Received: 7 December 2021 | Revised: 22 August 2023 | Accepted: 17 September 2023

DOI: 10.1002/smj.3569

RESEARCH ARTICLE

SMS | Strategic Management Journal

WILEY

Decision authority and the returns to algorithms

Hyunjin Kim[1] | Edward L. Glaeser[2] | Andrew Hillis[3] | Scott Duke Kominers[4] | Michael Luca[5]

[1]INSEAD, Fontainebleau, France

[2]Department of Economics, Harvard University, Cambridge, Massachusetts, USA

[3]Independent Scholar

[4]Entrepreneurial Management Unit, Harvard Business School, Boston, Massachusetts, USA

[5]Negotiation, Organizations, & Markets Unit, Harvard Business School, Boston, Massachusetts, USA

Correspondence
Hyunjin Kim, INSEAD, Bd de Constance, 77300 Fontainebleau, France.
Email: hyunjin.kim@insead.edu

Funding information
Alfred P. Sloan Foundation, Grant/Award Number: G-2021-16778; Ewing Marion Kauffman Foundation; Harvard Business School; Harvard Center of Mathematical Sciences and Applications; National Science Foundation, Grant/Award Numbers: CCF-1216095, DGE-1144152, SES-1459912, DMS-1928930; Radcliffe Exploratory Seminar Grant; The Harvard Milton Fund; The Ng Fund; The Rappaport Institute for Greater Boston; The Taubman Center for State and Local Government; Washington Center for Equitable Growth

Abstract

Research Summary: 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 often used their decision authority to override algorithmic recommendations, partly to consider other organizational objectives without improving outcomes. Interviews with 55 departments find that while some ran pilots seeking to prioritize inspections using data, all provided considerable decision authority to inspectors. These findings suggest that for algorithms to improve managerial decisions, organizations must consider both the returns to algorithms in the context and how decision authority is managed.

Managerial Summary: We evaluate a pilot in an Inspections Department to explore the returns to algorithms on decisions. We find that the greatest gains in this context come from integrating data into the decision process in the form of simple heuristics, rather than from increasing algorithmic sophistication or additional data. We also find that these improvements in prediction do not fully translate into improved decisions. Decision-makers were less likely to follow data-driven recommendations, partly in consideration of other organizational objectives, but without substantially improving on them overall. These findings suggest that organizations should consider the returns to technical sophistication in each context, and that the design and management of decision authority can be a key choice that impacts the value organizations can capture from using predictive analytics.

KEYWORDS

algorithms, decision-making, data-driven decision-making, future of work, organizational economics

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 et al., 2018; Cowgill, 2019; Choudhury et al., 2020). The potential for algorithms to improve prediction was demonstrated as early as the 1950s (e.g., Dawes, 1979; Grove & Meehl, 1996; Kahneman et al., 2016; Meehl, 1954), and firms today increasingly invest in data and algorithmic sophistication (Bajari et al., 2019; Brynjolfsson & McElheran, 2019). At the same time, the returns to such investments appear to be limited in many settings (Brynjolfsson et al., 2021; Ransbotham et al., 2019).

This raises a question on the extent to which investing in data and 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, 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 increased data and predictive analytics are limited, given the degree of uncertainty involved and the power of simple heuristics (e.g., Gigerenzer et al., 1999; 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, 2019; Agrawal et al., 2018; Choudhury et al., 2020; Raisch & Krakowski, 2021). 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 use their decision authority to make a decision that does not leverage potential prediction gains from algorithms—for instance, because they are balancing other organizational objectives beyond the predicted measure or have private information, 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 et al., 2020).

In this article, 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 rankings to algorithmic predictions, 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 historical data from within the organization and additional data from online platforms. We find that algorithms indeed appear to provide substantial gains over human rankings that may have considered broader organizational objectives. But the greatest gains come from integrating data into the decision process in the form of simple heuristics, rather than from increasing algorithmic sophistication or additional data. Moreover, these improvements in prediction do not fully translate into improved decisions. Decision-makers were less likely to follow data-driven recommendations, in part due to consideration of other organizational objectives, but overall without generally improving on these other dimensions—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 by using algorithms, understanding and managing decision authority is important.

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 at different points in time. This setting offers a number of compelling attributes to test the value of predictive algorithms: (i) identifying restaurants with health code violations is a key component of the role; (ii) inspectors have secondary organizational objectives to consider that are defined by the department and observable (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 predictions (Lehman, 2014); and (iv) inspectors retain ultimate decision authority and carry out the inspections themselves.

We compare three approaches to allocate inspectors: (1) “business-as-usual,” under which inspectors chose which restaurants to prioritize; (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 data from online ratings.[1] 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 2 weeks each. This design allows us to observe inspector rankings 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; indeed, algorithms identified restaurants with over 50% more violations than inspector rankings. Most gains came from integrating any data into the process, with the data-poor algorithm providing improvements as large as those from the data-rich algorithm. Given the difficulty of generalizing from this context, the main insight we draw is that even simple data was valuable in improving predictions, and there may be similar managerial contexts where this is the case.

However, the gains in prediction do not appear to have fully translated into improved decisions for the department. 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 gains from using algorithms. While inspectors varied in the extent to which they deviated from the algorithm, most inspected more restaurants that they prioritized versus those identified by algorithms. Given this behavior, we also explore the possibility that selection could

[1] 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 article. 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.

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 considering secondary organizational objectives, such as how overdue the inspection was. However, inspectors ultimately did not inspect restaurants that were substantially more overdue, did not identify more severe violations, and did not decrease travel time—suggesting that any improvements on secondary objectives from discretion were likely limited, at least based on the measurable outcomes.

While our analysis cannot fully pin down the mechanism driving inspectors' use of their decision authority in our context, we find some anecdotal and exploratory empirical evidence consistent with the interpretation that inspectors may have overridden algorithmic recommendations when those recommendations 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. Other potential explanations including algorithm aversion, private information, or social relationships with owners are also possible, although the department chose to not explicitly communicate that these recommendations were driven by algorithms, and inspectors were assigned to a different neighborhood every 2 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.[2]

To explore the broader context beyond this department, we contacted inspectional departments serving the largest 200 metropolitan areas in the United States to conduct unstructured interviews. We interviewed 55 departments for up to one hour to understand their processes and views around using algorithms and decision authority. We found that only a few departments had run pilots using sophisticated algorithms. Moreover, most believed that the historical data they had was not sufficient to provide useful predictions, or that they lacked the relevant 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 also reported dissipated informational gains from using algorithms.

Together, our findings suggest that managing decision authority is an important 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. Although measurement is challenging, organizations can evaluate the returns to algorithmic sophistication in each context and look for secondary measures to partially assess whether decision authority is being used to achieve other organizational goals.

[2] One important limitation of this analysis is that it takes the department's objective of identifying violations as given—from a policy perspective, this would assume that inspection decisions do not have a deterrence effect, that violations are a good measure of health risks, and that there are no other competing policy goals beyond the ones we were able to quantify. As these assumptions are likely to deviate in practice, one should not make direct policy prescriptions from this analysis. Rather, our focus is on understanding how the predictions were used.

This study contributes to emerging research on algorithms and decision-making in organizations. Studies have examined various organizational implications of advancements in AI (Brynjolfsson et al., 2021; Choudhury et al., 2020; Felten et al., 2021; Tong 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 et al., 2015; Logg et al., 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. Our findings highlight that decision authority can be a key organizational choice that impacts the value organizations can capture from using predictive analytics.

Our work is most closely related to that of Hoffman et al. (2018), who explore 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 similarly find that algorithmic recommendations improved performance on the measured primary outcome, and build on these findings in two respects. First, we explore the returns to algorithmic sophistication by comparing two algorithms with varying inputs, and find that algorithmic sophistication provided limited marginal returns at least in this context. Second, we directly explore how decision-makers use their discretion. 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 on the measured dimensions, suggesting that organizations need to better understand how decision-makers are using their discretion when using algorithms.

In addition to the literature on algorithms and decision-making, our analysis contributes to research on information technology investments and digitalization more broadly. A growing body of work has identified organizational practices that shape the returns to investments in information technology (Bresnahan et al., 2002; Bartel et al., 2007; Bloom et al., 2012; Brynjolfsson et al., 2021). Our findings point to organizational design challenges in deploying information technologies in practice. Our study highlights that the design and management of decision authority can be an important choice, in addition to technical sophistication.

2 | THE RETURNS TO ALGORITHMS ON DECISIONS IN ORGANIZATIONS

Starting with Meehl's (1954) review of forecasting studies showing that algorithms have the potential to outperform 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 (Berk, 2017; Kleinberg et al., 2018), medical diagnoses (Dawes et al., 1989; Grove et al., 2000), and many others (Goodwin & Fildes, 2007; Vrieze & Grove, 2009). A large literature has also documented limitations of algorithms, including challenges such as reinforcing bias or focusing on a narrow set of outcomes (Choudhury et al., 2020; Ludwig & Mullainathan, 2021).

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. Despite growing investment into algorithmic sophistication, evidence suggests that firms do not always see large returns to their investments in analytics (Brynjolfsson et al., 2021). In this article, we explore the returns to algorithms on decisions within an organizational context and propose two factors that may limit their returns.

2.1 | Algorithmic sophistication and improvements in prediction

One reason that the returns to algorithms for managerial decisions may be limited is that for 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., Dhami, 2003; Gigerenzer & Brighton, 2009; Gigerenzer & Goldstein, 1996). This research highlights the bias-variance tradeoff, proposing that heuristics may at times be more accurate than complex strategies in contexts with higher uncertainty and redundancy where simpler methods with fewer free parameters may reduce variance (Gigerenzer & Gaissmaier, 2011).

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.

2.2 | The role of decision authority

A second consideration is that decision-makers may use their decision authority to make a decision that does not leverage potential prediction improvements. Many organizations today provide algorithmic recommendations as decision aids for managers who make the ultimate decision. One rationale for this is that managers have private knowledge to inform the decision and correct problematic algorithmic recommendations.

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

Therefore, even when algorithms provide gains in prediction, decision-makers may not be able to take advantage of them to improve their decision, because doing so will make the decision worse on other dimensions. An alternative way of characterizing this issue is that current versions of algorithms may have misaligned objective functions relative to true organizational objectives, which managers with decision authority can help correct.

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. 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 in ways that result in a worse decision overall.

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—can be more likely to prefer their own forecasts over algorithmic predictions after seeing algorithms err, even when they are more accurate overall (Dietvorst et al., 2015; Logg et al., 2019). While a part of this effect may be driven by a negative reaction to algorithms, as found by Tong et al. (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 et al., 2015 for a review) and resistance to advice and change among professionals with specialized knowledge and strong norms (e.g., Greenwood et al., 2019; Kellogg, 2014). 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.

2.3 | Empirical evidence from organizational contexts

While evaluating these factors is important 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 non-organizational contexts such as the laboratory, classroom, or online settings (e.g., Choudhury et al., 2020; Dietvorst et al., 2015; Logg et al., 2019; Yeomans et al., 2019).

There is relatively little work exploring how decisionmakers in organizations leverage their contextual knowledge on organizational objectives to inform decisions when faced with algorithmic inputs.[3] Examining how decisionmakers in organizations use their contextual knowledge can provide insight on the role of decision authority in capturing the returns to algorithms.

In this article, we empirically examine the returns to algorithms on managerial decisions, focusing on the gains in prediction and the role of decision authority within a real organizational setting.

3 | EMPIRICAL CONTEXT

We explore the role of algorithms in decision-making within an Inspectional Services Department of a major metropolitan city in the United States ("the City"). The key decision we studied

[3] 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. Choudhury et al. (2020) 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.

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 ward or "neighborhood" and rotated across wards every 2 years.[4] Inspections took approximately 2–4 h, 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.

Inspections yielded substantial variation in the violations found across restaurants. For example, inspections conducted between 2007 and 2015 uncovered 0–60 weighted violations per restaurant (Figure B1). The department assigned weights for violations based on their severity as follows[5]: Level I violations (1 point) corresponded to noncritical violations such as building defects or standing water. Level II violations (2 points) were critical violations more likely to create food contamination, illness, or environmental hazard. Level III violations (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 that restaurant's operating 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 of the decision for this department was 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 identify and incapacitate establishments that posed the highest risk to public health.[6] Thus, decision quality in this department 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 at least in part observable: 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 or restaurant size), this could be interpreted as not improving the ultimate decision according to the organization's stated objectives.

A third advantage was accessibility of data with the potential to improve these predictions: both the City's historical administrative data, as well as external data (e.g., from platforms such as Yelp, TripAdvisor, and Twitter).

[4] Larger wards were assigned to multiple inspectors, who subdivided the ward geographically.

[5] This weighting scheme was designed and provided by the City.

[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 importance of identifying restaurants posing the highest risk to public health.

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, their performance factored into promotions and affected their workload. Inspection outcomes and customer complaints about unsanitary conditions were recorded, 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.

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 not able to reach that target. Thus, better prioritizing inspections could potentially improve the allocation of inspectors' scarce time.

Many of 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 (27% prioritized the timing since the last inspection, 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 design, the City made final design choices and executed the implementation.

Business-as-usual represented the status quo: relying on inspectors' own rankings to prioritize 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.[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, ratings, price range, hours, services (e.g., reservations), business ambience (e.g., children-friendly), and 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.

[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 main analysis.

[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 (Glaeser et al., 2016). 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.[9] 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 took each inspector's list, sourced equal numbers of the highest-ranked restaurants from the algorithm lists, removed any duplicates, and randomly sorted the remaining restaurants.[10] This meant that rather than assigning inspectors to different conditions, the experiment randomly sorted restaurants from the three methods on each 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 they would be supplemented by those prioritized using data processed by the data team, to help identify restaurants with more 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, this City was one of a few departments to use predictive algorithms for this purpose. The City's data team had 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, which would potentially drive them to use their discretion more often.

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, 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 in practice inspected substantially fewer restaurants than the 1042 assigned on the dockets: only 243 were inspected, which averaged to 14 restaurants per inspector. This occurred for several reasons:

- First, restaurant inspections were only one component of inspectors' jobs. They also had higher-priority inspections of hospitals, nursing homes, and schools, which had to be inspected at required intervals. Inspectors had many of these during the pilot, which meant that only approximately half of the period was dedicated to restaurant inspections. Second, two inspectors were sick and unable to undertake inspections for the full period.

[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.

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 algorithmically-ranked lists. (They also wanted to ensure that inspectors did not run out of restaurants in case many were closed.)

- 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 indicated 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.[11]

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. 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 overlaps between methods, especially algorithms, with 176 restaurants (26%) ranked in the top-20 by at least two methods. One hundred eight restaurants (16%) were ranked in the top-20 by the data-rich algorithm alone, 97 (14%) were ranked by the data-poor algorithm alone, and 293 (43%) were ranked 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:

$$Total\ Violations_i = \alpha + \beta DataRich_i + \gamma DataPoor_i + \sum_{k}^{4} \delta_k MultipleMethods_{k,i} + \eta_j + \epsilon_i \tag{1}$$

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; meanwhile, δ accounts for overlaps between methods and represents the mean 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).

[11] 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 highly ranked by the data-rich algorithm. We were missing a ranking from the data-poor algorithm as the restaurant had no historical inspections.

We estimate this model with and without inspector fixed effects (η_j). We explore the robustness of our results across the full sample of inspected restaurants, as well as across alternative subsamples, varying the threshold away from 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, inspectors were only about two-thirds as likely to follow algorithmic recommendations relative to their own lists, dissipating the informational gains from algorithms. Furthermore, we find little supportive evidence that inspectors substantially improved inspection decisions 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

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 slightly more than the combination of a Level II violation and a Level III violation. Our estimates of the gains from algorithms over inspectors, β and γ, are 4.94 ($p = .001$) and 5.17 ($p = .006$), 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 (but not inspectors) had 6.8 more violations on average compared to inspectors alone, and restaurants ranked highly by all three methods had 6.3 more violations. Being ranked highly by inspectors and one of the algorithms provided limited gains relative to inspectors alone.

In Columns 3 and 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 $p = .79$ 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 the result is imprecise: the confidence interval of the difference between the algorithms ranges from -5.8 to 7.5.

These results are qualitatively robust across the full sample of inspected restaurants, as well as alternative subsamples that vary the threshold of top-ranked restaurants (Table B1). 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 (Table B2).

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 marginal benefit of additional data may be limited in this case. This is consistent with findings in similar applications to problems with representative datasets,

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

Note: This table shows the violations found across restaurants prioritized by inspectors and the algorithms. 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 Level I, II, and III violations. (Level I violations received 1 point; Level II received 2 points; and Level III received 5 points). *Data-Rich Algorithm Only* and *Data-Poor Algorithm Only* are binary variables indicating restaurants that were respectively only ranked in the top-20 by the data-rich algorithm or the data-poor algorithm. *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. Standard errors are reported in parentheses.

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 mandate to prioritize restaurants with higher violations, we make the same assumption in our interpretations above.

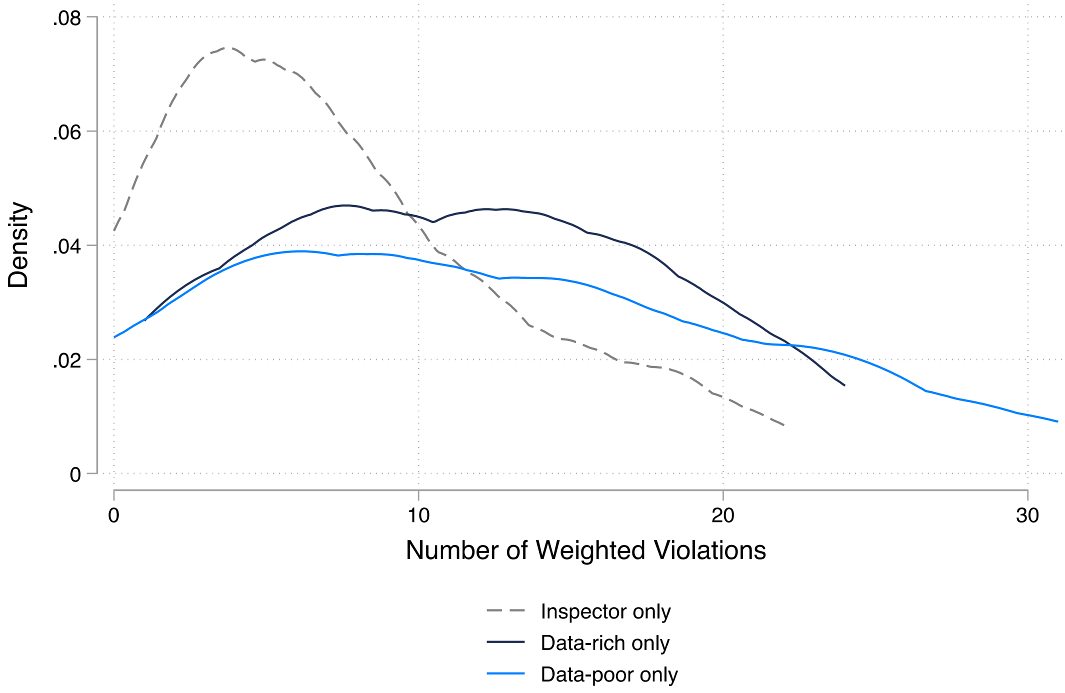

FIGURE 1 The number of violations across restaurants prioritized by each method. This figure shows kernel density plots of weighted violations found at restaurants ranked by inspectors alone compared to each of the algorithms alone.

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 pure prediction gains from algorithms compared to inspectors may be smaller than our estimates. We explore this by comparing whether inspector rankings performed better on secondary objectives relative to algorithms.[12] We find that inspectors listed restaurants in closer proximity to each other compared to algorithms if the rankings were followed in order (Figure 2), and were more likely to rank more overdue restaurants compared to algorithms (Table 2 Panel A). We do not observe that inspectors were substantially more likely to prioritize popular restaurants as proxied by Yelp reviews and ratings (Table 2, Panel A). This evidence raises the possibility that inspector rankings considered secondary organizational objectives beyond prediction alone, and that the prediction gains from algorithms may be smaller than estimated.

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

[12] We are not able to conduct this analysis for the number of violations, because violations are not observed prior to inspections. We show differences in violation severity across all inspected restaurants ranked in the top-20 by method in Panel C of Table 2.

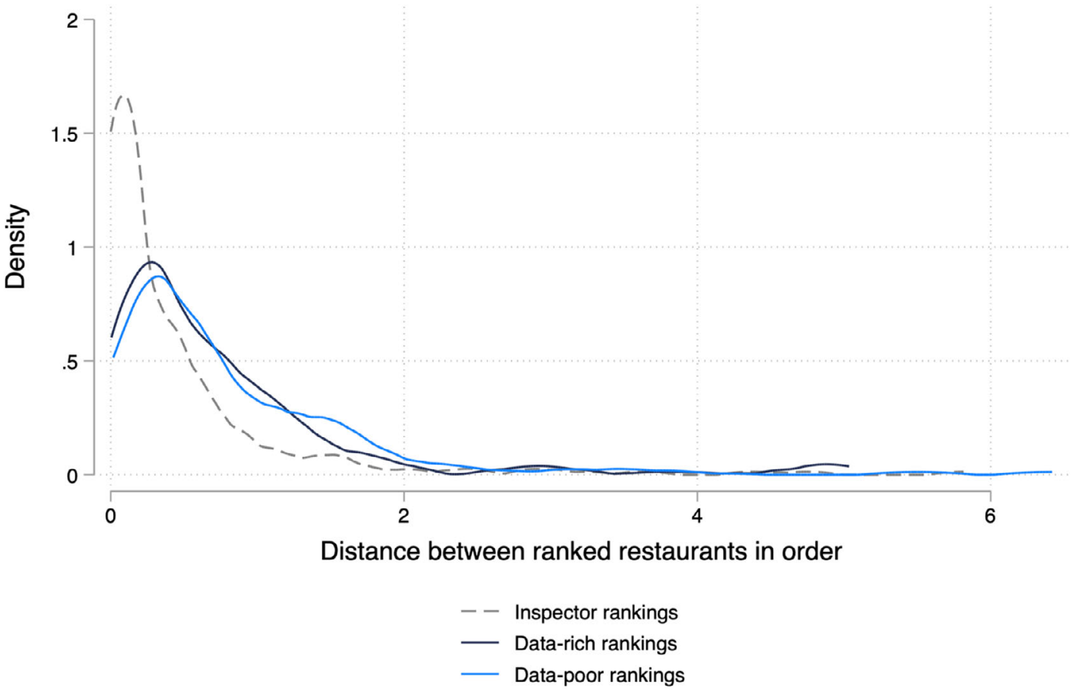

FIGURE 2 Distance between ranked restaurants in order by method. This figure shows kernel density plots of the distance between restaurants if traveled to in order based on inspector and algorithm rankings. Mean distance between restaurants ranked by inspectors is 0.45 miles (SE = 0.05), compared to 0.82 miles (SE = 0.05) for data-poor rankings and 0.77 miles (SE = 0.05) for data-rich rankings.

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. Moreover, some departments might want to use inspections as a deterrent, which would change the role of predictions in inspection decisions. Given these considerations and implementation costs, the net benefit of changing the targeting system is unclear. While many cities target inspections, these considerations should consider the dynamic nature of inspections and accuracy of violation data, 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 predictions of violations.

5.2 | Decision authority and the returns to algorithms

Despite the extent to which algorithms were found to be able to improve on inspector rankings, these potential gains did not fully translate into better decisions about which restaurants to inspect. Inspectors were less likely to inspect algorithmically-ranked restaurants compared to those that they themselves had ranked.

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	<.001	.28
Yelp rating	3.14	2.6	2.97	.33	.17
Review count	119.9	144.41	154.28	.19	.74
Seafood	0	0.05	0.06	.004	.66
Restaurant age	1.69	3.18	7.27	.003	.05
Price range	1.4	1.14	1.27	.17	.40
Accepts reservations	0.27	0.22	0.21	.20	.84
Table service	0.46	0.38	0.32	.03	.34
Days since last inspection	190.52	246.57	302.58	<.001	.09
N	108	97	293		

Panel B: Inspected versus non-inspected restaurants in Top 20

	Inspected	Not inspected	*p*-value
Chain	0.02	0.03	.28
Yelp rating	2.82	2.95	.45
Review count	123.04	138.05	.55
Seafood	0.03	0.05	.22
Restaurant age	4.92	5.27	.84
Price range	1.14	1.33	.04
Accepts reservations	0.18	0.26	.04
Table service	0.34	0.41	.05
Days since last inspection	270.97	247.68	.59
N	174	500	

Panel C: Inspected restaurants ranked in top-20 by method

	(1) Data-rich algorithm only	(2) Data-poor algorithm only	(3) Inspector only	*p*-value (1) = (3)	*p*-value (2) = (3)
Level I violation	5.58	6.17	4.2	.10	.10
Level II violation	0.63	0.48	0.29	.11	.13
Level III violation	1.47	1.7	0.85	.21	.02
N	19	23	91		

Note: 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 columns (4) and (5) display *p*-values of the difference between those columns. Panel B compares the attributes of inspected and non-inspected restaurants among all top-20 ranked restaurants. Panel C compares the number of violations by severity across inspected restaurants from each of the lists.

TABLE 3 Inspector compliance.

	(1)	(2)	(3)
	Number of restaurants inspected	%	% of restaurants inspected out of all top-20 ranked restaurants in category
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	

Note: This table shows the breakdown of inspected restaurants by ranking method. Columns (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 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 4.1, 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).

Figure 3 plots the percentage of restaurants inspected by each inspector, the red line indicating where this (in dark gray) would have ended if the inspector had fully followed the dockets. While there is substantial heterogeneity across inspectors in the extent to which they deviated from the algorithm, this figure shows that most 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.

5.2.1 | Robustness

While the potential of inspectors deviating from algorithmic recommendations 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

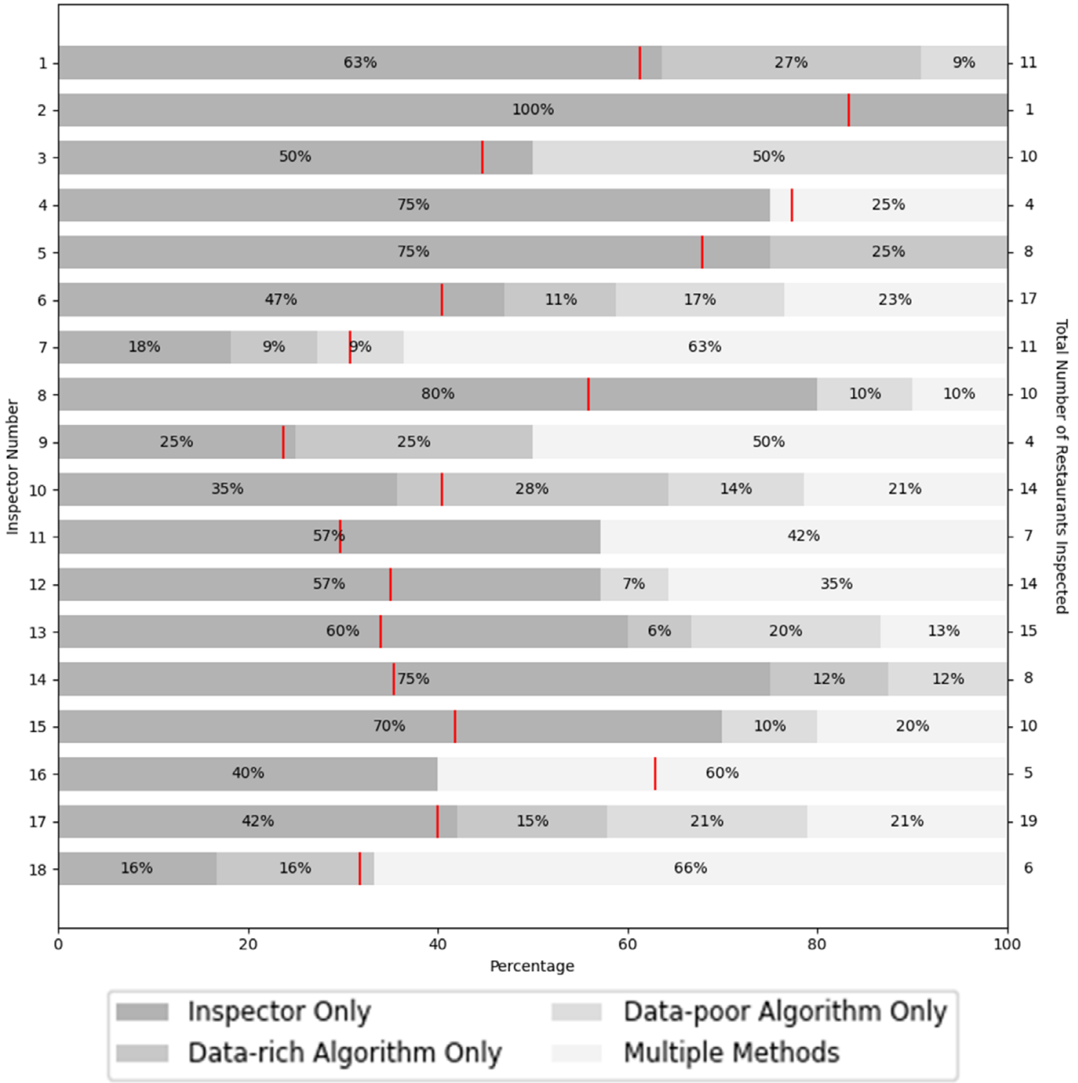

FIGURE 3 Percentage inspected by method across inspectors. This figure plots the percentage of inspected and top-20 ranked restaurants by method for each inspector. Each bar represents a single inspector, where the left axis indicates the inspector, and the right axis shows the number of restaurants 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 gray) should have ended if inspectors had fully complied.

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.

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 and imprecise ($\beta = 1.07$, $p = .433$).

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 × 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

Note: 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 Level I, II, and III violations. (Level I violations received 1 point; Level II received 2 points; and Level III received 5 points). Standard errors are reported in parentheses.

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 ($p = .418$) and 0.19 ($p = .562$) for data-rich and data-poor algorithms, 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 qualitatively consistent results, with gains of 2.3–3.3 more violations flagged by the data-rich algorithm and 4.3–5.4 by the data-poor algorithm, compared to 7.2–8.4 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 measured 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 algorithmically-ranked lists compared to their own. This suggests that inspectors may not have been making tradeoffs using private information to identify more violations, and makes it difficult to construct a clear alternative

TABLE 5 Robustness in gains from algorithms across inspections data up to 2022.

Outcome	(1) Total violations	(2) Total violations	(3) Total violations	(4) Total violations	(5) Total violations	(6) 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

Note: 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. (Level I violations received 1 point; Level II received 2 points; and Level III received 5 points). Standard errors are reported in parentheses.

explanation driven by selection. Second, the magnitude of the differences we observe between algorithms and inspectors is large, and does not differ significantly across rankings. Hence, it seems unlikely that selection would change these results directionally.

5.2.2 | How inspectors use decision authority

We explore whether inspectors used their decision authority to improve the decision in other respects by balancing secondary organizational objectives. We compare whether restaurants that they chose to inspect improved on these objectives relative to those that they chose not to inspect, examining each in turn: reducing travel costs, targeting more overdue inspections, prioritizing more popular restaurants that may pose a larger risk to public safety, and placing more weight on higher-level violations.

We do not find evidence that inspectors ultimately improved substantially upon the dimensions reflecting these other organizational objectives. Rather, we find that inspectors often did not follow algorithms to make decisions that generally worsened these objectives. First, 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 vs. 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

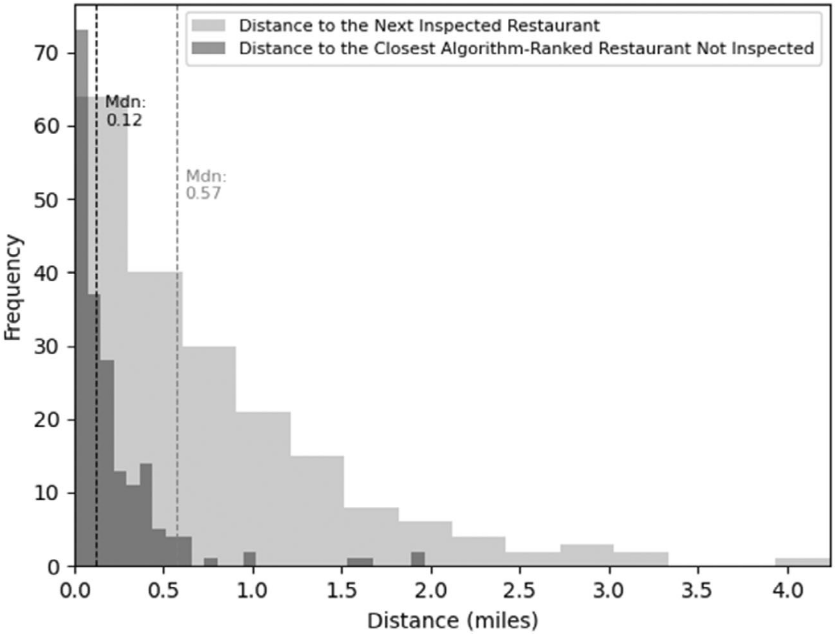

FIGURE 4 The distance inspectors traveled versus the closest algorithm-ranked restaurant Not Inspected. This figure plots the distribution of the distances inspectors traveled to their next restaurant, compared with the distance to the closest algorithm-ranked restaurant on the docket that was not inspected. Mean distance to the closest algorithm-ranked restaurant not inspected was 0.21 miles (SE = 0.02), while mean distance to the next inspected restaurant was 0.76 miles (SE = 0.05).

(Figure B2). While one possibility may be that inspectors traveled farther to conduct high-risk inspections (e.g., nursing homes, re-inspections), we find this result to be robust in subsamples of the data when inspectors only conducted regular restaurant inspections.[13] We focus on restaurant inspections on days when inspectors had no high-risk inspections, as well as all consecutive restaurant inspections, and find that across both subsamples, inspectors generally traveled farther than the closest algorithm-ranked restaurant (Figure B3). Moreover, inspectors had some flexibility in when they could do their high-risk inspections (e.g., within a particular week or time of day), which meant that they also made a decision on when to fit them in and change the order of restaurants that they inspected.

We also find little evidence that inspectors placed significantly higher weight on overdue inspections, restaurant popularity, or violation severity in their inspection decisions. We do not find substantial differences in the days overdue or the number of reviews and ratings between inspected and non-inspected restaurants on average (Table 2 Panel B).[14] While we cannot compare the number of violations found across inspected and non-inspected restaurants (since the number of violations are not observed for non-inspected restaurants), we find that algorithm-ranked restaurants that were inspected had more violations in all three risk levels compared to inspector-ranked restaurants that were inspected, although differences are marginal in magnitude (Table 2 Panel C). This helps us bound our estimates of prediction gains given the concern of what inspector rankings may represent (discussed in Section 5.1): if inspectors were indeed as good as algorithms in predicting violations and the estimated gains were simply arising from their consideration of secondary objectives, then we would not expect algorithms to be able to improve upon them as our findings suggest.

Although inspectors did not ultimately improve upon secondary objectives on average, we observe some suggestive evidence that they sought to improve on them, especially for overdue inspections. Figure 5 shows evidence consistent with the interpretation that inspectors were sensitive to how overdue restaurants were. It suggests that inspectors were marginally more likely to inspect restaurants from algorithm-ranked lists that were more overdue, with the distribution of non-inspected restaurants from algorithm lists shifted to the left. However, the difference on average is not substantial (17.96 days, $p = .617$), and there are many algorithm-ranked restaurants that were overlooked by inspectors with more days overdue than those inspected. Taken together, these results suggest that while it is possible that in some cases, inspectors were placing more weight on days overdue and willing to travel far and catch fewer severe violations in order to do so, the gains in days overdue may have been marginal, and we do not observe this to be the case on average.

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 whether they were chains, seafood restaurants, older, or lower-end—which may have helped them make decisions prior to using algorithms and driven how they applied their judgment. We find some evidence, though speculative, consistent with this interpretation. Table 2 Panel A shows that relative to algorithms, inspectors were more likely to

[13] This analysis is conducted on inspections located in wards assigned to a single inspector (15 out of 22 wards), because the data that we have on high-risk inspections only identify the ward of the inspection, and we cannot identify which high-risk inspection was assigned to which inspector in wards where multiple inspectors are assigned.

[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.

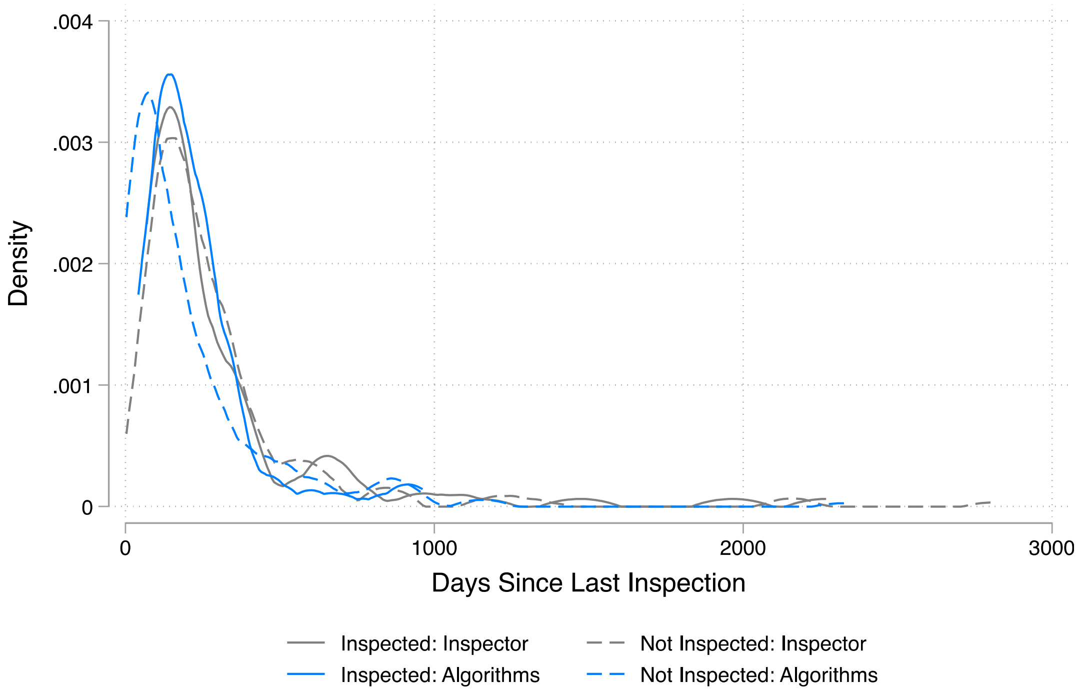

FIGURE 5 Days overdue by method and inspection status. This figure plots kernel density plots of the number of days overdue across inspected and non-inspected restaurants by whether they were ranked by inspectors or algorithms alone.

prioritize older businesses in their rankings, and chains and seafood restaurants to a lesser extent. Inspectors also placed higher priority on businesses with lower prices and without reservation 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 also 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: none of the inspectors held on to their original rankings in whole or even loosely in order, suggesting that there was some trust in the lists.

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 2 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 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. Nevertheless, 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., Greenwood et al., 2019; Kellogg, 2014; Logg 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 (Avan et al., 2019; Goodwin & Fildes, 1999). 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. In situations where discretion is sufficiently valuable and implementation is complicated or costly, algorithms might end up being less valuable in practice even if they have predictive power.

6 | SURVEY EVIDENCE FROM INSPECTIONAL DEPARTMENTS ACROSS THE UNITED STATES

Our findings suggest the following managerial implication: 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 private firms, our findings may be limited to the particularities of the department that ran the pilot.

To explore how these findings might generalize, we contacted 176 inspectional departments covering the largest 200 metropolitan areas[15] in the United States to conduct interviews on how they approach restaurant inspections and their perspectives on algorithmic sophistication and inspector discretion. We reached 55 departments that cover 45 US counties (392 cities, towns, and other territories)[16] for interviews that lasted up to 1 h on (1) how they prioritize their

[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 that was willing to be interviewed. Only one department that we reached refused to be interviewed.

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 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 to meaningfully integrate predictive algorithms. 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 fifty-five departments reported using some algorithmic rule or software to guide inspections, with three having run pilots using 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 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%) also ran behind their inspection targets, suggesting that they may have benefitted from using some data to prioritize inspections.

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 piloted using 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 private 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 common reason was that inspectors had organizational knowledge on other 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, some of those items may be there all day…so temperature issues [i.e., how serious the violation is] becomes much more important. 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 end of their tenure were less likely to use algorithmic recommendations, consistent with findings of Allen and Choudhury (2022). Indeed, 4 of 7 inspectors (57%) with 0–3 years of experience working in the department reported using algorithmic recommendations, as did 6 of 8 (75%) among those with 3–6 years of experience; 6 of 7 (86%) among those with 6–9 years of experience; 1 of 1 (100%) among those with 9–12 years of experience; and 5 of 10 (50%) among those with over 12 years of experience. Among those who used them, only 4 (17%) used them on a weekly basis, most using them 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% reported that algorithms were 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 in addition to algorithmic sophistication, decision authority is a relevant consideration for organizations seeking to use algorithms as decision aids. However, these interviews also raised other potential benefits to decision authority that are beyond the scope of our analysis such as improving recruiting, retention, and morale—highlighting that consideration of outcomes beyond the predictable outcome is important. While data and algorithms can play an important role in decision-making, these insights highlight that there are implementation costs and unintended consequences of using algorithms; hence, organizations should weigh the benefits of analytics against potential costs.

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 highlight the importance of managing decision authority in order for organizations to capture algorithms' value—potentially rather than improving algorithmic sophistication and/or expanding the range of input data. In our setting, even a simple algorithm based on historical inspection data did a better job at prioritizing restaurants for inspection relative to human rankings based on the primary outcome of interest. Nevertheless, the improvements in predicting violations did not fully translate into better inspection decisions, as inspectors often chose to prioritize restaurants based on their own rankings, without substantially improving the decision along the other key organizational objectives. While the City continued to explore the use of targeting following this pilot, they ultimately discontinued the program and returned to their old system, which did not use algorithms to prioritize inspections. Our findings suggest that managing decision authority may merit further consideration for many organizations, and that in some cases, organizations might choose not to use algorithms even if they have some predictive power over a relevant set of outcomes.

Our analysis has a number of limitations. First, our analysis takes the primary goal of the Inspectional Services Department as given, that is, to prioritize based on the number of violations. In practice, there may be other goals that departments could 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 or less productive than those we examined here, 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 (non-)compliance patterns we observe may not generalize to other settings with different communication and organizational dynamics, and exploring heterogeneity across conditions and types of decision-makers may provide a fruitful direction for future research.

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—which suggests that managers may have other tradeoffs in mind that may not be captured in a predictive algorithm with a narrow objective.

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.

ACKNOWLEDGMENTS

Following norms at the Strategic Management Journal, the lead author is listed first; other 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, numerous seminar participants, and the editor and referees. Data for this project was provided by the Inspectional Services Department and Yelp. Kim and Luca have consulted for tech companies, including Yelp. Kominers is a research partner at a16z crypto, and advises a number of firms engaged in marketplace development and design. 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, a Radcliffe Exploratory Seminar Grant, the Alfred P. Sloan Foundation, the Ewing Marion Kauffman Foundation, the Washington Center for Equitable Growth, and the Ng Fund and the Mathematics in Economics Research Fund of the Harvard Center of Mathematical Sciences and Applications. Part of this work was conducted during the Simons Laufer Mathematical Sciences Institute Fall 2023 program on the Mathematics and Computer Science of Market and Mechanism Design, which was supported by the National Science Foundation under Grant No. DMS-1928930 and by the Alfred P. Sloan Foundation under grant G-2021-16778. All errors are our own.

DATA AVAILABILITY STATEMENT

The data in this study are not publicly available due to privacy restrictions based on a nondisclosure agreement.

ORCID

Hyunjin Kim https://orcid.org/0000-0002-0296-8977
Scott Duke Kominers https://orcid.org/0000-0002-7608-6619
Michael Luca https://orcid.org/0000-0002-0747-7544

REFERENCES

Agrawal, A. (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, *33*(2), 31–50.

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press.

Allen, R., & Choudhury, P. (2022). Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organization Science*, *33*(1), 149–169.

Athey, S., Bryan, K., & 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., & Suzuki, J. (2019). The impact of big data on firm performance: An empirical investigation. *AEA Papers & Proceedings*, *109*, 33–37.

Bartel, A., Ichniowski, C., & 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., & 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., & 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., & 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., Jin, W., & McElheran, K. (2021). *The power of prediction*. Working Paper.

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

Choudhury, P., Starr, E., & 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.

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.

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

Dietvorst, B., Simmons, J., & 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.

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., & 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), 650–669.

Gigerenzer, G., Todd, P., & The ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Glaeser, E., Hillis, A., Kominers, S., & Luca, M. (2016). Crowdsourcing city government: Using tournaments to improve inspection accuracy. *AER Papers & Proceedings*, *106*(5), 114–118.

Goodwin, P., & 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., & 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), 19–30.

Hoffman, M., Kahn, L., & Li, D. (2018). Discretion in hiring. *Quarterly Journal of Economics*, *133*(2), 765–800.

Jin, G., & Lee, J. (2018). *A tale of repetition: Lessons from Florida restaurant inspections*. Working Paper.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & 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. (2021). *The impact of communicating multiple goals*. Working Paper.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Lehman, S. (2014). *Twitter helps Chicago find sources of food poisoning*. Reuters Health.

Logg, J., Minson, J. A., & 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., & Nickerson, J. (2011). Regulator heterogeneity and endogenous efforts to close the information asymmetry gap. *Journal of Law and Economics*, *54*, 25–54.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of judgment and decision making*. John Wiley & Sons, Ltd.

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, S., Khodabandeh, S., Fehling, R., Lafountain, B., & Kiron, D. (2019). Winning with AI: Pioneers combine strategy, organizational behavior, and technology. *MIT Sloan Management Review*, *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), 525–531.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

[Correction added on 31 January 2023, after first online publication: Appendix has been removed in the proof and published as an online only supporting information in this version.]

How to cite this article: Kim, H., Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2024). Decision authority and the returns to algorithms. *Strategic Management Journal*, *45*(4), 619–648. https://doi.org/10.1002/smj.3569