# How Does Compliance Affect the Returns to Algorithms? Evidence from Boston's Restaurant Inspectors[i]

Edward L. Glaeser[ii], Andrew Hillis[iii], Hyunjin Kim[iv], Scott Duke Kominers[v] and Michael Luca[vi]

June 2019

## Abstract

Algorithms have the potential to improve our ability to target services – but the returns are limited if employees make choices largely based on other considerations. Partnering with Boston's Inspectional Services Department, we compare the performance of three different methods of targeting restaurant hygiene inspections: (1) human judgment based on inspector discretion (the status quo); (2) a "data-poor" algorithm based on the average number of violations across historical inspections; and (3) a "data-rich" algorithm based on a random forest model trained on historical inspections and Yelp data. The "data-rich" algorithm slightly outperforms the "data-poor" algorithm, and both dramatically improve upon inspector discretion -- suggesting that the greatest gains come from using data to supplement inspectors' priors, rather than from sophisticated algorithm design. Yet, inspectors are only half as likely to comply with inspection directives based on either algorithm, relative to individual judgment. These findings suggest that the implementation gains from big data and machine learning are limited by employee compliance.

[i] The authors are grateful for the collaboration of the City of Boston (especially Ben Batorsky, Matthew Mayrl, and Commissioner William Christopher), Yelp (Artem Avdacev, Luther Lowe, and Aaron Schur), and DrivenData (Peter Bull and Greg Lipstein). Additionally, the authors gratefully acknowledge the helpful comments of Susan Athey, Avi Goldfarb, Shane Greenstein, Sendhil Mullainathan, Andrei Shleifer, and Mitchell Weiss. Data for this project was provided by Yelp and the City of Boston. Luca has done consulting for tech companies, including Yelp. Kim has been employed by Yelp. We are grateful for the support of the National Science Foundation (grants CCF-1216095, DGE-1144152, and SES-1459912), the Harvard Milton Fund, Yelp, the Taubman Center for State and Local Government, the Rappaport Institute for Greater Boston, the Sloan Foundation, the Ewing Marion Kauffman Foundation, and the Ng Fund and the Mathematics in Economics Research Fund of the Harvard Center of Mathematical Sciences and Applications.

[ii] Harvard University, eglaeser@harvard.edu
[iii] Persistent Systems, abhillis@gmail.com
[iv] Harvard Business School, hkim@hbs.edu
[v] Harvard University, skominers@hbs.edu
[vi] Harvard Business School, mluca@hbs.edu

## 1. Introduction

Since the 1970s, social scientists have emphasized that predictive algorithms can improve human judgment and decision-making (Dawes (1979), Dawes et al. (1989), Grove and Meehl (1996), Grove et al. 2000). Recent reductions in data and computing costs have further increased interest in putting predictive algorithms into practice (Chalfin et al. (2016), Kleinberg et al. (2017), Agarwal et al. (2018)), in contexts from hiring to medical diagnoses to bail determination (Dhami (2003), Kleinberg et al. (2017), Lakkaraju et al. (2017), Cowgill (2018)). While abundant research documents the statistical gains from improved algorithms, the practical implementation of algorithms has been less widely studied (Jung et al. 2018),[vii] and even the most sophistical algorithms will do little if workers just follow their own priors.

In this paper, we test the relative importance of algorithmic sophistication and worker compliance. We seek to compare, in the field, the gains from enhanced predictive power with the gains from inducing workers to comply with predictions of a simple predictive model.

We partnered with the City of Boston's inspectional services department to implement and test algorithms aimed at identifying restaurants that are at risk of health code violations. Our setting is a natural place to test the power of predictive algorithms, since inspectors' scarce time must be allocated with a fundamentally predictive objective: identifying restaurants that have health code violations.[viii] Moreover, the inspectional services department's history of past violations and consumer reviews from platforms like Yelp together yield strong data inputs for a sophisticated prediction algorithm.

We compare three different approaches of allocating inspectors: (1) human judgment relying on inspector discretion ("business-as-usual"); (2) a "data-poor" algorithm based on the average number of historical violations of each restaurant; and (3) a "data-rich" algorithm based on a random forest model trained on historical violation data and Yelp reviews. Our "data-rich" algorithm theoretically yields an approximately 40 percent improvement in efficiency relative to the "business-as-usual" model (see Glaeser et al. (2016a,b)). We take the restaurants that have the largest predicted risk of hygiene violations according to each approach, and provide these to inspectors to guide their inspections over four periods of roughly eleven days each.

By examining both "data-poor" and "data-rich" algorithms, we can assess how differences in algorithmic quality impact inspection efficiency in the field. Our design overcome a selective labels problem by observing outcomes across each method. We also observe counterfactual inspector decisions and (non-)compliance behavior. By asking inspectors which restaurants they would prioritize prior to providing algorithmic recommendations, we can explore whether choices change with access to algorithmic predictions.

We find that our predictive algorithms outperform business-as-usual by identifying restaurants with 50 percent more weighted violations. Most of the gains stem from integrating historical violations data – the data-poor algorithm results in improvements as large as those from the data-rich algorithm. Yet despite our algorithms' superior performance, inspectors are only half as likely to inspect

[vii] Recent empirical studies using regression discontinuity designs have shed light on the use of algorithmic tools in judicial decisions, but finds conflicting results across different contexts. For example, Stevenson (2017) finds only small, temporary impact of algorithms on pretrial release outcomes, while Berk (2017) finds large reductions in re-arrests.

[viii] While there are additional objectives in our setting, such as deterring restaurants from committing violations in the first place or ensuring fairness in the inspection allocation, our discussions with health departments highlighted the first-order importance of identifying and inspecting restaurants with the highest likelihood of health violations.

restaurants based on the algorithm, relative to those based on their own judgment. Thus, in the context we study, worker compliance appears to be far more important than algorithm sophistication.

Our work shows the potential for algorithms to increase the number of hygiene violations found. At the same time, our findings do not imply that hygiene inspectors are making poor decisions, as they may have objectives (e.g., fairness) other than maximizing the number of violations found. While we also cannot claim that our work generalizes to other settings, the case of Boston hygiene inspectors suggests that the basics of worker compliance can be more important than even significant refinements of predictive algorithms.

## 2. Empirical Context

The Boston Inspectional Services Department conducts inspections to "serve the public by protecting the health, safety, and environmental stability of Boston's business and residential communities" (Boston). The Division of Health Inspections within the department oversees food safety codes, which are regulated by the state-level Massachusetts State Sanitary Code. The department employs 20-30 inspectors at any given time, who are assigned to at least one of 22 "wards" that divide up Boston into neighborhoods. A few wards have multiple inspectors assigned to cover different areas, and inspectors' ward assignments are changed every two years.

Inspectors are responsible for inspecting all licensed establishments, with the objective of targeting the highest-risk establishments more frequently. Facilities like hospitals, nursing homes, day care centers, schools, and caterers require three inspections per year. These are followed by restaurant establishments that prepare, cook, and serve most products immediately. The department aims to conduct at least two inspections per restaurant per year, but in practice inspectors' time budget constraint binds, so that inspectors are sometimes not able to visit a restaurant more than once in an inspection cycle. Inspectors are encouraged to conduct inspections that are in close proximity to each other, but their main objective is to prioritize inspections of establishments with the highest risk to public health.[ix]

Inspections vary in the number of violations found: between 2007 and 2015, restaurant inspectors found anywhere from 0 to 60 weighted violations per inspection (Figure 1). Weights are assigned based on the severity of the violation: Level I corresponds to non-critical violations such as building defects or standing water. Level II violations are "Critical Violations," such as the presence of fruit flies, which are more likely than Level I violations to create food contamination, illness or environmental hazard. Level III violations are considered "Food-borne Illness Risk Factor[s]"; examples include insufficient refrigeration or a lack of allergen advisories on menus. A Level I violation has a weight of 1, Level II a weight of 2, and Level III a weight of 5. When critical violations are found in a restaurant, the City reinspects that restaurant within 30 days. [x]

## 3. Research Design

[ix] Furthermore, other priority situations such as re-inspections are prioritized above geographic proximity.

[x] If violations are deemed to pose an imminent public health risk, the City may temporarily suspend the restaurant's food permit.

Between February 1 and March 25, 2016, we partnered with the City of Boston to compare three methods of allocating inspectors. At the beginning of each of the four inspection periods in our study window (February 1-12, February 15-26, February 29-March 11, and March 14-25), Boston's Head Inspector asked all inspectors to rank the restaurants in their ward by the order in which they felt those restaurants should be inspected.[xi]

For each inspection period, inspectors received a docket of restaurants to inspect. This docket was presented as a "new way of doing inspections" to guide inspector decisions.[xii] The docket consisted of the top fifteen ranked restaurants from each of three approaches to targeting inspections, sorted in random order.

The first method for determining the dockets – "business-as-usual" – relied on inspectors' own rankings. The second method, which we call a "data-poor" algorithm, used the average number of violations per historical inspection to rank restaurants in each ward from most to least likely to have violations. The third method ranked restaurants according to a "data-rich" algorithm, which used a random forest model trained on both historical violations and Yelp data including the number of Yelp reviews, Yelp rating, price range, hours, services available (e.g., wifi, alcohol, take-out, appointments), business ambience (e.g., noise level, children-friendly, ages allowed), and business neighborhood.[xiii]

In subsequent periods, restaurants from the previous period's docket that were not inspected were added to the docket first, followed by a new crop of restaurants, chosen through the same mechanism as before. The docketing system exogenously influenced which restaurants were inspected when, allowing us to separate confounding effects of when inspectors decided to visit certain restaurants and tease out whether algorithmically-ranked restaurants indeed had a higher number of violations. Our approach also provided us with a way to observe which restaurants – of those listed on their dockets – inspectors chose to prioritize.

Not all restaurants were ranked, as some inspectors listed fewer restaurants than others. Inspectors were also not able to inspect all restaurants that had been ranked during the study period. Across all three targeting methods, we observed a total of 1,042 restaurants that were ranked, and 361 restaurants that were inspected. To ensure consistent availability of rankings across the three docket formats, we limit our analysis to the top 20 restaurants ranked by each condition for each ward, which comprise of 245 restaurants out of the full set of 361 that were inspected, and result in 372 restaurant-method level observations.[xiv]

We use linear regression to test for the difference in weighted violations across the three methods, with the following model for the number of weighted violations $Y_{im}$ associated with restaurant $i$ ranked by method $m$:

$$Y_{im} = \alpha + \beta T_{data-rich,im} + \gamma T_{data-poor,im} + \varepsilon_{im}.$$

[xi] In order to isolate confounds, we excluded from rankings any high-risk establishments (e.g., hospitals and nursing homes), as well as restaurants that had a mandated priority to re-inspect, such as prior inspections that yielded major violations requiring a re-inspection in 30 days.

[xii] The Health Commissioner chose this wording in order to avoid calling attention to the new docket practice.

[xiii] This method was the second-place winner in a tournament that the City of Boston ran to source algorithms for predicting hygiene violations (described in detail in Glaeser et al. (2016a,b)). This option provided theoretical efficiency gains of around 40% relative to using inspector discretion; it was chosen above the first-place winner because the City of Boston felt it would be it substantially easier to implement in-house.

[xiv] Our results are robust to other definitions of this threshold, including focusing on restaurants in the top 5, 10, and 15 of each method's ranking.

Here, $\alpha$ represents the mean number of weighted violations for restaurants ranked by human judgment; $\beta$ and $\gamma$ represent the mean expected increase in weighted violations for a restaurant drawn from the ranking by the data-rich and data-poor algorithms relative to a restaurant drawn from ranking by human judgment, respectively. For each restaurant inspected, these coefficients inform how many more weighted violations the City may expect to find if they use the data-rich or data-poor algorithm relative to business-as-usual.

## 4. Results

### 4.1. Performance

We begin by comparing the performance of the three methods over the study period. Our outcome of interest is the weighted sum of violations found during an inspection.

Table 1 presents our estimates of $\alpha, \beta,$ and $\gamma$ under two specifications. Column 1 shows a data-poor comparison of mean weighted violations according to the ranking method. The mean number of weighted violations for restaurants ranked by human judgment is 8.19. This is equivalent to having one each of a Level I, II, and III violation. Our estimates of $\beta$ and $\gamma$ are 4.51 and 4.02 respectively. They are not statistically distinguishable, though we lack the statistical power to reject meaningful differences. The difference between the algorithm and human judgment equates to targeting a restaurant with on average one more Level III violation and one more Level I violation.

In Column 2 of Table 1, we explore whether the improvements in performance differ depending on the ranking position in the list. We find that the impact of rank is not statistically significant, and coefficients are both small and relatively precise around 0, suggesting that improvements are spread across the ranking distribution.

We draw three conclusions from Table 1. First, the data-poor and data-rich algorithms outperform human judgment for predicting weighted violations. These performance improvements are on the order of 50% and statistically significant. Second, the performance of the data-poor and data-rich algorithms are statistically indistinguishable, suggesting that the marginal benefit of additional data may be limited in this case. Third, the improvements are spread across the ranking distribution.

### 4.2. Ranking Differences

We next explore differences between inspector discretion and algorithmic methods to better understand what might be driving the observed differences in performance. To explore which restaurants each method prioritizes, Table 2 reports summary statistics for restaurants ranked in the top 20 for each method, with Columns 1-3 presenting the mean of each variable for each method and Columns 4 and 5respectively, showing the *p*-value for a *t*-test for the equality of means between (4) the data-rich algorithm and data-poor algorithm and (5) the data-rich algorithm and human-judgment ranked restaurants. We examine characteristics that may be of interest to policymakers and consumers: chain status, Yelp ratings, Yelp review count, type of cuisine, delivery, price range, availability of takeout, and years of operation. Notably, the data-rich and data-poor algorithms are less likely to prioritize inspecting chain restaurants and more likely to prioritize inspecting restaurants with ethnic cuisine and restaurants that deliver. The targeting of ethnic cuisine suggests that implementations of algorithms optimized for efficiency also requires managers to explicitly identify and engineer the fairness constraints that we have in place (based on interviews we conducted, some inspection departments are sensitive to this type of targeting).

### 4.3. Threats to Validity

We also found that inspectors were roughly twice as likely to inspect restaurants from dockets based on their own priors, relative to dockets suggested by our algorithms. This non-compliance issue highlights an important implementation concern: in practice, management and incentives may be key to helping organizations realize the gains from using algorithms.

Additionally, non-compliance in our study created a potential problem of attrition: we are able to observe violation results for only a subset of the restaurants on each docket – and the results described in the prior two sections depend upon having a representative sample from each method's rankings.

As shown in Table 3, there are two sources of missing data related to non-compliance. Some inspectors did not provide full rankings across all restaurants in their wards. Due to challenges in implementation, we observe 94%, 93% and 72% of the rankings for the data-rich algorithm, data-poor algorithm, and inspectors, respectively (Column 1 of Table 3). In addition, inspectors did not inspect all restaurants assigned to them from each method (Column 2 of Table 3). Among restaurants for which we observe a ranking in the top 20, 31%, 32% and 61% were inspected from the data-rich algorithm, data-poor algorithm, and inspector list, respectively.

Attrition poses two potential threats to our results. The first concern arises in comparing performance across the algorithms. It may be, for instance, that inspectors were more likely to inspect restaurants ranked highly by the data-rich and data-poor algorithms. The differences we observe across methods could then be driven primarily by not observing the outcomes for restaurants ranked lower by the data-rich and data-poor algorithms; we test against this concern in Column 1 of Table 4 by looking for differences in average ranking by targeting method for restaurants that were inspected.

The point estimates suggest that there is a slight bias in favor of restaurants ranked highly by the data-rich and data-poor algorithms. Only the data-poor algorithm shows a statistically significant difference in average ranking, coming in 1.5 positions higher on average than under business-as-usual. There is no statistically significant difference in comparing average rankings among inspected restaurants targeted by the data-rich and business-as-usual methods. Overall, the differences in ranking of inspected restaurants are small. Moreover, Column 2 of Table 1 suggests that the violation levels did not differ significantly by ranking position. Thus, the performance differences observed in Column 1 of Table 1 are unlikely to be driven primarily by differences in the ranking positions of restaurants inspected under each method.

A second potential threat to validity is differences in characteristics across restaurants ranked by each method in Table 2. If we observe different parts of the ranking distribution for each docketing method, we may misattribute differences in ranking for differences in characteristics. Column 2 of Table 4 tests against this concern by looking for differences in average ranking by docketing method for restaurants that were ranked. Here we see a similar pattern to Column 1 of Table 4, albeit with even smaller magnitudes. Given the high magnitude of differences in restaurant characteristics observed in Tables 2 and 3, we do not see Column 2 of Table 4 as a major threat to validity.

## 5. Discussion

Our study shows that at least for the case of restaurant hygiene inspections, using predictive algorithms can significantly improve efficiency. Even a simple algorithm based on internal data improved predictions relative to inspector discretion; moreover, a simple algorithm provided gains nearly as large as those from a more data-rich algorithm. But we also found compliance to be a first-order issue: inspectors frequently chose to prioritize restaurants based on their own judgment, rather than dockets based on our algorithms – potentially because of algorithm aversion (Dietvorst et al. 2015), differing objectives, or simply strong priors.

Our findings show a clear role for algorithms, but also highlight that improving organizational efficiency is rarely just a matter of better prediction.  Academics are often interested in intellectually challenging problems, such as improving predictive accuracy or policy nuance, but in many practical settings, basic capacity – such as the willingness of workers to follow directions – is far more important. Workers with discretion will not comply with algorithms that ignore workers' priors and preferences, which means that algorithms must either cater to those preferences or managers must tether worker discretion with targeted incentives.

## References

Agarwal, A., Gans, J., and Goldfarb, A. Prediction Machines: The Simple Economics of Artificial Intelligence

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

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. American Economic Review, 106 (5), 124–127.

City of Boston (2017). Inspectional Services. Retrieved 2017-04-01 from https://www.boston.gov/departments/inspectional-services

Cowgill, B. (2018). Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening. Working Paper.

Dawes, R. M. (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), 146–147.

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

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. Journal of Experimental Psychology: General, 144 (1), 114–126.

Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2016a). Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. American Economic Review, 106 (5), 114–118.

Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2016b). Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. NBER Working Paper No. 22124.

Grove, W., & Meehl, P. (1996). Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy. Psychol Public Policy 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.

Jung, J, Shroff, R., Feller, A., & Goel, S. (2018). Algorithmic Decision Making in the Presence of Unmeasured Confounding. Working Paper.

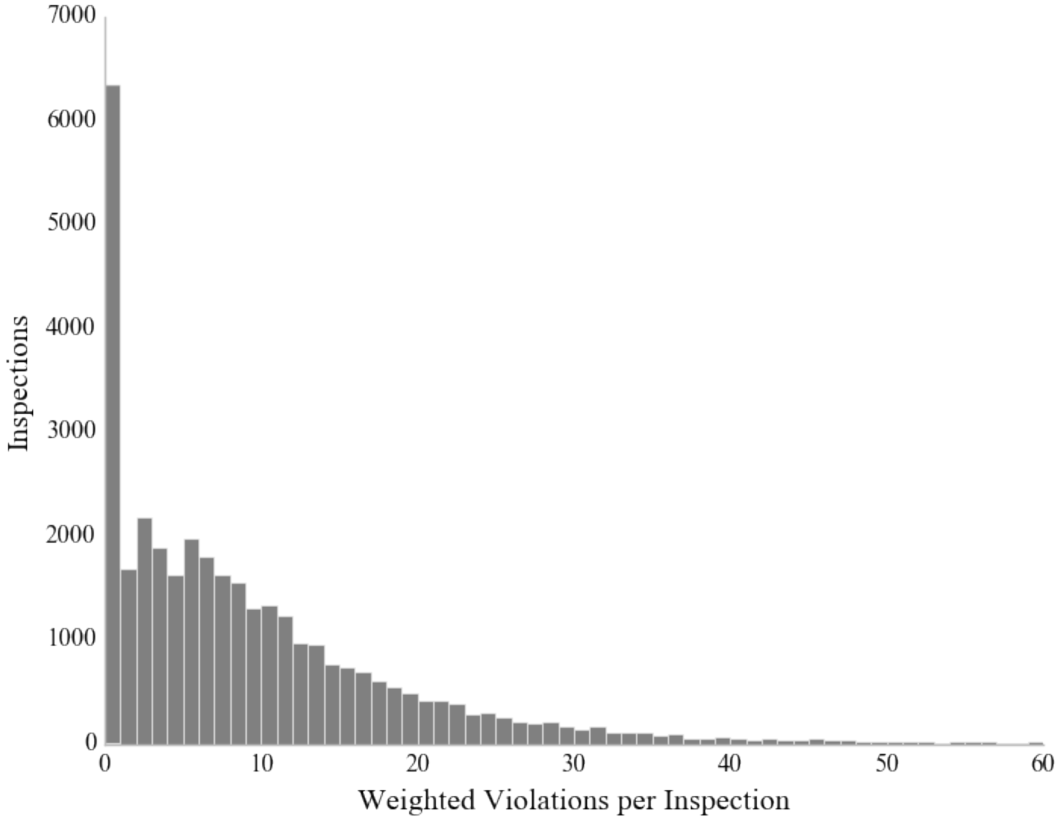
Kleinberg, J., Lakkaraj, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and Machine Predictions. NBER Working Paper No. 23180.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 275–284.

Stevenson, Megan (2017). "Assessing Risk Assessment in Action," George Mason Law & Economics Research Paper, (17-36), 4.

**Figures**

Figure 1: Distribution of Violations



The figure shows the distribution of weighted violations across inspections from January 2007 through June 2015. (Level I has a weight of 1; Level II, a weight of 2; and Level III, a weight of 5.)

## Tables

Table 1: Performance of Rankings

| Outcome: | (1)<br>Total Violations<br>b/se | (2)<br>Total Violations<br>b/se |
|---|---|---|
| Ranked by Data-rich Algorithm | $4.508^{***}$ | $3.365^{**}$ |
| | (0.792) | (1.354) |
| Ranked by Data-poor Algorithm | $4.023^{***}$ | 3.741 |
| | (0.928) | (2.698) |
| Data-rich Algorithm x Rank | | 0.107 |
| | | (0.120) |
| Data-poor Algorithm x Rank | | 0.024 |
| | | (0.229) |
| Rank | | -0.029 |
| | | (0.113) |
| Constant | $8.190^{***}$ | $8.519^{***}$ |
| | (0.942) | (0.982) |
| Observations | 372 | 372 |
| Including Ranking Up To: | 20 | 20 |

The sample consists of 372 restaurant-condition observations. Only restaurants ranked within the top 20 by any condition are included. Total violations are a weighted sum of one, two, and three star violations.
* p< 0.1, ** p<0.05, *** p<0.01.

Table 2: Mean Characteristics of Restaurants Ranked by Each Approach

| | (1)<br>Data-rich<br>Algorithm | (2)<br>Data-poor<br>Algorithm | (3)<br>Inspector<br>Discretion | (4)<br>p: 1=2 | (5)<br>p: 1=3 |
|---|---|---|---|---|---|
| Chain | 0.321 | 0.365 | 0.494 | 0.301 | 0.011** |
| Review Average | 3.438 | 3.352 | 3.494 | 0.092* | 0.409 |
| % Trusted Reviews | 0.786 | 0.798 | 0.795 | 0.014** | 0.285 |
| Review Count | 141.041 | 175.232 | 172 | 0.026** | 0.208 |
| Ethnic Cuisine | 0.538 | 0.542 | 0.319 | 0.133 | 0.000*** |
| Restaurant Delivers | 0.612 | 0.529 | 0.41 | 0.002*** | 0.009*** |
| Price Range | 1.632 | 1.596 | 1.573 | 0.5 | 0.462 |
| Take Out Offered | 0.967 | 0.974 | 0.945 | 0.835 | 0.413 |
| Restaurant Age | 18.938 | 20.527 | 28.645 | 0.433 | 0.157 |

Columns (1)-(3) show the mean value of each characteristic across restaurants ranked by each approach. Column (4) shows p-values testing whether the mean in column (1) is significantly different from that in column (2); column (5) shows p-values testing whether the mean in column (1) is significantly different from column (3).

Table 3: Inspector Compliance

| | % of Top 20 Rankings Observed | % of Observed Rankings Inspected |
|---|---|---|
| Data-rich Algorithm | 94.44 | 31.18 |
| Data-poor Algorithm | 92.78 | 32.33 |
| Business-as-Usual | 71.94 | 61.01 |

Column 1 shows the percent of restaurants that were ranked by each method. Column 2 shows the percent of restaurants inspected among those that were ranked by each method.

Table 4: Differences in Ranking

| | (1) | (2) |
|---|---|---|
| Outcome: | Ranking | Ranking |
| | b/se | b/se |
| Data-rich Algorithm | -0.750 | -0.285 |
| | (0.630) | (0.185) |
| Data-poor Algorithm | -1.501** | -0.584** |
| | (0.600) | (0.243) |
| Constant | 11.278*** | 10.629*** |
| | (0.429) | (0.150) |
| Observations | 372 | 933 |

The sample consists of restaurant-condition observations ranked within the top 20 by any condition. Column 1 further restricts the sample to restaurants that were inspected.

* p< 0.1, ** p<0.05, *** p<0.01