risk and Insurance* 70 (4): 759–80.

Blundell, Wesley, Gautam Gowrisankaran, and Ashley Langer. 2020. "Escalation of Scrutiny: The Gains from Dynamic Enforcement of Environmental Regulations." *American Economic Review* 110 (8): 2558–85.

Boden, Leslie I., Nicole Nestoriak, and Brooks Pierce. 2010. Using Capture-Recapture Analysis to Identify Factors Associated with Differential Reporting of Workplace Injuries and Illnesses. Alexandria, VA: American Statistical Association.

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

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2020. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments." NBER Working Paper No. 24678.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1): C1–C68.

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

Cohen, Mark A. 2000. "Empirical Research on the Deterrence Effect of Environmental Monitoring and Enforcement." *Environmental Law Reporter* 30: 10245–52.

Davis, Jonathan M. V., and Sara B. Heller. 2020. "Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs." *Review of Economics and Statistics* 102 (4): 664–77.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2018. "The Value of Regulatory Discretion: Estimates from Environmental Inspections in India." *Econometrica* 86 (6): 2123–60.

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

Foley, Michael, Z. Joyce Fan, Eddy Rauser, and Barbara Silverstein. 2012. "The Impact of Regulatory Enforcement and Consultation Visits on Workers' Compensation Claims Incidence Rates and Costs, 1999–2008." *American Journal of Industrial Medicine* 55 (11): 976–90.

Glaeser, Edward L., Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy." *American Economic Review* 106 (5): 114–18.

Gonzalez-Lira, Andres, and Ahmed Mushfiq Mobarak. 2019. "Slippery Fish: Enforcing Regulation under Subversive Adaptation." IZA Working Paper No. 12179.

Gray, Wayne B., and John M. Mendeloff. 2005. "The Declining Effects of OSHA Inspections on Manufacturing Injuries, 1979–1998." *ILR Review* 58 (4): 571–87.

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

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

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

Haviland, Amelia M., Rachel M. Burns, Wayne B. Gray, Teague Ruder, and John Mendeloff. 2012. "A New Estimate of the Impact of OSHA Inspections on Manufacturing Injury Rates, 1998–2005." *American Journal of Industrial Medicine* 55 (11): 964–75.

Hino, Miyuki, Elinor Benami, and Nina Brooks. 2018. "Machine Learning for Environmental Monitoring." *Nature Sustainability* 1 (10): 583–88.

Ibanez, Maria R., and Michael W. Toffel. 2020. "How Scheduling Can Bias Quality Assessment: Evidence from Food Safety Inspections." *Management Science* 66 (6): 2396–2416.

Jacob, Daniel. 2020. "Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects." Unpublished

Johnson, Matthew, David Levine, and Michael Toffel. 2023. "Replication data for: Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.38886/E168101V1.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Kleven, Henrik Jacobsen, Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez. 2011. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." *Econometrica* 79 (3): 651–92.

Lee, Jonathan M., and Laura O. Taylor. 2019. "Randomized Safety Inspections and Risk Exposure on the Job: Quasi-experimental Estimates of the Value of a Statistical Life." *American Economic Journal: Economic Policy* 11 (4): 350–74.

Leigh, J. Paul. 2011. "Economic Burden of Occupational Injury and Illness in the United States." *Milbank Quarterly* 89 (4): 728–72.

Levine, David I., Michael W. Toffel, and Matthew S. Johnson. 2012. "Randomized Government Safety Inspections Reduce Worker Injuries with No Detectable Job Loss." *Science* 336 (6083): 907–11.

Li, Ling, and Perry Singleton. 2019. "The Effect of Workplace Inspections on Worker Safety." *ILR Review* 72 (3): 718–48.

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

McKenzie, David. 2012. "Beyond Baseline and Follow-up: The Case for More T in Experiments." *Journal of Development Economics* 99 (2): 210–21.

Musick, Tom, Sarah Trotto, and Kyle Morrison. 2016. "Compliance Assistance—Not Fines—Should Be Priority, Senators Tell OSHA." *Safety and Health*, February 12. https://www.safetyandhealthmagazine.com/articles/13662-compliance-assistance-not-fines-should-be-priority-senators-tell-osha.

NETS. 2016. Special Extract of the "National Establishment Time Series." Walls and Associates, obtained August 2016. For information, contact dwalls2@earthlink.net.

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

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

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

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

Ruser, John W., and Robert S. Smith. 1991. "Re-estimating OSHA's Effects: Have the Data Changed?" *Journal of Human Resources* 26 (2): 212–35.

Shimshack, Jay P. 2014. "The Economics of Environmental Monitoring and Enforcement." *Annual Review of Resource Economics* 6 (1): 339–60.
Shimshack, Jay P., and Michael B. Ward. 2005. "Regulator Reputation, Enforcement, and Environmental Compliance." *Journal of Environmental Economics and Management* 50 (3): 519–40.
Short, Jodi L., Michael W. Toffel, and Andrea R. Hugill. 2016. "Monitoring Global Supply Chains." *Strategic Management Journal* 37 (9): 1878–97.
Slemrod, Joel, Marsha Blumenthal, and Charles Christian. 2001. "Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota." *Journal of Public Economics* 79 (3): 455–83.
Smith, Robert Stewart. 1979. "The Impact of OSHA Inspections on Manufacturing Injury Rates." *Journal of Human Resources* 14 (2): 145–70.
Spector, Julian. 2016. "Chicago Is Predicting Food Safety Violations. Why Aren't Other Cities?" *CityLab*, January 7. https://www.citylab.com/solutions/2016/01/chicago-is-predicting-food-safety-violations-why-arent-other-cities/422511/.
Stigler, George J. 1971. "The Theory of Economic Regulation." *Bell Journal of Economics and Management Science* 2 (1): 3–21.
Telle, Kjetil. 2013. "Monitoring and Enforcement of Environmental Regulations: Lessons from a Natural Field Experiment in Norway." *Journal of Public Economics* 99: 24–34.
US Bureau of Labor Statistics. 2007. *Nonfatal Occupational Injuries and Illnesses Requiring Days Away from Work, 2005*. Washington, DC: US Bureau of Labor Statistics.
US Bureau of Safety and Environmental Enforcement. 2018. "BSEE Launches Risk-Based Inspection Program." https://www.bsee.gov/newsroom/latest-news/statements-and-releases/press-releases/bsee-launches-risk-based-inspection (accessed March 12, 2019).
US Department of Health and Human Services. 2011. *Fiscal Year 2012 Food and Drug Administration, Justification of Estimates for Appropriations Committees*. Washington, DC: US Department of Health and Human Services.
US Department of Labor. 2008. *Congressional Budget Justification: Occupational Safety and Health Administration, FY 2009*. Washington, DC: US Department of Labor.
US Food and Drug Administration. 2016. *2016 Annual Report on Inspections of Establishments in FY 2015*. Silver Spring, MD: US Food and Drug Administration.
US Government Accountability Office. 2016. *FDA's Targeting Tool Has Enhanced Screening, but Further Improvements Are Possible*. Washington, DC: US Government Accountability Office.
US Occupational Safety and Health Administration. 2004. *Nationwide Site-Specific Targeting (SST) Inspection Program Request for Comments*. Washington, DC: US Occupational Safety and Health Administration.
US Occupational Safety and Health Administration. 2008. *OSHA Announces its Site-Specific Targeting Plan for 2008*. Washington, DC: US Occupational Safety and Health Administration.
US Occupational Safety and Health Administration. 2011. "Annual Site-Specific Targeting (SST) Target Lists, 2001—2010." US Occupational Safety and Health Administration. (accessed October 2014).
US Occupational Safety and Health Administration. 2013. Annual OSHA Data Initiative (ODI) Survey Responses, 1996—2011. (accessed October 2014 via a Data Sharing Agreement with OSHA's Office of Statistical Analysis).
US Occupational Safety and Health Administration. 2014. OSHA Enforcement Data. https://enforcedata.dol.gov/views/data_summary.php (accessed December 2014).
US Occupational Safety and Health Administration. 2017a. "Commonly Used Statistics." US Occupational Safety and Health Administration. https://www.osha.gov/oshstats/commonstats.html (accessed February 2017).
US Occupational Safety and Health Administration. 2017b. *OSHA Fact Sheet: OSHA Inspections*. Washington, DC: US Occupational Safety and Health Administration.
US Occupational Safety and Health Administration. 2021. "About OSHA." US Department of Labor. https://www.osha.gov/aboutosha (accessed November 2021).
van der Laan, Mark J., and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer Science and Business Media.
van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1): Article 25.
Viscusi, W. Kip. 1986. "The Impact of Occupational Safety and Health Regulation, 1973–1983." *Bell Journal of Economics* 17 (4): 567–80.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.

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