Decision Authority and the Returns to Algorithms

Edward L. Glaeser, Andrew Hillis, Hyunjin Kim, Scott Duke Kominers, and Michael Luca*

December 12, 2022

PRELIMINARY AND NOT FOR DISTRIBUTION

Abstract

We evaluate a pilot in an Inspections Department to explore the returns to a pair of algorithms that varied in their sophistication. We find that both algorithms provided substantial prediction gains, suggesting that even simple data may be helpful. However, these gains did not result in improved decisions. Inspectors used their decision authority to often override algorithmic recommendations, partly to consider other organizational objectives without improving outcomes. Interviews with 55 departments find that while many ran similar pilots, all provided considerable decision authority to inspectors, and those with sophisticated pilots transitioned to simpler approaches. These findings suggest that for algorithms to improve managerial decisions, organizations must consider the returns to algorithmic sophistication in each context, and carefully manage how decision authority is allocated and used.

* Authors are listed alphabetically. Ruijing Chen, Fabian Konig, and Gabrielle Lamont-Dobbin provided excellent research assistance. The authors gratefully acknowledge the helpful comments of Susan Athey, Raj Choudhury, Felipe Csaszar, J.P Eggers, Avi Goldfarb, Shane Greenstein, Jorge Guzman, Kristina McElheran, Sendhil Mullainathan, Abhishek Nagaraj, Phanish Puranam, Andrei Shleifer, Mitchell Weiss, and seminar participants. Data for this project was provided by the Inspectional Services Department and Yelp. Kim and Luca have consulted for tech companies, including Yelp. Kominers advises firms engaged in marketplace design and development. We are grateful for the support of the National Science Foundation (grants CCF-1216095, DGE-1144152, and SES-1459912), the Harvard Milton Fund, the Taubman Center for State and Local Government, Harvard Business School, the Rappaport Institute for Greater Boston, the Alfred P. Sloan Foundation, the Ewing Marion Kauffman Foundation, the Ng Fund and the Mathematics in Economics Research Fund of the Harvard Center of Mathematical Sciences and Applications. All errors are our own.

1. Introduction

Organizations are increasingly interested in using algorithms to support decision-making, in contexts as diverse as selecting applicants to hire, identifying promising innovations, and making resource investment decisions (e.g. Agrawal, Gans, & Goldfarb, 2018; Cowgill, 2019; Choudhury, Starr, & Agarwal, 2020). The potential for algorithms to improve prediction was demonstrated as early as the 1950s (e.g., Meehl, 1954; Dawes, 1979; Grove & Meehl, 1996; Kahneman et al, 2016), and firms today increasingly invest in data and algorithmic sophistication (Brynjolfsson & McElheran, 2019; Bajari, Hortaçsu, & Suzuki, 2019). Even so, evidence suggests that many firms may not be seeing the returns to their investments (Brynjolfsson, Jin, & McElheran, 2021; Ransbotham et al., 2019).[1]

This raises a question on the extent to which the use of algorithms ultimately translates into improvements in decisions in organizational contexts. While a general assumption is that more information and computation will lead to more accurate decisions (Gigerenzer et al., 1999; Agrawal et al., 2018), there are at least two key reasons why this may not be the case. First, it may be that for certain managerial problems, the returns to algorithmic sophistication in terms of prediction are limited, given the degree of uncertainty involved and the power of simple heuristics (e.g. Sull & Eisenhardt, 2015).

Second, when leveraging data and algorithms, organizations decide not only whether to use them, but how. While some decisions can be fully automated, many managerial decisions involve judgment beyond prediction (Agrawal et al., 2018; 2019; Cowgill, 2019; Choudhury et al., 2020; Raisch & Krakowski, 2020). In such cases, algorithms provide predictions that decision-makers may take as inputs, as they choose to use algorithmic recommendations as a decision aid, rather than a decision rule. Decision-makers may thus use their decision authority

[1] Across a survey of more than 2,500 executives, "seven out of 10 companies surveyed report minimal or no impact from artificial intelligence (AI) so far. Among the 90% of companies that have made some investment in AI, fewer than 2 out of 5 report business gains from AI in the past three years… this means 40% of organizations making significant investments in AI do not report business gains from AI" (Ransbotham et al., 2019).

to make a decision that does not leverage potential prediction gains from algorithms—either because they are balancing other organizational objectives beyond the predicted measure, or because they end up dissipating informational gains through their discretion. Evaluating how decision-makers use their discretion when faced with data-driven inputs is thus an important step in understanding the impact of algorithms for organizations (Athey, Bryan, & Gans, 2020).

In this paper, we evaluate the returns to algorithms on managerial decisions within a real organizational context, and explore these two factors. We compare the performance of human predictions to algorithmic ones, and further compare two algorithms with varying degrees of sophistication: one based on simple historical averages, the other based on a random forest model trained on both historical data from within the organization and additional data from online platforms. We find that algorithms indeed appear to provide substantial gains over human prediction. But, the greatest gains come from simply integrating data into the decision process, rather than from algorithmic sophistication. Moreover, these improvements in prediction do not translate into improved decisions. Decision-makers often reject data-driven recommendations, in part in consideration of other organizational objectives, but on average without balancing or improving upon them—suggesting that they may have used their decision authority to diminish potential gains from algorithms. These findings suggest that while organizations can improve decision-making from using algorithms, managing decision authority may be at least as important as investments in algorithmic sophistication or enhanced data collection, at least in these early days of putting algorithms to into practice.

We evaluate the returns to algorithms on managerial decisions through an intervention implemented by an Inspectional Services department, where inspectors rely on their judgment to decide which restaurants to inspect. This setting offers a number of compelling attributes to test the power of predictive algorithms: (i) inspectors' scarce time must be allocated with an uncertain but primarily predictive objective: identifying restaurants with more health violations; (ii) inspectors have secondary organizational objectives to consider that are clearly defined by

the department and measurable (i.e. reducing travel distance, targeting more overdue inspections, prioritizing more serious violations and popular restaurants); (iii) while inspectors possess informative experience and insight, there are historical administrative records and external data that raise the possibility of improving algorithmic sophistication (Lehman, 2014); and (iv) inspectors retain ultimate decision authority and carry out the inspections themselves.

We compare three approaches to allocate inspectors: (1) inspector prediction ("business-as-usual"), (2) a "data-poor" algorithm based on the average number of historical violations for each restaurant; and (3) a "data-rich" algorithm based on a random forest model trained on historical violations and Yelp data.[2] Restaurants with the highest predicted likelihood of violations according to each approach were randomly sorted and provided as lists to inspectors to guide their inspections over four periods of two weeks each. This design allows us to observe counterfactual inspector predictions and their ultimate decisions, and offers insights into the gains from algorithmic sophistication in the field by comparing the two algorithms.

Our results suggest substantial gains from predicting violations using algorithms, which identified restaurants with over 50 percent more violations than inspectors. Most gains came from integrating data into the process, with the data-poor algorithm providing improvements nearly as large as those from the data-rich algorithm. Given the difficulty of generalizing from this context, the main insight we draw from is that even simple data was valuable in improving predictions, and there may be similar managerial contexts where this is the case.

However, these gains in prediction do not appear to have translated into improved decisions. Inspectors were only about two-thirds as likely to inspect algorithm-recommended restaurants relative to those that they had ranked highly, thereby dissipating much of the prediction gains. While inspectors varied in the extent to which they deviated from the algorithm, most inspected

[2] We use the term "data-rich" in a relative sense to the other algorithm. One can imagine using a vast set of other data that may yield higher-quality insights, which is beyond the scope of this paper. The motivation behind this treatment was to explore the extent to which richer data modeled in a more sophisticated way adds any marginal gain, given rising interest and investment in data and advanced technologies.

more restaurants that they prioritized compared to those by algorithms. Given this behavior, we also explore the possibility that selection could be driving the estimated gains from algorithmic inputs, but find little evidence that this can explain the full magnitude of the observed effects.

We find some evidence that inspectors sought to improve the decision by leveraging their private knowledge on secondary organizational objectives. In particular, inspectors appear to have been sensitive to how overdue restaurants were. However, they did not ultimately make economically and statistically significant improvements on these secondary objectives when overriding algorithmic inputs. Inspectors did not on average inspect restaurants that were more overdue, and instead increased travel costs and targeted less severe violations – suggesting that they used their discretion to dissipate the gains from algorithms rather than improving on secondary objectives.

While our analysis cannot fully pin down the mechanism, we find some anecdotal and exploratory empirical evidence consistent with the interpretation that inspectors may have rejected algorithmic recommendations when they conflicted with their priors on restaurant attributes that drive violations. Our findings thus raise the possibility that simple rules-of-thumb developed in the presence of uncertainty can work against the introduction of algorithms to support decision-making. While other potential explanations such as algorithm aversion or social relationships with owners are possible, they are less likely to explain our results because of the particularities of this context: the department chose to not explicitly communicate that these recommendations were driven by algorithms, and inspectors were assigned to a different neighborhood every two years, providing little opportunity to build relationships. Nevertheless, these explanations are also broadly consistent with our finding that inspectors did not use their decision authority to improve the decision.

To explore the extent to which our findings might generalize beyond our pilot department, we contacted inspectional departments serving the largest 200 metropolitan areas in the U.S. to conduct unstructured interviews. We interviewed 55 departments for up to one hour to

understand their decisions on using algorithms and decision authority. We found that only a few departments had run pilots using sophisticated algorithms, as most believed that they did not have sufficient data or technical capability, consistent with similar statements from C-level executives on data availability being their greatest challenge for using artificial intelligence (AI) (CognitiveScale, 2021). Furthermore, all departments gave inspectors considerable decision authority, citing the private knowledge they had that would help them better predict violations and balance other organizational objectives. However, departments using algorithms to guide their decisions faced similar issues as our pilot city, where many inspectors appeared to dissipate the informational gains from using algorithms.

Together, our findings suggest that while organizations place much value in algorithmic sophistication relative to managing decision authority, the latter may merit a more serious consideration when seeking to use algorithms as decision aids. As firms increasingly make investments in data and AI, estimated at over $40 billion USD in 2020 and projected to double in the next few years, our findings offer relevant practical implications. While data and algorithms can provide substantial improvements in decision-making, the returns to algorithmic sophistication may be limited in some contexts, and potential gains may be dissipated by managers who are intended to oversee and improve algorithmic recommendations. These findings suggest that organizations may need to think carefully about the returns to algorithmic sophistication in each context, and explore how decision-making processes can be redesigned to make use of managers' private contextual knowledge.

This study contributes to emerging research on algorithms and decision-making in organizations. Studies have examined various implications of advancements in AI on organizations (Felten, Raj, & Seamans, 2021; Choudhury et al., 2020; Tong, Jia, Luo, & Fang, 2021; Brynjolfsson et al., 2021). While much work especially in psychology has explored how algorithms improve on human predictions and how individual preferences to rely on algorithms evolve (e.g. Dietvorst, Simmons, & Massey, 2015; Logg, Minson, & Moore, 2019) there has been

less insight from managerial contexts on how decision-makers within organizations use their decision authority and leverage their private contextual knowledge when using algorithms as decision aids. Given that the default arrangement across many organizations is to allocate final decision authority to managers so they can correct problematic algorithmic recommendations or better inform the decision according to broader organizational objectives, this study highlights a key factor that organizations may need to consider in realizing gains from using algorithms.

Our work is most closely related to Hoffman et al (2018), which explores the impact of adopting a job testing technology on hiring and finds that managers who are more likely to hire against test recommendations make worse average hires. We build on these findings in two key respects. First, we explore the returns to algorithmic sophistication by comparing two algorithms with varying inputs, and find that algorithmic sophistication provides limited returns at least in this context. Second, we separate these prediction gains from improvements in decisions by explicitly examining how decision-makers use their discretion. While Hoffman et al (2018) infer whether managers may be more likely to hire against test recommendations, we directly observe and evaluate the extent to which decision-makers reject algorithmic recommendations and whether they are balancing other organizational objectives when doing so. Our findings suggest that while decision-makers override algorithms at least in part to consider other goals, they are not ultimately able to substantially improve upon them, suggesting that organizations need to better understand how decision-makers can apply their judgment usefully to improve decisions when working with algorithmic inputs.

In addition to the literature on algorithms and decision-making, our analysis contributes to research on information technology investments and digital transformation more broadly. A growing body of work has identified organizational practices that shape the returns to investments in information technology (Bresnahan, Brynjolfsson, & Hitt, 2002; Bartel, Ichinowski, & Shaw, 2007; Bloom, Sadun, & Van Reenen, 2012; Brynjolfsson et al., 2021). Our findings point to organizational design challenges faced by organizations in deploying

information technologies in practice. While no single context fully generalizes to other settings, our study highlights that the design and management of decision authority can at least in some cases be more important than the sophistication of the data and algorithms themselves.

2. The returns to algorithms on decisions in organizations

While much research has examined the potential for algorithms to improve predictions, there has been less insight on how the use of algorithms ultimately translates into improvements in decisions within organizations. We explore the returns to algorithms on decisions within a real organizational context, and propose that the returns may be limited for managerial decisions due to the limitations of algorithmic sophistication and the role of decision authority.

2.1 Algorithms as decision aids

Recent developments in machine learning have enabled the use of algorithms in many contexts that were not previously possible, and organizations are increasingly interested in using algorithms to support their decision-making (Kleinberg et al., 2017; Cowgill, 2018; Choudhury et al., 2020). Machine learning algorithms can work with far more complex functional forms and data inputs, which has fueled growing interest in algorithmic sophistication via investment in more data and model complexity (Ludwig & Mullainathan, 2021).

Much research highlights that algorithms can improve upon human prediction. Starting with Meehl's (1954) review of forecasting studies showing that algorithms outperformed human experts, a long line of research has provided evidence that algorithmic predictions can reduce bias and increase consistency relative to human predictions (e.g., Dawes, 1979; Grove & Meehl, 1996; Kahneman et al, 2016). The accuracy of algorithmic predictions over human predictions has since been documented across a large variety of domains, such as recidivism (Thompson, 1952; Stevenson, 2017; Berk, 2017; Kleinberg et al, 2017), medical diagnoses (Dawes et al., 1989; Grove et al., 2000), and many others (Goodwin & Fildes, 2007; Vrieze and Grove, 2009).

However, there has been less insight on the extent to which the use of algorithms as decision aids ultimately translates into improvements in managerial decisions. There is a general

assumption that more information and computation will lead to more accurate decisions (Gigerenzer et al., 1999). Yet despite growing investment into algorithmic sophistication, evidence suggests that firms may not be seeing the returns to their investments (Brynjolfsson et al., 2021), raising the possibility that this may not be the case. We propose that the returns to algorithms for managerial decisions may be limited, due to two possible mechanisms.

2.2 Algorithmic sophistication and improvements in prediction

One reason that the returns to algorithms may be limited for managerial decisions is that for many such decisions, more information and computation may not necessarily improve predictions. Research on heuristics finds an inverse-U-shaped relationship between accuracy and the amount of information or computation used (see Gigerenzer & Gaissmaier, 2011 for a review)—suggesting that in some cases, more sophistication may in fact be harmful for improving predictive accuracy (e.g. Czerlinski et al., 1999; Gigerenzer & Goldstein, 1996; Dhami, 2003; Wubben & Wangenheim, 2008; Gigenrenzer & Brighton, 2009). This research highlights the bias-variance tradeoff, proposing that heuristics may be more likely to be accurate than complex strategies in contexts with higher uncertainty and redundancy where simpler methods with fewer free parameters may reduce variance (Gigerenzer & Gaissmaier, 2011).[3]

Qualitative research in strategy has provided evidence consistent with this idea, suggesting that the use of simple rules can help organizations make better decisions (Bingham & Eisenhardt, 2011; Sull & Eisenhardt, 2015). This work highlights that organizations that develop and employ simple rules can reduce mental costs for decision-makers, increase clarity in the decision, and improve coordination across the organization. While this research does not examine the impact of data-driven decision tools or evaluate the returns to algorithmic

[3] A Bayesian interpretation of Occam's razor, the principle that unnecessarily complex models should not be preferred to simpler ones, also provides quantitative support (MacKay, 1992). The basic underlying intuition is that while complex models can always fit the data better, simpler models can be the more probable model. While a more complex model with more free parameters can predict a higher variety of datasets compared to a simple model, Bayes rule rewards models in proportion to how much they predicted the data that occurred. Assuming that equal prior probabilities were assigned to the two models, the more complex model does not predict data sets in a given region as strongly as the simpler model that fits well. Rather, the simpler model is the more probably model, with higher posterior probabilities.

sophistication, their findings raise the possibility that simpler algorithmic decision aids may have the potential to help decision-makers as much as complex ones.

As many decisions in managerial contexts often involve high uncertainty and redundancy, algorithmic sophistication may not provide substantial improvements in prediction. For example, resource allocation decisions are made regularly but with uncertainty about its impacts on performance, especially in the context of dynamically changing environmental factors. Similarly, integrating some data into the process as simple (though data-driven) rules of thumb may provide similar returns compared to additional technical sophistication.

Thus, we propose that despite the increasing interest in technical investment, there may be limited returns to algorithmic sophistication for prediction in managerial decision contexts.

2.3 The role of decision authority

A second mechanism is that decision-makers may use their decision authority to make a decision that does not leverage potential prediction gains from algorithms. Most organizations today provide algorithmic recommendations as decision aids for managers who make the ultimate decision. The common reasoning for this is that managers have private knowledge to inform the decision and correct any problematic algorithmic recommendations, however rare.

One reason why decision-makers' ultimate decisions may not leverage potential prediction gains from algorithms is that decision-makers in organizations often balance multiple objectives to make their decisions (e.g. Obloj & Sengul, 2020). Even if algorithms provide better predictions, there may be other objectives to consider in making the decision that constrain their choices. Many decisions in organizations are complex with much richer objective functions than what can be captured by most algorithms (Ludwig & Mullainathan, 2021), and decision-makers have contextual knowledge on these various objectives and the relative weights placed across them, which may vary dynamically (e.g. Gaba & Greve, 2019; Kim, 2021).

This may mean that even when algorithms provide gains in prediction, decision-makers are not able to take advantage of them to improve their decision, because doing so may make the

decision worse on other dimensions. Even when algorithms provide better predictions, we may thus not observe any changes in the ultimate decisions, because doing so may be detrimental to other dimensions valued by the organization. An alternative way of characterizing this issue is that current versions of algorithms may have misaligned objective functions from true organizational objectives, which managers with decision authority can help correct.

For example, in the context of resource allocation decisions for inspections, as examined in this paper, the primary objective is to identify businesses with the highest likelihood of violations. However, there are secondary objectives in place stated by the department that are clearly defined and measurable: reducing the costs associated with geographic travel distance, and targeting restaurants with more overdue inspections, as well as those that are more popular and committing more serious violations. While balancing these objectives may not translate into choosing businesses with the highest likelihood of violations as enabled by better prediction, it may ultimately be a better decision according to broader organizational objectives.

Another way in which the same outcome may manifest is that managers actively dissipate informational gains using their discretion: they try to inform the decision with their private knowledge, but in doing so make a worse decision across the various objectives than algorithms. In this case, it is not that the manager chooses to reject an algorithmic recommendation because it would be detrimental to other objectives. Rather, it is that managers use their decision authority to make a worse decision across objectives than the one recommended by algorithms.

It may be that this operates through some aversion to algorithms or to external advice. Growing research in psychology suggests that individuals—especially those with experience—are more likely to prefer their own forecasts over algorithmic predictions after seeing them err, even when algorithms are more accurate overall (Dietvorst et al., 2015; Logg et al., 2019).[4] While a part of this effect may be driven by a negative reaction to algorithms, as found by Tong et al

[4] There has also been work on the implications of algorithmic usage or improvements in operational contexts like inventory and supply chain management. However, these generally do not focus on the interaction of algorithms with human decision-makers, as algorithms in these contexts have generally been automated.

(2021) across employees who were informed that performance feedback was provided by an algorithm, other work proposes that this may also reflect a preference for one's opinions over others' advice (Logg et al., 2019). This is broadly consistent with extensive research on overconfidence (e.g. see Moore & Healy, 2008; Moore, Tenney, & Haran, 2015 for a review) and resistance to advice and change among professionals with specialized knowledge and strong norms (e.g. Kellogg, 2014; Greenwood et al., 2019). These insights suggest that even when algorithmic sophistication improves predictions, managers may not recognize their improvements and use their discretion to dissipate any potential informational gains.

More broadly, this latter channel would suggest that it may be important for organizations to more seriously consider how to manage and allocate decision authority as a key factor for capturing any potential gains from algorithms, in addition to technical sophistication.

2.4 Limited empirical insight from organizational contexts

While these mechanisms are important to evaluate to understand how organizations can productively use algorithms as decision aids, there has been limited empirical insight. Much prior work on algorithms and decision-making has examined individual decision-makers in the laboratory, classroom, or online settings (e.g. Dietvorst et al., 2015; Yeomans et al., 2019; Logg et al., 2019; Choudhury et al., 2020). This work has tended to compare predictions between humans and algorithms, or individual preferences for algorithmic predictions, in non-organizational contexts (a notable exception is Allen & Choudhury, 2021). While this work provides important insights on individual responses to algorithms, how the use of algorithms improves predictions and ultimate decisions in organizational contexts is less clearly addressed.

Furthermore, prior studies have not explored how decisionmakers in organizations leverage their contextual knowledge on organizational objectives to inform decisions when faced with algorithmic inputs.[5] This is striking, given that the default arrangement for working with

[5] Choudhury et al (2020) examine one source of private knowledge that also manifests in non-organizational contexts, individual domain expertise, which operates by enabling decision-makers to make better predictions, rather than how they make the ultimate decision. They provide evidence that graduate students examining patents who were provided with domain

algorithms in organizations is to allocate final decision authority to managers. Examining how decisionmakers in organizations use their contextual knowledge can thus provide valuable insights on the role of managing decision authority in capturing the returns to algorithms.

In this paper, we empirically examine the returns to algorithms on managerial decisions, focusing on these two mechanisms. We evaluate them in a real organizational setting by experimentally testing the returns to algorithmic sophistication on prediction and observing the extent to which any prediction gains translate into improved decisions.

3. Empirical Context

We explore the role of algorithms in decision-making within the Inspectional Services department of a major metropolitan city in the United States ("the City"), which provided a compelling context to study decision-making in administrative organizations (Simon, 1947).

The key decision we studied involved resource allocation, where inspectors used their judgment to decide which restaurants to inspect. The City employed approximately 20-30 inspectors at any given time, who were assigned to at least one of 22 wards or "neighborhoods" and rotated across wards every two years.[6] Inspections took approximately two to four hours, which limited the number of restaurants that could be inspected in a day. During this process, inspectors had a formal list of practices that they checked to evaluate the restaurant's compliance with food safety regulations. For example, they used thermometers to check temperatures of cold food storage areas and evaluated employee hygiene (e.g., use of gloves, thawing practices). Checking each of these across all areas of the establishment, recording notes, and discussing with management to fix any immediate small issues could take substantial time.

knowledge through advice from an experienced patent examiner were better able to identify applications that were strategically using new words and references to enhance the perceived novelty of their art—which algorithms were less able to identify based on the patent application text. This finding suggests that when machines are less able to make good predictions, individuals with domain expertise can correct them through their superior predictions.

[6] As there were 22 wards and no inspectors with more than 40 year of tenure, every inspector was similarly new to the ward that they were assigned to during the time of the experimental pilot. Larger wards were assigned to multiple inspectors, who subdivided the ward geographically.

Inspections yielded substantial variation in the violations found across restaurants. For example, inspections conducted between 2007 and 2015 uncovered 0 to 60 weighted violations per restaurant (Appendix Figure 1). The department assigned weights for violations based on their severity: Level I (1 point) corresponded to non-critical violations such as building defects or standing water. Level II (2 points) were critical violations more likely to create food contamination, illness, or environmental hazard. Level III (5 points) were considered "food-borne illness risk factor[s]" such as insufficient refrigeration or a lack of allergen advisories on menus. When critical violations were found in a restaurant, the City temporarily suspended its food permit if the violations were perceived as representing an imminent public health risk.

This context provided several research advantages. First, while the strategy of which restaurants to inspect may be complex, a key component involved predicting which ones will have violations, thereby raising the potential for algorithms to enhance decision-making. The main objective defined by the Head Inspector was to incapacitate establishments that posed the highest risk to public health.[7] Thus, decision quality depended on inspectors' ability to prioritize restaurants according to their likelihood of violation, flagging those that posed the greatest risk.

Second, the department had secondary objectives beyond this primary prediction that were clearly stated and measurable: reducing travel distance, targeting more overdue inspections, and prioritizing more serious violations and popular restaurants. The department wanted inspectors to prioritize their inspections accordingly whenever these secondary objectives could be improved without substantially reducing the targeting of restaurants with the highest number of violations. This provided clarity in interpreting inspector behavior and how they used their discretion: if they did not improve upon these objectives, their behavior could be interpreted as using their discretion to diminish the gains from algorithms rather than improve the decision. This meant that even if inspectors were prioritizing other potential objectives (e.g. chain status

[7] While there are additional possible objectives, such as deterring restaurants from committing violations or ensuring fairness in the inspection allocation, our discussions with this department highlighted the primary importance of identifying restaurants posing the highest risk to public health.

or restaurant size), this could be interpreted as not improving the ultimate decision according to the organization's stated objectives.

A third advantage was the accessibility of data to potentially improve these predictions, from both historical administrative data at the City as well as external data (e.g., from platforms such as Yelp, TripAdvisor, and Twitter). The data used in this implementation were similar to those used in the work of Glaeser et al. (2016), which found that algorithms could in principle help city governments to identify a much larger number of health code violations.

Fourth, inspectors possessed experience to inform their decisions, and were motivated to prioritize higher-risk restaurants. There was no direct monetary incentive based on the number of violations found. However, inspectors were instructed to do so, as these factored into promotions and had direct implications for their workload. Their inspection outcomes and customer complaints about unsanitary conditions were tracked, and any complaints triggered immediate inspections as they generally stemmed from restaurants with a high number of violations. This meant that inspectors could avoid uncompensated increases in their workload by prioritizing inspections with a greater likelihood of violation. It is worth noting that if there were large penalties or strong incentives based on the number of violations found, then this would limit the extent to which we could explore how decision-makers use their discretion.

Fifth, improving the targeting of inspections had a direct impact on organizational performance. Inspectors were responsible for inspecting all establishments in their ward at least twice a year. However, inspectors were time-constrained and only reached 40% as required. Thus, better prioritizing inspections could improve the allocation of inspectors' scarce time.

Together, these attributes provided a compelling setting to evaluate the returns to algorithms. These attributes were also shared by other inspectional departments across the US. Interviews of 55 departments (detailed in Section 6) revealed that most departments (67%) also ran behind their target number of inspections. 47% of departments also prioritized inspections

based on the likelihood of violations alone, while 27% prioritized the recency of inspection over likelihood of violation, and 30% prioritized both equally.

4. Empirical Design

Between February 1 and March 25, 2016, the City evaluated three methods to predict restaurant violations: (1) business-as-usual, (2) a "data-poor" algorithm, and (3) a "data-rich" algorithm. While we advised on the empirical design, the City made final design choices and executed the implementation.

The first method (business-as-usual) represented the status quo: relying on inspectors' own predictions to rank restaurants. The Head Inspector asked all inspectors to rank the restaurants in their ward in the order that they intended to inspect them, as a natural way to obtain rankings as inspectors had a clear mandate to prioritize restaurants with a higher predicted likelihood of violation.[8] The second method (a "data-poor" algorithm) used the average number of violations across historical inspections to rank restaurants in each ward from most to least likely to have violations. The third method (a "data-rich" algorithm) ranked restaurants using a random forest model trained on both historical violations and Yelp data—including Yelp reviews, ratings, price range, hours, services (e.g., reservations), business ambience (e.g., children-friendly), and neighborhood (details in Appendix A).[9] This method was one of the winning algorithms from a crowdsourced tournament across machine learning engineers. Although more sophisticated approaches may have yielded higher-quality insights, this algorithm used a comparatively more sophisticated model and richer data than the "data-poor" algorithm, emulating common practices by firms that invested in upgraded technologies and more complex data.

[8] This wording was chosen by the City as the most natural way to obtain inspector rankings, given that inspectors were mandated to prioritize restaurants with a higher likelihood of violation. There were also establishments with a required urgent priority to inspect, which were treated separately from regular inspections. These included high-risk establishments (e.g., hospitals and nursing homes), re-inspections, and restaurants flagged by complaints; these were excluded from the pilot and our analysis in order to assess how inspectors themselves prioritized restaurants.

[9] This method was one of the winning algorithms from a crowdsourced tournament across machine learning engineers, and provided theoretical efficiency gains of 40% relative to inspectors. Data scientists at the City maintained and ran the algorithm to generate rankings for this pilot. Appendix A provides further details.

Each inspector received a docket of restaurants to inspect in each period, which listed the top-ranked restaurants from each method in randomly sorted order.[10] The City determined the number of restaurants on each docket based on the number of restaurants that each inspector ranked for that period, which typically ranged from 15 to 21. The City's data team sourced equal numbers of the highest rankings from the other two methods, removed any duplicates, and randomly sorted all restaurants to create a docket.[11] This meant that rather than randomly assigning inspectors to different conditions, the experiment randomly sorted restaurants from the three methods on every inspector's docket to provide a mechanism that exogenously influenced which restaurants were inspected when, while also allowing us to observe how each inspector used their decision authority. These dockets were presented as a "new way of doing inspections". Inspectors were informed when they submitted their rankings that these would be supplemented by those prioritized using data processed by the City's data team, to help identify restaurants more likely to have violations. They were thus asked to not hold on to their own lists and to work down the docket using their judgment. This was a new approach, as inspectors generally did not plan out their work and adjusted which restaurants to visit across the day.

At the time of the pilot in 2016, this City was one of the early departments to use predictive algorithms for this purpose. The City's data team had successfully hired PhD-trained scientists with relevant experience in policy and technology. Based on our discussions with city personnel, the team seemed well respected within the City, and viewed as competent. Moreover, this change did not threaten inspectors' main task and expertise in conducting inspections, and was rather explained as a way to provide more support. Nonetheless, it is possible that inspectors would view any external guidance with skepticism.

In this design, because inspectors were asked to first rank their own choices, it was easier to understand what their counterfactual decisions would have been without algorithms. Moreover,

[10] Each inspection period covered approximately 2 weeks, and rankings were processed prior to the inspection periods.
[11] The City made this decision in order to include all restaurants inspectors had prioritized.

variation in the degree of algorithmic sophistication could shed light on how features of different algorithms may impact outcomes. Lastly, randomly ordering restaurants made it possible to identify whether algorithmic methods identified restaurants with more violations.

4.1 Data and Empirical Approach

The resulting data we observed was anonymized data on rankings and inspection results. However, several important implementation issues led to empirical challenges.

First, inspectors inspected substantially fewer restaurants than the 1,042 assigned on the dockets: only 243 were inspected, averaging 14 per inspector. This was due to a few reasons. One was that restaurant inspections were only one component of inspectors' jobs. They also had higher-priority inspections of hospitals, nursing homes, and schools and scheduled at required intervals. Inspectors had many of these scheduled during the pilot, which meant that almost half of the period was blocked and only the remaining weeks assigned to restaurant inspections. Second, two inspectors were sick and unable to undertake inspections for the full period. Lastly, the City's data team listed more restaurants than could be inspected on the docket. They wanted to include all inspector-ranked restaurants, and thus sourced an equal number of restaurants from the algorithm lists. They also wanted to ensure that inspectors did not run out of restaurants in case many were closed and there were no higher-priority inspections scheduled.

Second, the City modified the docket generation process for the last two periods. Dockets were filled with restaurants that had not been inspected from previous dockets, capped at 47, which made it possible that each docket no longer sourced an equal number of restaurants from each method if there was an imbalance in restaurants inspected across methods in prior weeks.

Third, rankings from all three methods were not available for all restaurants. Inspectors ranked only their highest-ranked restaurants in each period, so those that were listed on the dockets because they were ranked highly only by algorithmic methods did not have an inspector

ranking. Some restaurants ranked highly by inspectors also lacked rankings from algorithmic methods if there were no data from historical inspections or Yelp.[12]

To address these issues, we take the following steps. First, we focus our analysis on evaluating whether inspected restaurants ranked in the top 20 by algorithms have a higher number of violations than those ranked in the top 20 by inspectors. Restricting to this subsample ensures a more consistent availability of rankings, and allows us to compare inspection outcomes across comparable rankings under each method. Furthermore, since inspectors ranked their highest-priority restaurants, comparing the top 20-ranked restaurants provides insight into how the top-ranked restaurants under each method differ, and whether restaurants ranked highly by algorithms have a higher number of violations.

This subsample consists of 174 out of all 243 restaurants inspected, which represents a subset of 674 restaurants ranked in the top 20 by any method. Across the full set, we found substantial overlaps between methods, especially algorithms, with 176 restaurants (26%) ranked in the top 20 by at least two methods. 108 (16%) were ranked in the top 20 by the data-rich algorithm alone, 97 (14%) by the data-poor algorithm alone, and 293 (43%) by inspectors alone.

We assess the gains from using algorithms by examining the number of violations found across restaurants ranked in the top 20 by algorithmic methods compared to inspectors. We use the following model as our main specification for restaurant i inspected by inspector j:

$$Total\ Violations_i = \alpha + \beta DataRich_i + \gamma DataPoor_i + \sum_k^4 \delta_k MultipleMethods_{k,i} + \gamma_j + \epsilon_i \quad (1)$$

Here, α represents the mean number of weighted violations for restaurants ranked in the top 20 by inspectors; β and γ represent the mean expected difference in weighted violations for a restaurant ranked by the data-poor and data-rich algorithms relative to a restaurant ranked by inspectors, respectively; δ accounts for overlaps between methods and represents the mean

[12] We note that there was only one new restaurant that was new in our sample (i.e., not yet inspected by the time of the experiment). This restaurant was not highly ranked by the inspectors, but top-ranked by the data-rich algorithm. We were missing a ranking from the data-poor algorithm as the restaurant had no historical inspections.

expected difference in weighted violations for a restaurant ranked highly by multiple methods (i.e., inspector and each of the algorithms alone, both algorithms, or all methods). We estimate this model with and without inspector fixed effects (γ_j). We then explore the robustness of our results across the full sample of inspected restaurants, as well as across alternative subsamples, varying the threshold of the top 20. We also account for changes in the docket generation process by restricting our sample to the first two periods before the modification occurred.

5. Results

Our findings suggest large gains from predicting violations using algorithms: algorithms identified restaurants with over 50% more violations on average compared to those prioritized by inspectors. The largest gains appear to stem from integrating any data, rather than algorithmic sophistication. However, these predictions gains from algorithms did not seem to translate into improved decisions. Inspectors were only about two-thirds as likely to follow algorithmic recommendations compared to their own lists, dissipating the informational gains from algorithms. Furthermore, we find little supportive evidence that inspectors significantly improved the decision with respect to other organizational objectives such as reducing travel costs, inspecting more overdue restaurants, or targeting more serious violations.

5.1 The gains in prediction from using algorithms

[INSERT TABLE 1 and FIGURE 1 HERE]

Algorithms identified restaurants with more violations than those prioritized by inspectors. Table 1 Column 2 shows that restaurants ranked by inspectors alone had 7.2 violations on average, equivalent to a Level II and a Level III violation. Our estimates of the gains from algorithms over inspectors, β and γ, are 4.94 (p=0.001) and 5.17 (p=0.006) respectively, which represents a difference of targeting a restaurant with one more Level III violation. Figure 1 plots the kernel density estimates, which show that these differences emerge across the distribution.

We separate out restaurants that were ranked by more than one method. Restaurants ranked highly by both algorithms flag 6.8 more violations on average compared to inspectors alone, and

restaurants ranked by all three methods flag 6.3 more violations. Restaurants ranked highly by inspectors and one of the algorithms do not provide gains relative to inspectors alone.

In Column 3-4 of Table 1, we explore the difference between the two algorithms, examining only the subsample of restaurants ranked by each algorithm alone. The difference between the two algorithms (approximately 0.8 with p=0.79 in both specifications) suggests that the gains in our setting came from integrating any data into the process, rather than using more data or sophisticated algorithms. However, we interpret this with caution as we are underpowered to detect larger differences: the confidence interval of the difference between the algorithms ranges from -5.8 to 7.5, and we therefore cannot rule out large effects.

These results are robust across the full sample of inspected restaurants, as well as alternative subsamples that vary the threshold of top-ranked restaurants (Appendix Table 1). We also find consistent results across subsamples that restrict to the first one or two inspection periods prior to the modification in the docket generation process (Appendix Table 2).

Based on these results, we draw two conclusions. First, both algorithms outperformed inspector rankings on violations, and these performance improvements were on the order of over 50%. Second, the performance of the data-poor and data-rich algorithms was statistically indistinguishable, suggesting that the marginal benefit of additional data may be limited in this case. This is consistent with findings in similar applications to problems with representative datasets, especially when the scale of the dataset is smaller (Ng (2018))—although this result is particularly likely to be context-specific. This result suggests that in some cases algorithmic sophistication may not lead to substantially larger gains in prediction, and reinforces that simple heuristics can go a long way—but when driven by data, rather than human decision-makers.

A key consideration in interpreting the gains from algorithms relative to inspectors is what the inspector-ranked method represents. Inspectors were asked to rank the restaurants in the order they intended to inspect them, raising the possibility that inspector rankings may not reflect their predictions of violations. Because this wording was chosen by the City as the most

natural way to obtain inspector predictions given the clear mandate to prioritize restaurants with higher violations, we make the same assumption in our interpretations above.

However, if this assumption did not hold, and rather represented which restaurants inspectors planned to inspect, the interpretation of these results would change to reflect the gains from algorithmic predictions over inspectors' prioritization – which may have considered secondary organizational objectives beyond violations alone. While this would not affect our interpretation of the returns to algorithmic sophistication (i.e., the relative gains from the data-rich and data-poor algorithms), it raises the possibility that predictions gains from algorithms compared to inspectors may be smaller than our estimates. We explore this by examining whether inspector rankings performed better on secondary objectives relative to algorithms. We find that inspectors were more likely to rank more overdue restaurants compared to algorithms (Table 2 Panel A), and listed restaurants in closer proximity in order compared to algorithms (Figure 2). We do not observe that inspectors were substantially more likely to prioritize popular restaurants as proxied by Yelp reviews and ratings. This evidence raises the possibility that inspector rankings considered secondary organizational objectives beyond prediction alone, and that the estimated prediction gains from algorithms may be smaller than estimated.

[INSERT TABLE 2 and FIGURE 2 HERE]

Furthermore, while these results suggest that prior violations play an important role in predicting current violations, there may be important reasons why a city might not want to use them to guide inspection decisions. For example, if heterogeneity is driven by variation in inspector stringency rather than true variation in violations, as found by Macher et al. (2011) and Jin and Lee (2018), there may be concerns about relying heavily on past data. Moreover, as with any simple algorithm, using historical violations to guide decisions may facilitate strategic behavior that might lead to regulatory capture, eventually reducing the efficacy of this approach. To implement this in an ongoing basis, a city would need to think about the dynamic nature of inspections, which could be quite different from a temporary algorithm used to help with short-

run prioritization. Lastly, while predicting violations are part of the managerial problem, they are clearly not the whole problem. To the extent that inspections are meant to do more than rectify existing problems, it may be unwise to prioritize them solely based on such predictions.

5.2 Decision authority and the returns to algorithms

Despite the estimated prediction gains from algorithms compared to inspector rankings, these gains did not fully translate into improvements in decisions. Inspectors were less likely to inspect algorithmically-ranked restaurants compared to those that they themselves had ranked.

Table 3 shows that inspector-only ranked restaurants accounted for 52% of all inspected restaurants, whereas either of the algorithm-only categories each accounted for only 11-13% of all inspected restaurants. Mapping these to the numbers of top-20 ranked restaurants by each method as detailed in Section 3.1, inspectors were only two-thirds as likely to inspect restaurants based on the algorithms relative to their own rankings. They inspected 31% of the 293 restaurants that they alone ranked in the top 20, but only 18% and 24% of the 108 and 97 restaurants that the data-rich and data-poor algorithms alone ranked (20% overall across both) .

[INSERT TABLE 3 HERE]

We observe some heterogeneity across inspectors. Figure 3 plots the percentage of restaurants inspected by each inspector, the red line indicating where this would have ended if the inspector had fully followed the dockets.[13] While there is substantial heterogeneity across inspectors in the extent to which they deviated from the algorithm, this figure shows that most (72%) inspected more restaurants that they prioritized compared to those ranked by algorithms. This suggests that algorithms may provide limited improvements for managerial decisions in some contexts, as managers may use their discretion to dissipate any informational gains.

[INSERT FIGURE 3 HERE]

[13] As shown in Table 3, the absolute share of restaurants inspected that were ranked by inspectors alone does not provide an accurate breakdown of the extent to which inspectors prioritized their own restaurants. Because there are restaurants ranked by multiple methods, and the number of restaurants inspected vary substantially across inspectors, what identifies whether inspectors over-prioritized their own rankings is indicated by whether the dark gray bar exceeds the red line.

Robustness in main results While this highlights an important challenge for organizations in capturing gains from algorithms in practice, it also poses a potential threat to our results, because we observe inspection results for a subset of restaurants—which raises the concern that inspectors may have selected algorithm-ranked restaurants with higher likelihoods of violation. The performance differences we observe across methods could then be driven by a selection effect of not observing outcomes for restaurants ranked lower by algorithms.

We explore this concern in two ways. First, we test whether inspectors chose higher-ranked restaurants on the algorithm lists, and whether the gains from algorithms stem from the top of the rankings. Second, we use inspection records up to spring 2022 to obtain inspection results on all restaurants in the sample inspected after the pilot.

We first test for differences in average ranking by method for inspected restaurants, excluding any that were ranked by multiple methods (41 out of 174 restaurants). If inspectors cherry-picked higher-ranked restaurants on algorithmic lists, then the average ranking of restaurants on algorithmic lists should be smaller than those on the inspector-generated lists.

[INSERT TABLE 4 and TABLE 5 HERE]

The point estimates in Table 4 (Column 1) suggest that there is a slight bias in the opposite direction, with restaurants ranked by inspectors alone occupying higher ranking positions compared to those by the algorithms, although differences are small ($\beta = 1.07$) and statistically insignificant (p=0.433). This suggests that the results are unlikely to be driven by observing different parts of the ranking distribution for each method. We also find little evidence that the gains from algorithms emerge from a particular part of the ranking distribution. In Table 4 (Column 2), we explore whether the gains from algorithms vary across rank. We find that coefficients on interactions with rank are fairly small (0.28 and 0.19 for data-rich and data-poor algorithms, respectively) and statistically insignificant (p=0.418 and p=0.562, respectively).

We further explore this selection issue by obtaining data on restaurant inspections since the pilot up to spring 2022, which covers nearly 85% of all restaurants—allowing us to more directly

evaluate whether our results are driven by selection. This analysis provides consistent results with our findings, with gains of 2.3 to 3.3 more violations flagged by the data-rich algorithm and 4.3 to 5.4 by the data-poor algorithm, compared to 7.2 to 8.4 violations flagged by inspectors alone (Table 5). One issue is that restaurant inspections occurring after the pilot period may estimate what the unobserved earlier outcome would have been with some error, due to the passage of time. The passage of time is unlikely to differentially affect restaurants ranked by methods, so we expect any bias to be on average downward due to measurement error.

In context of our broader findings, these results suggest that the prediction gains from algorithms over inspector rankings are unlikely to be fully explained by selection alone. First, while there may be selection in the restaurants that inspectors chose to inspect, they do not appear to have chosen substantially more violation-prone restaurants from the algorithm lists compared to their own. This suggests that inspectors may not have been making sophisticated tradeoffs, and makes it difficult to construct a clear alternative explanation driven by selection. Second, the magnitude of the differences we observe between algorithms and inspectors is quite large, and does not differ significantly across rankings. Hence it seems unlikely that selection would change these results directionally.

How inspectors use decision authority We explore whether inspectors used their decision authority to improve the decision in other respects. Even if they used their discretion to dissipate gains from algorithmic prediction, they may have done so to balance secondary organizational objectives as discussed in prior sections. We thus examine each objective in turn: reducing travel costs, targeting more overdue inspections, placing more weight on higher level violations, and prioritizing more popular restaurants that may pose a larger risk to public safety.

We do not find supportive evidence that inspectors ultimately substantially improved upon these secondary organizational objectives, rather finding that inspectors often overrode algorithms to make decisions that worsened these objectives. We compare the distance inspectors traveled to the next restaurant with the distance from the closest algorithm-ranked

restaurant that they did not inspect. We find the latter to be a subset of the first—suggesting that inspectors often had an algorithmically-ranked restaurant in closer proximity (a median distance of 0.1 versus 0.6 miles) than the next restaurant they traveled to (Figure 4), as well as the closest inspector-ranked restaurant that they did not inspect (Appendix Figure 2).

[INSERT FIGURE 4 and FIGURE 5HERE]

We also find little evidence that inspectors placed significantly higher weight on severe violations, overdue inspections, or popular restaurants in their inspection decisions. Algorithm-ranked restaurants had more violations in all three risk levels compared to inspectors, although differences are marginally significant (Table 2 Panel B).We also do not find economically or statistically significant differences in the days overdue or the number of Yelp reviews and ratings between inspected and non-inspected restaurants on average (Table 2 Panel C).[14] This helps us bound our estimates of prediction gains regarding the concern of what inspector rankings may represent (as discussed in Section 5.1): if inspectors were indeed as good as algorithms in predicting violations, and the estimated prediction gains were simply arising from their consideration of secondary objectives, then we would not expect algorithms to be able to improve upon these objectives as these findings suggest.

Nevertheless, we find some suggestive evidence that inspectors sought to improve upon secondary objectives, even if they did not ultimately do so on average. Figure 5 shows evidence consistent with the interpretation that inspectors were sensitive to how overdue restaurants were. However, the plots also highlight that 23% of the algorithm-ranked restaurants overlooked by inspectors (59 out of 254 restaurants) had more than the average 271 days elapsed across inspected restaurants, with 90% of these (53 out of 59) having more than the average 303 days elapsed across restaurants ranked by inspectors. In comparison, 65% of restaurants (59 out of 91) ranked and inspected by inspectors had fewer than 271 days elapsed

[14] We note that Yelp ratings and review numbers are an imperfect measure of restaurant popularity, so to the extent that they are a poor proxy, it is possible that inspectors improved upon this objective more than we can observe empirically, though on average non-inspected restaurants had a slightly higher number of reviews.

since the last inspection, with 68% (62 out of 91) at fewer than 303 days. Given that algorithm-ranked restaurants were on average closer and flagged more serious violations, these results raise the question of why inspectors deviated from algorithmic recommendations in these cases.

While we cannot fully empirically pin down this mechanism, we find some suggestive evidence that inspectors deviated from algorithmic predictions due to their own priors. Discussions with the department indicated that inspectors viewed certain restaurant features as being correlated with violations, such as chains, seafood restaurants, and older, more lower-end businesses—which may have helped them make decisions prior to using algorithms and driven how they applied their judgment. We find some suggestive evidence consistent with this interpretation. Table 2 Panel A shows that inspectors were more likely to prioritize older businesses in their rankings relative to algorithms, as well as chains and seafood restaurants to a lesser extent. Inspectors also placed higher priority on businesses with lower prices and reservations or table service offerings, although some of these differences are small, making it difficult to draw any clear conclusions.

Other potential explanations appear less likely to explain the results, and are broadly consistent with the interpretation that inspectors partly used their discretion to dissipate gains from using algorithms. For example, one potential alternative explanation is algorithm aversion, which has been found to play a role in some settings (e.g., Dietvorst et al., 2015). However, in this case, the department chose to not explicitly communicate that these recommendations were driven by algorithms to reduce the likelihood of triggering algorithm aversion. Rather, the implementation only communicated that they supplemented inspectors' lists with restaurants prioritized using data. This meant that the use of algorithms in this setting only added restaurants to inspect on their dockets. This implementation approach, however, raises another potential concern that inspectors may have perceived the city's data team as not competent to provide reliable information, leading them to not follow the dockets. While we cannot fully rule out this explanation, we do not observe supportive evidence. Our understanding is that the

City's data team was highly regarded both within and outside the department at the time, and if inspectors perceived the data team to be incompetent, we would expect that they would have been more likely to stick to their original lists. However, none of the inspectors stuck to their original rankings in whole or even loosely in order, suggesting that this may be less likely.

Another possibility is that inspectors may have deviated from algorithmic recommendations due to regulatory capture, reducing inspections of owners with whom they had social relationships. However, this appears unlikely, as inspectors were assigned to a different ward every two years and often did not meet the target of inspecting restaurants twice a year—providing them with little opportunity to build relationships.

It is also possible that inspectors were prioritizing other potential objectives than those defined by the organization, such as prioritizing personal preferences or objectives. For example, it is possible that inspectors believed that deviating from their own rankings would be perceived as a lack of competence, or that it would lead the organization to maintain this change, which they were against. It is also possible that while the docket was presented as a guide to apply their judgment, it was perceived as too ambitious a goal to complete, demotivating them as a result. While we cannot fully rule out these alternatives, they appear to be less consistent with the contextual details and the evidence we observe, as dockets were presented as a guide rather than a goal, and none of the inspectors held on to their lists. These potential explanations are also broadly consistent with our interpretation that inspectors did not use their discretion to improve the decision according to the department's stated objectives.

Together, our analysis suggests that an important consideration for organizations in using algorithms as decision aids may be managing decision authority. In principle, it may not be clear how decision-makers can apply their judgment to enhance the decision, and simple rules-of-thumb that supported decision-making in the past may become an impediment when using discretion. This is consistent with evidence found by Hoffman et al. (2018), as well as broader evidence on the challenges of managing professional workforces with specialized knowledge and

strong norms who resist advice (e.g. Logg et al., 2019; Kellogg, 2014; Greenwood et al., 2019). Our findings are also consistent with lab evidence that given statistical forecasts, participants may not always sufficiently update their beliefs, and this behavior can persist even after participants are informed that their predictions are far less accurate than the forecasts (Lim & O'Connor, 1995; Goodwin & Fildes, 1999; Avan et al., 2019). As theorized by Athey et al. (2020), how to allocate decision authority to decision-makers may depend on various factors relating to the organizational context, and the value of discretion may be highly dynamic if decision-makers become more likely to rely on algorithms as they observe their performance.

6. SURVEY EVIDENCE FROM INSPECTIONAL DEPARTMENTS ACROSS THE U.S.

Our findings suggest a clear managerial implication: organizations should be careful not to over-invest in algorithmic sophistication, and decision authority may deserve a deeper consideration in addition to technical investment. However, this evidence stems from a single context, which raises questions on how generalizable these findings may be and the extent to which this may be an important issue for organizations more broadly. While inspectional departments are part of many organizations across both government agencies and companies, our findings may be limited to the particularities of the department that ran the pilot.

To explore the extent to which these findings might generalize, we contacted 176 inspectional departments covering the largest 200 metropolitan areas[15] in the U.S. to conduct interviews on how they approach restaurant inspections and their thoughts on algorithmic sophistication and inspector discretion. We reached 55 departments that cover 45 counties (392 cities, towns, and other territories)[16] for interviews that lasted up to one hour on (1) how they prioritize their inspections (2) whether they have used data to prioritize inspections and details on how or why not; and (3) how important they considered inspector discretion to be and why

[15] Departments vary in whether they cover a city, county, or parts of counties.
[16] Some departments were organized at the county level that covers more than one city. We conducted interviews with any department that we reached who were willing to be interviewed. Only one department that we reached refused to be interviewed.

or why not (Appendix B). These interviews were conducted between August 2021 and February 2022, allowing us to document insights from more departments that had attempted using data and algorithms to guide their inspections relative to 2016 when we ran the pilot.

These interviews provided two key insights. First, although simple data can provide large returns, many departments did not use it because they believed they would need more data or technical capability. Second, most departments saw retaining managerial decision authority as crucial, suggesting that our findings may have wider practical implications beyond our pilot.

Nine of the 55 departments reported using some algorithmic rule or software to guide inspections, with three having run pilots using rich external data, and one having run multiple pilots leveraging machine learning algorithms using data from Google and Twitter. However, many departments that had attempted using more sophisticated solutions reported eventually having abandoned those approaches for a simpler model.

A key barrier mentioned by those who had not used data to guide their decisions was the lack of data and technical capability. Although historical inspections data were available for all departments, most believed that they would need far more data as well as technical talent to be able to improve their decisions—echoing similar survey responses from C-level executives who list data availability as their greatest challenge for using AI (CognitiveScale, 2021). However, most departments (67%) severely ran behind their inspection targets, suggesting that they may have benefitted from using historical data to improve targeting like our pilot city.

We also found that most departments placed high value on allocating full decision authority to inspectors. All departments that we interviewed gave inspectors ultimate discretion in prioritizing inspections, with 69% of departments rating inspector discretion as being very important (4 or 5 on a scale of 1-5). Departments that had used algorithms to guide inspections had also all provided inspectors with ultimate decision authority. The reasoning behind this generally fell into the two categories we explored in our pilot. The most common reasoning was that inspectors possessed deep knowledge of businesses that would enable them to better

predict violations, as we explored in the first part of our empirical analysis. One manager explained, *"Inspectors have the most information about the food establishments that they are going to inspect. They know which ones tend to do well on inspections and which ones tend to do poorly."* Another corroborated that inspectors had *"training and experience"* that provided them with "*first-hand knowledge of businesses, and especially the frequent violators*". Another manager elaborated on this with an example:

"There are things that inspectors as humans can ascertain better than an algorithm….like for instance restaurants near a baseball stadium may need more inspections closer to baseball season because that is when they're more busy and they're more likely to fall behind on ensuring that they're following health procedures. This is the kind of information that a software might not take into account but that human judgment can."

The second most common reason was that inspectors had organizational knowledge on multiple objectives and how to balance them. Most frequently mentioned was travel costs based on geographic distance, as we explored in the second part of our analysis. One department explained: *"It's important that [inspectors] are not driving across the county to do inspections."* Another emphasized, "*It doesn't make sense to just go to high-violating restaurants. Inspectors should pick high-violating restaurants in one area."* One department mentioned another objective that we examined in our analysis, the severity of the violation:

"A mom-and-pop restaurant that serves hamburgers probably serves 4-500 hamburgers a day compared to a fast food restaurant like McDonald's that serves 1,000 hamburgers in a day. So those temperature issues are not nearly as much of an issue in the big chain because [food] is not going to be sitting out there for more than 15-20 minutes at a time. Whereas at a mom-and-pop it may be there all day…so temperature issues and [how serious the violation is] becomes much more important [at the mom-and-pop]. There's your discretion."

Yet despite the value that most departments placed on providing inspectors with decision authority, those that attempted using data and algorithms to guide their decisions faced similar issues with discretion as our pilot city. One department elaborated that "*only about a third of their inspectors actually utilized the software [regularly]*". Another highlighted that *"inspectors do not really access [the data] that often."* In fact, a department that had used machine learning algorithms in collaboration with tech companies reported that 39% of 36 inspectors never used

the algorithmic recommendations. This variation in usage persisted across inspector tenure, although inspectors at either ends of tenure were less likely to use algorithmic recommendations, consistent with findings in Allen and Choudhury (2021). 4 of 7 inspectors (57%) with 0 to 3 years of experience working in the department reported using algorithmic recommendations; 6 of 8 (75%) among those with 3-6 years; 6 of 7 (86%) among those with 6-9 years; 1 of 1 (100%) among those with 9-12 years; and 5 of 10 (50%) among those with over 12 years. Among those who used them, only 4 (17%) used them on a weekly basis, most using them on monthly, quarterly, or when their assignments changed. When surveyed by the department, 44% of the inspectors reported feeling neutral about the usefulness of the algorithms, and 6% not useful. This low usage limited the gains from algorithmic recommendations, which the department also reported as identifying more violations compared to inspectors as in our pilot (e.g. an algorithm leveraging data from Twitter was 64% more effective compared to inspectors).

Together, these interviews provide insights consistent with our empirical findings, and highlight that while organizations commonly place high value in algorithmic sophistication relative to managing decision authority, the latter may merit a more serious consideration from organizations seeking to use algorithms as decision aids.

7. DISCUSSION AND CONCLUSION

In a world where organizations are increasingly investing in technologies to support decision-making, our findings speak to the potential as well as the challenges involved in implementing such approaches at scale. Our results indicate a clear role for algorithms in improving decisions, but also highlight the importance of managing decision authority for organizations to capture their value. In this case, even a simple algorithm based on internal historical data better prioritized restaurants relative to human prediction. Nevertheless, these prediction gains did not translate into better decisions, as inspectors chose to prioritize restaurants based on their own predictions, without ultimately improving the decision on other organizational objectives. While the City continued to explore the use of targeting for a few years

following this pilot, they ultimately discontinued the program and returned to their old system, which did not use any data to prioritize inspections. While comparing the relative gains from algorithmic sophistication and improving decision authority is beyond the scope of this study, our findings suggest that decision authority may merit a more serious consideration for many organizations, especially relative to their interest in algorithmic sophistication.

Our analysis has a number of limitations. First, our analysis takes the primary goal of the department as given, i.e., to prioritize based on the number of violations. In practice, there may be other goals that departments pursue. For example, if inspections deter future violations, then a department may want to change its approach to prioritization over time. Second, our analysis assumes that inspections accurately capture actual violations. To the extent that violations are inaccurate or biased, then predictions based on them would also be biased. Third, we examined one specific data set in one particular context. Other datasets or algorithms may be more productive than these approaches, and organizations need to carefully consider the quality of their data, and the noise and bias present. This decision context is characterized by moderate complexity, with higher costs for mistakes that make some degree of human supervision valuable, and our findings may be most generalizable to similar contexts. In settings with greater complexity and richer data, the benefits of algorithmic inputs to decision-making may be higher than those found here. Similarly, the compliance patterns we observe may not generalize to other settings with different communication and organizational dynamics.

Our results highlight the importance of carefully considering how decision authority is allocated and managed. However, the solution is rarely as simple as removing decision authority from human decision-makers. In many managerial contexts, removing humans from the decision process may involve substantial risks, and some degree of human supervision may remain necessary for edge cases. Furthermore, discretion may be important for other reasons beyond decision quality. For example, two departments we interviewed highlighted that

discretion may be important to maintain the well-being of inspectors, by providing flexibility to reduce "burnout" and improve job satisfaction.

More work remains to be done to further understand when and how organizations can effectively capture value from algorithms without removing managerial discretion. In addition to understanding how organizations can train decision-makers to better apply their private knowledge to improve decisions when using data, exploring how the decision process can be redesigned (e.g., Puranam, 2021) may provide a promising direction for future work. While organizations commonly default to providing decision-makers with algorithmic recommendations, other possibilities such as incorporating human preferences into algorithms may provide better options for decision-making in some contexts.

REFERENCES

Agrawal, A., Gans, J., and Goldfarb, A (2018). Prediction Machines: The Simple Economics of Artificial Intelligence. Harvard Business Review Press.

------ (2019). Artificial intelligence: the ambiguous labor market impact of automating prediction. Journal of Economic Perspectives, 33(2):31-50.

Allen, R., and P. Choudhury (2021). "Algorithm-Augmented Work and Domain Experience: The Countervailing Forces of Ability and Aversion." Organization Science, forthcoming.

Athey, S., Bryan, K., and Gans, J. (2020). The allocation of decision authority to human and artificial intelligence. AEA Papers & Proceedings 110:80-84.

Avan, M., Fahimnia, B., Reisi, M., Siemsen, E. (2019). Integrating human judgment into quantitative forecasting methods: A review. Omega 86:237-252.

Bajari, P., Chernozhukov, V., Hortaçsu, A., and Suzuki, J. (2019). The impact of big data on firm performance: An empirical investigation. AEA Papers & Proceedings 109: 33-37.

Bartel, A., Ichniowski, C., and Shaw, K. (2007). How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. Quarterly Journal of Economics 122(4):1721-1758.

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. Journal of Experimental Criminology, 13(2), 193-216.

Bingham, C., and Eisenhardt, K. (2011). Rational heuristics: the 'simple rules' that strategists learn from process experience. Strategic Management Journal 32(13):1437-1464.

Bloom, N., Sadun, R., and Van Reenen, J. (2012). Americans do IT better: US multinationals and the productivity miracle. American Economic Review 102(1):167-201.

Bresnahan, T., Brynjolfsson, E., and Hitt, L. (2002). Information technology, workplace organization and the demand for skilled labor: Firm-level evidence. Quarterly Journal of Economics 117(1):339-376.

Brynjolfsson, E., and McElheran, K. (2019). Data in action: data-driven decision making and predictive analytics in US manufacturing. Rotman School of Management Working Paper.

Brynjolfsson, E., Jin, W., and McElheran, K. (2021). The Power of Prediction. Working Paper.

Camuffo, A., Cordova, A., Gambardella, A., and Spina, C. (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. Management Science 66(2):564-586.

Choudhury, P., Starr, E., and Agarwal, R. (2020). Machine learning and human capital: Experimental evidence on productivity complementarities. Strategic Management Journal 41(8): 1381-1411.

CognitiveScale (2021). Uncovering the Drivers of Enterprise AI Adoption.

Cowgill, B. (2019). Bias and Productivity in Humans and Algorithms. Working Paper.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics?. In *Simple heuristics that make us smart* (pp. 97-118). Oxford University Press.

Dhami, M. K. (2003). Psychological models of professional decision making. Psychological Science, 14(2), 175-180.

Dawes, R. (1979). The Robust Beauty of Improper Linear Models in Decision Making. American Psychologist 34(7):571–582.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243(4899), 1668-1674.

Dietvorst, B., Simmons, J., and Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. Journal of Experimental Psychology: General 144(1):114–126.

Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. Strategic Management Journal, 42 (12), 2195– 2217.
Hoffman, M., Kahn, L., and Li, D. (2018). Discretion in hiring. Quarterly Journal of Economics 133(2):765-800.
Gaba, V., & Greve, H. R. (2019). Safe or profitable? The pursuit of conflicting goals. Organization Science, 30(4), 647-667.
Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. Topics in cognitive science, 1(1), 107-143.
Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. Annual Review of Psychology 62:451-482.
Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. Psychological review, 103(4), 650.
Gigerenzer, G., Todd, P., & the ABC Research Group. (1999). Simple heuristics that make us smart. New York, NY: Oxford University Press
Glaeser, E., Hillis, A., Kominers, S., and Luca, M. (2016). Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. AER Papers & Proceedings 106(5):114–118.
Goodwin, P., and Fildes, R. (1999). Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy? Journal of Behavioral Decision Making 12:37-53.
Greenwood, B., Agarwal, R., Agarwal, R, Gopal, A (2019) The role of individual and organizational expertise in the adoption of new practices. Organization Science 30(1):1526-5455.
Grove, W., and Meehl, P. (1996). Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy. Psychology, Public Policy, and Law, 2(2):293–323.
Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. Psychological assessment, 12(1), 19.
Jin, G. and Lee, J. (2018). A Tale of Repetition: Lessons from Florida Restaurant Inspections. Working Paper.
Kahneman, D., Rosenfield, A. M., Gandhi, L., and Blaser, T. (2016). NOISE: How to overcome the high, hidden cost of inconsistent decision making. Harvard Business Review 94(10):38-46.
Kellogg, K. (2014). Brokerage professions and implementing reform in an age of experts. American Sociological Review. 79(5):912–941.
Kim, H. (2020). The Value of Competitor Information: Evidence from a Field Experiment. Working Paper.
Kim, H. (2021). Multiple Goals and Learning. Working Paper.
Kleinberg, J., Lakkaraj, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. NBER Working Paper No.23180.
Lehman, S. (2014). Twitter helps Chicago find sources of food poisoning. Reuters Health.
Lim, J. and O'Connor, M. Judgmental adjustment of initial forecasts: its effectiveness and biases. Journal of Behavioral Decision Making 8:149-168.
Logg, J., Minson, J.A., and Moore, D.A. (2019). Algorithm Appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151:90-103.
Ludwig, J., & Mullainathan, S. (2021). Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System. Journal of Economic Perspectives, 35(4), 71-96.
Macher, J., Mayo, J., and Nickerson, J. (2011). Regulator Heterogeneity and Endogenous Efforts to Close the Information Asymmetry Gap. Journal of Law and Economics 54:25-54.
MacKay, D. J. (1992). Bayesian interpolation. *Neural computation*, *4*(3), 415-447.
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. Psychological review, 115(2), 502.
Moore, D.A., Tenney, E.R. and Haran, U. (2015). Overprecision in Judgment. In The Wiley Blackwell Handbook of Judgment and Decision Making (eds G. Keren and G. Wu).
Ng, A. (2018). Machine Learning Yearning: Technical Strategy for AI Engineers.
Obloj, T., & Sengul, M. (2020). What do multiple objectives really mean for performance? Empirical evidence from the French manufacturing sector. Strategic Management Journal, 41(13), 2518-2547.
Puranam, P. (2021). Human-AI Collaborative Decision-Making as an Organization Design Problem. Journal of Organization Design 10: 5–80.
Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192-210.
Ransbotham, Sam, Shervin Khodabandeh, Ronny Fehling, Burt Lafountain, and David Kiron (2019). "Winning with AI: Pioneers Combine Strategy, Organizational Behavior, and Technology." MIT Sloan Management Review, October 15, 2019.
Sull, D. N., & Eisenhardt, K. M. (2015). Simple rules: How to thrive in a complex world. Houghton Mifflin Harcourt.
Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. Strategic Management Journal, 42(9), 1600– 1631.
Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. Professional Psychology: Research and Practice, 40(5), 525.
Wübben, M., & Wangenheim, F. V. (2008). Instant customer base analysis: Managerial heuristics often "get it right". Journal of Marketing, 72(3), 82-93.
Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. Journal of Behavioral Decision Making, 32(4), 403-414.

TABLE 1: The Informational Gains from Algorithms

	Comparing All Methods			***Comparing Algorithms***	
	(1)	(2)		(3)	(4)
Outcome:	Total Violations	Total Violations		Total Violations	Total Violations
Data-rich Algorithm Only	3.96**	4.94***		0.86	0.79
	(1.65)	(1.48)		(3.23)	(2.99)
Data-poor Algorithm Only	4.91**	5.17***			
	(1.96)	(1.87)			
Both Algorithms	6.75***	6.81***			1.22
	(1.67)	(1.64)			(3.05)
Inspector + Data-rich Algorithm Only	0.19	0.18			
	(1.59)	(2.49)			
Inspector + Data-poor Algorithm Only	-0.45	-0.71			
	(1.46)	(1.49)			
All Methods	6.54***	6.29**			
	(2.26)	(2.92)			
Constant	7.31***	7.19***		11.40***	11.83***
	(0.60)	(0.59)		(2.12)	(2.10)
Inspector FE	No	Yes		Yes	Yes
R-squared	0.16	0.33		0.3	0.3
Observations	174	174		42	59
Including Ranking Up To:	20	20		20	20

Only restaurants ranked within the top 20 by any condition are included. Columns (1) and (2) compare all three methods across the full sample. Columns (3) and (4) restrict the sample to restaurants in the top 20 ranked by the algorithms, not the inspectors, to compare the difference between the two algorithmic approaches. Total violations is a weighted sum of one, two, and three star violations. *Data-rich Algorithm Only* and *Data-poor Algorithm Only* are binary variables indicating restaurants that were ranked in the top 20 by the data-rich algorithm or the data-poor algorithm only, respectively. *Both Algorithms* indicates restaurants ranked in the top 20 by both data-rich and data-poor algorithms, but not the inspectors. *All Methods* indicates restaurants ranked in the top 20 by all three methods.

TABLE 2: Characteristics of Ranked and Inspected Restaurants

Panel A: Restaurants Ranked in Top 20 by Each Method

	(1) Data-Rich Algorithm Only	*(2) Data-Poor Algorithm Only*	*(3) Inspector Only*	*p-value (1)=(3)*	*p-value (2)=(3)*
Chain	0	0.03	0.05	<0.001	0.28
Yelp Rating	3.14	2.6	2.97	0.33	0.17
Review Count	119.9	144.41	154.28	0.19	0.74
Seafood	0	0.05	0.06	0.004	0.66
Restaurant Age	1.69	3.18	7.27	0.003	0.05
Price Range	1.4	1.14	1.27	0.17	0.40
Accepts Reservations	0.27	0.22	0.21	0.20	0.84
Table Service	0.46	0.38	0.32	0.03	0.34
Days Since Last Inspection	190.52	246.57	302.58	<0.001	0.09
N	*108*	*97*	*293*		

PANEL B: Inspected Restaurants Ranked in Top 20 by Method

Level I Violation	5.58	6.17	4.2	0.10	0.10
Level II Violation	0.63	0.48	0.29	0.11	0.13
Level III Violation	1.47	1.7	0.85	0.21	0.02
N	*19*	*23*	*91*		

PANEL C: Inspected vs. Non-Inspected Restaurants in Top 20

	Inspected	*Not Inspected*	*p-value*
Chain	0.02	0.03	0.28
Yelp Rating	2.82	2.95	0.45
Review Count	123.04	138.05	0.55
Seafood	0.03	0.05	0.22
Restaurant Age	4.92	5.27	0.84
Price Range	1.14	1.33	0.04
Accepts Reservations	0.18	0.26	0.04
Table Service	0.34	0.41	0.05
Days Since Last Inspection	270.97	247.68	0.59
N	174	500	

Panel A compares the attributes of restaurants ranked in the top 20 by each method, excluding any restaurants ranked by multiple methods. Columns (1)-(3) show means and (4)-(5) display p-values of the difference between those columns. Panel B compares the number of violations by severity across inspected restaurants from each of the lists. Panel C compares the attributes of inspected and non-inspected restaurants among all top-20 ranked restaurants.

TABLE 3: Inspector Compliance

	(1)	(2)	(3)
	Number of Restaurants Inspected	%	% of Restaurants Inspected Out of All Top-20 Ranked Restaurants
Data-rich Algorithm Only	19	10.92	17.59
Data-poor Algorithm Only	23	13.22	23.71
Inspector Only	91	52.3	31.06
Inspector+Data-poor Algorithm	7	4.02	29.17
Inspector+Data-rich Algorithm	4	2.3	23.53
Both Algorithms	17	9.77	15.6
All Methods	13	7.47	50
Total	174	100	

This table shows the breakdown of inspected restaurants by ranking method. Column (1) and (2) respectively show the number of restaurants that were inspected in each category and the corresponding percentages. Column (3) shows the percentage of restaurants inspected out of all top-20 ranked restaurants in that category.

TABLE 4: Differences in Rankings and Performance across the Ranking Distribution

	(1)	(2)
Outcome:	Rank	Total Violations
Data-rich Algorithm Only	1.07	0.76
	(1.36)	(4.29)
Data-poor Algorithm Only	1.50	2.65
	(1.26)	(4.56)
Data-Rich Algorithm x Rank		0.28
		(0.35)
Data-Poor Algorithm x Rank		0.19
		(0.33)
Rank		-0.02
		(0.11)
Constant	10.24	7.49
	(0.57)	(1.26)
R-squared	0.01	0.1
Observations	133	133

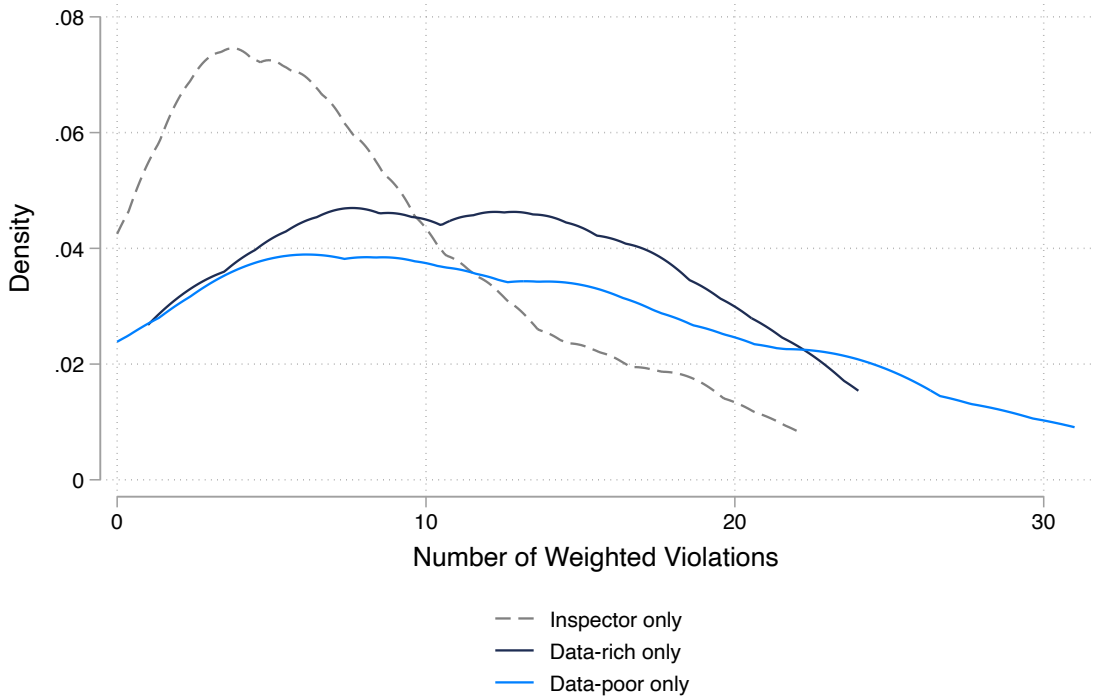
These regressions are run across the subsample of restaurants ranked in the top 20 by one of the methods alone, excluding any restaurants ranked by multiple methods. Column (1) analyzes differences in rankings across inspected restaurants, where *Rank* indicates the ranking position using the method that ranked the restaurant in the top 20. Column (2) analyzes whether the performance of algorithmic methods differs depending on the ranking position, where *Total violations* is a weighted sum of one, two, and three star violations.

TABLE 5: Robustness in Gains from Algorithms Across Inspections Data up to 2022

	(1)	(2)	(3)	(4)	(5)	(6)
Outcome:	Total Violations	Total Violations	Total Violations	Total Violations	Total Violations	Total Violations
Data-rich Algorithm Only	3.32***	2.61***	2.46***	2.31**	2.33***	2.40***
	(0.77)	(0.87)	(0.82)	(0.80)	(0.80)	(0.78)
Data-poor Algorithm Only	5.43***	4.70***	4.49***	4.28***	4.30***	4.34***
	(0.71)	(0.68)	(0.73)	(0.69)	(0.68)	(0.65)
Both Algorithms	4.85***	4.17***	3.97***	3.83***	3.88***	3.95***
	(1.01)	(1.05)	(1.09)	(1.01)	(1.00)	(0.96)
Inspector + Data-rich Algorithm Only	0.46	-0.09	-0.41	-0.61	-0.60	-0.55
	(1.64)	(1.66)	(1.45)	(1.52)	(1.55)	(1.56)
Inspector + Data-poor Algorithm Only	3.09*	2.55	2.33	2.24	2.30	2.42
	(1.77)	(1.74)	(1.71)	(1.62)	(1.61)	(1.62)
All Methods	4.82**	4.34**	4.14**	4.14**	4.21**	4.28***
	(1.68)	(1.58)	(1.62)	(1.49)	(1.47)	(1.46)
Constant	7.21***	7.90***	8.18***	8.38***	8.35***	8.27***
	(0.37)	(0.33)	(0.29)	(0.24)	(0.24)	(0.22)
Observations	586	688	762	817	843	875
R-squared	0.29	0.24	0.22	0.2	0.19	0.2
Inspector Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Including Ranking Up To:	20	25	30	35	40	All

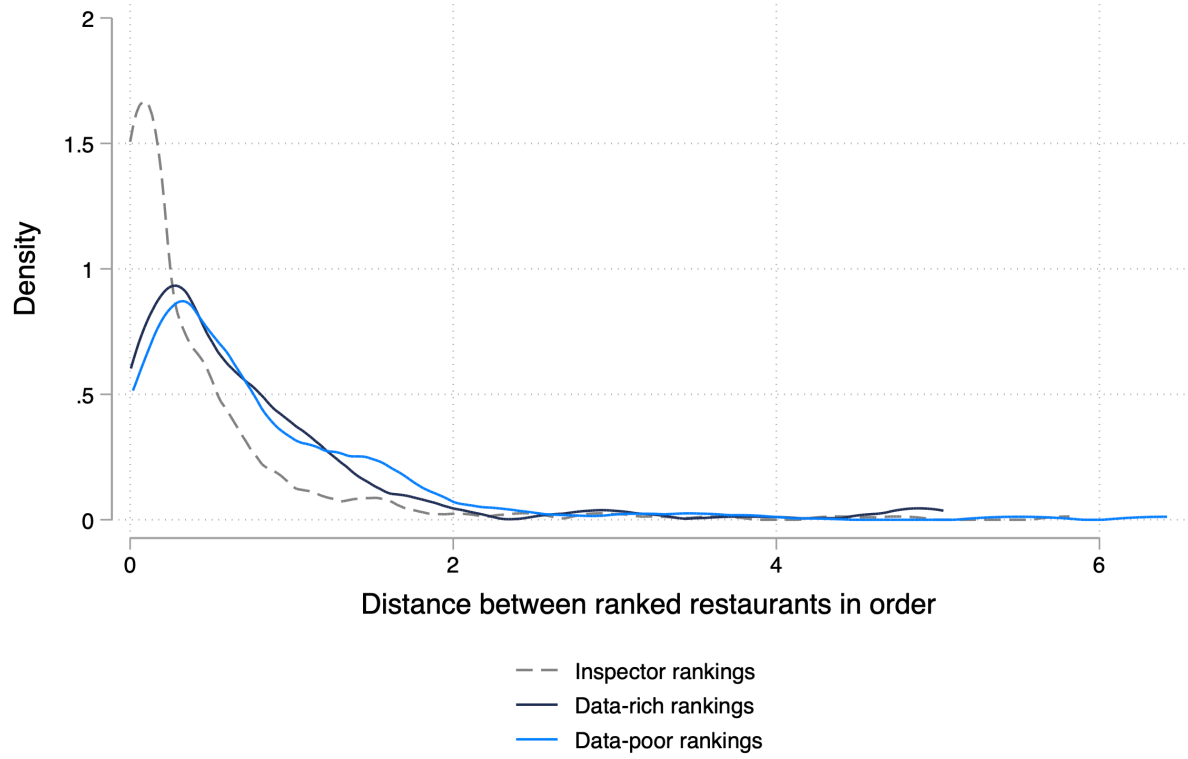
Each column shows the robustness of results across different sample restrictions in the full sample of available inspections data up to Spring 2022. Column (1) restricts the sample to restaurants ranked within the top 20 by any method, column (2) within the top 25, column (3) within the top 30, column (4) within the top 35, column (5) within the top 40, and column (6) across all inspected restaurants. *Total violations* is a weighted sum of level I, II, and III violations.

FIGURE 1: The Number of Violations Across Restaurants Prioritized by Each Method



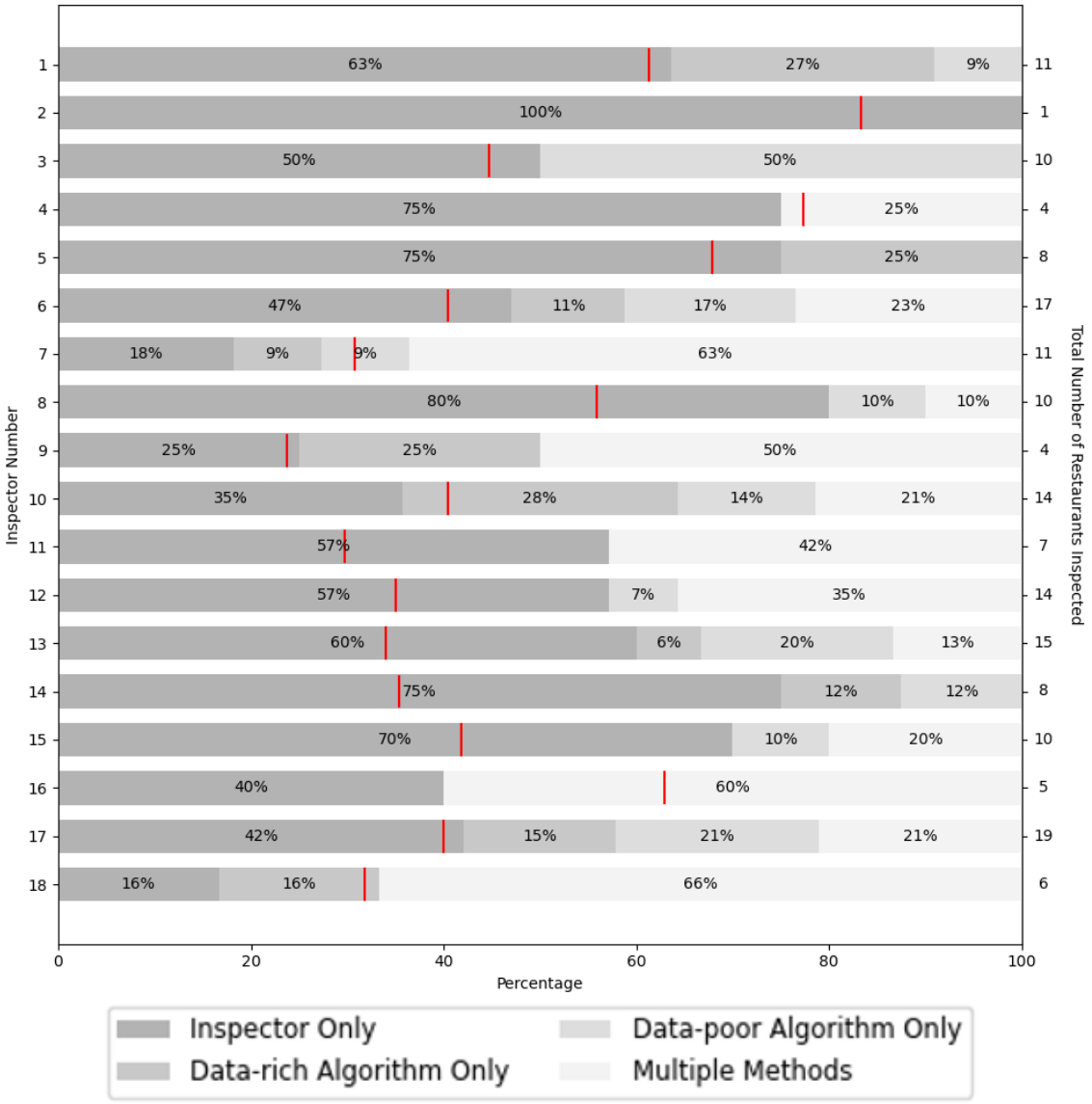
This figure shows kernel density plots of weighted violations found at restaurants ranked by inspectors alone compared to each of the algorithms alone.

FIGURE 2: Distance Between Ranked Restaurants in Order by Method



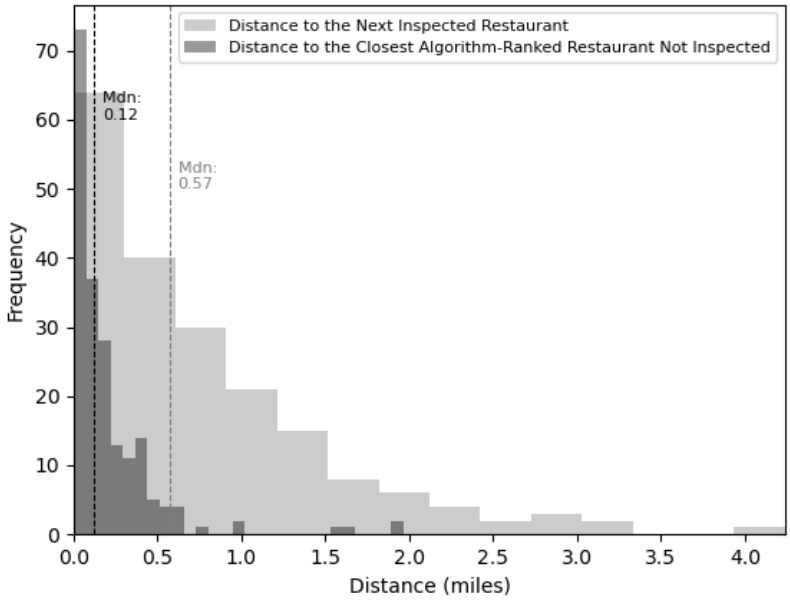
This figure shows kernel density plots of the distance between restaurants if travelled to in order based on inspector and algorithm rankings.

FIGURE 3: Percentage Inspected by Method across Inspectors



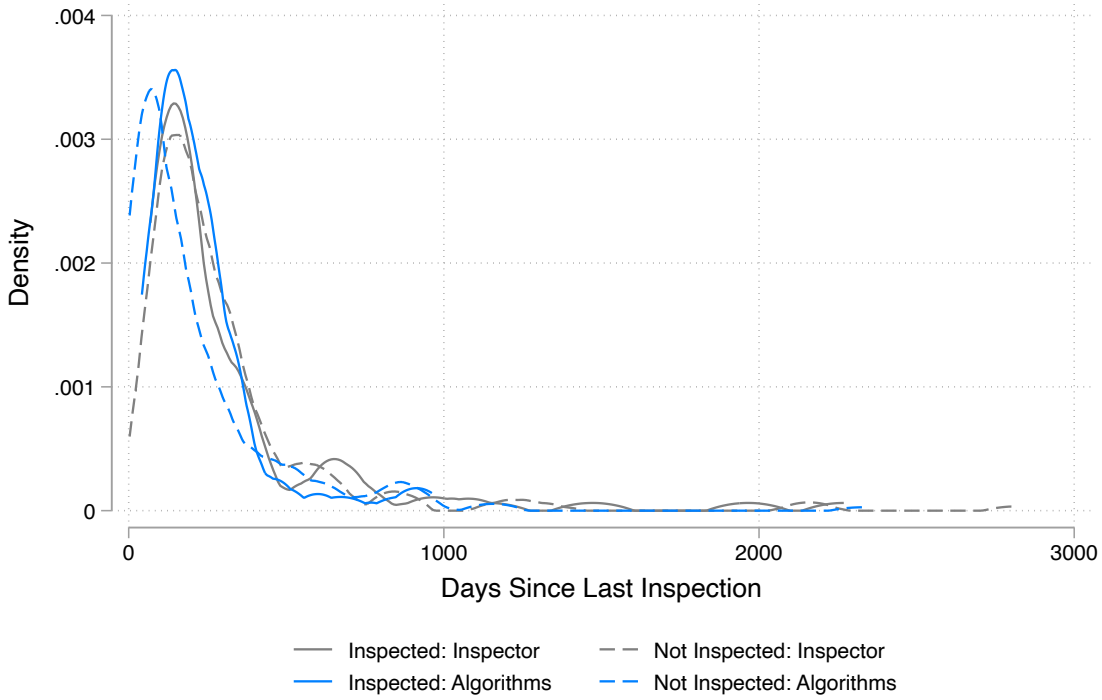
This figure plots the percentage of inspected and top-20 ranked restaurants by method for each inspector. Each bar represents a single inspector, where the left axis indicates the inspector, and the right axis shows the number of restaurants that the inspector inspected. The red line indicates the percentage of inspector-only ranked restaurants in the full sample of top 20-ranked restaurants, which is where the Inspector-Only bar (in dark grey) should have ended if inspectors had fully complied.

FIGURE 4: The Distance Inspectors Travelled vs. the Closest Algorithm-ranked Restaurant Not Inspected



This figure plots the distribution of the distance inspectors travelled to their next restaurant, compared with the distance to the closest algorithm-ranked restaurant on the docket that was not inspected.

FIGURE 5: Days Overdue by Method and Inspection Status



This figure plots the distribution of how overdue an inspection was across inspected and non-inspected restaurants by whether they were highly ranked by inspectors compared to algorithms alone.

APPENDIX A: Details on the "Data-rich" Algorithm

The "data-rich" algorithm was sourced from an open contest run across machine learning engineers, which awarded financial prizes for algorithms that most effectively used Yelp data to predict restaurant health violations. Designers of the winning algorithms received a minimum of USD $1,000, provided by Yelp, and over 700 people signed up for the tournament.

Participants were given access to a dataset recording more than 30,000 historical inspections data from the City (spanning approximately 9 years), and a linked data set of all available Yelp reviews, ratings, and business attributes for restaurants over the same time period. They were given three months to work on developing algorithms. Algorithm submissions were evaluated for their effectiveness in predicting inspection outcomes out-of-sample, based on restaurant inspections conducted over a period of two months after the close of submissions. Specifically, algorithm performance was measured by root mean squared logarithmic error (RMSLE).

The algorithm chosen by the City was one of the winning solutions, which used Python to implement a random forest model across the full data set. Features used were constructed from:

- All historical inspections data for each restaurant, which included the number of violations found across inspections, the type of violations found, notes from the inspection, the order of inspections (i.e. whether it was the first inspection for the restaurant or subsequent ones), the number of days since the prior inspection, the number of businesses within a 1.2km distance of the business and the number of inspections across neighbors, and the day of the most recent inspection
- All Yelp data, which included the text of every review available for the restaurant, the average star rating, the most recent star ratings, the date, total and average review length, counts of health-related terms in the review text (e.g. "unclean", "health"), attributes of the user who provided each review, and all available business attributes on Yelp for that restaurant (e.g. business category, business neighborhood, business ambience, business location, noise level, price range, whether it provides delivery services)

The random forest model was implemented using RandomForestRegressor in scikit-learn, and predicted the number of violations, log transformed. The parameters of the model were as follows: n_estimators = 1001, max_features, 600, n_jobs=8, oob_score=True. This model was adapted by data scientists at the City and re-run prior to the pilot.

APPENDIX B Key questions for interviews with departments

Hello, I am a researcher conducting research on how health departments prioritize and target food establishments for inspection. I would love to speak to the relevant person in your department. Could you transfer me?

Everything is anonymous, and no individual city responses will be presented.

1. How does your city currently target and prioritize inspections? What is your department's primary goal with carrying out inspections?

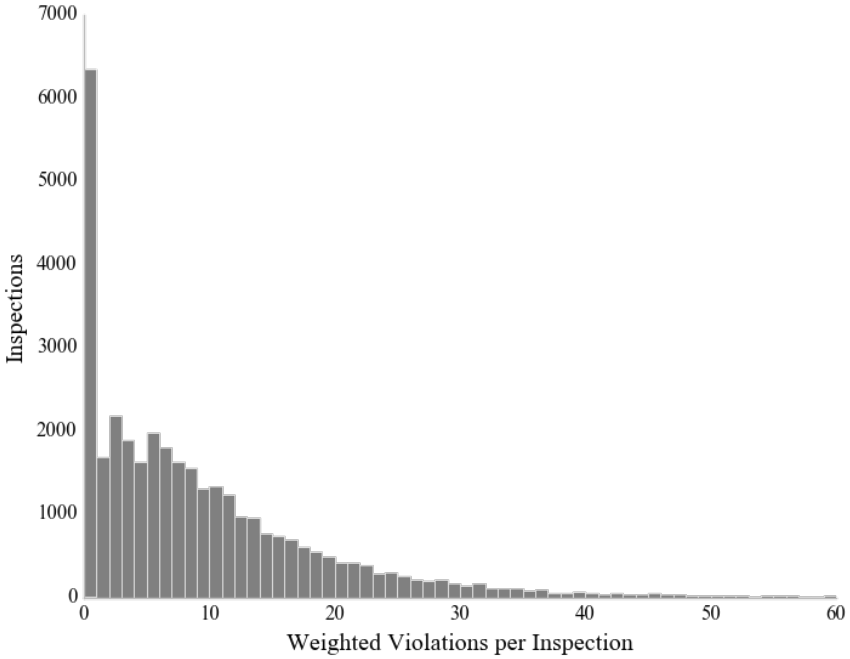
2. How often do you currently inspect restaurants, and are you able to meet that that goal?

3. When there are more restaurants to inspect relative to capacity, what formal guidance or rules do you have to determine how to prioritize them?

4. Have you thought about using data to target or prioritize inspections?
   (a) If no, why not?
   (b) If yes, ask for details on the implementation (e.g. the kind of data and algorithms involved, how it was implemented across inspectors, whether inspectors had decision authority, how successful it was, whether they are still using this)

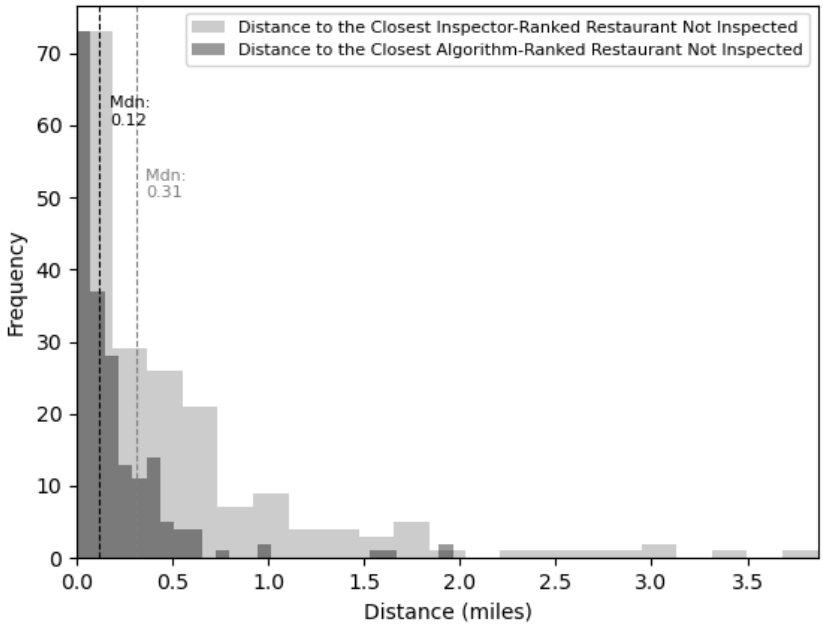
5. On a scale of 1-5, 5 being very important, how important do you think inspector discretion is when deciding what business to inspect on a given day?
   - If important or very important: why is it important for inspectors to have discretion?

APPENDIX FIGURE 1: Distribution of Violations



This figure shows the distribution of weighted violations across inspections from January 2007 through June 2015.

APPENDIX FIGURE 2: Distance to the Closest Algorithm-Ranked Restaurant Not Inspected vs. the Closest Inspector-Ranked Restaurant Not Inspected



This figure compares the distance to the closest inspector-ranked restaurant not inspected to the closest algorithm-ranked restaurant not inspected on inspector dockets.

APPENDIX TABLE 1: Robustness across sample restrictions

	(1)	(2)	(3)	(4)	(5)
Outcome:	Total Violations	Total Violations	Total Violations	Total Violations	Total Violations
Data-rich Algorithm Only	4.11	5.05**	3.92*	4.17*	3.72*
	(2.76)	(2.07)	(2.15)	(2.04)	(2.06)
Data-poor Algorithm Only	4.66	4.54**	4.31**	4.41**	4.13**
	(2.71)	(1.89)	(1.75)	(1.78)	(1.72)
Multiple Methods	5.62**	4.92**	3.94**	4.13**	3.69**
	(2.47)	(1.74)	(1.65)	(1.48)	(1.50)
Constant	7.12***	7.07***	8.01***	7.89***	8.23***
	(1.00)	(0.61)	(0.54)	(0.45)	(0.41)
Observations	101	138	201	222	243
Inspector Fixed Effect	Yes	Yes	Yes	Yes	Yes
Including Ranking Up To:	10	15	25	30	All

Each column shows the robustness of results across different sample restrictions. Column (1) restricts the sample to restaurants ranked within the top 10 by any method, column (2) within the top 15, column (3) within the top 25, column (4) within the top 30, and column (5) across all inspected restaurants. *Total violations* is a weighted sum of level I, II, and III violations.

APPENDIX TABLE 2: Robustness across time periods

	(1)	(2)
Outcome:	Total Violations	Total Violations
Data-rich Algorithm Only	2.83	4.11
	(1.81)	(2.11)
Data-poor Algorithm Only	4.88	4.78
	(2.05)	(2.07)
Multiple Methods	6.66	5.30
	(1.92)	(1.87)
Constant	7.17	7.50
	(0.61)	(0.63)
Observations	113	130
Inspector Fixed Effect	Yes	Yes
Including Periods Up To:	1	2

Only restaurants ranked within the top 20 by any method are included. *Total violations* are a weighted sum of level I, II, and III violations.

APPENDIX TABLE 3: Overlaps across methods by inspector

Inspector #	Data-Rich Algorithm Only(%)	Data-Poor Algorithm Only(%)	Inspector Only(%)	Inspector + Data-Poor(%)	Inspector +Data-Rich(%)	Both Algs (%)	All Methods (%)	Total (N)
1	25.81	6.45	61.29	0	3.23	3.23	0	31
2	8.33	8.33	83.33	0	0	0	0	24
3	13.16	31.58	44.74	2.63	5.26	2.63	0	38
4	0	9.09	77.27	4.55	9.09	0	0	22
5	28.57	0	67.86	3.57	0	0	0	28
6	7.14	11.9	40.48	0	4.76	33.33	2.38	42
7	17.95	10.26	30.77	7.69	0	20.51	12.82	39
8	11.76	20.59	55.88	0	0	8.82	2.94	34
9	18.42	10.53	23.68	13.16	5.26	18.42	10.53	38
10	17.02	14.89	40.43	2.13	0	25.53	0	47
11	10.81	8.11	29.73	8.11	5.41	27.03	10.81	37
12	15	15	35	0	0	20	15	40
13	21.28	17.02	34.04	6.38	2.13	19.15	0	47
14	25	22.92	35.42	4.17	2.08	10.42	0	48
15	11.63	11.63	41.86	0	0	30.23	4.65	43
16	3.7	11.11	62.96	3.7	7.41	11.11	0	27
17	20	17.78	40	2.22	2.22	17.78	0	45
18	20.45	18.18	31.82	4.55	2.27	15.91	6.82	44

This table shows the breakdown of restaurants ranked in the top-20 across methods by inspector.

APPENDIX TABLE 4: Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
(1) Number of Historical Inspections	1041	13.83	5.19	1	27
(2) Mean Historical Inspections	1041	8.40	3.75	0	25
(3) Days since Last Inspection	1041	280.36	366.50	3	4979
(4) Chain	1042	0.03	0.16	0	1
(5) Yelp Rating	671	2.93	1.39	0	5
(6) Review Count	671	144.97	228.73	0	1832
(7) Price Range	671	1.29	0.86	0	4
(8) Restaurant Age	671	5.01	14.40	0	140
(9) Table Service	671	0.40	0.49	0	1
(10) Seafood	671	0.05	0.22	0	1
(11) Highest Rank By Any Method	1042	17.48	12.40	1	61
(12) Inspection status	1042	0.23	0.42	0	1
(13) Numbers of Level-1 Violation	243	5.19	3.91	0	21
(14) Numbers of Level-2 Violation	243	0.40	0.70	0	5
(15) Numbers of Level-3 Violation	243	1.18	1.35	0	8
(16) Total Number of Violations	243	9.53	6.99	0	38

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(1)	1.00															
(2)	0.07	1.00														
(3)	-0.37	-0.06	1.00													
(4)	0.03	-0.13	-0.03	1.00												
(5)	-0.04	-0.10	-0.08	-0.13	1.00											
(6)	0.13	0.00	-0.09	-0.11	0.31	1.00										
(7)	0.00	-0.01	-0.08	-0.11	0.59	0.52	1.00									
(8)	0.12	-0.03	-0.01	-0.02	0.14	0.27	0.19	1.00								
(9)	0.04	0.07	-0.12	-0.17	0.38	0.49	0.63	0.11	1.00							
(10)	-0.03	-0.03	0.04	-0.05	0.06	0.07	0.07	0.08	0.10	1.00						
(11)	-0.15	-0.19	0.12	0.02	0.06	0.05	0.03	-0.02	0.01	0.02	1.00					
(12)	0.08	-0.07	-0.03	-0.02	-0.01	-0.02	-0.06	0.00	-0.06	-0.01	-0.11	1.00				
(13)	0.13	0.46	-0.08	-0.10	0.05	0.01	-0.03	0.09	0.10	-0.04	-0.11		1.00			
(14)	0.13	0.30	0.03	-0.08	-0.01	0.25	0.03	-0.01	0.18	0.04	-0.08		0.32	1.00		
(15)	0.01	0.19	-0.02	-0.04	0.08	0.07	0.02	-0.04	0.15	-0.09	0.08		0.30	0.22	1.00	
(16)	0.10	0.43	-0.05	-0.10	0.08	0.09	0.00	0.03	0.18	-0.07	-0.03		0.79	0.51	0.79	1.00