Consumer Reviews and Regulation: Evidence from New York City Restaurants

December 28, 2024

Abstract

We investigate the informational content of online reviews regarding hygiene standards. Using data from Yelp and New York City restaurant inspections, we show that online reviews are informative about dimensions of hygiene that consumers directly experience, but provide limited information on other dimensions. We present causal evidence that restaurant demand responds to hygiene signals in online reviews and that restaurants with higher visibility on review platforms violate less along dimensions for which reviews are more informative. These findings highlight the evolving role of consumer ratings for regulators and firms.

1 Introduction

Both online reviews and government regulation can inform consumers about the quality of service providers. Government regulation, such as occupational licensing and health and safety inspections, has historically been used to screen providers in settings in which choosing a low-quality provider can put consumers' health and safety at risk (Arrow, 1963; Akerlof, 1970). But as consumers increasingly rely on online reviews, a new question arises: can online reviews inform consumers about the dimensions of quality traditionally monitored by regulators? The answer depends both on whether the reviews contain such information and on whether consumers pay attention to it.

We explore this question in the context of New York City restaurant health inspections. We use detailed inspection records from the New York City Department of Health and Mental Hygiene, which we combine with consumer reviews from Yelp and reservation data from OpenTable. We find that reviews are more informative about hygiene violations that consumers directly experience, such as pests and food temperature, than about less-visible violations, such as worker hygiene and facility maintenance. We also show that both consumers and restaurants take hygiene information from reviews into account when deciding, respectively, where to eat or how much to invest in hygiene. Our results highlight both the potential and the limitations of online reviews with respect to informing consumers about dimensions of quality also monitored by regulators.

Our empirical analysis proceeds in two steps. First, we use a novel measure of restaurant hygiene to extract interpretable signals of hygiene violations from Yelp reviews. Second, we causally link these signals with changes in a restaurant's demand and its hygiene efforts.

In the first step, rather than measuring hygiene based on the score assigned by inspectors to each violation, as has typically been done in the literature, we propose a measure based on violation incidence. We show that our measure is less susceptible to inspector discretion. We then use machine learning methods to extract interpretable signals of hygiene violations from review text and to assess how well these signals predict the violations found by health inspectors. Key to our methodology is interpretability; we aim to extract signals that consumers can plausibly use to identify hygiene violations. We find substantial variation in the predictive accuracy of online reviews for different types of violation. For more accurately predicted violations, our algorithms identify keywords that serve as intuitive predictors, such as the word "nauseous" for violations related to food temperature.

In the second step, we estimate the impact of these hygiene signals on a restaurant's demand and on its hygiene efforts. Due to data limitations, our analysis of restaurant demand is restricted to the smaller subset of restaurants listed on OpenTable that sell out at least once during the sample period. Using an event study approach around the submission time of reviews containing signals of poor hygiene, we find that restaurants are significantly less likely to sell out in the weeks following such reviews. We estimate that about half of the impact of these negative reviews on restaurant demand can be attributed to the poor hygiene signals. Finally, we use an instrumental variable approach to provide evidence that restaurants take into account the signaling role of online reviews when choosing their hygiene level; restaurants more visible on Yelp tend to violate less along hygiene dimensions for which reviews contain more informative signals.

Overall, our findings suggest that online reviews are informative about a subset of hygiene violations that regulators monitor, above and beyond the effect that regulation already has in reducing information asymmetries in this market. Consumers consult reviews regularly and reviews are submitted more frequently than regulatory inspections, at least for certain businesses. Reviews are also cheap to collect, making them a valuable addition to the regulatory tools historically used to protect consumers from risky transactions.

However, our findings also suggest that, for many violations monitored through regulation, online reviews do not inform the public. This is an important consideration amidst recent debates over the ability of online platforms to self-regulate, which Cohen and Sundararajan (2015) define as "the reallocation of regulatory responsibility to parties other than the government."

Taken together, our results have practical implications for various stakeholders. First, we highlight the possibility of extracting specific quality signals from review text. Although we focus on how consumers use this information, others can apply this approach to quality dimensions other than hygiene. Regulators for instance, can use online information to better target their monitoring efforts, e.g., with respect to working conditions. Firms can use the information in online reviews to monitor the compliance of staff or suppliers with specific standards. For example, a franchiser could use the signals we have constructed to identify franchisees falling short of hygiene standards. Online platforms that collect reviews could apply our methods to automatically extract signals of violations and alert consumers (Dai and Luca, 2020).

Second, the very fact that both restaurants and their customers respond to hygiene information contained in online reviews implies that, despite existing health and safety regulations, important information asymmetries remain. We show that online reviews, being easy for consumers to evaluate, are a valuable addition to existing regulatory monitoring for hygiene dimensions. However, we also show how limited consumer monitoring can be, particularly for hygiene dimensions that consumers cannot easily observe or lack the expertise to assess. We discuss these issues in the conclusion.

We organize the paper as follows. Section 2 describes our contributions relative to existing work. Section 3 describes our empirical context and the data used for analysis. Section 4 presents our approach to predicting health violations from the text of Yelp reviews. Section 5 estimates the effect of review-based hygiene signals on demand and supply incentives. Section 6 discusses the limitations and implications of our work.

2 Related work

Our work brings together three strands of literature largely separate until now. The first focuses on using online reviews to predict the aggregate outcomes of regulatory inspections (i.e., the total health scores assigned by inspectors after tallying points assigned to specific violations). Prior research finds that online reviews and health inspection scores are broadly correlated (Kang et al., 2013; Harrison et al., 2014; Mejia et al., 2019), but does not account for the possibility that inspection scores may themselves deviate from the true hygiene of a restaurant. We offer a new approach to deriving hygiene conditions from health inspections that is less susceptible to inspector discretion than previously used measures.

Within the first strand, previous applications of algorithms predicting inspection scores from reviews have focused on enhancing expert decision-making (Glaeser et al., 2016; Kim et al., 2024) or detecting moral hazard between infrequent health inspections (Mejia et al., 2019). Our objective differs. First, instead of taking restaurant hygiene as one-dimensional, we try to identify distinct dimensions (e.g., pests, worker hygiene) for which reviews contain informative signals. We thus identify violations that are easier to predict using consumer reviews and therefore better candidates for online quality disclosure. Second, whereas prior work uses machine learning as a black box, we focus on interpretable methods to uncover hygiene signals in reviews that consumers can plausibly associate with restaurant hygiene. While broadly negative words might be predictive of worse hygiene scores, these words are unlikely to inform consumers (or regulators) about specific violations. Worse yet, consumers might associate these broadly negative words with aspects of poor quality that regulators are not concerned with (e.g., staff attitude towards customers), rather than unsafe hygiene conditions. Our approach allows us to construct hygiene signals that are orthogonal to the presence of generally negative words. The results reveal that violations for which Yelp is informative are predicted by words related to the substance of the specific violation. For example, we find the word "roach" to be a key predictor of violations pertaining to cockroaches, whereas the word "nauseous" is a key predictor of violations related to not keeping food at safe temperatures. By using review text as a signal of restaurant quality, our work also follows a more recent research trend that uses text as data in a broad set of applications (Taddy, 2013, 2015; Gentzkow et al., 2019a,b; Greenstein et al., 2021).

The second strand of literature to which we contribute focuses on the market effects of the quality signals in reviews. Dating back to Chevalier and Mayzlin (2006), a large body of work has shown the role that online reviews play in consumer choices. In particular, Luca (2016) and Fang (2022) show that high online ratings increase restaurant revenues. Firms therefore have strong incentives to respond to online reviews by changing prices (Lewis and Zervas, 2016), advertising strategies (Hollenbeck et al., 2019), and entry and exit decisions (Bao et al., 2024). Research has especially focused on one side-effect of online reputation: fake reviews (Mayzlin et al., 2014; Luca and Zervas, 2016; He et al., 2022).

Research has so far been agnostic as to what type of information reviews actually disclose to consumers, focusing instead on the aggregate numeric rating that reviewers assign to service providers. But such ratings reflect reviewers' overall satisfaction on several quality dimensions; the extent to which they capture restaurant hygiene will therefore depend on the relative weight reviewers place on that dimension.[1] Using the text of reviews allows us to break down a reviewer's overall assessment of a restaurant and separate out the causal effect of hygiene information on demand, an approach that can be extended to other quality dimensions. We show that hygiene is important to restaurant patrons and that its effect, when present, is a large component of the effect of negative ratings.

The third strand of literature our work contributes studies the role of government in reducing asymmetric information and moral hazard. Pioneering work by Jin and Leslie (2003, 2009) has shown the value of mandated hygiene disclosure in shifting consumer demand towards more hygienic restaurants and in incentivizing restaurants to improve their hygiene. Similar benefits have been identified in an online context (Jin et al., 2022). But there is also evidence that, in equilibrium, regulatory monitoring may neither achieve the anticipated quality improvements (Kugler and Sauer, 2005; Barrios, 2022) nor push demand towards certified providers (Farronato et al., 2020).

Our work adds to this literature by testing whether online reputation systems reduce information asymmetries for quality dimensions that are already monitored by inspectors. This is a first step in a larger debate around the need, given alternative incentive mechanisms such as online reputation, for regulatory restrictions to protect consumers (Shapiro, 1986). Exploring additional mechanisms of regulatory monitoring is important given the substantial discretionary power of inspectors (Ibanez and Toffel, 2020), the fact that inspectors' grades are often coarse, inflated, and infrequently updated (Mejia et al., 2019), and the risk that regulation may have anti-competitive effects (Kleiner and Soltas, 2019; Blair and Fisher,

[1] For example, Lehman et al. (2014) show that ratings are less susceptible to unsanitary conditions for restaurants perceived as more "authentic."

2022).

The analyses in this paper help inform the debate over the regulation of online marketplaces that rely heavily on consumer reviews to screen and monitor providers. Recent papers have focused on the welfare benefits of flexible labor for both providers (Chen et al., 2019; Farronato and Fradkin, 2022) and consumers (Cohen et al., 2016; Farronato et al., 2020; Farronato and Fradkin, 2022). Still, little is known about the role of reviews in disclosing information about the quality provided by provider already subject to regulatory monitoring (Einav et al., 2016). We find that reviews can provide useful information about a small subset of dimensions of regulated quality. When that information is available, consumers and restaurants take it into account, above and beyond the effects of regulation.

3 Data

This section describes the three data sources we rely on for our analyses: health inspections, consumer reviews, and restaurant availability.

Our first dataset contains detailed records of health inspections and was obtained from the New York City Department of Health and Mental Hygiene (DoH) through a Freedom of Information Act request. The DoH conducts unannounced inspections of food-serving establishments in the five boroughs of New York City, with each establishment inspected at least once a year. Inspectors check for compliance across many hygiene dimensions, including food handling, employees' personal hygiene, and vermin control. There are over 100 violation codes. Inspectors assign points to each violation cited, with more points assigned to more severe violations. These points are tallied to compute an inspection score,[2] which is used to assign a letter grade, A through C (Appendix Figure A1). The restaurant must post the letter grade at its entrance for customers to see.[3]

We have data at the level of each violation code for all inspections conducted between July 2010, when the most recent overhaul of restaurant inspections was implemented, and September 2016. The DoH inspection dataset also contains restaurant-level characteristics, such as cuisine type, whether the restaurant is part of a chain, date of entry and exit, and anonymous inspector identifiers for each inspection. We confirmed with the DoH that

[2] The rules for assigning points during a health inspection are available at `http://www1.nyc.gov/site/doh/services/restaurant-grades.page`.

[3] The restaurant can temporarily post a "Grade Pending" card if it decides to dispute a B or C grade at an administrative tribunal. More details on inspection regulation and grading can be found at `http://www1.nyc.gov/assets/doh/downloads/pdf/rii/inspection-cycle-overview.pdf`. In principle, consumers could access the list of violations found during an inspection by visiting the DoH website. Anecdotally, this rarely happens.

assignment of inspectors is random.[4]

The second dataset includes consumer reviews from Yelp.com. Yelp provides business information, such as address and phone number, and a historical record of reviews, including text, time of submission, and an identifier of the reviewer.[5] We also collect the entire set of reviews by each user who ever submitted a review for a New York City restaurant. In Section 5.2, we use this information to construct instruments for the number of reviews restaurants receive, based on user-level propensities to review businesses online.

Our third dataset comes from OpenTable, an online platform that allows consumers to make restaurant reservations. We have daily information between 2013 and 2016 on whether each restaurant on OpenTable had a table available for two people between 6:30 pm and 7:30 pm. Each morning, we checked for table availability for the same evening.[6]

We match the three datasets based on restaurant name, address, and phone number. Overall, of the 49,034 individual restaurants present in the DoH data, 61.3% were present on Yelp[7] and 4.5% were present on OpenTable. Table 1 presents restaurant-level descriptive statistics for the three samples. Relative to all restaurants inspected by the DoH, restaurants with Yelp reviews tend to be more concentrated in Manhattan, are less likely to be fast-food restaurants, and are less likely to go out of business within our sample period. That is even more true for restaurants on OpenTable.

3.1 Identifying Hygiene Conditions from Restaurant Inspections

To evaluate whether online reviews contain information about restaurants' hygiene means predicting hygiene conditions from online reviews. We therefore need a credible measure of the ground truth. Because hygiene conditions are not observable to academic researchers, research has relied on inspection outcomes as a proxy for the ground truth. In this section, however, we show that such proxies (Mejia et al., 2019; Kim et al., 2024) are likely to be unreliable due to sizable bunching around grade thresholds. We propose a novel measure of hygiene and show that it is more likely to reflect true hygiene conditions.

New York restaurants undergo inspections in cycles. A cycle starts with an initial inspection, which is sometimes followed by a reinspection and compliance inspections that can lead

[4] See Appendix B for an analysis of inspectors' assignment to restaurants.

[5] Reviews that Yelp deems fake are not displayed online and do not count towards a restaurant's average rating (Luca and Zervas, 2016).

[6] Over the course of three years, our data collection occasionally failed due to technical issues outside our control (e.g., network outages, website updates.) This resulted in a few short gaps in our OpenTable panel. We do not have any reason to believe that these gaps are systematically correlated with the outcomes we study.

[7] This is not because it is difficult to match restaurants, but rather because the food-serving establishments that the DoH inspects include some that are less likely to be on Yelp, such as workplace cafeterias.

Table 1: Restaurant Characteristics

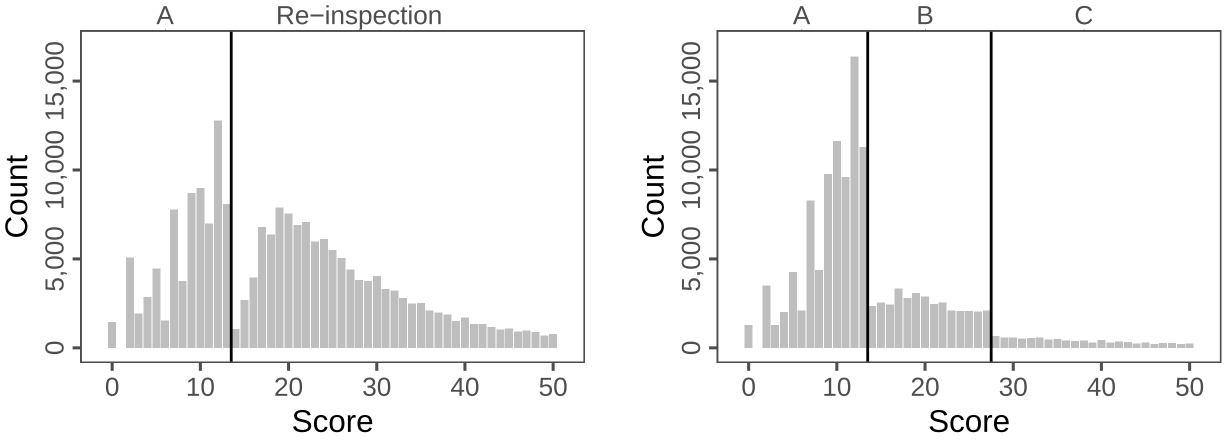
Characteristics	Restaurants All (1)	 On Yelp (2)	 On OpenTable (3)
Cuisine - American	22.8%	22.3%*	31.2%*
Cuisine - Cafe/Bakery	7.3%	7.4%	0.4%*
Cuisine - Chinese	11.3%	10.9%*	0.8%*
Cuisine - Italian	3.5%	4.8%*	19.2%*
Cuisine - Latin/Mexican	6.8%	6.1%*	6.2%
Cuisine - Pizza	6.7%	7.6%*	1.9%*
Boro - Bronx	10.0%	6.5%*	1.0%*
Boro - Brooklyn	25.3%	25.3%	11.8%*
Boro - Manhattan	37.3%	43.8%*	84.2%*
Boro - Queens	23.4%	20.5%*	2.3%*
Boro - Staten Island	3.9%	3.8%	0.7%*
Venue - Bar/Pub	5.4%	5.4%	2.9%*
Venue - Fast Food	9.0%	8.9%	0.1%*
Venue - Restaurant	54.5%	64.0%*	94.8%*
Share Chain	10.8%	11.4%*	1.0%*
Share Closed	47.6%	37.7%*	12.9%*
Share Newly Opened	49.5%	48.9%*	53.0%*
N Restaurants	49,647	30,447	2,215

Summary of restaurant characteristics for the three samples: all restaurants inspected by the New York City Department of Health, restaurants with Yelp reviews, and restaurants on OpenTable. The stars in Columns 2 and 3 indicate statistically significant differences in means compared to Column 1 at the 5% confidence level.

to a letter-grade update. Figure A1 in the Appendix explains how inspection cycles work. The time between cycles depends on the hygiene conditions of the restaurant during its prior inspection, with more frequent monitoring for restaurants with poorer hygiene.[8] In practice, there is substantial variability in the time between inspections (Appendix Figures A2 and A3), due to inspectors' wish to show up unannounced. There is also substantial variation in the grades that restaurants receive in consecutive cycles (Appendix Table A1). For example, only 41% of restaurants with an A-grade maintain the A at the following initial inspection; the other 59% are reinspected. This fact makes it difficult to predict a given restaurant's hygiene conditions at a given time.

Figure 1a shows the distribution of scores at initial inspection. On average, 36% of restaurants receive an A-grade on initial inspection (corresponding to a score of 13 points

[8] If a restaurant scores fewer than 14 points at initial inspection, it receives an A-grade and will be inspected again in about a year. If a restaurant scores 14–27 points at initial inspection and gets a A- or B-grade at reinspection, it will be inspected 5–7 months after the most recent reinspection. If the restaurant scores 28 points or more at initial inspection or gets a C-grade at reinspection, it will be inspected 3–5 months after the most recent compliance inspection—that is, a follow-up inspection to check that specific critical violations have been resolved. A list of critical violations is at `http://www1.nyc.gov/assets/doh/downloads/pdf/rii/blue-book.pdf`.



(a) Initial inspection

(b) Reinspection

Figure 1: Distribution of Inspection Scores

or less), whereas the rest require reinspection. Reinspections occur within a few weeks and an inspector (likely a different one) shows up unannounced.[9] Looking at the distribution of scores, we observe significant bunching at the score of 13, the threshold above which restaurants are reinspected.

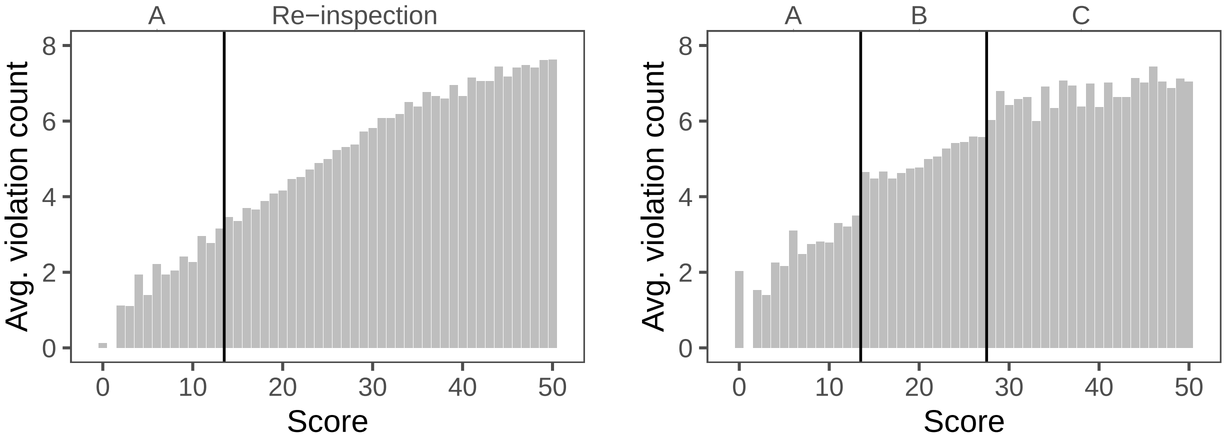
Figure 1b shows the distribution of scores for reinspected restaurants. After re- inspection, the vast majority receive an A-grade. Indeed, 77% end an inspection cycle with an A-grade, 17% with a B-grade, and only 6% with a C-grade. Compared to initial inspection scores, reinspection scores exhibit even more pronounced bunching at 13 and 27 points—the thresholds between different letter grades.

The bunching cannot be explained by restaurant characteristics. With the exception of chain affiliation, the effects of observables are relatively small and explain only about 8% of the variation in inspection scores (see Appendix Table A2). Inspector fixed effects alone explain more of that variation than restaurant characteristics do (R-squared in Columns 2 and 4 versus Columns 1 and 3 in Appendix Table A2). Even though the random allocation of inspectors to restaurants does not lead to systematic differences in inspection outcomes across restaurants, inspectors still have significant discretionary power during individual inspections.

These results suggest that inspection scores, whether at initial inspection or at reinspection, can deviate from true hygiene conditions when scores are near a threshold.[10] Thus,

[9] See Appendix Figure A2 for a distribution of the time lag between inspections and reinspections.

[10] While we do not know for sure that the distributions in Figure 1 deviate from the true underlying hygiene quality, the substantial bunching to the left of each threshold is consistent with research on inspectors' discretionary power (Ibanez and Toffel, 2020).



(a) Initial inspection

(b) Reinspection

Figure 2: Average Number of Violations per Inspection

machine learning models trained to predict total inspection scores may fail to accurately predict true hygiene conditions for inspections near letter-grade thresholds, even if they do predict hygiene scores well.

To construct a more accurate measure of hygiene, we rely on the observation that inspectors have control over the total inspection score on both the extensive and intensive margins. On the former, they have control over which violations they cite; on the latter, they choose a score for each violation they cite based on the extent of the violation. For instance, when an inspector cites violation 02B, the number of points depends on the number of food items outside the correct temperature range. A single item results in fewer points than two items. The range of permissible scores varies by violation, the typical range being four points. Some violations, such as not holding required paperwork, are binary in nature and can only be assigned a single score.

Our proposed measure of hygiene is based on initial inspections outcomes on the extensive margins. Rather than considering the scores, we simply focus on whether the inspector cites a restaurant for specific violations. Violation counts are less likely to be distorted in general and especially for initial inspections, as Figure 2 shows. For each inspection, we count the number of violations cited and plot the average number of cited violations by total inspection score. Figure 2a displays these results for initial inspections. Unlike Figure 1a, we see no bunching at the 13-point threshold. Figure 2b, which repeats the same analysis for reinspections, shows some bunching at both thresholds, though less pronounced than the bunching in Figure 1b.[11]

[11] The bunching in Figure 2b is to the right of the A-B threshold, whereas it is to the left in Figure 1. While the cause is unclear (for instance, inspectors may cite additional violation codes to justify scores

Overall, our results suggest that restaurant inspectors can exercise considerable discretion when evaluating hygiene. This makes total inspection and reinspection scores less accurate representations of true hygiene. Violation incidence—especially for initial inspections—is a more credible measure of hygiene conditions. In the next section, we use this measure as ground truth to extract hygiene signals from online reviews.

4 Do Online Reviews Contain Signals of Hygiene?

A first step to assess the value of online reviews in informing consumers about restaurant hygiene is extracting hygiene signals from review text. To do that, we need a measure of a restaurant's true underlying hygiene quality. In the previous section, we argued that violation incidence at initial inspection is the measure of restaurant hygiene least affected by inspector discretion. Building on this observation, we use machine learning methods to predict specific violations during initial inspections from the text of online reviews. The better Yelp reviews can predict a particular violation, the more informative we define Yelp to be about that particular dimension of hygiene. We focus our analysis on the 20 most frequent violation codes, which are listed in Appendix Table A3 and constitute over 80% of all violations cited during initial inspections.

To incorporate review text in our subsequent analysis, we need to reduce the text's dimensionality. Yelp reviews contain hundreds of thousands of unique words; using each as a covariate to predict inspection outcomes is impossible as we would have more covariates than observations. We solve this problem with an approach recently applied to analyze congressional speech by Gentzkow et al. (2019b), whose results were also used by Greenstein et al. (2021) to evaluate political slant in Wikipedia articles. We first extract signals of hygiene from text, then measure their informativeness.

4.1 Extracting Violation-specific Signals of Hygiene from Reviews

In the first step, we develop a model that learns what reviewers say when hygiene violations occur, then use this model to construct low-dimensional, violation-specific signals of hygiene from review text.

We begin by associating each initial inspection with reviews that were submitted up to three months before the inspection. There are two reasons for choosing this threshold.

just to the right of the threshold), this bunching does not affect our analyses as we focus solely on initial inspections, which display a smooth distribution around the threshold (Figure 2a). Conditional on total inspection scores, violation counts do not impact re-inspections or letter grades, so they likely reflect hygiene quality more accurately.

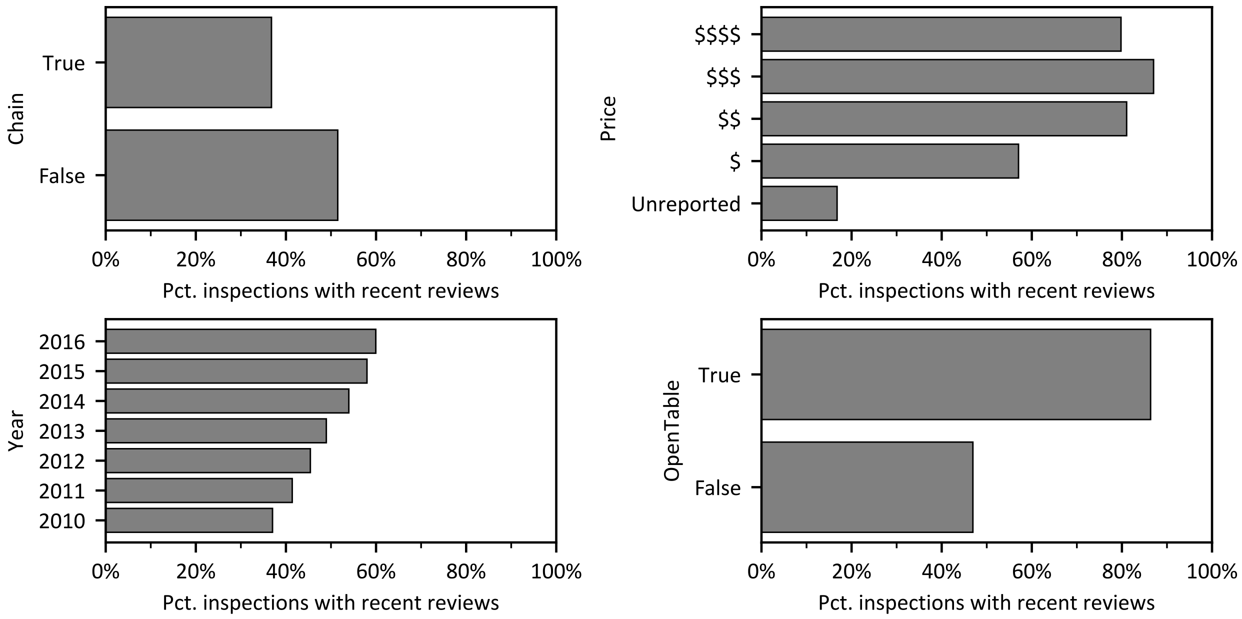


Figure 3: Percentage of Inspections with Recent Reviews by Restaurant Type

First, online reviews are not very frequent, so a longer time interval can capture more heterogeneity across restaurants. Second, the minimum time between inspection cycles is about three months, which allows us to associate each review with at most one inspection.

Since our methodology relies on recent reviews, we first check whether recent reviews are more likely to be associated with certain types of restaurants. Figure 3 shows the percentage of inspections with recent reviews by restaurant type. Inspections of chain restaurants are less likely to be associated with recent reviews than inspections of independent restaurants, consistent with earlier findings that chain restaurants are less likely to be reviewed on Yelp. The figure also shows that inspections of more-expensive restaurants are more likely to be associated with recent reviews. Restaurant on OpenTable are also more likely to have recent reviews, possibly because OpenTable lists popular restaurants. Finally, the figure shows that the share of inspections with recent reviews has been increasing over time, suggesting that our methodology will become more broadly applicable as review platforms become more popular. At the same time, these descriptives highlight the fact that our methodology is not applicable to a significant fraction of restaurants (about 40% as of 2016) and that the results we obtain are likely to be more representative of a selected group of restaurants that are popular online.

To construct our vocabulary of words associated with health inspections, we do not use every word as it appears on Yelp. First, we take the raw text of the reviews and eliminate punctuation and numbers. We then replace each word with its stem, using the Porter stemming algorithm (Porter, 1980). Finally, to exclude both common and rare words, we exclude stems that appear in fewer than five reviews or in more than 50% of the reviews. We end up with a vocabulary of 12,176 words.

In constructing our vocabulary, we perform a final preprocessing step best illustrated with an example. The word "clean" has a different meaning depending on the context in which it appears. Indeed, the presence of the word "clean" in a 1-star review likely implies "dirty." We therefore separately count word frequencies in three rating groups: 1- and 2-star reviews, 3-star reviews, and 4- and 5-star reviews. This triples the size of our vocabulary. In the rest of the paper, unless otherwise noted, we consider each word-rating-group combination as a separate *word* in our vocabulary.

The combined text of reviews submitted in the three months before each initial inspection constitutes a document, which is simply a collection of word counts in no particular order. We let $\boldsymbol{c}_i$ denote the observed vector of word counts in reviews associated with inspection i. We assume that $\boldsymbol{c}_i$ is drawn from a multinomial distribution

$$\boldsymbol{c}_i \sim \mathrm{MN}(\boldsymbol{q}_i, m_i), \tag{1}$$

where m_i is the document length—the total number of words in reviews linked to inspection i—and $\boldsymbol{q}_i$ is a vector of probabilities with length equal to the number of distinct words that consumers could use. The element q_{ij} is the probability of occurrence of word j in document i. Given the distributional assumption, we have $q_{ij} = \frac{e^{\eta_{ij}}}{\sum_k e^{\eta_{ik}}}$, where

$$\eta_{ij} = \mu_j + \boldsymbol{\alpha}_j \boldsymbol{r}_j + \boldsymbol{\phi}_j(\boldsymbol{r}_j \cdot \boldsymbol{v}_i) + \boldsymbol{\beta}_j \boldsymbol{x}_i + \epsilon_{ij}. \tag{2}$$

In the above equation, the coefficients of interest are contained in the vectors $\boldsymbol{\phi}_j$ (one vector per word); they tell us by how much the frequency of word j changes when each violation in $\boldsymbol{v}_i$ occurs. The vector $\boldsymbol{v}_i$ contains one indicator variable per violation code that is set to 1 in the presence of that violation and 0 otherwise. The terms $\boldsymbol{r}_j$ are rating-group dummies that allow word probabilities in positive (4-5 stars) and neutral (3 stars) reviews to vary compared to the baseline intensity of negative review (1-2 stars) word probabilities, which are captured by the intercept μ_j. For instance, we expect the word "great" to be more frequent in positive than in negative reviews. We interact $\boldsymbol{v}_i$ with the rating-group dummies $\boldsymbol{r}_j$ to flexibly capture changes in word frequency by rating group and violation code. For example, if violation code `04M` occurs (presence of roaches), we expect the frequency of the word "roach" to increase in negative reviews but not in positive reviews. The vector $\boldsymbol{x}_i$ includes various controls: year-month fixed effects, cuisine fixed effects, zipcode fixed effects, and a dummy for whether the restaurant is part of a chain.

Including a rich array of controls is important to isolate the direct impact of violations on word frequencies. Without such controls, the coefficients in $\boldsymbol{\phi}_j$ can pick up correlations between a restaurant's propensity to commit a specific violation and the restaurant's characteristics. For example, consider violation `02G`, which pertains to food not being kept cool enough. Sushi restaurants are more prone to violation `02G` because they serve raw food. Without controls for restaurant characteristics, we might infer that when violation `02G` occurs, the frequency of the word "sushi" goes up. Nevertheless, the word "sushi" does not, in and of itself, suggest that during a specific inspection the restaurant in question was more likely to be cited for violation `02G`. Including restaurant-specific controls helps avoid such spurious correlations.

Estimating this multinomial logit model is prohibitively expensive because the coefficients associated with each word in c_i depend on the coefficients of all other words (via the denominator of q_{ij}). We approximate the multinomial logit model with as many independent Poisson regressions as we have words, following Taddy (2015)'s distributed multinomial regression framework. This approximation makes the estimation tractable.

We estimate one regression per word by minimizing a penalized log-likelihood:

$$\min_{\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\phi}_j} \frac{1}{N} \sum_{i=1}^{N} l(c_{ij}, \eta_{ij}; \mu_j) + \lambda_j (\|\boldsymbol{\alpha}_j\|_1 + \|\boldsymbol{\beta}_j\|_1 + \|\boldsymbol{\phi}_j\|_1), \tag{3}$$

where l is the Poisson log-likelihood function and $\mu_j = \log \sum_i c_{ij}$ is an offset term that controls for the baseline intensity of each word as described in Taddy (2015). We apply a lasso penalty to enforce sparsity and avoid overfitting.[12] Lasso is natural in our setting, as we expect many coefficients for our controls to be zero. For example, we expect the dummy for Japanese cuisine to have a zero coefficient for the word "pizza."

We tune the word-specific penalty parameters λ_j using five-fold cross-validation. To avoid data leakage due to correlations within each restaurant's inspections, we divide our data in folds using blocked sampling by restaurant. This way, each restaurant's entire set of inspections ends up in the same fold. Then, for each word, we select the penalty that minimizes cross-validation error:

$$\lambda_j^{min} = \arg\min_{\lambda} \mathrm{CV}_j(\lambda) = \arg\min_{\lambda} \frac{1}{5} \sum_{k=1}^{5} \mathrm{CV}_{jk}(\lambda_j),$$

where $\mathrm{CV}_{jk}(\lambda)$ is the cross-validation error of fold k for word j evaluated at λ.

[12] We also apply a small penalty to the intercept term to aid convergence.

Results

The matrix of estimated coefficients $\hat{\mathbf{\Phi}}$ with entries $\hat{\phi}_{jv}$ tells us by how much the frequency of word j changes when violation v occurs. While $\hat{\mathbf{\Phi}}$ is already relatively sparse due to the lasso penalty, it contains too many non-zero entries to comfortably summarize. To further aid interpretation, we use a heuristic approach to extract the strongest predictive relationships between words and violations. Intuitively, our approach entails increasing the value of the lasso penalty until only very few non-zero entries remain in $\hat{\mathbf{\Phi}}$. These remaining entries correspond to the strongest predictors of word frequencies.

One complication we must deal with is that each word is estimated using a different lasso path. Thus, the same increase in penalty will induce different amounts of sparsity for different words. To solve this problem, we use a heuristic inspired by the "one standard error rule" (Hastie et al., 2009), which selects the most parsimonious model whose error is within one standard error of the minimum cross-validation error.

For each violation code v and word j, we compute by how many standard errors we would have to increase the minimum cross-validation error in order to make the coefficient on that violation-code dummy zero. Specifically, we compute the quantity

$$\gamma_{jv} = \arg\min_{\gamma} \mathrm{CV}_j(\lambda_j^{min}) + \gamma \mathrm{SE}_j(\lambda_j^{min}) \ s.t. \ \hat{\phi}_{jv} = 0,$$

where $\mathrm{SE}_j(\lambda) = \sqrt{\mathrm{Var}(\mathrm{CV}_i(\lambda), \dots, \mathrm{CV}_5(\lambda))/5}$ is the cross-validation standard error. Then, for each violation code v, we sort words in descending order γ_{jv}, which provides a ranking of the most predictable changes in word frequency when violation v occurs.

Table 2 displays the top-10 strongest relationships between violations and increases in negative review-word frequency as ranked by γ_{jv}. A few interesting patterns emerge. Looking at violation code `04M`, which pertains to roaches, we see that Yelp reviewers tend to increase their use of words like *roach* and *filth* in the three months leading up to the violation being uncovered by a health inspector. A similar pattern appears for violations `02B` and `02G`, which pertain to keeping food at an appropriate temperature and are predicted by words like *sick*, *nauseous*, and *poison.* If a consumer were to read reviews containing these words, we might expect them to correctly predict that a restaurant has roaches or that food has gone bad. By contrast, consider violation code `10F`, towards the bottom of the list, which relates to surfaces that do not directly come into contact with food, but are still part of the environment where food is prepared. A priori, we might not expect the average Yelp reviewer to know what materials or methods are permitted for the construction of these surfaces. Looking at the words that increase in frequency prior to this violation occurring, we observe changes in generally negative words that are unlikely to alert a consumer that

Table 2: Most Predictive Words for Each Violation Code

02B	poison	dept	hung	sick	nauseou	grubhub	tasteless	overcook	phone	smh
04H	racist	bouncer	gratuiti	incompet	lame	smh	tab	she	atroci	bartend
04M	roach	filth	filthi	risk	homeless	health	diarrhea	disgust	cook	waiter
04A	driver	groupon	phone	deliveri	smh	call	hung	refund	order	horribl
02G	ined	dept	bland	poison	wors	tasteless	apolog	gross	downhil	refus
10H	salvag	moron	mandatori	remak	confront	insult	dept	threaten	unwarr	coat
06D	bouncer	gratuiti	disrespect	downhil	manag	terribl	rude	apolog	horribl	overcook
04N	tourist	blvd	cashier	she	manag	ask	hire	smh	filthi	overpr
06E	tgifriday	inexcus	limp	taim	sambal	cockroach	horrend	driver	deplor	seamlessweb
06C	slimi	dept	ined	zero	tasteless	bland	nerv	hung	gross	phone
10B	dissatisfi	insult	gratuiti	bland	gross	poison	worst	overpr	incompet	disgust
08A	gratuiti	incompet	refus	gross	nasti	ined	unaccept	terribl	downhil	groupon
06F	mortifi	irat	scrap	blandest	health	mush	violent	undercook	argu	audac
04L	dimsum	fraud	nickel	demean	sanitari	calmli	abomin	spa	grubhub	hung
09C	eat24	inconvenienc	fraud	spa	microwav	quinn	chipotl	hostil	mash	seamless
10F	incompet	smh	refus	disrespect	rudest	nasti	refund	wors	unaccept	attitud
05D	groupon	coat	vomit	phone	bouncer	apologet	tomato	inconvenienc	deliveri	health
06A	dept	smh	stench	drove	trap	hung	unsanitari	avoid	inattent	unhygien
08C	debacl	snide	ghetto	spanish	cop	session	disrespect	smh	threaten	groupon
04J	spa	wack	hostil	unsanitari	townhous	aggrav	indiffer	sandwich	wors	blatantli

For a description of the violation codes, see Appendix Table A3. Violation codes are sorted as in Figure 4.

non-food contact surfaces are improperly constructed. We may thus expect that Yelp would contain more informative signals for violations such as 04M and 02B at the top of Table 2, for which the text is descriptive of the actual violation, compared to 10F at the bottom, for which text is much less specific. We confirm these differences in Section 4.2.

Constructing Low-dimensional Signals of Hygiene

For each inspection and violation code, we map the text contained in all negative reviews occurring up to 90 days before the inspection to a one-dimensional index, which will be higher when the reviews preceding the inspection contain many words typically associated with the violation in question.

To compute these indices for a single inspection i, we multiply the matrix of estimated coefficients $\hat{\boldsymbol{\Phi}}$ with the vector of word frequencies $\boldsymbol{c}_i$ to obtain

$$\boldsymbol{z}_i = \hat{\boldsymbol{\Phi}} \boldsymbol{c}_i,$$

which is a vector with 20 entries, one for each violation code. These indices are known in the literature as sufficient reduction (SR) projections (Taddy, 2015) because they project text onto attributes of interest, which in our application are violation code dummies. Intuitively, the SR projections $\boldsymbol{z}_i$ are weighted sums of word counts with higher weights associated with words that are more predictive of specific violations. A key property of these SR projections is that they are sufficient statistics for the violation codes, i.e., $\boldsymbol{v}_i \perp \boldsymbol{c}_i | \boldsymbol{z}_i$. In words, given the low-dimensional SR projections $\boldsymbol{z}_i$, the high-dimensional vector of text $\boldsymbol{c}_i$ is orthogonal to the violation code dummies (and can thus be conditionally ignored). This property of SR projections allows us to reduce the dimensionality of text from thousands of words down to a single index for each violation code that captures variation in review text specifically pertaining to the occurrence of that violation.[13]

4.2 Evaluating the Informativeness of Yelp Hygiene Signals

Next, we evaluate the informativeness of Yelp hygiene signals constructed as SR projections. Our basic approach is to compare the predictive power of two classifiers predicting violations: a *baseline classifier* that relies exclusively on DoH hygiene signals and a *review-augmented classifier* that uses both DoH and Yelp hygiene signals. The objective of these classifiers is to approximate what a consumer might learn about a restaurant's hygiene from Yelp reviews, above and beyond what regulatory monitoring already signals through letter grades. To achieve this goal, we maximize prediction accuracy subject to the constraint of using interpretable machine learning methods.[14] Algorithm 1 describes in detail the steps we take to build and evaluate these two classifiers. Next, we discuss a few key components of our algorithm.

A key decision we have to make is which features to include in each classifier. This decision requires assumptions regarding the information sets to which consumers have access when choosing where to eat. For the baseline classifier, we use the letter grade posted at the restaurant door at the time of the inspection, which is what a customer would see.[15] The review-augmented classifier adds two signals from Yelp: the restaurant's average star-rating on the day of the inspection and the SR projections that we constructed from recent reviews.

[13] Our approach differs substantially from that of Mejia et al. (2019) in that: (a) we use a more accurate measure of hygiene ground truth, (b) words are not selected by the researcher to be predictive of hygiene or not, (c) we allow for different words to predict different violations (e.g., the word "roach" predicts pests violations, whereas "nausea" predicts violations related to food temperature), (d) instead of relying on word frequencies to measure hygiene, we construct an index that weighs words differently depending on their individual importance in predicting hygiene violations while controlling for a rich set of confounders, and (e) we estimate the causal effect of our index on the outcomes of interest—namely, a restaurant's sold-out probability and hygiene violation incidence—whereas Mejia et al. (2019) use word count as a proxy to identify moral hazard by restaurants.

[14] It is not our sole objective to maximize prediction accuracy. If it were, we would (a) use more flexible but less interpretable methods such as neural networks and (b) include predictors to which consumers are unlikely to have access (e.g., identifiers for health inspectors, which are very predictive of inspection results as per Appendix Table A2).

[15] This may be an unrealistic assumption for sophisticated consumers who rely on richer information sets to evaluate restaurant hygiene. For example, certain consumers might rely on information from prior visits or from friends. Or, they might look up a restaurant's entire history of inspections in the DoH database. Anecdotal evidence suggests that most consumers do not use the DoH database.

When computing SR projections for each inspection, we avoid data leakage by excluding all inspections of the restaurant whose violations we are trying to predict. To see how data leakage can arise, recall that SR projections associate violations with changes in word frequencies via the learned projection matrix $\hat{\mathbf{\Phi}}$. If we included the focal inspection to learn the projection matrix $\hat{\mathbf{\Phi}}$, we would be peeking at the outcome we are trying to predict, resulting in data leakage and overstated classifier accuracy. To prevent leakage, we make careful use of cross-validation. We divide our data into five folds, with each restaurant's entire set of inspections assigned to the same fold. For a restaurant in, say, the first fold, we use the other four folds to estimate $\hat{\mathbf{\Phi}}$ and rely on the first fold to construct the SR projections and evaluate their predictive power. In other words, the projection matrix $\hat{\mathbf{\Phi}}$ for each restaurant is only learned from inspections and reviews of other restaurants.

To predict violation incidence, we train gradient-boosted tree classifiers (Ke et al., 2017), as described in Algorithm 1.[16] We evaluate the performance of the baseline and review-augmented classifiers using the AUC (area under the curve) metric.[17] The lowest possible AUC value, 0.5, means that our classifier performs as well as a random guess.

Figure 4a displays AUCs for each violation code and for each of the two classifiers separately.[18] All AUC metrics range between 0.51 and 0.68 suggesting that it is relatively difficult for consumers to predict the incidence of individual hygiene violations. This can be due to Yelp reviews not being able to capture hygienic conditions, but it is also possible that the inspection itself is a noisy signal of hygiene.

Although letter grades are in general a poor predictor of specific violations—all but one AUC are below 0.55 for the baseline classifier (white bars in Figure 4a)—we observe more variation in the performance of the review-augmented classifier (grey bars), with some violations being easier to predict than others. Figure 4b displays the improvement in AUC of the review-augmented classifier relative to the baseline classifier. We use this improvement as the measure of informativeness of Yelp reviews for each violation code. The vast majority of the improvement comes from review text, rather than from star-rating (see Appendix Figure A7.)

Comparing the violation codes ranked higher in Figure 4b with those ranked lower, it becomes apparent that Yelp reviews tend to be better predictors of violations such as vermin,

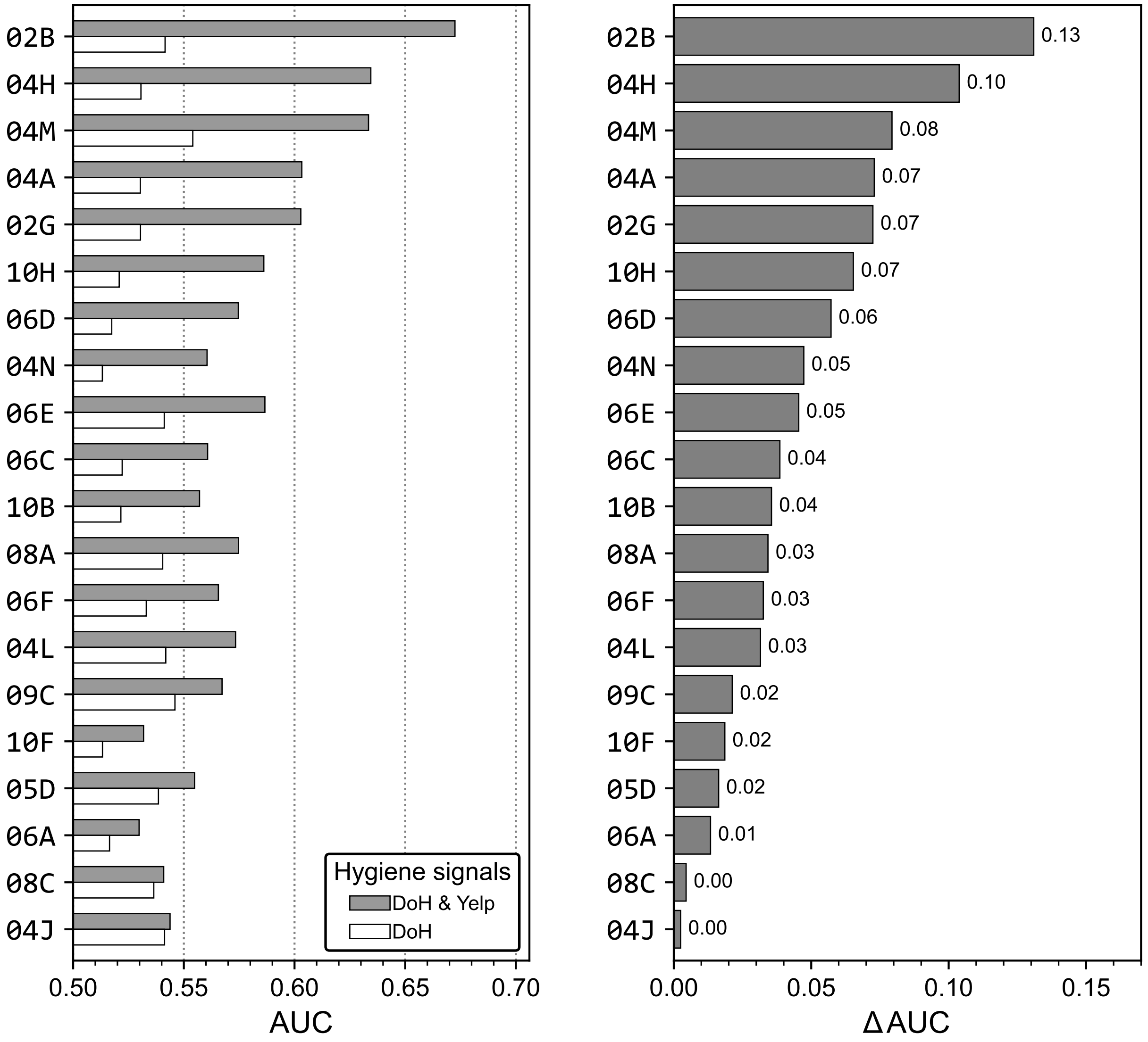
[16] We obtain similar results with a penalized logistic regression.

[17] AUC is a ranking metric: given a pair of inspections belonging to different classes (in our case, an inspection in which the violation occurred and another in which it did not), we assign the value 1 to the pair if the predicted probability of the positive case is higher than that of the negative case, and 0 otherwise. AUC averages these values over all possible positive-negative pairs.

[18] The AUC of a classifier trained only on Yelp reviews and excluding DoH grades is only slightly lower than the AUC of the same classifier trained on both Yelp reviews and DoH grades. This suggests that, given Yelp reviews, the DoH letter grades add little in terms of predictive performance.

Algorithm 1: Nested cross-validation to compare out-of-sample performance for predicting violation code v with and without hygiene signals constructed from Yelp reviews.

```
Input: Data $\mathcal{D} = [\mathcal{D}_V, \mathcal{D}_H, \mathcal{D}_R, \mathcal{D}_C]$ with one row per inspection, where $\mathcal{D}_V$ are violation
       dummies, $\mathcal{D}_H$ are health grades assigned by the DoH, $\mathcal{D}_R$ are Yelp average ratings, and
       $\mathcal{D}_C$ are Yelp review word counts.
Output: $\mathrm{AUC}_0, \mathrm{AUC}_1$: CV AUC without and with Yelp review hygiene signals
/* outer loop to evaluate performance                                              */
Divide $\mathcal{D}$ in 5 folds using block-sampling by restaurant;
for each fold $k_1 \leftarrow 1$ to 5 do
    /* We use $\mathcal{D}^k$ to denote data belonging to fold $k$, and $\mathcal{D}^{-k}$ for data belonging to
       all other folds                                                               */
    $\mathcal{G} \leftarrow \mathcal{D}^{-k_1}$ ; /* outer loop train folds                                    */
    Divide $\mathcal{G}$ in 5 folds using block-sampling by restaurant;
    /* inner loop to tune hyper-parameters                                           */
    for each fold $k_2 \leftarrow 1$ to 5 do
        for each set of hyper-parameters $h \in H$ do
            /* To avoid data leakage, the SR projection matrix is estimated using
               train folds, and then used to construct SR projections for both train
               and test folds                                                        */
            Estimate SR projection matrix $\hat{\Phi}(\mathcal{G}_C^{-k_2})$ on inner train folds $\mathcal{G}_C^{-k_2}$ using methodology
              described in Section 4.1;
            $Z^{-k_2} \leftarrow \hat{\Phi}(\mathcal{G}_C^{-k_2})\mathcal{G}_C^{-k_2}$ ; /* SR projections of inner train folds      */
            $Z^{k_2} \leftarrow \hat{\Phi}(\mathcal{G}_C^{-k_2})\mathcal{G}_C^{k_2}$ ; /* SR projections of inner test fold         */
            Train violation classifier without Yelp signals $\hat{g}_0(\mathcal{G}_V^{-k_2}, \mathcal{G}_H^{-k_2}; h)$;
            Compute AUC of $\hat{g}_0$ for test fold $k_2$;
            Train violation classifier with Yelp signals $\hat{g}_1(\mathcal{G}_V^{-k_2}, \mathcal{G}_H^{-k_2}, \mathcal{G}_R^{-k_2}, Z^{-k_2}); h)$;
            Compute AUC of $\hat{g}_1$ for test fold $k_2$;
        Compute average CV AUC of each classifier (with and without text signals) for
          hyper-parameters $h$;
    Select $h_0^*$ and $h_1^*$ that minimize the average CV AUC of the two classifiers;
    Estimate SR projection matrix $\hat{\Phi}(\mathcal{D}^{-k_1})$ using outer train folds $\mathcal{D}^{-k_1}$;
    $Z^{-k_1} \leftarrow \hat{\Phi}(\mathcal{D}_C^{-k_1})\mathcal{D}_C^{-k_1}$ ; /* SR projections of outer train folds          */
    $Z^{k_1} \leftarrow \hat{\Phi}(\mathcal{D}_C^{-k_1})\mathcal{D}_C^{k_1}$ ; /* SR projections of outer test fold             */
    Train violation classifiers $\hat{f}_0(\mathcal{D}_V^{-k_2}, \mathcal{D}_H^{-k_2}; h)$ and $\hat{f}_1(\mathcal{D}_V^{-k_2}, \mathcal{D}_H^{-k_2}, \mathcal{D}_R^{-k_2}, Z^{-k_2}); h)$;
    $\mathrm{AUC}_{0,k_1} \leftarrow$ AUC of $\hat{f}_0$ for test fold $k_1$;
    $\mathrm{AUC}_{1,k_1} \leftarrow$ AUC of $\hat{f}_1$ for test fold $k_1$;
/* Compute average CV AUC of each classifiers                                        */
$\mathrm{AUC}_0 \leftarrow \frac{1}{5}\sum_{k_1=1}^{5} \mathrm{AUC}_{0,k_1}$; /* AUC without Yelp hygiene signals                     */
$\mathrm{AUC}_1 \leftarrow \frac{1}{5}\sum_{k_1=1}^{5} \mathrm{AUC}_{1,k_1}$; /* AUC with Yelp hygiene signals                        */
```



(a) AUC by violation code for two classifiers (with and without signals extracted from Yelp reviews). For a comparison with two other classifiers, which independently use star-ratings and review text, see Appendix Figure A7.

(b) Difference in AUC between the two classifiers in Figure 4a. For a full description of the violation codes, see Appendix Table A3.

Figure 4: Prediction Performance

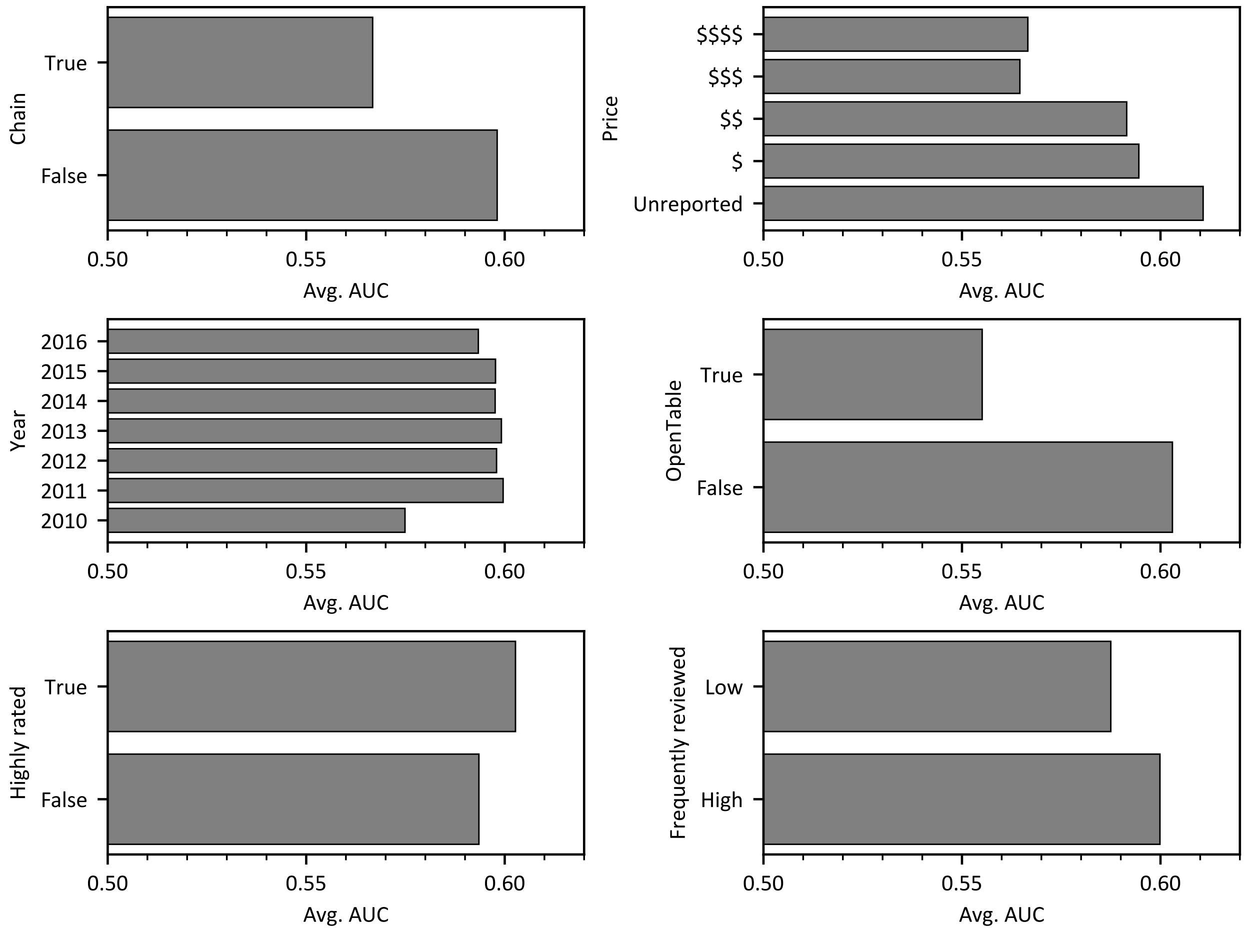


Figure 5: Prediction Performance by Restaurant Type

food temperature, and food handling than of violations relating to pesticides, construction materials, and certifications. It is reassuring to see that the violations that we can predict more accurately are the violations we would intuitively expect consumers to be most likely to notice.

Higher prediction accuracy across hygiene dimensions does not automatically imply that Yelp readers can infer the same information. However, the interpretability of our predictors, as shown in Table 2, suggests that our algorithm may approximate what readers glean from reviews. The words most predictive of the top violations are often descriptive of the actual infractions. For example, roach is a strong predictor of `04M` (live roaches present), and nauseous is highly associated with `02B` (Hot food item not held above 140^o F). Importantly, while the association between hygiene violations and changes in word frequency is correlational, our event study provides causal evidence that changes in the frequency of these predictive words do impact restaurant demand.

Next, we evaluate the performance of the review-augmented classifier by restaurant type.

Figure 5 shows that the review-augmented classifier performs better for independent restaurants, cheaper restaurants, and restaurants that are more highly rated and more frequently reviewed. The performance of the classifier does not vary significantly over time. It performs better for restaurants not listed on OpenTable, a fact we return to at the end of Section 5.1.

Overall, the results in this section point to one main conclusion: consumers discuss restaurant hygiene on Yelp, but not all dimensions of hygiene are equally captured by their reviews. Indeed, reviews tend to better capture violations that consumers have a direct experience with, such as pests or food handling. In the next section, we study whether the information about restaurant hygiene contained in Yelp reviews affects restaurant demand and restaurants' incentives around hygiene.

5 Effects of Hygiene Signals on Demand and Supply

To confirm that our exercise from Section 4 picks up information that consumers and providers actually take into account when choosing where to eat, we provide evidence that Yelp hygiene information affects consumer choices and restaurant incentives.

A simple demand-and-supply framework with asymmetric information helps motivate the hypotheses we test below. Imagine that consumers are uncertain about restaurants' hygiene and online reviews serve as informative signals for some of those hygiene dimensions. Bayesian consumers will use online reviews to update their beliefs about hygiene and choose where to eat. Subsection 5.1 explores whether consumer choices are indeed affected by the hygiene signals in Yelp reviews.

If demand is affected by hygiene information on Yelp, restaurants will take that into account and invest in hygiene quality. But not all restaurants are equally exposed to scrutiny through online reviews, so we should expect more-exposed restaurants to invest more in hygiene quality. Also, since online reviews are only informative about certain dimensions of hygiene, restaurants will only need to adjust those dimensions.[19] These observations lead to the difference-in-differences approach we implement in Subsection 5.2.

[19] We abstract away from restaurants' pricing decisions; i.e., prices are assumed to be fixed and not a function of restaurant quality. This is not an innocuous assumption, but one that is needed to avoid multiplicity of equilibria. It is a reasonable one in a context in which restaurants make short-term quality decisions—such as whether to install mousetraps or how frequently to clean kitchen counters—that are unlikely to affect menu prices. Our assumption is supported by the fact that the National Restaurant Association does not include hygiene maintenance as an important factor when considering whether to raise prices (http://www.restaurant.org/Manage-My-Restaurant/Marketing-Sales/Food/Is-it-time-to-raise-your-prices). Other factors, such as food ingredients and labor, constitute larger shares of overall costs and, despite the volatility of those costs, restaurants tend to have stable menu prices (http://smallbusiness.chron.com/restaurant-food-pricing-strategies-14229.html).

5.1 Consumer Demand

We use the probability of being sold out on OpenTable as our demand proxy to analyze the effect of hygiene information in online reviews on demand. The advantage of using sold-out probability is that it is a measure of restaurant success that changes daily and thus allows us to look at changes in demand immediately following a particular review. Even if OpenTable is only one channel through which consumers reserve a restaurant, being sold out on OpenTable is an indicator of increased demand from all channels. The drawback is that we have this outcome for only a small subset of restaurants (see Table 1 for selection on observables).[20]

To estimate demand, one may be tempted to regress sold-out probability on each quality signal—the letter grade posted at the restaurant door, average ratings, and our hygiene signal—to compare its relative influence on demand. However, all three quality signals are likely endogenous. To estimate the causal impact of our hygiene signal on demand, we take advantage of the submission time of Yelp reviews. Assuming that the *timing* of a negative review is exogenous, we can use an event-study approach and compare the probability that a restaurant sells out in the days just before and after a review that discusses its poor hygiene.

To identify reviews discussing poor hygiene, we begin by ranking violations by how informative Yelp reviews are about them, as described in the previous section (Figure 4b). We then restrict attention to the top five violations for which Yelp is most informative:

- `02B` (hot food item not kept at or above 140^o F),
- `04H` (raw, cooked, or prepared food is adulterated, contaminated, cross-contaminated, or not approriately discarded),
- `04M` (live roaches in food and/or non-food areas),
- `04A` (Food Protection Certificate not held by supervisor),
- `02G` (cold food item kept above 41^{o}F except during preparation).

While the choice to focus on the top five violations is arbitrary, our results do not depend on it. Appendix Table A5 presents estimates for other sets of violations.

We construct a hygiene signal for restaurant i on a given day by summing the sufficient reductions of these five violation codes contained in 1-, 2-, and 3-star reviews (which we refer to as low-star reviews) that were submitted for that restaurant that day. Recall that each sufficient reduction approximates the probability that the corresponding violation occurs.

[20] Another minor drawback is that 1.8% of restaurants never sell out and, given the presence of restaurant fixed effects, must be excluded from estimation.

The sum of multiple sufficient reductions collectively highlights the corresponding violations. A higher hygiene signal constructed this way means worse conditions. It can originate from one review discussing one hygiene dimension from the list above in a very negative way, one review discussing several of the five hygiene dimensions in the list, or even multiple reviews submitted on the same day, each discussing one or more hygiene dimensions.

We define focal events as days on which (a) a restaurant receives at least one low-star review and (b) the aggregate hygiene signal contained in that day's reviews is among the 20% most negative among all low-star reviews. Because low-star reviews are relatively rare, the time windows around focal events do not tend to overlap. We consider a 30-day window around focal events to estimate the following regression:

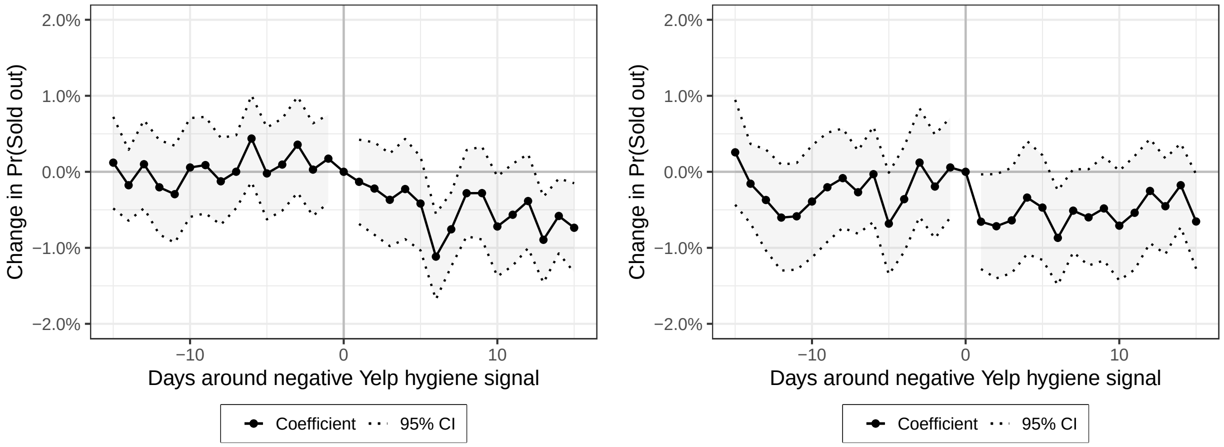
$$\text{sold\_out}_{ijt} = \sum_{t=-15}^{15} \beta_t + \boldsymbol{X}_{ijt} \cdot \boldsymbol{\alpha} + \epsilon_{ijt}. \tag{4}$$

The subscript i denotes a restaurant, j denotes the set of reviews submitted on a given day, and t denotes the number of days since the event. The outcome, $sold\_out_{ijt}$, equals 1 if restaurant i, which received review(s) j at $t = 0$, is sold out between 6:30pm and 7:30pm t days since the focal event. We look at a month around the focal event, so t goes from -15 days to +15 days. The coefficient on the day of the submission β_0 is normalized to zero. The control variables, $\boldsymbol{X}_{ijt}$, include restaurant-review fixed effects (which translate to a unique fixed effect for each focal event), day-of-week fixed effects, Yelp average star ratings, and the hygiene card displayed at the door. This vector is designed to account for demand fluctuations across seasons and restaurants, systematic differences by day of the week, and variations driven by time-varying quality signals, such as star ratings and hygiene cards. We cluster standard errors at the restaurant level.[21]

We present results in the left plot of Figure 6, which shows a decrease in the probability of selling out that is gradual at first and reaches its minimum around a week later. We can aggregate the days before the focal event and those after it to estimate a single "post-review" coefficient. We present these results in the first column of Table 3. We find that following the negative hygiene signal, restaurants experience a 0.6 percentage-point decline in their probability of being sold out. This is equivalent to a 2.9% reduction in the average sold out probability of 0.19.[22]

[21] Given the presence of high-dimensional fixed effects, we estimate linear probability models as opposed to logistic regressions. Linear probability models are reasonable approximations to logistic models when the relationship between the probability of the outcome (a reservation slot being sold out) and the log odds of the same outcome is approximately linear, which we have confirmed to be the case for our data.

[22] Note that the estimated effect relies on the assumption that consumers evaluating a restaurant read the most recent reviews. But not all consumers read reviews and even those who do may only read the



Left panel plots the β_t estimates from Equation 4. Right panel plots the γ_t estimates from Equation 5. Standard errors are clustered at the restaurant level. Table 3 provides regression results in which the day-level estimates are replaced by a dummy variable for whether the day is after the submission of the focal review(s).

Figure 6: Yelp Hygiene Signals and Sold-out Probability—Event Study

To control for the possibility that this effect is driven by characteristics of the low-star reviews other than the poor hygiene signal they contain, we perform the event study analysis as a difference-in-differences, comparing the probability of selling out for restaurants receiving a low-star review with an especially poor hygiene signal (as we defined the focal events above) against a baseline of all other restaurants receiving a low-star review:

$$\text{sold\_out}_{ijt} = \sum_{t=-15}^{15} \beta_t + \sum_{t=-15}^{15} \gamma_t \cdot \text{bad\_hygiene\_signal}_{ij} + \boldsymbol{X}_{ijt} \cdot \boldsymbol{\alpha} + \epsilon_{ijt}. \tag{5}$$

Relative to Equation 4, the new specification includes all days when restaurants receive low-star reviews as focal events, but we also interact the fixed effects for each day since the focal event with a dummy for whether the reviews on the focal day contain any of the 20% most negative hygiene signals. The coefficients of interest are γ_t, which measure the change in sold-out probability relative to a restaurant experiencing a low-star review without such a negative hygiene signal. We plot the γ_t coefficients in the right panel of Figure 6. While these estimates are noisier, we continue to find a decrease in the probability of selling out.

Aggregating all days before the event and all days after the event leads to a single difference-in-differences coefficient, displayed in Column 2 of Table 3. This estimate confirms

top-ranked ones. Yelp determines the order in which reviews are shown as a function of "recency, user voting, and other review quality factors, which is why an older review may appear before a newer one" (see `https://www.yelp-support.com/article/How-is-the-order-of-reviews-determined`). The effect we estimate is therefore mediated by Yelp's role in determining a review's position. It is thus conceivable that the effect could be even larger if reviews were ranked only by recency.

Table 3: Yelp Hygiene Signals and Sold-out Probability

	Sold Out on OpenTable	
	(1)	(2)
After Review	-0.006^{***}	-0.002^{***}
	(0.001)	(0.0005)
Bad Yelp Hygiene Signal*After Review		-0.003^{***}
		(0.001)
Day of Week FE	Yes	Yes
Restaurant-Review FE	Yes	Yes
Observations	534,511	2,644,279
Adjusted R^2	0.500	0.498
Note:	*p<0.1; **p<0.05; ***p<0.01	

Estimates of modified Equations 4 and 5. In Column (1), the β_t coefficients from Equation 4 are replaced by a coefficient on a dummy for whether the day occurs after the focal review(s); i.e., $t > 0$. In Column (2), the β_t and γ_t coefficients from Equation 5 are similarly aggregated. Standard errors clustered at the restaurant level.

that restaurants receiving a negative hygiene signal in a low-star review are 0.3 percentage points less likely to be sold out in the two weeks following the review. The difference-in-differences coefficient estimate is 58% of the coefficient estimate from Column 1 (0.0032/0.0055), implying that about half of the reduction in sold-out probability following bad reviews is attributable to the hygiene information they contain.

To confirm that this result is not simply due to online reviews capturing information already disclosed by letter grades, we can restrict attention to reviews submitted in a period in which the letter grade displayed at the door is "*A*," either because the restaurant experienced no inspections during that period or because the restaurant could continue to display an A-grade after an inspection occurring around the same time. Of the review submissions considered in the baseline specification, 83% are in this category. Appendix Table A4 confirms that the coefficients of interest and their standard errors remain the same, suggesting that hygiene signals in online reviews contain information that is not captured—or at least not as frequently—by the formal inspection process.

Although we have ruled out that the effect of review hygiene signals is due to concurrent drops in letter grade, it is still possible that letter grades have a similar or even larger effect on demand. To test that, we can use a similar event-study approach around the time when a restaurant's grade drops below "*A*." Results (presented in Appendix Table A8) fail to reject the hypothesis of a null effect of letter grades on sold-out probability.

The estimated effects of review signals on demand are robust to a number of additional checks. First, we verify that our results do not change if, instead of using all low-star reviews as the control group, we use low-star reviews with the 20% least-negative hygiene signals. Second, because reviews are more likely for restaurants with high demand, we remove five days around the review date (from two days before to two days after) to avoid bias due to mean reversion. Third, we use more- and less-stringent definitions of bad hygiene signals, by selecting the 10% and the 30% worst signals, respectively, as our treated groups. Appendix Table A4 shows that the difference-in-differences coefficients do not change for these different specifications. Finally, because the decision to sum the sufficient reductions of the five violation codes for which Yelp is most informative is somewhat arbitrary, we progressively add the sufficient reductions of violation codes for which Yelp is less and less informative. We start from a sufficient reduction based on a single violation and end with the sum of the sufficient reductions of all 20 most-frequent violation codes. Each time, we reestimate the difference-in-differences specification. The coefficients of interest are presented in Appendix Table A5. A couple of coefficients are statistically indistinguishable from 0; all others are around -0.003 or -0.004, as in the baseline presented in Column 2 of Table 3.

Overall, our results confirm that, when choosing restaurants, demand responds to specific and interpretable hygiene signals in the text of online reviews. This result adds to the existing evidence focusing on how demand is affected by numeric review scores (Chevalier and Mayzlin, 2006) and by mandated inspections and public disclosure (Jin and Leslie, 2003). We note, however, two limitations of our results. First, the need for a demand proxy that varies daily constrained us to the small subset of New York City (restaurants available on OpenTable. On one hand, to the extent that OpenTable users are more likely to rely on online information when selecting restaurants, we would expect our results to be an overestimate of the effect of hygiene signals across all New York City restaurants. On the other hand, as we have shown in Figure 5, Yelp reviews are more informative regarding hygiene for non-OpenTable restaurants, suggesting that our results may instead underestimate the effect of hygiene signals across all New York City restaurants. Second, our outcome focuses on the extensive margins (sold out or not), limiting our ability to convert these effects into revenues or number of customers. Nonetheless, our result is useful in that it sheds new light on how consumers make choices and which dimensions of quality they care about (in this case, hygiene). New data-tracking efforts, such as Safegraph,[23] make it possible to extend our approach to more restaurants, more local businesses, and more geographies.

[23] https://www.safegraph.com/. This data is not available for our period of study.

5.2 Restaurants' Hygiene Incentives

In this section, we explore whether the negative hygiene information present in consumer reviews and consumers' response to it influence restaurants' hygiene efforts.

We take advantage of our finding from Section 4 that the informativeness of Yelp signals differs across violation codes and of the fact that Yelp's search algorithm controls how visible a restaurant is to consumers, directly affecting how exposed it is to information contained in Yelp reviews. When consumers search for restaurants in a given location, Yelp identifies which are to be displayed and in what order. Even if a restaurant is on Yelp, it can be difficult to find it if it does not appear in search results or if it is ranked low. So we expect consumers to be more responsive to reviews for restaurants that are ranked higher by Yelp's search algorithm.

We do not have access to Yelp's ranking algorithm, but the more recently reviewed a restaurant is, the more likely it is to be ranked higher in search results. (Appendix C shows that these correlations are sizable.) We therefore proxy for a restaurant's Yelp visibility with a dummy for whether it has been reviewed in the last 90 days. As long as restaurants know that Yelp reviews make them more visible, this variable has the advantage of being salient to restaurants. Yelp can automatically notify restaurants of new reviews, making review submission a practical and informative indicator of their visibility online.[24] We want to test the following hypothesis: restaurants that are more visible on Yelp violate less along hygiene dimensions for which Yelp provides a more informative signal, compared to restaurants less visible on Yelp and compared to hygiene dimensions for which Yelp is less informative. We consider the hygiene conditions at initial inspections and run difference-in-differences regressions of the following form:

$$\begin{split}\text{violation}_{vit} = \alpha \cdot \text{has\_recent\_reviews}_{it} + \beta \cdot \text{yelp\_informative}_{v} + \\ \gamma \cdot (\text{has\_recent\_reviews}_{it} \cdot \text{yelp\_informative}_{v}) + \boldsymbol{X}_{vit} \cdot \boldsymbol{\delta} + \epsilon_{vit}\end{split} \tag{6}$$

where violation$_{vit}$ equals 1 if restaurant i was found violating code v during initial inspection t. The dummy $has\_recent\_reviews_{it}$ is equals 1 if the restaurant has received any Yelp reviews in the 90 days before inspection t. The variable $yelp\_informative_v$ equals 1 for the top five violation codes for which Yelp is most informative, as defined in Section 5.1. We provide three specifications. The first has no controls, so the vector $\boldsymbol{X}_{vit}$ is empty. In the second, $\boldsymbol{X}_{vit}$ includes inspection fixed effects and violation-code fixed effects, which imply that $\alpha \cdot has\_recent\_reviews_{it} + \beta \cdot yelp\_informative_v$ are dropped from Equation 6 because of multi-collinearity. The two sets of fixed effects control for (a) inspection-level

[24] See `https://docs.developer.yelp.com/docs/reviews-webhooks-v2`.

characteristics—such as the identity of the inspector, seasonality, and recent efforts by the restaurant to be clean—that affect all violation codes equally; and (b) violation-code-level determinants of cleanliness that are common across restaurants and time periods. In the final specification, we add violation-code–restaurant fixed effects to control for the possibility that each restaurant has a different propensity to comply with each violation code. Appendix Table A6 provides summary statistics of the variables used in this analysis. In particular, it shows that there remains substantial variation in both the outcome and the main explanatory variable after controlling for the most stringent set of fixed effects.

In Equation 6, the coefficient of interest is the difference-in-differences coefficient γ, which measures the propensity to violate along dimensions of hygiene for which Yelp is more informative by recently reviewed restaurants compared to other restaurants and other hygiene dimensions. In the specification with the most stringent set of controls, γ measures a restaurant's propensity to violate on specific hygiene dimensions, conditional on its overall hygiene level and on inspector effort on that particular day and conditional on restaurant-specific time-invariant factors that make it easier or harder to comply with a specific hygiene code. We expect γ to be negative: recent reviews on Yelp should decrease a restaurant's violation rate on codes for which Yelp reviews are informative.

We present coefficient estimates in Panel *A* of Table 4. The first column displays results with no controls; the last displays results with inspection fixed effects and violation-code–restaurant fixed effects. The estimate in Column 1 implies that restaurants with recent reviews on Yelp are also 0.6 percentage points less likely to be found violating codes for which Yelp reviews are informative. Controlling for inspection-specific and violation-specific factors does not change the estimate of interest (Column 2). Adding violation-code–restaurant fixed effects in Column 3 reduces the coefficient to 0.2 and makes it indistinguishable from zero.

Despite the stringent controls, unobservable characteristics may affect both a restaurant's propensity to receive reviews and its hygiene level during an inspection. For example, cleanliness and Yelp reviews may both result from the restaurant's effort to increase its appeal. Our specification would suffer from omitted-variable bias if this effort affected hygiene dimensions for which Yelp is more informative more than others. This seems possible: a restaurant could temporarily increase advertising to attract consumers while making sure that the front of the house is clean to impress new customers who then submit Yelp reviews. In such cases, advertising (which could also lead to new reviews) and cleanliness would be co-determined by a restaurant's strategy to improve its appeal. Another plausible threat to causal identification is reverse causality: a restaurant could temporarily improve its hygiene (visible in the inspection outcomes), which in turn could attract customers who then leave Yelp reviews.

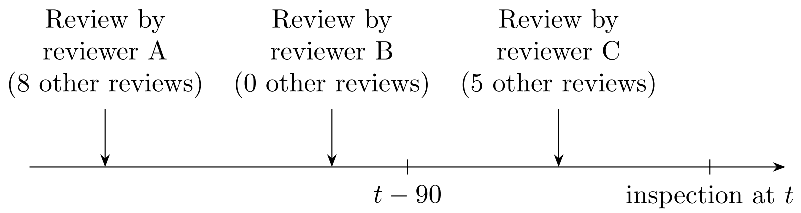


Figure 7: Timeline of Yelp Reviews and Health Inspection

To address these types of endogeneity, we take advantage of reviewers' behavior across the entire Yelp platform and construct an instrument based on *past* reviews. We leverage the idea that if a restaurant has been reviewed by users who frequently review businesses on Yelp, it is more likely to also receive reviews around the time of a health inspection. Specifically, we identify each reviewer of a focal restaurant and define their propensity to rate businesses online as the average number of Yelp reviews they have submitted, excluding the focal review.

To provide some intuition, Figure 7 shows three reviewers of a focal restaurant: reviewer A, who submitted eight other Yelp reviews; reviewer B, who submitted no other reviews; and reviewer C, who submitted five other reviews. We isolate the 90-day period preceding a health inspection to the focal restaurant. The endogenous variable $has\_recent\_reviews_{it}$ equals 1 if the restaurant has received reviews in this 90-day period. In Figure 7, $has\_recent\_reviews_{it} = 1$ because reviewer C submitted one review. We instrument for $has\_recent\_reviews_{it}$ with the average rating propensity of all reviewers of the focal restaurant who have rated it up until time $t - 90$. In Figure 7, the instrument is equal to 4; that is, the average review propensity of reviewers A and B. Because we interact recent reviews with whether Yelp is informative for a particular violation code, we effectively have two endogenous variables: $has\_recent\_reviews_{it}$ and $has\_recent\_reviews_{it} \cdot yelp\_informative_v$. We use the main instrument and its interaction with $yelp\_informative_v$. When the main instrument cannot be constructed because there are no reviews up to 90 days before the inspection (this happens for 25% of the inspections, as per Appendix Table A6), we add a dummy for the event indicating that there are no previous reviewers and interact it with $yelp\_informative_v$. When there are prior reviewers, only 0.2% of the inspections have a review propensity of 0. Given the skewed distribution of review propensity (Appendix Table A6), we use the log of review propensity plus 1 in our main specifications and demonstrate the robustness of our results to this transformation in the Appendix.

This instrument is valid if two conditions hold. First, the relevance condition, which we can test, requires that past reviewers' behavior is correlated with the propensity of current customers to leave a review. A likely mechanism driving this correlation is one in which certain restaurant characteristics (e.g., type of cuisine, distinctive ambiance) consistently attract customers more inclined to share their experiences online. Second, the exclusion restriction requires that the review propensity of *past* reviewers is not correlated with *current* hygiene efforts except through its effect on the likelihood that a restaurant receives new reviews on Yelp. Essentially, the propensity to review of past reviewers (going as far back as Yelp allows us to go and up until 90 days before the focal inspection) affects whether the restaurant is more visible on Yelp today, but otherwise neither affects nor is affected by the restaurant's concurrent hygiene efforts. This assumption appears valid, as restaurant hygiene frequently varies between inspections (see Appendix Table A1, discussed in Section 3), while customer types tend to remain more stable. Still, to the extent that hygiene conditions display some persistence—which may correlate with customers' characteristics (particularly, their propensity to use Yelp)—our stringent list of covariates allow us to control for this risk. In particular, violation-code–restaurant fixed effects control for the possibility that, for some restaurants, certain dimensions of hygiene are intrinsically easier to comply with than others for reasons that may be correlated with demand characteristics. Inspection fixed effects control for the fact that a particular restaurant may be more likely to comply with hygiene at certain times than at others because of its own efforts, the time of the inspection, or the identity of the health inspector.[25]

We present the IV results in Panel *B* of Table 4. The first stage for our IV estimates is in Appendix Table A7; the Kleiberger-Paap Wald F statistic, which tests whether instruments are weak for both our endogenous variables while adjusting for clustered standard errors, allows us to reject the null hypothesis of weak instruments (Stock and Yogo, 2005). The IV coefficients are all larger in absolute value and statistically different from the OLS estimates in Panel *A*. The IV estimates imply a 1.2 percentage-point reduction in the propensity to violate, or a 7% decrease off the baseline probability. This effect occurs while holding regulatory monitoring constant, so it should be interpreted as an effect above and beyond any role that regulation may play in restaurants' hygiene incentives. This result provides support for our hypothesis that exposure on Yelp makes restaurants clean up along hygiene

[25] We use the *average* number of reviews submitted by reviewers of the focal restaurant, rather than the *total* review count to avoid potential endogeneity. The total number of reviews submitted by these reviewers up to 90 days before an inspection is closely tied to the focal restaurant's own review volume, creating a mechanical relationship with restaurant demand and, potentially, its performance. This relationship may be confounded by events that jointly affect demand and hygiene, such as management changes, which can increase customer interest while also improving operational standards over time. In contrast, the average propensity of reviewers to post reviews is less directly tied to the focal restaurant's demand. It reflects the general reviewing behavior of those who chose to review the restaurant, independent of their review count for this specific restaurant. By using this average, we avoid the risk of capturing variation in review count that may be driven by the restaurant's popularity or operational changes. Therefore, it offers a more stable measure that aligns with the relevance and exclusion restrictions of the instrument.

Table 4: Yelp Signals and Restaurants' Hygiene Compliance—OLS and IV

	Panel A: Violation Found—OLS		
	(1)	(2)	(3)
Has Recent Reviews × Informative	-0.006***	-0.006***	-0.002
	(0.001)	(0.001)	(0.001)
Constant	0.175***		
	(0.001)		
Adjusted R^2	0.000	0.120	0.171
F stat.	2.018	257.986	14,901.928
Wald	180.106	25.238	1.266
	Panel B: Violation Found—IV		
Has Recent Reviews × Informative	-0.012***	-0.012***	-0.012***
	(0.002)	(0.002)	(0.003)
Constant	0.173***		
	(0.001)		
Adjusted R^2	0.000	0.120	0.171
F stat.	2.076	257.996	14,902.180
Wald	183.417	39.138	14.879
Inspection fixed effects		Yes	Yes
Violation Code fixed effects		Yes	
Violation Code-Restaurant fixed effects			Yes
Observations	2,904,680	2,904,680	2,904,680
Note:	*p<0.1; **p<0.05; ***p<0.01		

Coefficient estimates of Equation 6. Column 1 does not include any controls. Column 2 includes inspection fixed effects and violation-code fixed effects. Column 3 includes the most stringent set of controls: inspection fixed effects and violation-code–restaurant fixed effects. Standard errors clustered at the restaurant level. Panel A presents OLS estimates. Panel B presents IV estimates; corresponding first-stage results are in Appendix Table A7.

dimensions for which Yelp is more informative.

Taken together, our demand- and supply-side analyses point to a consistent story: consumers take into account hygiene information contained in online reviews and restaurants seem to be aware of this information and respond to it by cleaning up more when their hygiene quality is exposed online. Note that this effect occurs *above and beyond* any effect on supply and demand induced by regulatory inspections. In an ideal setting, one would want to com- pare the separate effects of health inspections and reviews on demand and supply incentives. Our setting does not allow such a comparison. In the data we use, we do not have clean changes in inspection rules or monitoring efforts, which would be needed to evaluate whether hygiene conditions change as a function of those rules. This makes our prediction problem from Section 4 easier, but also limits our ability to evaluate the role of inspections in affecting a restaurant's hygiene.

6 Conclusion

Regulation, especially for consumer protection, is often designed to address asymmetric information and moral hazard. But the same is true for online reviews. Our key empirical insight is to identify where the monitoring capabilities of online reviews add value beyond existing regulation. While reviews miss certain hygiene dimensions that inspectors are both trained and legally able to evaluate, they do contain information about other hygiene dimensions that are monitored by inspectors and that affect consumer demand. We tease these dimensions apart and evaluate the consequences of disclosing hygiene information through online review for firms, consumers, and regulators.

In the context of New York City restaurants, we have shown that there are differences in the degree to which Yelp reviews can be an informative signal of various dimensions of hygiene already monitored through mandated inspections. Yelp reviews contain relevant information on dimensions of hygiene that consumers directly experience, such as pests and food handling violations, but not on other violations, such as facility maintenance. We have also shown that the hygiene signals contained in Yelp reviews affect consumers' choices of where to eat, above and beyond the information contained in the aggregate Yelp rating and in the city-mandated letter grade. Finally, we find that hygiene signals in Yelp reviews may drive restaurants that are more exposed to Yelp to better comply with those hygiene standards for which Yelp is most informative.

Our work highlights the importance of separating the effect of aggregate online reviews into the dimensions of quality that reviews can capture. It is possible for aggregate reviews to matter for consumers, as the literature has repeatedly shown, but for specific dimensions of those reviews to be unimportant to them. Recent advances in machine learning have made it easier to decompose consumer reviews into their various dimensions, so firms can better identify and address what customers care about. We expect such approaches to be an important avenue of future research and to be of managerial relevance.

On one hand, our results indicate that government monitoring cannot be replaced by online reviews. In particular, consumers lack either the ability or the willingness to rate many of the dimensions of restaurant hygiene that regulators monitor. Those dimensions of hygiene tend to relate to back-of-the-house hygiene conditions which customers are unlikely to see or directly experience.

On the other hand, our results suggest that government monitoring as currently designed does not completely remove information asymmetries related to restaurant hygiene. Consumers receive information about certain hygiene conditions through Yelp and both supply and demand respond to such disclosure. In particular, we find that supply and demand are responsive to signals of hygiene contained in online reviews *above and beyond* the effect of health inspections (Jin and Leslie, 2003).

We have focused on how consumers use hygiene signals in online reviews, but regulators and firms, too, can leverage this information. Regulators can use online information to use resources more efficiently by conducting targeted inspections (Kim et al., 2024). Firms can leverage the information contained in online reviews to monitor their own hygiene standards. For example, franchisers often perform inspections of franchisees to ensure compliance with corporate hygiene standards. Using review-based signals like those we have constructed could help them more quickly identify low-performing establishments.

Our work offers valuable insights for policy contexts where regulatory monitoring is either not required or applies only to a subset of service providers. For instance, while inspections are not mandated for home rentals, online reputation systems can still hold providers accountable. These systems can capture and reflect quality standards that are regulated for larger or more traditional competitors, such as hotels.

When online reviews are used to monitor and maintain certain quality dimensions, a new set of questions merits investigation. Research has shown that reviews can be strategically submitted (Resnick and Zeckhauser, 2002; Cabral and Hortaçsu, 2010) or positively selected (Nosko and Tadelis, 2015), especially when consumers and providers interact on a personal basis (Fradkin et al., 2018). Providers obviously have incentives to manipulate reviews (Mayzlin et al., 2014; Luca and Zervas, 2016). Even if platforms address review bias, they must also consider how to display the information contained in consumer reviews. Aggregate rat- ings can be too coarse a measure of quality and individual reviews can be too idiosyncratic. In addition, if a provider's quality changes over time or if certain reviews are more useful than others, aggregation needs to incorporate those features (Dai et al., 2018). Platform incentives to provide current and unbiased signals of quality deserve future research.

Finally, as our descriptive evidence emphasizes, not every business is reviewed online frequently enough for reviews to be useful. And despite the wide popularity of online review platforms, some customers lack access to them. These gaps in coverage underscore the importance of regulation to protect all consumers. Nonetheless, our research has shed light on the possibility of complementing regulation with information provided by consumers themselves.

The results presented here show that online reviews contain useful hygiene information for consumers and restaurants, but we are unable to evaluate the extent to which the monitoring role of online reviews is a substitute for existing regulation. To answer this question, future research will need access not only to plausibly exogenous variation in reviewing behavior (such as the variation we exploit here), but also to analogous variation in inspection monitoring.

Our work highlights the multidimensional nature of hygiene and the varying effectiveness of reviews, inspections, and their combination as monitoring devices across these dimensions. Understanding the relative strengths of inspections and reviews can inform strategic resource allocation, particularly where reputation systems are weak or violations are unobservable to consumers. These policy implications extend beyond restaurant hygiene, offering valuable insights for resource allocation in diverse regulatory contexts.

7 Funding and Competing Interests

All authors certify that they have no affiliation with or involvement in any organization or entity with any financial or nonfinancial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

References

G. A. Akerlof. The market for ”lemons”: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

K. J. Arrow. Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973, 1963.

Y. Bao, L. Fang, and M. Osborne. The effect of quality disclosure on firm entry and exit dynamics: Evidence from online review platforms. *Working Paper*, 2024.

J. M. Barrios. Occupational licensing and accountant quality: Evidence from the 150-hour rule. *Journal of Accounting Research*, 60(1):3–43, 2022.

P. Q. Blair and M. Fisher. Does occupational licensing reduce value creation on digital platforms? *NBER Working Paper No. 30388*, 2022.

L. Cabral and A. Hortaçsu. The dynamics of seller reputation: Evidence from ebay. *The Journal of Industrial Economics*, 58(1):54–78, 2010.

J. Chen and J. Roth. Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, 139(2):891–936, 2024.

M. K. Chen, J. A. Chevalier, P. E. Rossi, and E. Oehlsen. The value of flexible work: Evidence from uber drivers. *Journal of Political Economy*, 127(6):2735–2794, 2019.

J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2006.

M. Cohen and A. Sundararajan. Self-regulation and innovation in the peer-to-peer sharing economy. *University of Chicago Law Review Dialogue*, 82:116, 2015.

P. Cohen, R. Hahn, J. Hall, S. Levitt, and R. Metcalfe. Using big data to estimate consumer surplus: The case of uber. *NBER Working Paper No. 22627*, 2016.

J. B. Cohn, Z. Liu, and M. I. Wardlaw. Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2):529–551, 2022.

W. Dai and M. Luca. Digitizing disclosure: The case of restaurant hygiene scores. *American Economic Journal: Microeconomics*, 12(2):41–59, 2020.

W. Dai, G. Jin, J. Lee, and M. Luca. Aggregation of consumer ratings: an application to Yelp.com. *Quantitative Marketing and Economics*, 16:289–339, 2018.

L. Einav, C. Farronato, and J. D. Levin. Peer-to-peer markets. *Annual Review of Economics*, 8(1):615–635, 2016.

L. Fang. The effects of online review platforms on restaurant revenue, consumer learning, and welfare. *Management Science*, 68(11):8116–8143, 2022.

C. Farronato and A. Fradkin. The welfare effects of peer entry: The case of Airbnb and the accommodation industry. *American Economic Review*, 112(6):1782–1817, 2022.

C. Farronato, A. Fradkin, B. J. Larsen, and E. Brynjolfsson. Consumer protection in an online world: An analysis of occupational licensing. *NBER Working Paper No. 26601*, 2020.

A. Fradkin, E. Grewal, and D. Holtz. The determinants of online review informativeness: Evidence from field experiments on Airbnb. *SSRN Electronic Journal*, 41:1–12, 2018.

M. Gentzkow, B. Kelly, and M. Taddy. Text as data. *Journal of Economic Literature*, 57 (3):535–74, September 2019a.

M. Gentzkow, J. M. Shapiro, and M. Taddy. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340, July 2019b.

E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review*, 106(5): 114–18, 2016.

S. Greenstein, G. Gu, and F. Zhu. Ideology and composition among an online crowd: Evidence from wikipedians. *Management Science*, 67(5):3067–3086, 2021.

C. Harrison, M. Jorder, H. Stern, F. Stavinsky, V. Reddy, H. Hanson, H. Waechter, L. Lowe, L. Gravano, and S. Balter. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness – new york city, 2012–2013. *CDC Morbidity and Mortality Weekly Report*, 63(20):441–5, 2014.

T. Hastie, R. Tibshirani, and J. Friedman. Unsupervised learning. In *The Elements of Statistical Learning*, pages 485–585. Springer, 2009.

S. He, B. Hollenbeck, and D. Proserpio. The market for fake reviews. *Marketing Science*, 41 (5):896–921, 2022.

B. Hollenbeck, S. Moorthy, and D. Proserpio. Advertising strategy in the presence of reviews: An empirical analysis. *Marketing Science*, 38(5):793–811, 2019.

M. R. Ibanez and M. W. Toffel. How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science*, 66(6):2396–2416, 2020.

G. Z. Jin and P. Leslie. The effect of information on product quality: Evidence from restaurant hygiene grade cards. *The Quarterly Journal of Economics*, 118(2):409–451, 2003.

G. Z. Jin and P. Leslie. Reputational incentives for restaurant hygiene. *American Economic Journal: Microeconomics*, 1(1):237–267, 2009.

G. Z. Jin, Z. Lu, X. Zhou, and C. Li. The effects of government licensing on e-commerce: evidence from alibaba. *The Journal of Law and Economics*, 65(S1):S191–S221, 2022.

J. S. Kang, P. Kuznetsova, M. Luca, and Y. Choi. Where not to eat? improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1443–1448, 2013.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30:3146–3154, 2017.

H. Kim, E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca. Decision authority and the returns to algorithms. *Strategic Management Journal*, 45(4):619–648, 2024.

G. King. Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential poisson regression model. *American Journal of Political Science*, pages 838–863, 1988.

M. M. Kleiner and E. J. Soltas. A welfare analysis of occupational licensing in us states. *NBER Working Paper No. 26383*, 2019.

A. D. Kugler and R. M. Sauer. Doctors without borders? relicensing requirements and negative selection in the market for physicians. *Journal of Labor Economics*, 23(3):437–465, 2005.

D. W. Lehman, B. Kovács, and G. R. Carroll. Conflicting social codes and organizations: Hygiene and authenticity in consumer evaluations of restaurants. *Management Science*, 60(10):2602–2617, 2014.

G. Lewis and G. Zervas. The welfare impact of consumer reviews: A case study of the hotel industry. *Working Paper*, 2016.

M. Luca. Reviews, reputation, and revenue: The case of yelp. com. *Harvard Business School Working Paper*, 2016.

M. Luca and G. Zervas. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science*, 62(12):3412–3427, December 2016.

D. Mayzlin, Y. Dover, and J. Chevalier. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–2455, August 2014.

J. Mejia, S. Mankad, and A. Gopal. A for effort? using the crowd to identify moral hazard in new york city restaurant hygiene inspections. *Information Systems Research*, 30(4): 1363–1386, 2019.

C. Nosko and S. Tadelis. The limits of reputation in platform markets: An empirical analysis and field experiment. *NBER Working Paper No. 20830*, 2015.

R. O'Hara and J. Kotze. Do not log-transform count data. *Nature Precedings*, pages 1–1, 2010.

M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

P. Resnick and R. Zeckhauser. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *Advances in applied microeconomics*, 11:127–157, 2002.

C. Shapiro. Investment, moral hazard, and occupational licensing. *The Review of Economic Studies*, 53(5):843–862, 1986.

J. Stock and M. Yogo. Testing for Weak Instruments in Linear IV Regression. In *Identification and Inference for Econometric Models*. Cambridge University Press, 2005.

M. Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

M. Taddy. Distributed multinomial regression. *The Annals of Applied Statistics*, 9(3):1394–1414, 2015.

Appendix A Additional Figures and Tables

Table A1: Grade Transitions

		Score at		
Prior		Initial Inspection		
Card	N	0-13	14-27	28+
A	126,540	0.41	0.36	0.22
B	27,345	0.21	0.43	0.36
C	6,367	0.18	0.42	0.40

		Card Posted at		
Prior		End of Inspection Cycle		
Card	N	A	B	C
A	126,540	0.85	0.13	0.02
B	27,345	0.64	0.30	0.06
C	6,367	0.59	0.29	0.13

For every inspection cycle with a previous grade, the left panel shows the card displayed before the cycle starts (rows) and the distribution over initial inspection scores (columns). For example, of the 126,540 restaurant-inspections obtaining an A-grade during the previous inspection cycle, 41% scored between 0 and 13 points during the initial inspection, 36% scored between 14 and 27 points, and 22% scored 28 or more points. The right panel shows the card displayed before the cycle starts (rows) and the distribution over letter grades at the end of the inspection cycle (columns). For example, of the 126,540 restaurant-inspections starting a new inspection cycle with an A-grade, 85% kept it, 13% dropped to B-grade, and 2% dropped to C-grade. The rows do not always sum to 100% due to rounding.

Table A2: Restaurant Characteristics and Initial Inspections

	Initial Score	Initial Score	Score>13	Score>13
	(1)	(2)	(3)	(4)
Not on Yelp	−0.533	−0.516	−0.023	−0.021
Inexpensive	−0.538	−0.370	−0.007	−0.002‡
Moderate	0.030‡	0.028‡	0.016	0.014
Pricey	−1.453	−1.931	−0.030	−0.050
High End	−2.365	−2.814	−0.055	−0.074
Bronx	−0.218‡	0.881	−0.012	0.021
Brooklyn	−0.318	−0.166	−0.006	−0.003‡
Queens	0.462	−0.323	0.026	−0.003‡
Staten Island	−1.228	−0.698	−0.044	−0.027
Unknown Borough	−7.199	−7.864	−0.219‡	−0.218
Bar/Pub	0.401	−0.467	0.060	0.029
Fast Food	1.299	1.310	0.081	0.085
Restaurant	3.255	3.147	0.127	0.124
American	−1.299	−1.314	−0.043	−0.043
Cafe/Bakery	−2.659	−2.553	−0.089	−0.083
Chinese	1.219	1.376	0.048	0.051
Italian	−0.833	−0.826	−0.026	−0.028
Latin/Mexican	1.275	1.398	0.038	0.040
Pizza	−0.395	−0.208‡	0.003‡	0.009
Chain Restaurant	−6.327	−6.318	−0.219	−0.216
Month-Year FE	Yes	Yes	Yes	Yes
Inspector FE	No	Yes	No	Yes
Mean Dep. Var.	21.71	21.71	0.64	0.64
Observations	206,160	206,160	206,160	206,160
R^2	0.077	0.181	0.078	0.148

For every initial inspection, the total violation score is regressed against observable characteristics of the restaurant. In Columns (1)-(2), the outcome is the total score, with higher scores denoting worse hygiene. In Columns (3)-(4), the outcome is whether or not the score is 14 points or more, which is the threshold past which the restaurant is not assigned a A-grade and is reinspected within a few weeks. Controls include month-year fixed effects in all columns and inspector fixed effects in even-numbered columns. The left-out category refers to restaurants in Manhattan that are listed on Yelp, with unknown price category, venue, or cuisine. The average score is 21.7 points, with a standard deviation of 14.5. Standard errors clustered at the restaurant level. To improve readability, standard errors are excluded and the symbol ‡ denotes a coefficient that is ***not*** *statistically significant at the 5% confidence level.*

Table A3: Top 20 Violation Codes

Code	Description	Share of Inspections
02B	Hot food item not held at or above 140^{o} F.	19.9%
04H	Raw, cooked or prepared food is adulterated, contaminated, cross-contaminated, or not discarded in accordance with HACCP plan.	11.4%
04M	Live roaches present in facility's food and/or non-food areas.	7.8%
04A	Food Protection Certificate not held by supervisor of food operations.	9.4%
02G	Cold food item held above 41^{o}F (smoked fish and reduced oxygen packaged foods above 38^{o}F) except during necessary preparation.	33%
10H	Proper sanitization not provided for utensil ware washing operation.	7.4%
06D	Food contact surface not properly washed, rinsed and sanitized after each use and following any activity when contamination may have occurred.	27%
04N	Filth flies or food/refuse/sewage-associated (FRSA) flies present in facility's food and/or non-food areas. Filth flies include house flies, little house flies, blow flies, bottle flies and flesh flies. Food/refuse/sewage-associated flies include fruit flies, drain flies and Phorid flies.	13.4%
06E	Sanitized equipment or utensil, including in-use food dispensing utensil, improperly used or stored.	11.2%
06C	Food not protected from potential source of contamination during storage, preparation, transportation, display or service.	23.1%
10B	Plumbing not properly installed or maintained; anti-siphonage or backflow prevention device not provided where required; equipment or floor not properly drained; sewage disposal system in disrepair or not functioning properly.	23.9%
08A	Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.	41.8%
06F	Wiping cloths soiled or not stored in sanitizing solution.	8.4%
04L	Evidence of mice or live mice present in facility's food and/or non-food areas.	25.9%
09C	Food contact surface not properly maintained.	7.6%
10F	Non-food contact surface improperly constructed. Unacceptable material used. Non-food contact surface or equipment improperly maintained and/or not properly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.	45.8%
05D	Hand washing facility not provided in or near food preparation area and toilet room. Hot and cold running water at adequate pressure to enable cleanliness of employees not provided at facility. Soap and an acceptable hand-drying device not provided.	6.4%
06A	Personal cleanliness inadequate. Outer garment soiled with possible contaminant. Effective hair restraint not worn in an area where food is prepared.	7.9%
08C	Pesticide use not in accordance with label or applicable laws. Prohibited chemical used/stored. Open bait station used.	5.1%
04J	Appropriately scaled metal stem-type thermometer or thermocouple not provided or used to evaluate temperatures of potentially hazardous foods during cooking, cooling, reheating and holding.	7.3%

The 20 violation codes that occur most frequently in initial inspections. The last column shows the share of initial inspections during which the inspector found a particular violation. Violation codes are ordered as in Figure 4, based on how well Yelp reviews can predict that specific violation (best at the top).

Table A4: Yelp Hygiene Signal and Sold-out Probability—Robustness

	Sold Out on OpenTable					
	(1)	(2)	(3)	(4)	(5)	(6)
After Review	-0.002^{***}	-0.002^{***}	-0.002^{**}	-0.003^{***}	-0.003^{***}	-0.002^{***}
	(0.0005)	(0.001)	(0.001)	(0.001)	(0.0004)	(0.0005)
Bad Yelp Hygiene Signal*After Review	-0.003^{***}	-0.003^{**}	-0.004^{***}	-0.003^{**}	-0.004^{***}	-0.003^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Day of Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Restaurant-Review FE	Yes	Yes	Yes	Yes	Yes	Yes
Specifications	Baseline	A-Grade Card Only	Different Control	No Days Around Event	Worst 10% Signal	Worst 30% Signal
Observations	2,644,279	2,212,002	1,056,122	2,217,553	2,644,279	2,644,279
Adjusted R^2	0.498	0.499	0.500	0.495	0.498	0.498
Note:						*p<0.1; **p<0.05; ***p<0.01

Robustness checks of the difference-in-differences specification in Column 2 of Table 3. Column 1 reproduces the baseline results. The other columns each change one parameter or sample-selection criterion at a time. Column 2 focuses on events when the restaurant displayed an A-grade card in the month surrounding the focal Yelp review. Column 3 only uses the 1-, 2-, and 3-star reviews with the 20% best hygiene signals as control group. Column 4 removes 5 days surrounding the review from the observations (from 2 days before to 2 days after the focal review). Column 5 defines the treated group as restaurants receiving a Yelp review hygiene signal from among the top 10% worst signals. Column 6 defines the treated group as restaurants receiving a Yelp review hygiene signal from among the top 30% worst signals.

Table A5: Yelp Hygiene Signal and Sold-out Probability—Robustness

Violation Codes	Interaction Coefficient
02B	-0.002** (9e-04)
02B, 04H	-0.001 (0.001)
02B, 04H, 04M	-0.0014 (0.001)
02B, 04H, 04M, 04A	-0.0023** (0.001)
02B, 04H, 04M, 04A, 02G	-0.0032*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H	-0.004*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D	-0.004*** (0.001)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N	-0.0032*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E	-0.004*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C	-0.0033*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B	-0.0036*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A	-0.004*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A, 06F	-0.0039*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A, 06F, 04L	-0.0039*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A, 06F, 04L, 09C	-0.0035*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A, 06F, 04L, 09C, 10F	-0.0038*** (9e-04)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A, 06F, 04L, 09C, 10F, 05D	-0.0039*** (0.001)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A, 06F, 04L, 09C, 10F, 05D, 06A	-0.0036*** (0.001)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A, 06F, 04L, 09C, 10F, 05D, 06A, 08C	-0.0036*** (0.001)
02B, 04H, 04M, 04A, 02G, 10H, 06D, 04N, 06E, 06C, 10B, 08A, 06F, 04L, 09C, 10F, 05D, 06A, 08C, 04J	-0.0035*** (9e-04)

Note: *p<0.1; **p<0.05; ***p<0.01

Robustness checks of the difference-in-differences specification presented in Column 2 of Table 3. Each row displays the difference-in-differences coefficient from a different regression. The first row uses only the sufficient reduction from violation code 02B (hot food not kept at or above 140^oF)—the violation code for which Yelp is most informative—to define the event of a bad hygiene signal. Subsequent rows add the sufficient reduction of violation codes for which Yelp is progressively less informative. The fifth row reproduces the difference-in-differences coefficient estimate from Column 2 of Table 3.

Table A6: Supply Side—Descriptive Statistics

Variable	Mean	Standard Deviation	25^{th} Percentile	Median	75^{th} Percentile
Violation Found	0.17	0.38	0.00	0.00	0.00
Violation Found (Residual)	0.00	0.30	-0.16	-0.03	0.06
Has Recent Reviews	0.65	0.48	0.00	1.00	1.00
Informative	0.25	0.43	0.00	0.00	0.00
Has Recent Reviews × Informative	0.16	0.37	0.00	0.00	0.00
Has Recent Reviews × Informative (Residual)	0.00	0.14	-0.04	0.00	0.04
Average Review Propensity	198	197	100	165	242
No Previous Reviewers	0.25	0.44	0.00	0.00	1.00

Summary statistics for the variables in the supply-side analysis (Section 5.2). For the dependent variable and main endogenous variable, we also present summary statistics of the residuals of regressions on inspection fixed effects and violation-code–restaurant fixed effects.

Table A7: Yelp Signal and Restaurants' Hygiene Compliance—IV First Stage

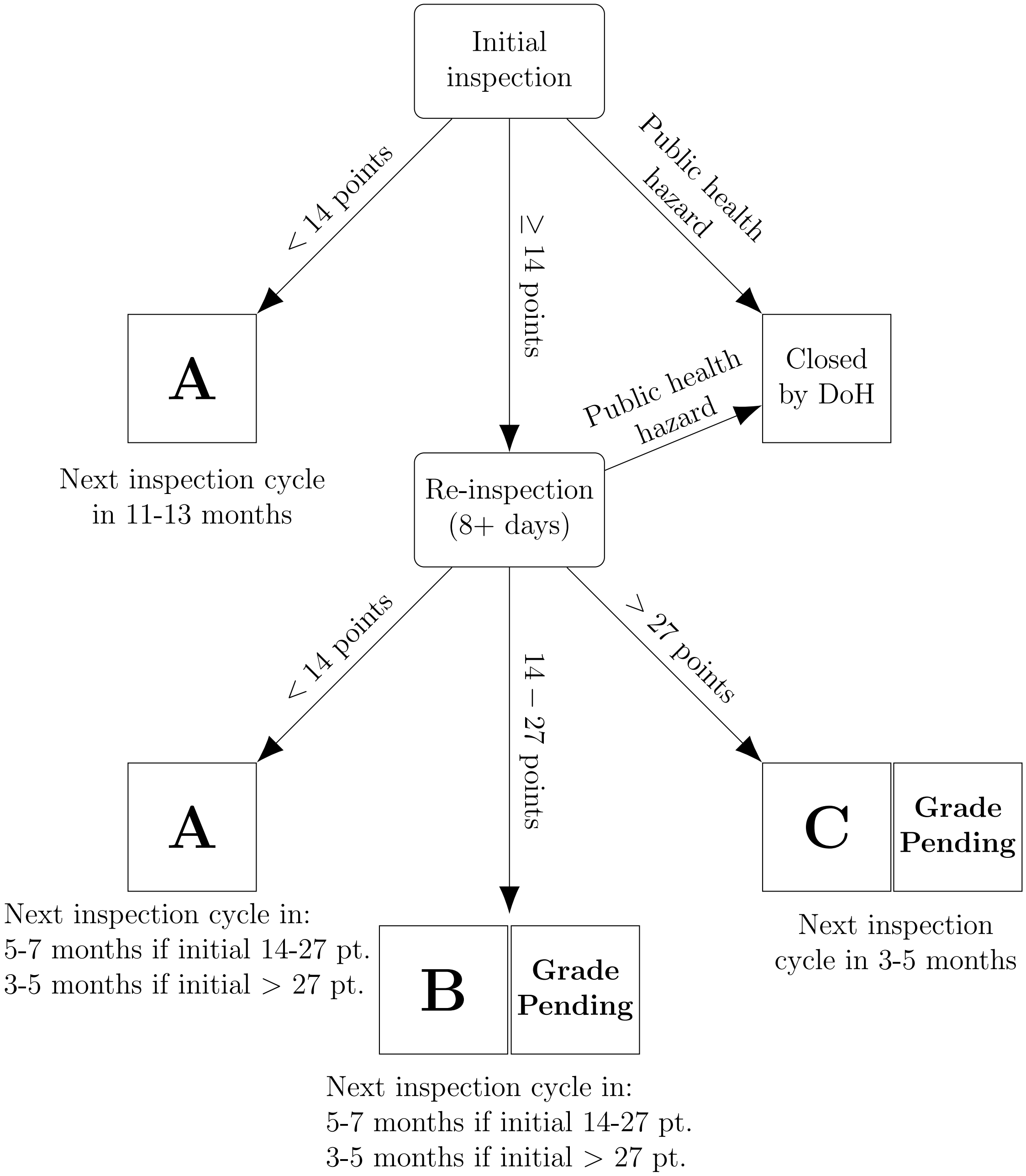
	Panel A: Has Recent Reviews – First-Stage		
	(1)	(2)	(3)
Log(Avg. Review Propensity + 1)	0.029***		
	(0.002)		
No Previous Reviewers	-0.500***		
	(0.013)		
Adjusted R-Squared	0.350		
F stat.	2,821.908		
Wald	14,739.365		
	Panel B: Has Recent Reviews*Informative – First-Stage		
Log(Avg. Review Propensity + 1)*Informative	0.029***	0.029***	-0.002
	(0.002)	(0.002)	(0.002)
No Previous Reviewers	-0.500***	-0.500***	-0.494***
	(0.013)	(0.013)	(0.012)
Adjusted R-Squared	0.729	0.786	0.843
F stat.	14,124.767	4,900.688	95,275.101
Wald	43,383.958	29,478.568	7,665.681
Inspection Fixed Effects		Yes	Yes
Violation Code Fixed Effects		Yes	
Violation Code-Restaurant Fixed Effects			Yes
Observations	2,904,680	2,904,680	2,904,680
Note:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$		

First-stage estimates of Panel B in Table 4.

Table A8: Yelp Signals and Restaurants' Hygiene Compliance—Different IV

	Violation Found - IV		
Has Recent Reviews × Informative	-0.011***	-0.011***	-0.013***
	(0.002)	(0.002)	(0.003)
Constant	0.173***		
	(0.001)		
Adjusted R^2	0.000	0.120	0.171
F stat.	2.070	257.993	14,902.206
Wald	182.780	34.884	16.291
Inspection fixed effects		Yes	Yes
Violation Code fixed effects		Yes	
Violation Code-Restaurant fixed effects			Yes
Observations	2,904,680	2,904,680	2,904,680
Note:	*p<0.1; **p<0.05; ***p<0.01		

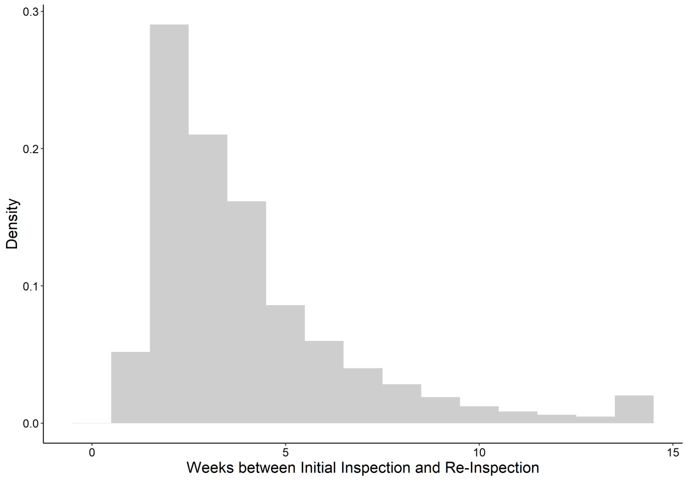
IV coefficient estimates of Equation 6 as in Panel B of Table 4. In Panel B of Table 4, we present estimates in which the main instrument is log-transformed; here, we present estimates in which the instrument is in levels, given the possible biases of log transformation (King, 1988; O'Hara and Kotze, 2010; Cohn et al., 2022; Chen and Roth, 2024).



Initial inspection
< 14 points
≥ 14 points
Public health hazard
A
Next inspection cycle in 11-13 months
Closed by DoH
Public health hazard
Re-inspection (8+ days)
< 14 points
14 – 27 points
> 27 points
A
Next inspection cycle in:
5-7 months if initial 14-27 pt.
3-5 months if initial > 27 pt.
B
Grade Pending
Next inspection cycle in:
5-7 months if initial 14-27 pt.
3-5 months if initial > 27 pt.
C
Grade Pending
Next inspection cycle in 3-5 months

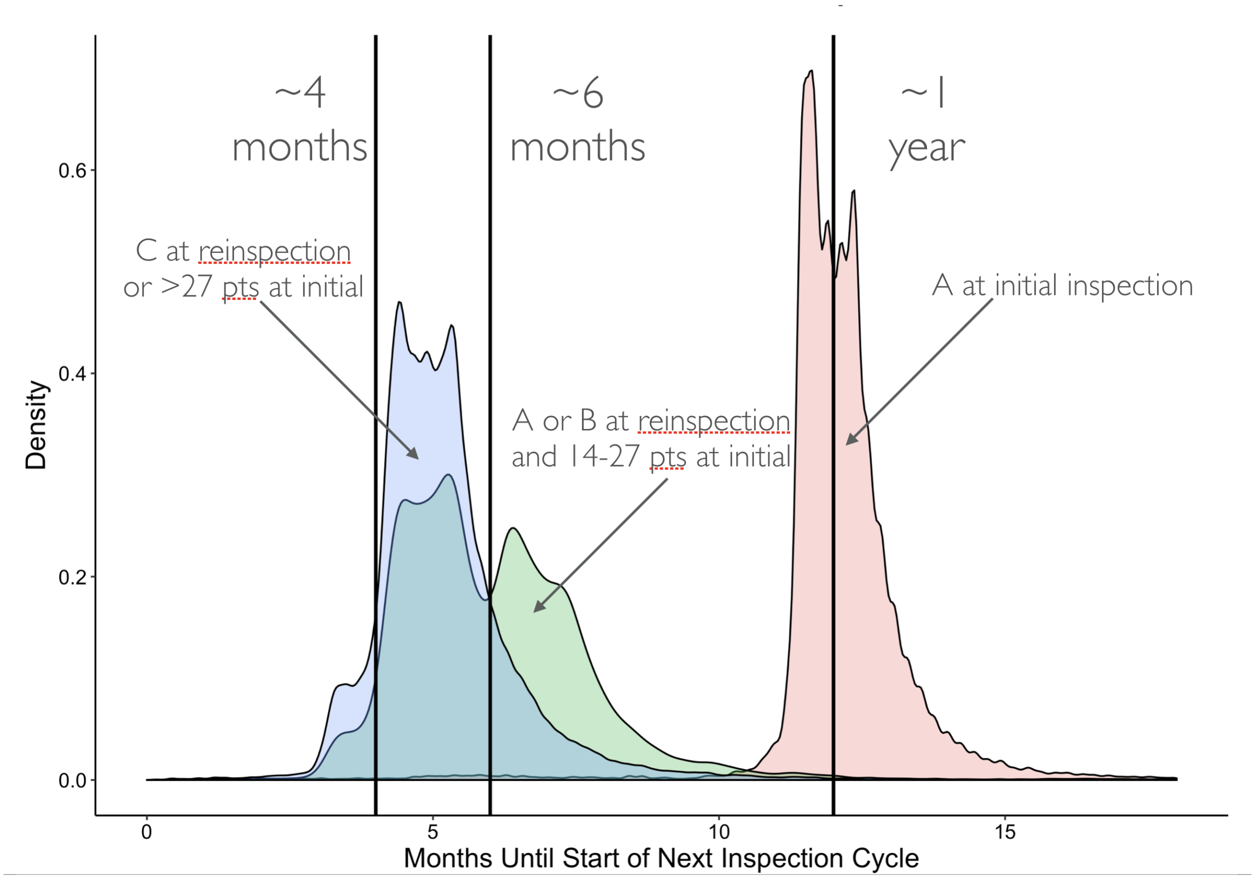
Restaurant inspection cycle conducted by the New York City Department of Health and Mental Hygiene (adapted from `https://www1.nyc.gov/assets/doh/downloads/pdf/rii/inspection-cycle-overview.pdf`*).*

Figure A1: Inspection Cycle



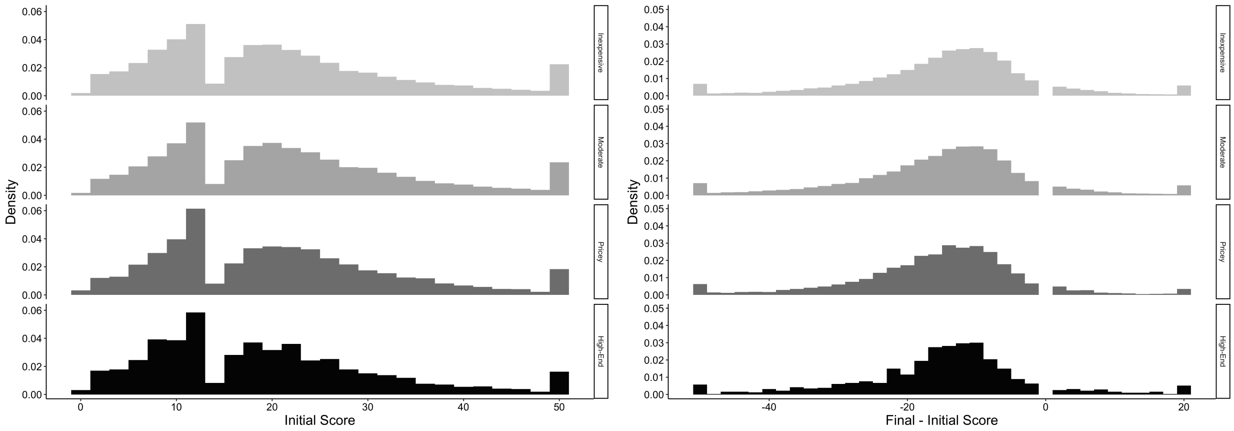
Distribution of the time lag, in weeks, between an initial inspection and a reinspection for restaurants that received 14 or more points during the initial inspection.

Figure A2: Time between Initial Inspection and Reinspection



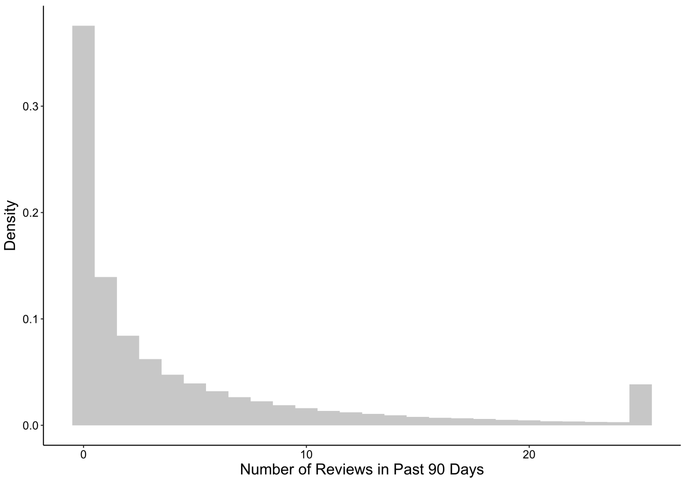
Distribution of time between the last inspection of the current inspection cycle and the first inspection of the next cycle. For restaurants obtaining an A-grade at initial inspection during the current inspection cycle (pink), the expected time is 12 months since the last inspection. For restaurants scoring 14-27 points at initial inspection and obtaining A- or B-grades at reinspection, the expected time is 5-7 months since the last inspection. For restaurants scoring 28+ points at initial inspection or obtaining a C-grade at reinspection, the expected time is 3-5 months since the last inspection. The plot shows substantial variation in the time between inspections.

Figure A3: Time between Inspection Cycles



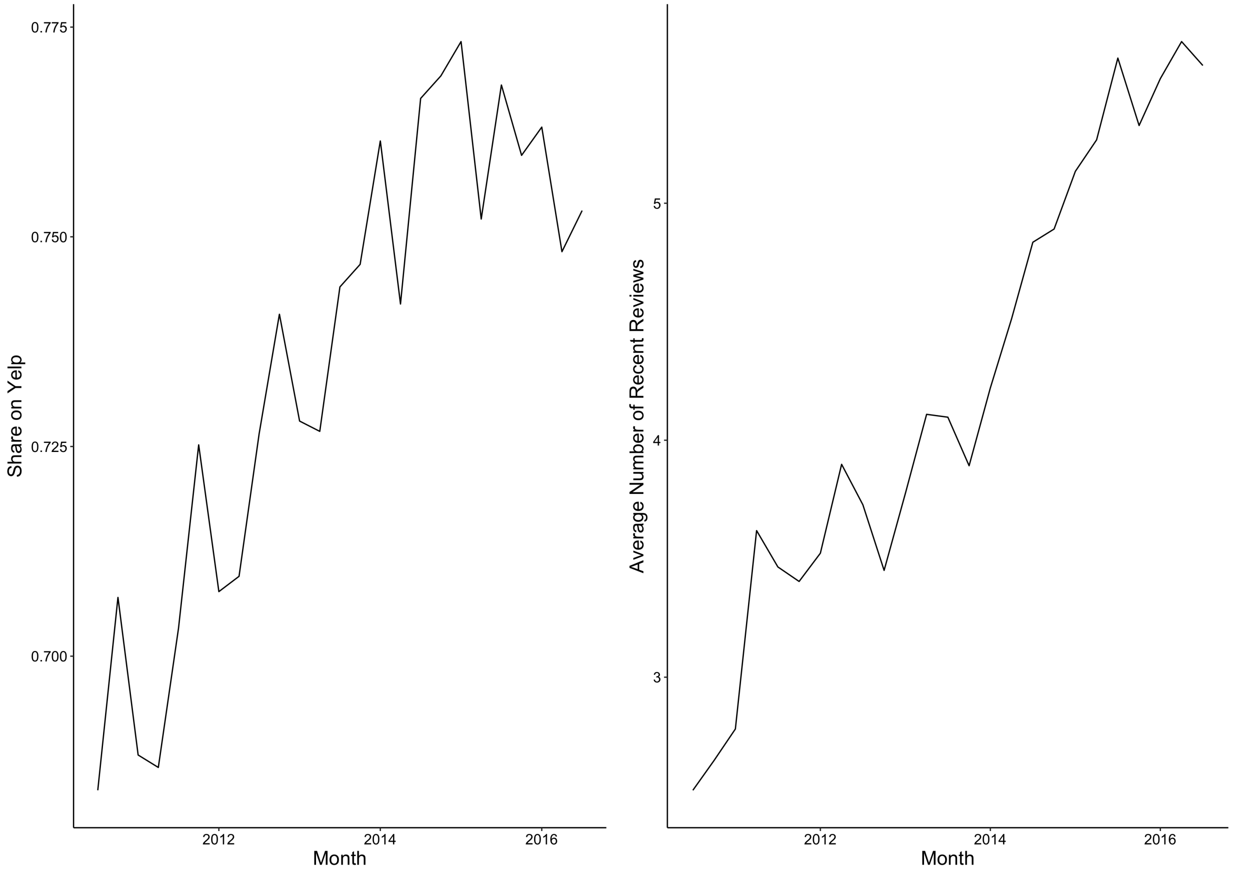
These figures are similar to Figure 1, except that restaurants are divided into four price groups, from "inexpensive" at the top to "high-end" at the bottom. For each inspection cycle, the left column shows the distribution of violation scores that restaurants obtain during the initial inspection. For the purpose of these plots, inspection scores are capped at 50. For restaurants that undergo a reinspection, the right column displays the difference between the reinspection score and the initial inspection score (a negative number means that hygiene improved). For the purpose of these plots, the difference in inspection scores is bounded between -20 and 20 (i.e., higher differences in absolute value are capped at +/-20).

Figure A4: Inspection Outcomes by Price Group



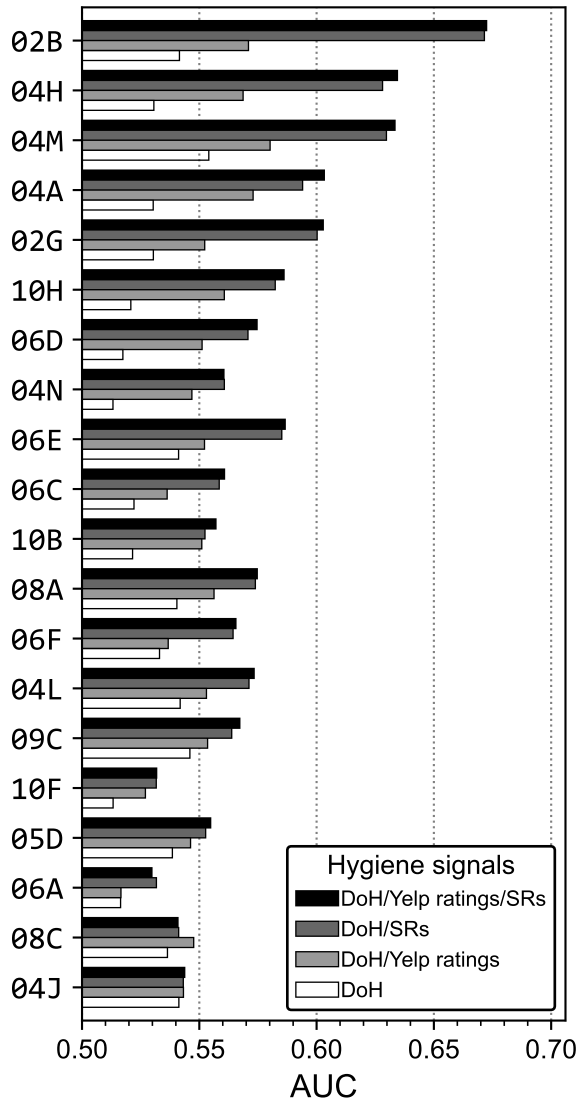
Distribution of the number of reviews that a restaurant on Yelp obtains in the 90 days preceding an initial inspection. The median number of reviews received before an initial inspection is 1; the mean is 5. For the purpose of this plot, the number of reviews is capped at 25.

Figure A5: Number of Reviews in Previous 90 Days



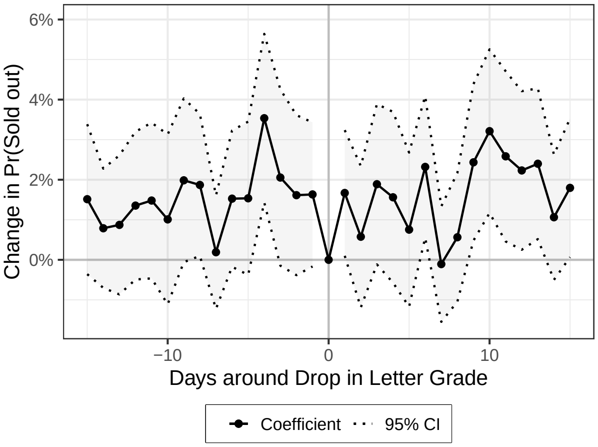
The left figure plots the share of initial inspections during a given quarter for restaurants listed on Yelp. The right figure plots how the average number of reviews that a restaurant on Yelp obtains in the 90 days preceding an initial inspection changes over time. Initial inspections, which are the unit of observation, are aggregated at the quarterly level in the time plots.

Figure A6: Review Frequency over Time



This figure plots the area under the curve (AUC) of the prediction of the 20 most-frequent violation codes, separately for four classifiers: the baseline classifier (white) and the review-augmented classifiers (black), both of which are in the main paper; a classifier that uses letter grades and average star-ratings (light grey); and a classifier that uses letter grades and the sufficient reduction projections (dark grey).

Figure A7: Comparing Prediction Accuracy across Classifiers



Results from event study regressions (Equation 4), with the event defined as a day when the letter grade drops from "A" to anything lower. There are 1,261 such events for which we have data on sold-out probability (compared to 19,627 events when a Yelp review is submitted with poor hygiene signals, in Figure 6). Despite the fluctuating estimates, which may be due to the small sample size, we cannot reject the hypothesis that the probability of being sold out is the same after the change in letter grade as before the event.

Figure A8: Letter-grade Drop and Sold-out Probability—Event Study

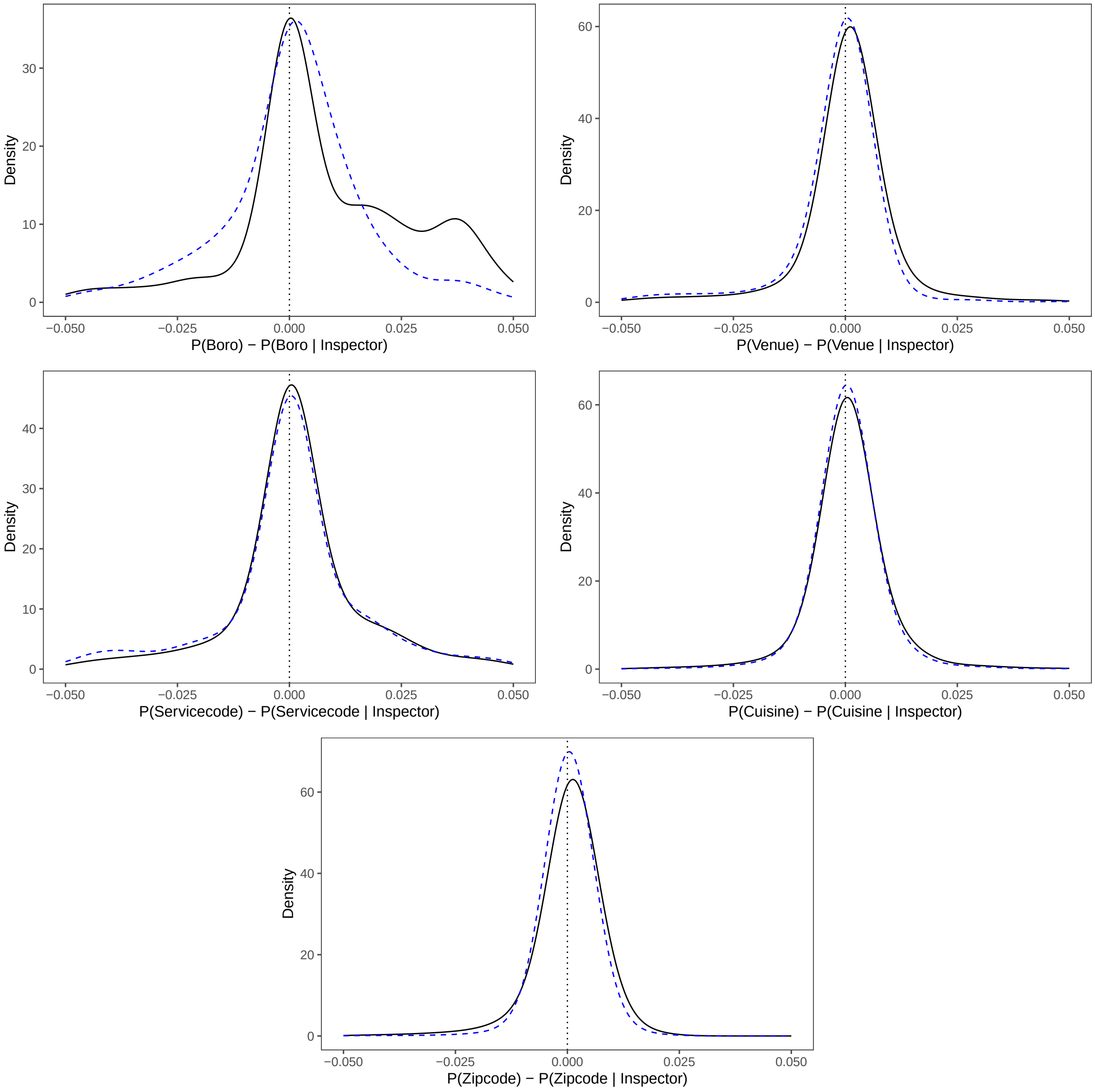
Appendix B How Are Inspectors and Reviewers Assigned to Restaurants?

Here we verify that, conditional on observables, we cannot reject the hypothesis that inspectors and reviewers are randomly assigned to evaluate restaurants. We describe the procedure for inspectors, but use it also for reviewers. We compute two probability distributions. First, we compute the unconditional distribution of a particular restaurant characteristic, denoted $P(X)$. Second, we compute the distribution of X conditional on a particular inspector Z, denoted $P(X|Z)$. We then take the difference $P(X)-P(X|Z)$ across all possible values of X and across all inspectors. We compare the distribution of this difference to $P(X)-P(X|Z')$. The only difference between $P(X|Z)$ and $P(X|Z')$ is that $P(X|Z)$ is based on the actual allocation of inspectors to restaurants, while $P(X|Z')$ is a random permutation.

Figure A9 displays the distributions of $P(X)-P(X|Z)$—solid line—and $P(X)-P(X|Z')$—dotted line—across all inspectors and for different observable characteristics. If inspectors were randomly assigned to restaurants conditional on observable X, the dotted and solid density functions would be indistinguishable. The figures show that inspectors tend to specialize by geography, inspecting restaurants in one New York City borough or a few zip codes more than in other boroughs or zip codes. There does not seem to be any specialization of inspectors across observable restaurant characteristics other than geography.

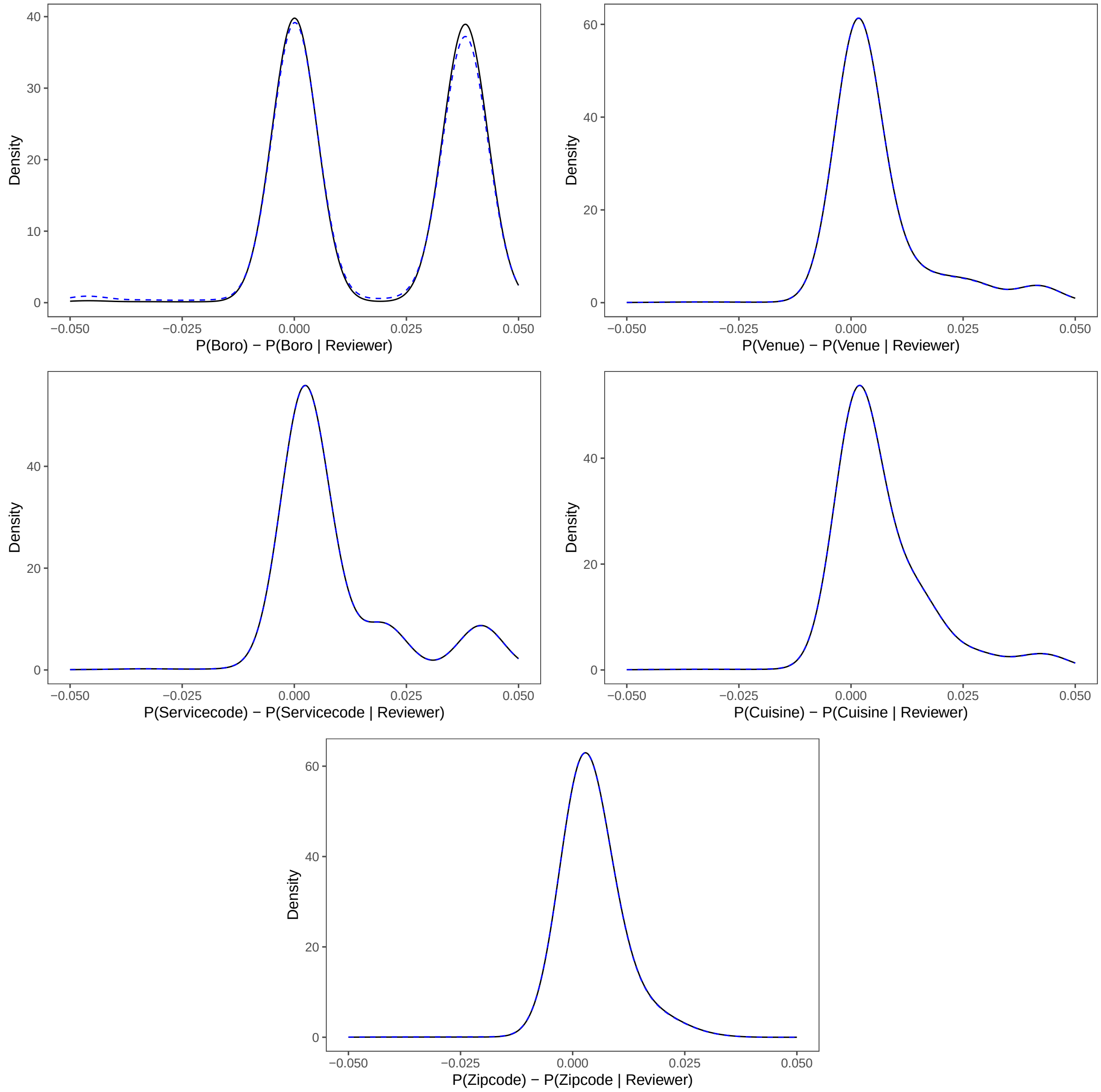
As for how reviewers are "assigned" to restaurants, Figure A10 shows that there is no particular pattern for how users choose to submit reviews of restaurants.

Figure A9: Independence of Inspectors and Restaurant Characteristics



Difference in the distribution of observables unconditional and conditional on inspectors. The solid line plots $P(X) - P(X|Z)$*, where the conditional distribution is a function of the actual allocation of inspectors to restaurants. The dotted line computes the conditional distribution after a random permutation of inspectors to restaurants. The average number of restaurants inspected by each inspector is 807 in our data, with a large standard deviation of 822.*

Figure A10: Independence of Reviewers and Restaurant Characteristics



Difference in the distribution of observables unconditional and conditional on reviewers. The solid line plots $P(X) - P(X|Z)$, where the conditional distribution is a function of the actual allocation of Yelp reviewers to restaurants. The dotted line computes the conditional distribution after a random permutation of reviewers to restaurants. The average number of restaurants reviewed by each Yelp reviewer is 4.3 in our data, with a large standard deviation of 14.5.

Appendix C Timing of Reviews and Ranking of Restaurants on Yelp

We want to verify that restaurants with more recent reviews are ranked higher in Yelp search results. To do this, we pulled data from the Yelp API. We submitted the query "Find: Restaurants | Near: New York, NY" on April 8, 2019, at midnight. Yelp places a limit of 1,000 results to be returned and the order in which they are returned reflects the order shown on the webpage if a user were to perform the same search on Yelp. The restaurants returned in this list are the *ranked* restaurants out of all New York City restaurants. Whether a restaurant shows up at all in this list and whether it shows up at the top or at the bottom of the search results will be our outcomes of interest.

We also compile the list of all Yelp restaurants in New York by performing a similar query as before, but separately for each zip code.[26] Given the limit to the number of results returned by the Yelp API, a zip code is further disaggregated if the returned results are 1,000. Specifically, if a query for a given zip code returns fewer than 1,000 restaurants, results are recorded and we move to the next zip code. Otherwise the zip code is split into four quadrants, and we conduct four searches, one for each quadrant, using its center and half its diagonal as the search radius. We continue splitting geographies until the results returned are fewer than 1,000 for each search. After dropping duplicates and businesses outside New York,[27] we are left with 23,387 restaurants, which constitute the population of New York City restaurants on Yelp on April 8, 2019.

For each restaurant, we scrape additional information from Yelp using the URLs obtained from the API. This information includes the date of the most recent review—our treatment variable of interest—and additional controls such as restaurant category, price, Yelp stars, and total number of reviews.

We run regressions of the following type:

$$y_i = \alpha \log \text{days\_since\_last\_review}_i + \boldsymbol{X}_i \beta + \epsilon_i, \tag{7}$$

where i denotes a restaurant, and y is one of two outcomes: first, a dummy for whether the

[26] The following website contains the list of zip codes—compiled by the New York State Department of Health (DoH)—used in the grid search: `https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm`.

[27] The search algorithm may pick up restaurants outside New York City, or in neighboring zip codes, since the search radius is conservatively set to cover larger areas than the quadrant of interest. To determine whether a zip code is located in New York City, we check whether it can be found on the DoH list used in the grid search or whether it is located in New York City according to the US Postal Service's ZIP Code Lookup data.

restaurant is *ranked* in the general query for "restaurants in New York City" and, second, conditional on being ranked, the rank in the search results (where 1 denotes the top result and 999 denotes the last). We expect that the shorter the time since the last review, the more likely a restaurant is to be ranked and the lower—that is, closer to the top of the page—its rank will be. The variable $days\_since\_last\_review_i$ is measured relative to April 7, 2018, the day before our data pull. The vector X_i includes the number of reviews (log) and fixed effects for restaurant category, price grouping ($, $$, $$$, or $$$$), Yelp stars, and zip code. To estimate the relationships of interest, we use OLS, logistic, and probit regressions when the outcome is a dummy for being ranked and OLS, ordered logistic, and ordered probit when the outcome is the exact ranking.

Summary statistics are presented in Table A9 and regression results are presented in Table A10. The results show that businesses with more recent reviews are indeed more likely to be ranked (Columns 1–3) and, if so, placed higher in search results (Columns 4–6). The effects are sizable. When looking at the probability of being in the top 1,000 results, we discuss the logistic regression estimates, given that only 4.3% of New York City restaurants are actually ranked (probit estimates imply similar magnitudes). The estimates from the logistic regression (Column 2) suggest that, all else held constant, doubling the age of the most recent review is associated with a -0.382 decrease in log odds of being ranked, implying an odds ratio of $exp(-0.382) = 0.682$. This suggests significantly reduced odds of being ranked for businesses with older reviews.

When we focus on the sample of restaurants that are ranked (Columns 4–6), the age of the most recent review predicts the actual rank in the search results. All else held constant, OLS results suggest that doubling the age of the most recent review is associated with a *decrease*in rank by 35 positions; for example, dropping from the top result to the 36^{th} result. Similarly, the odds ratio of $exp(-0.265) = 0.767$ from Column 5 confirms the reduced odds of a business moving up the rank the older its most recent review.

Table A9: Summary Statistics

	1^{st} Quartile	Median	3^{rd} Quartile	Mean	Std. Dev.	N
Ranked	0	0	0	0.043	0.202	23,385
Days since last review	7	32	185	322.5	754.8	23,383
Days since last review | ranked	0	2	7	5.6	11.7	999

Summary statistics for Yelp restaurants in New York City. Data were obtained from the Yelp API on April 8, 2019.

Table A10: Review Age and Search Ranking Outcomes

	Ranked 0/1			Rank | Ranked		
	Normal	*Logistic*	*Probit*	*OLS*	*Ordered Logistic*	*Ordered Probit*
	(1)	(2)	(3)	(4)	(5)	(6)
Days since last review + 1 (log)	−0.001 (0.001)	−0.382*** (0.043)	−0.218*** (0.023)	34.880*** (9.403)	−0.265*** (0.073)	−0.165*** (0.037)
Observations	23,383	23,383	23,383	999	999	999
Log Likelihood	6,945.889	−1,545.434	−1,550.226			
Akaike Inf. Crit.	−12,829.780	4,152.869	4,162.451			
R^2				0.370		
Adjusted R^2				0.223		
Note:				*p<0.1; **p<0.05; ***p<0.01		

Regression results of Equation 7. In Columns 1–3, the outcome of interest is a dummy for whether a restaurant is in the top 1,000 search results for the query "Find: Restaurants | Near: New York, NY". In Columns 4–6, the outcome is the actual rank in the search results, conditional on being in the top 1,000 results. We drop one observation from the search results because the most recent review date was missing. Controls include number of reviews (in logs) and fixed effects for restaurant category, price grouping ($, $$, $$$, or $$$$), Yelp stars, and zip code.