Decision Authority and the Returns to Algorithms

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

December 9, 2021

PRELIMINARY AND NOT FOR DISTRIBUTION

Abstract

We evaluate a pilot in an Inspections Department to test the returns to a pair of algorithms that varied in their sophistication. We find that both algorithms provided substantial prediction gains over inspectors, suggesting that even simple data may be helpful. However, these gains did not result in improved decisions. Inspectors used their decision authority to override algorithmic recommendations, without improving decisions based on other organizational objectives. Interviews with 29 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. Fabian Konig 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, and managers choose to use algorithmic recommendations as a decision aid, rather than a decision rule. Decision-makers may thus use their decision

[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).

authority to make a decision that does not leverage potential prediction gains from algorithms—either because of their private contextual knowledge on organizational objectives beyond the primary prediction, or because they end up dissipating any 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 in practice (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 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, and we find little supportive evidence that they improved the decision according to other organizational objectives—suggesting that they may have used their decision authority to dissipate 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 impact of algorithms on decision-making 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 at least partially predictive objective: identifying restaurants with health code

violations; (ii) while inspectors possess informative experience and insight, historical administrative records and external data may help to improve predictions (Lehman, 2014).

We compare three approaches implemented by the department 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 both 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 "data-poor" and "data-rich" algorithms.

We find substantial gains from predicting violations using algorithms: algorithmic methods identified restaurants with over 50 percent more violations compared to inspectors. Most gains came from simply integrating historical violations data, with the data-poor algorithm resulting in improvements nearly as large as those from the data-rich algorithm. Given the difficulty of generalizing from this specific context, the main insight we draw from this finding 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 did not translate into improved decisions. Inspectors were only half as likely to inspect algorithm-recommended restaurants relative to those that they had ranked highly, thereby dissipating much of the informational gains. We find little heterogeneity across inspectors in the extent to which they followed algorithmic recommendations. Given this behavior, we also explore the possibility that selection due to non-

[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.

compliance in our data could be driving the estimated gains from algorithmic inputs, but find little evidence that this can explain the full magnitude of the observed effects.

Furthermore, we find little supportive evidence that inspectors used their decision authority to improve the decision in other respects by leveraging their private knowledge on organizational objectives. We examine the extent to which inspectors took decisions that reduced travel costs, targeted more overdue inspections, and prioritized more serious violations or more popular restaurants, but find little supportive evidence that inspectors made economically or statistically significant improvements in any of these dimensions.

While our analysis cannot fully pin down the mechanism, we find some anecdotal and exploratory empirical evidence that inspectors rejected algorithmic recommendations when these conflicted with their predictions on restaurant attributes that drove violations. Our findings thus suggest 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, they are also broadly consistent with our broader 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 across the largest 100 cities in the U.S. to conduct unstructured interviews. We conducted interviews with 29 departments covering 37 cities that lasted up to one hour to understand their decisions on whether to use algorithms and how they thought about decision authority. We found that while a few departments have run pilots using algorithms, none of the departments continued to use them and reported having abandoned them for simpler approaches. Nonetheless, the primary barrier mentioned by departments that

had not used algorithms to guide their decisions was that they believed that they did not have sufficient data or technical sophistication, consistent with similar statements from C-level executives that data availability is their greatest challenge for using artificial intelligence (AI) (CognitiveScale, 2021). Furthermore, all departments—including those that used algorithms to guide their inspections—gave inspectors considerable decision authority in prioritizing inspections. 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 AI, estimated at over $40 billion USD in 2020 and projected to double in the next few years, our findings offer valuable practical implications. While 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 how algorithms impact decision-making in organizations. Growing studies have examined various implications of advancements in AI technologies on organizations, such as labor, employee management, and business performance (Felten, Raj, & Seamans, 2021; Choudhury et al., 2020; Tong, Jia, Luo, & Fang, 2021; Brynjolfsson et al., 2021). While much work 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 the default arrangement across many organizations 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.

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, both across research and practice (Gigerenzer et al., 1999). Yet despite growing investment into algorithmic sophistication, evidence suggests that many 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 computations used (see Gigerenzer & Gaissmaier, 2011 for a review)—suggesting that in some cases, more sophistication may in fact be harmful in 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 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

[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.

A second mechanism we propose 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 human managers to make the ultimate decision. The common reasoning behind this is so that decision-makers can use their private contextual knowledge to inform the decision or correct any problematic algorithmic recommendations, however rare.

One reason why decision-makers' ultimate decisions may not leverage any 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, it may be that there are other objectives to consider in making the decision that constrains 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 across the organization and the relative weights placed across them (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. As most algorithms today optimize for a single objective, decision-makers would need to balance predictions on this primary objective along with other secondary objectives, leveraging their private knowledge on the organizational context. 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, a key secondary priority is to reduce the costs associated in terms of the geographic distance traveled. The organization also places value on targeting restaurants with more overdue inspections, as well as those that are committing more egregious types of 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 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 (2021) across employees upon being informed that performance feedback was provided by an algorithm, recent work proposes that this may 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). While this research does not examine

[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.

ultimate decisions that are made within the context of organizational objectives, these insights suggest that even when algorithmic sophistication improves predictions, managers may not recognize their improvements and use their decision authority to dissipate any potential informational gains for the ultimate decision.

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 in realizing 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 human or algorithmic predictions, in non-organizational contexts (a notable exception is Allen & Choudhury, 2021). While this work provides important insights on how individuals respond to algorithms, how the use of algorithms improves predictions and ultimately translates into improvements in decisions in organizational contexts is less clearly addressed.

Furthermore, prior studies have not explored how decision-makers in organizations leverage their private contextual knowledge on organizational objectives to inform decisions when faced with algorithmic decision aids.[5] This is striking, given that the default arrangement across many organizations is to allocate final decision authority to managers when working with algorithmic

[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 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.

decision aids. How decisionmakers in organizations use their contextual knowledge can thus provide valuable insights on the extent to which how organizations manage decision authority may be a key determinant of 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 impact of algorithms and decision authority in 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, assigned to at least one of 22 wards or "neighborhoods", whose inspections uncovered from 0 to 60 weighted violations per restaurant between 2007 and 2015 (Appendix Figure 1). Weights were assigned based on the severity of the violation: 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.[6] Thus, decision quality depended on inspectors' ability to prioritize restaurants according to their likelihood of violation, flagging those that posed the greatest risk to public health as early as possible. While they were encouraged to conduct inspections in geographic proximity whenever possible, this was considered as secondary.

A second advantage was the accessibility of data to potentially improve these predictions – such as historical data held by the City and external data (e.g., from platforms such as Yelp, TripAdvisor, or 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.

Third, inspectors possessed experience to inform their decisions, and were motivated to prioritize higher-risk restaurants—providing a meaningful estimate for human predictions. Many inspectors had worked with the City for several years and developed expertise. They were assigned to a particular ward for two years, in order to balance learning about restaurants with avoiding the possibility of regulatory capture. Furthermore, as complaints about unsanitary conditions or illness required inspections within a specific time period, inspectors could avoid uncompensated increases in their workload by prioritizing inspections with a greater likelihood of violation. Their inspection targeting and quality could also play a role in career progression.

Fourth, 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% of them. Thus, better prioritizing inspections could improve the allocation of inspectors' scarce time.

[6] 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.

Together, these attributes provided a compelling setting to evaluate the returns to algorithms. Furthermore, we found that many of these attributes were broadly similar to other inspectional departments across the country. Surveying 29 departments covering 37 cities (details in Section 6), we found that the majority of departments (76%) were also running behind their target number of inspections, and that most (72%) also prioritized inspections based on the likelihood of violations. Four departments prioritized the recency of inspection over likelihood of violation, and four departments 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 the 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 were mandated to prioritize restaurants with a higher predicted likelihood of violation.[7] 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, Yelp ratings, price range, hours, services (e.g., reservations), business ambience (e.g., children-friendly), and

[7] 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.

neighborhood (details in Appendix A).[8] 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.

Each inspector received a docket of restaurants to inspect in each period, which listed the top-ranked restaurants from each of the methods in randomly sorted order.[9] The City determined the number of restaurants to list 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.[10]

These dockets were presented as a "new way of doing inspections" to guide decisions, with inspectors asked to work down the sequence on the docket in each period. Inspectors were informed that the list of restaurants that they had ranked were supplemented with those prioritized using data that the City's data team had processed, which sought to identify which restaurants were more likely to have violations. They were not explicitly informed that these restaurants were algorithmically generated, so that the fact that recommendations came from an algorithm was not salient and thus unlikely to trigger algorithm aversion (Dietvorst et al., 2015).

The reason this was a new approach was that inspectors generally did not plan out their work in advance and flexibly adjusted which restaurants to visit across the day. They were thus

[8] 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.

[9] Each inspection period covered approximately 2 weeks, and rankings were processed prior to the inspection periods.

[10] The City made this decision in order to include all restaurants inspectors had prioritized.

unlikely to be attached to any given order, and were also aware when they submitted their lists that they would receive some mixture of those and additional ones.

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, the 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 in practice than the 1,042 assigned on the dockets. Only 361 were inspected, averaging 20 per inspector.

Second, the City modified the docket generation process for the last two periods, after observing that inspectors could not complete the dockets. Dockets were filled with restaurants that had not been inspected from previous dockets, capped at a maximum of 47, which made it possible that each docket no longer sourced an equal number of restaurants from each method if there had been 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.

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 280 out of all 361 restaurants inspected, and 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 the 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 those ranked by inspectors. We use the following model as our main specification for restaurant i:

$$Total\ Violations_i = \alpha + \beta DataRich_i + \gamma DataPoor_i + \delta MultipleMethods_i + \varepsilon_i \qquad (1)$$

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 expected difference in weighted violations for a restaurant ranked by multiple methods.

We then explore the robustness of our results across the full sample, as well as across alternative subsamples, varying the threshold of the top 20 across the top 10-30. We also account for changes in the docket-generation process by restricting our sample to the first two periods before the modification occurred. Lastly, given the selection in our data due to only a subset of restaurants being inspected, we evaluate how selection bias might impact our estimates of the gains from algorithms.

5. Results

We find 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 stem from integrating any data, rather than algorithmic sophistication. However, these informational gains from algorithms did not translate into improved decisions. Inspectors were half as likely to follow algorithmic recommendations compared to restaurants that they ranked themselves, 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, prioritizing more overdue inspections, or targeting more serious violations.

5.1 The gains in prediction from using algorithms

Algorithms identified restaurants with more violations than those prioritized by inspectors. Table 1 Column 1 shows that restaurants ranked by inspectors alone had 6.8 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 5.03 ($p<0.001$) and 4.88 ($p=0.003$) respectively, which represents a difference of targeting a restaurant with one more Level III violation.

[INSERT TABLE 1 HERE]

In Column 2 of Table 1, we explore a specification that accounts for restaurants ranked by inspectors that were also ranked by one of the algorithms. The constant term shows the mean number of weighted violations for both restaurants ranked by inspectors alone and those that overlapped with one of the algorithms. Accounting for these overlaps increases the average number of violations found at inspector-ranked restaurants to 7.4. When we separate out restaurants ranked by both algorithms and all three methods, the violations increase twofold.

The difference between the point estimates for the two algorithms in both specifications (a difference of 0.15 with $p=0.94$ in both cases) 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 -3.76 to 4.06, and we therefore cannot rule out large effects.

These results are robust across the full set 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 human predictions 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 these results 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 inspectors may not have been prioritizing restaurants with more violations. Because this wording was chosen by the City as the most natural way to obtain inspector rankings—given the clear mandate to prioritize restaurants with higher violations, we make the same assumption in our interpretations. As described in Section 2, inspectors were trained to prioritize restaurants by their likelihood of violations, and had some incentive to do so as any high-risk restaurants later flagged through complaints would increase their workload.

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. It is possible that with a limited budget, investing in how to conduct inspections might be more valuable than investing in targeting strategies. Furthermore, 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 put this into practice 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

Although algorithms provided better predictions than inspectors, these gains did not ultimately translate into improvements in decisions. Inspectors were less likely to inspect algorithmically-ranked restaurants compared to those that they themselves had ranked, thereby dissipating much of the informational gains.

Table 2 shows that inspector-only ranked restaurants accounted for 61% of all inspected restaurants, whereas either of the algorithm-only categories each accounted for only 10% of all inspected restaurants. Mapping these numbers to all top-20 ranked restaurants as detailed in Section 3.1, inspectors were only half as likely to inspect restaurants based on algorithms relative to their own predictions. They inspected 58% of the 293 restaurants that they alone ranked in the top 20, but only 27% and 29% of the 108 and 97 restaurants that the data-rich and data-poor algorithms alone ranked.

[INSERT TABLE 2 HERE]

We observe little heterogeneity across inspectors. Figure 1 plots the percentage of restaurants inspected by each inspector, the red line indicating what the percentage breakdown would have been if the inspector had followed the dockets. The plot shows that nearly all inspectors inspected more restaurants they prioritized compared to those ranked by algorithms. This suggests that algorithms may provide limited improvements for managerial decisions, as managers may use their decision authority to dissipate any informational gains.

[INSERT FIGURE 1 HERE]

Robustness in main results While this highlights an important challenge for organizations in realizing gains from algorithms in practice, it also poses a potential threat to our results, because we observe inspection results for only a subset of restaurants. In particular, it 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 primarily by a selection effect of not observing outcomes for restaurants ranked lower by algorithms, rather than a treatment effect.

We test this concern in Table 3 (Column 1) by looking for differences in average ranking by method for inspected restaurants, excluding any that were ranked by multiple methods. If inspectors inspected higher-ranked restaurants on algorithmic lists, then the average ranking of restaurants on algorithmic lists should be higher than those on the inspector-generated list.

[INSERT TABLE 3 HERE]

The point estimates suggest that there is a slight bias in the opposite direction, with restaurants ranked by inspectors alone occupying higher ranking positions compared to those ranked by the algorithms, although differences are small ($\beta = 1.29$) and statistically insignificant (p=0.231). This suggests that the results are unlikely to be driven by observing different parts of the ranking distribution for each method.

Furthermore, we find little evidence that the gains from algorithms emerge from a particular part of the ranking distribution. In Table 3 (Column 2), we explore whether the gains from

algorithms vary across rank. We find that the gains appear to be spread across the ranking distribution, as coefficients on interactions with rank are fairly small (0.04 and 0.14 for data-rich and data-poor algorithms, respectively) and statistically insignificant (p=0.844 and p=0.582, respectively).

In context of our broader findings, these results suggest that the performance differences observed between the algorithms and inspector prediction 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 have chosen substantially more violation-prone restaurants from the algorithmic approaches compared to their own list. 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. However, one key limitation to our analysis is that we do not know the cleanliness of the restaurants that inspectors did not visit. Although inspectors were not systematically prioritizing restaurants predicted to have the most violations, it remains possible that uninspected restaurants are much cleaner than the inspected ones, which would affect the magnitude of gains from algorithms one could expect in a higher-compliance world.

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 in prediction from using algorithms, they may have done so based on their private knowledge to balance other organizational objectives. We thus examine other possible organizational objectives in turn, such as reducing travel costs, targeting more overdue inspections, placing higher weight on more serious violations, and prioritizing more popular restaurants that may pose a larger risk to public safety.

We find little supportive evidence that inspectors made economically or statistically significant improvements in each of these dimensions. First, we compare the distance inspectors traveled to their next restaurant with the distance from the closest algorithm-ranked restaurant that they did not inspect. However, we find the latter to be a subset of the first—suggesting that inspectors often had an algorithmically-ranked restaurant in closer proximity than the next restaurant they traveled to (Figure 2).

[INSERT FIGURE 2 HERE]

Second, we explore whether inspectors may have been more sensitive to overdue inspections, by examining the number of days elapsed since the last inspection. Inspectors on average prioritized restaurants that had received their last inspection 181 days prior, relative to 165 and 153 days for data-poor and data-rich algorithms, respectively (Table 4 Panel A). However, these differences in the number of days elapsed across restaurants are not statistically significant ($p=0.39$ and $p=0.23$ relative to data-poor and data-rich algorithms). Furthermore, we find little differences in the number of days elapsed across inspected versus non-inspected restaurants—181 versus 172 days on average—suggesting that decisions may not have substantially improved the targeting of more overdue inspections (Table 4 Panel B).

[INSERT TABLE 4 HERE]

We also find little evidence that inspectors placed higher weight on high-risk violations or more popular restaurants compared to algorithms. Algorithms identified restaurants with slightly more violations in all three risk levels compared to inspectors, although differences are small and marginally significant. Similarly, we find little significant differences in the number of Yelp reviews and ratings between restaurants ranked highly by the inspectors relative to the algorithms, or between inspected and non-inspected restaurants.

These findings provide little supportive evidence that inspectors used their private knowledge to improve the decisions with respect to other objectives. They also raise the question of why inspectors deviated from algorithmic recommendations. When asked after the pilot by

individuals at the city, inspectors provided many reasons why they did not adhere to the list. The most common reason was that algorithmic recommendations did not take into account distance, and they tried their best to go to the closest ones while prioritizing those that they believed to have the highest likelihood of violations.

While we cannot fully empirically pin down the mechanism, we find some suggestive evidence that inspectors deviated from algorithmic predictions due to their own priors on prediction. Discussions with individuals involved with the implementation indicated that inspectors viewed certain restaurant features as being correlated with violations, such as chains, seafood restaurants, older, and more lower-end businesses—which may have helped them make decisions prior to using algorithms.

We find some suggestive evidence consistent with this interpretation. Inspectors appear to have been more likely to prioritize older businesses (an average of 7.3 years) relative to algorithms (an average of 1.7 or 3.2 years by data-rich and data-poor algorithms, respectively). Inspectors also appear to have placed slightly higher priority on chain businesses, seafood restaurants, and businesses with lower price ranges and likelihood of offering reservations or table service compared to algorithms. However, these differences are small, making it difficult to draw any clear conclusions.

Other potential explanations such as algorithm aversion or social relationships are less likely to explain the results in this context. While aversion to algorithms has been found to play a role in some settings (e.g., Dietvorst et al., 2015), 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 to improve targeting those with high violations. This meant that the use of algorithms in this setting only added restaurants to inspect on their dockets. Furthermore, the department indicated that inspectors reportedly

expressed enthusiasm at the prospect of using data to aid their targeting, making it less likely that algorithm aversion might explain the effects we observe.

Another possibility is that inspectors may have deviated from algorithmic recommendations due to regulatory capture, reducing inspections of restaurants of owners with whom they had social relationships. However, this appears unlikely in this setting, 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. Furthermore, inspectors were more likely to target businesses that had been in operation for a longer time, which runs counter to this explanation. Nevertheless, if inspector decisions were driven by such considerations, it broadly supports the finding that inspectors did not use their private knowledge to improve the decision.

While our analysis provides little conclusive evidence on mechanisms, it suggests that an important consideration for organizations in using algorithms as decision aids is how to manage and allocate decision authority. In principle, simple rules-of-thumb that supported decision-making in the past may become an impediment to decision-makers when using decision authority to leverage their private knowledge. This is consistent with evidence found by Hoffman et al. (2018), where managers who appear to hire against job test scores ended up with worse average hires—as well as broader evidence on the challenges of managing the professional workforce with specialized knowledge and strong norms who resist advice and organizational change (e.g. Logg et al., 2019; Kellogg, 2014; Greenwood et al., 2019). Our findings are also consistent with lab evidence that show that given statistical forecasts, participants may not always sufficiently update their beliefs, which has found that 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), whether to allocate decision authority to human decision-makers or algorithms depends on a number of factors, including how much private

information decision-makers have, how aligned their incentives are with the objective at hand, how biased they may be, and how their predictions perform compared to algorithms. Furthermore, the value of discretion may be highly dynamic, if decision-makers become more likely to rely on algorithms as they observe their performance.

However, our findings—and the extent to which decision-makers use their discretion to reject algorithmic recommendations—are specific to the organizational context and accompanying practices. One key choice variable may be how to communicate about the algorithm being implemented; further explanations about the algorithm and the motivation for using it may help decision-makers use their discretion, although prior work has found that individuals may be less likely to follow algorithmic recommendations when feedback is provided (Dietvorst et al., 2015). Clarity on organizational objectives and higher-powered incentives may also help improve the alignment of decision-makers.

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

Our findings have 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 experimental 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 inspectional departments across the largest 100 cities in the U.S. to conduct interviews on how they approach inspections and their thoughts on algorithmic sophistication and inspector discretion. We

reached 29 departments that covered 37 cities for interviews that lasted up to one hour.[11] While the interviews were unstructured, all featured the following open-ended questions: (1) how they prioritize their inspections (2) whether they have used data to prioritize inspections and further details on how or why not; and (3) whether they considered inspector discretion to be important and why or why not (see Appendix B for more details).

These interviews provided two key insights: first, that while using even simple data can provide large returns, many departments do not use them because they believe that more data or sophistication is needed; and second, that most departments see managerial decision authority as being important to pair with using algorithms—suggesting that our findings may have wider practical implications beyond our pilot department.

First, seven of the 29 departments reported using some data-driven algorithms to guide inspections, with two having run pilots using rich external data with sophisticated analytics. One of the departments had run multiple pilots, one using data from Google and another using data from Twitter. However, all departments that had attempted using more sophisticated solutions reported eventually having abandoned those approaches for a simpler model.

However, the primary barrier mentioned by those who did not use algorithms to guide their decisions was that they believed that they did not have sufficient data or technical sophistication, which would require additional resources. Although historical inspections data were available for all departments, most departments believed that they would need more data and more technical complexity to be able to improve their decisions—echoing similar survey responses among C-level executives who listed data availability as their greatest challenge for using AI (CognitiveScale, 2021). However, the majority (76%) of departments were severely running behind their inspection targets, suggesting that they may have benefitted from using historical data to improve targeting, much like our pilot city.

[11] 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.

Second, we 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. Moreover, over 70% of the departments rated 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 decision authority on how to use them.

The reasoning behind this generally fell into the two categories we explored in the experiment. The most common reasoning was that inspectors possessed deep knowledge of businesses that would enable them to better predict which one might be more important to prioritize, 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 the following 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 raised reason was that inspectors had contextual organizational knowledge on multiple objectives and how to balance them. The most frequently mentioned factor 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. This is called 'clustering.' "* One of the departments mentioned another possible objective that we also examined in our analysis, the gravity of the violation committed:

"A mom-and-pop restaurant that serves hamburgers probably serves 400 hamburgers a day compared to a fast food restaurant like McDonald's that serves 4,000 hamburgers in a day. Food will be sitting out for 15-20 minutes at a time at McDonald's, but at a mom-and-pop it may be there all day, so temperature issues and [how serious the violation is] becomes much more important. There's your discretion."

However, despite the value that most departments placed on allocating decision authority to inspectors, those that attempted using data and algorithms to guide their decisions faced issues with discretion similar to our pilot city. One of the departments 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."*

One of the departments using algorithmic recommendations 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 at the department 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 a monthly or quarterly basis or when their district assignment changed. When surveyed by the department, 44% of the inspectors reported feeling neutral about the usefulness of the algorithms, 35% reported that they found the tool useful, 12% extremely useful, and 6% not useful to have available. This low usage limited the gains from algorithmic recommendations, which the department reported as identifying more violations compared to inspectors (e.g. one of the algorithms they had used previously leveraging data from Twitter was shown to be 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 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 realize these improvements. Even a simple algorithm based on internal historical data can more accurately prioritize restaurants relative to human prediction. Yet in our context, most of the gains from algorithmic prediction came from simply integrating data into the decision process rather than from using a more sophisticated algorithm. Nevertheless, these improvements did not translate into better decisions, as inspectors frequently chose to prioritize restaurants based on their own predictions, without using their private knowledge on organizational objectives to improve the decision. 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 beyond the number of required inspections and inbound complaints about specific incidents.

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 than organizations currently allocate, especially compared to their focus on algorithmic sophistication.

Our analysis has a number of limitations. First, our analysis takes the goal of the department as given, i.e., to prioritize based solely on the severity of violations. In practice, there may be other goals that departments might want to incorporate, and a simple predictive algorithm may not be fully aligned with such objectives. More broadly, if inspections deter future violations,

then a department may want to change its approach to prioritization. To the extent that behavior changes over time (through deterrence or other mechanisms), the effect of implementing different targeting strategies could vary. 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 particular 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, suggesting that decision authority may be better allocated to algorithms. 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. In some cases, increasing investments in human capital (e.g. hiring more inspectors, or increasing training) might be more valuable at the margin. Furthermore, discretion may be important for other reasons beyond decision quality. For example, two of the departments we interviewed additionally highlighted that discretion may be important to maintain the well-being of inspectors, by providing flexibility that can reduce "burnout" and improve job satisfaction.

More work remains to be done to further understand when and how organizations can effectively implement algorithms for decision-making without removing managerial discretion.

In addition to understanding the organizational practices needed to help decision-makers learn to better inform their decisions 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: Theory and Evidence from Resume Screening. 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
	b/se	b/se		b/se	b/se
Data-rich Algorithm Only	5.03	4.43		0.15	0.15
	(1.24)	(1.26)		(1.96)	(1.96)
Data-poor Algorithm Only	4.88	4.28			
	(1.62)	(1.64)			
Multiple Methods	7.66				
	(1.31)				
Both Algorithms		7.46			3.18
		(1.20)			(1.92)
All Methods		8.37			
		(2.12)			
Constant	6.80	7.40		11.68	11.68
	(0.45)	(0.51)		(1.57)	(1.58)
Observations	280	280		57	78
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. *Multiple Methods* indicates restaurants that were ranked in the top 20 by at least two or all three methods. *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: Inspector Compliance

	(1)	(2)	(3)
	Number of Restaurants Inspected	(%)	% of Restaurants Inspected Out of All Top-20 Ranked Restaurants
Data-rich Algorithm Only	29	10.36	26.85
Data-poor Algorithm Only	28	10	28.87
Inspector Only	171	61.07	58.36
Multiple Lists	52	18.57	29.55
Total	280	100%	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 3: Differences in Rankings and Performance across the Ranking Distribution

	(1)	(2)
Outcome:	Rank	Total Violations
	b/se	b/se
Data-rich Algorithm Only	1.29	4.55
	(1.07)	(2.85)
Data-poor Algorithm Only	1.06	3.28
	(1.09)	(3.40)
Data-Rich Algorithm x Rank		0.04
		(0.23)
Data-Poor Algorithm x Rank		0.14
		(0.26)
Rank		-0.03
		(0.08)
Constant	10.44	7.15
	(0.41)	(0.94)
Observations	228	228

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.

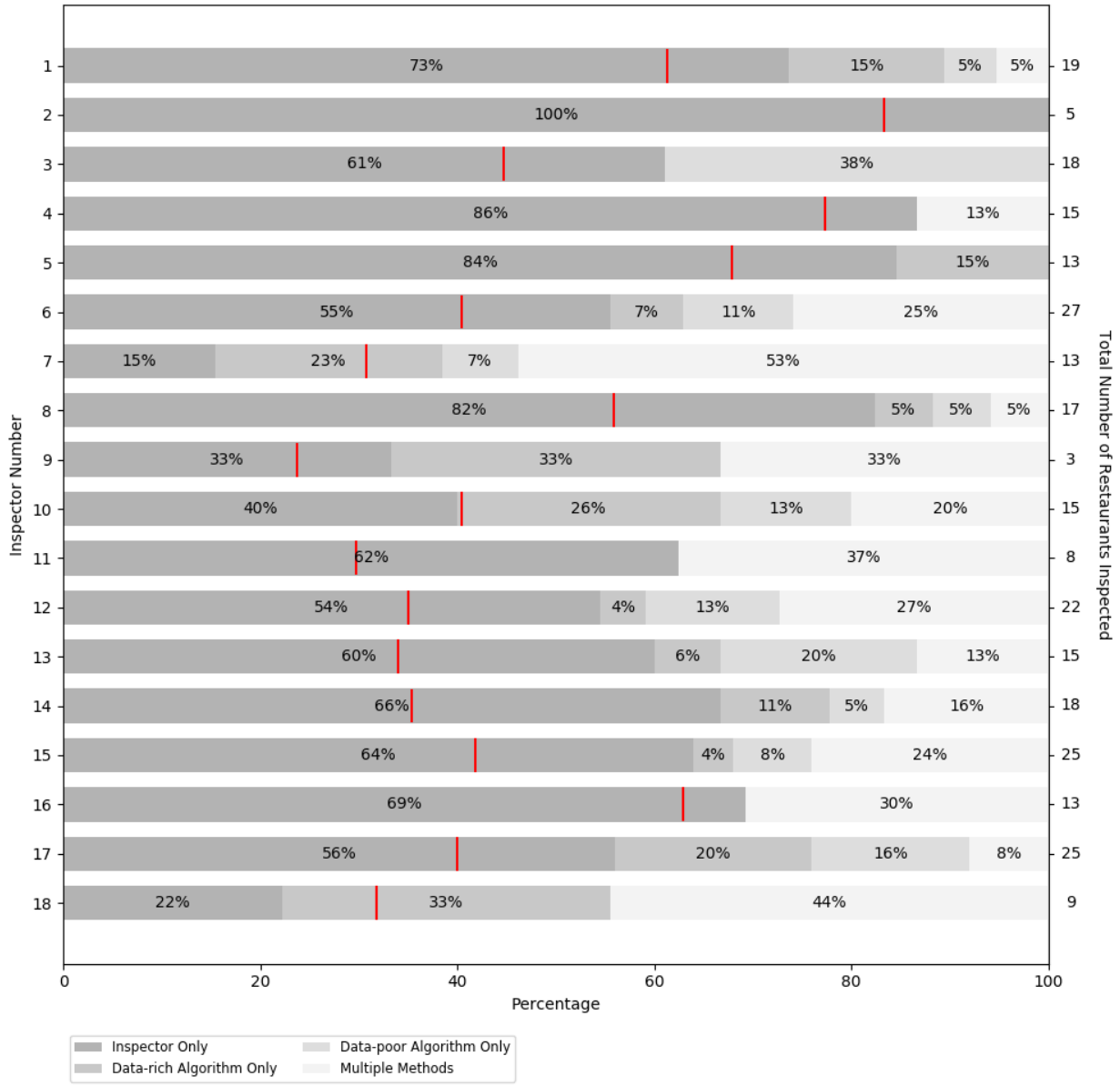
TABLE 4: Characteristics of Ranked and Inspected Restaurants

Panel A: Restaurant Ranked in the 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	0.04	0.1	<0.001	0.07
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	0.4
Accepts Reservations	0.27	0.22	0.21	0.2	0.84
Table Service	0.46	0.38	0.32	0.03	0.34
Days Since Last Inspection	153.43	164.91	181.21	0.23	0.39
Level I Violation	5.8	6.14	4.48	0.2	0.08
Level II Violation	0.45	0.46	0.3	0.33	0.1
Level III Violation	1.45	1.54	1.12	0.46	0.09

Panel B: Inspected vs. Non-Inspected Restaurants		
	Not Inspected	*Inspected*
Chain	0.05	0.06
	(0.01)	(0.02)
Yelp Rating	2.94	2.93
	(0.07)	(0.11)
Review Count	158.27	140.85
	(13.46)	(16.16)
Seafood	0.06	0.04
	(0.01)	(0.01)
Restaurant Age	4.34	5.47
	(0.58)	(1.17)
Price Range	1.31	1.23
	(0.05)	(0.06)
Accepts Reservations	0.28	0.16
	(0.02)	(0.03)
Table Service	0.42	0.32
	(0.03)	(0.03)
Days Since Last Inspection	172.46	180.74
	(9.08)	(6.68)

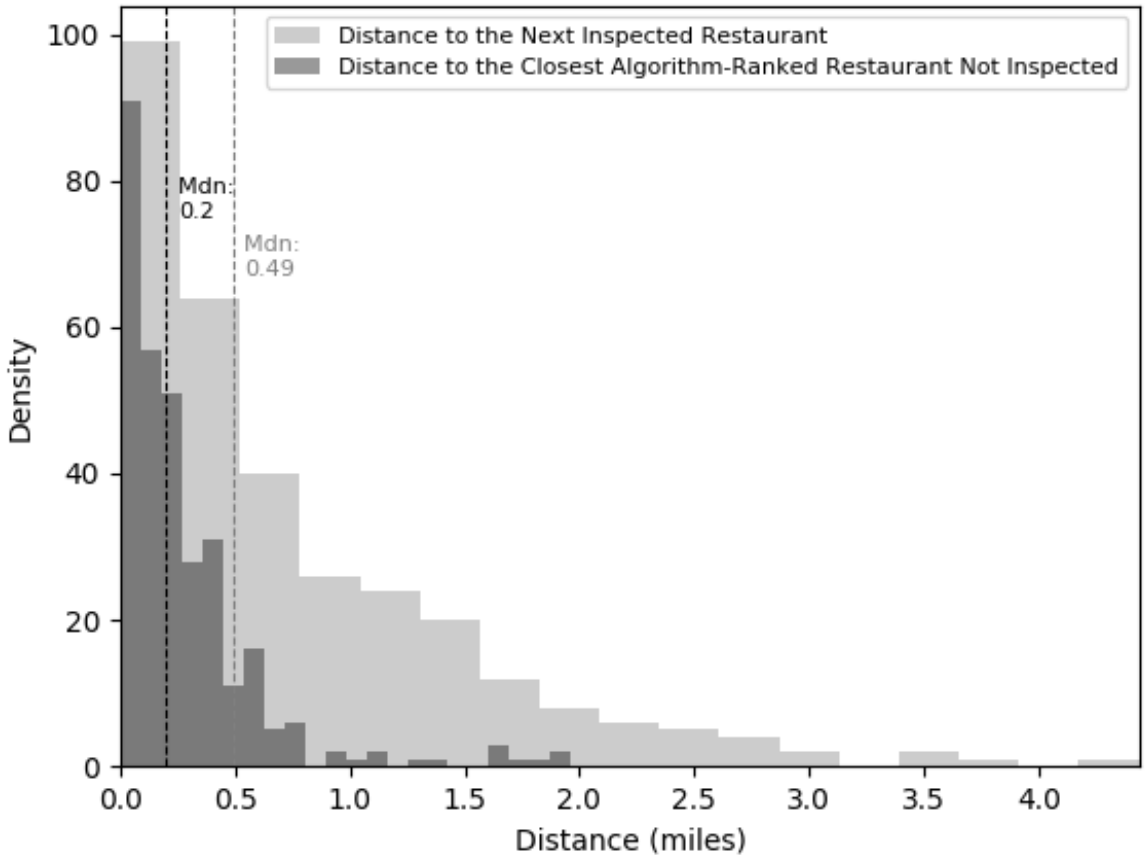
Panel A compares the attributes of restaurants ranked in the top 20 by each method, excluding any restaurants ranked by multiple methods. Columns (1)-(3) show means of each variable, and columns (4)-(5) display the p-value of the difference between restaurants ranked in the top 20 by the *Data-Rich Algorithm Only* and *Inspector Only*, and the *Data-Poor Algorithm Only* and *Inspector Only*, respectively, from a regression of the restaurant attribute on an indicator for being ranked by one of the algorithmic methods. Panel B compares the attributes of inspected and non-inspected restaurants.

FIGURE 1: Percentage Inspected by Method across Inspectors



This figure plots the percentage of inspected restaurants by ranked 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 2: 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.