NBER WORKING PAPER SERIES

CONSUMER REVIEWS AND REGULATION: EVIDENCE FROM NYC RESTAURANTS

Chiara Farronato
Georgios Zervas

Working Paper 29715
http://www.nber.org/papers/w29715

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2022

Yi Fung and Hirotaka Miura have provided outstanding research assistance. We thank Liran Einav, Shane Greenstein, Jonathan Levin, Jesse Shapiro, Steven Tadelis, and numerous seminar and conference participants for feedback. We thank the New York City Department of Health and Mental Hygiene for sharing data and valuable insights. Neither author has any material or financial interest in the entities that are the subject of this research. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Consumer Reviews and Regulation: Evidence from NYC Restaurants
Chiara Farronato and Georgios Zervas
NBER Working Paper No. 29715
February 2022
JEL No. D18,D22,D82,K2,K20,L15,L51,L80,L86

ABSTRACT

We investigate the informativeness of hygiene signals in online reviews, and their effect on consumer choice and restaurant hygiene. We first extract signals of hygiene from Yelp. Among all dimensions that regulators monitor through mandated restaurant inspections, we find that reviews are more informative about hygiene dimensions that consumers directly experience - food temperature and pests - than other dimensions. Next, we find causal evidence that consumer demand is sensitive to these hygiene signals. We also find suggestive evidence that restaurants that are more exposed to Yelp are cleaner along dimensions for which online reviews are more informative.

Chiara Farronato
Harvard Business School
Morgan Hall 427
Soldiers Field
Boston, MA 02163
and NBER
cfarronato@hbs.edu

Georgios Zervas
Boston University
zg@bu.edu

A data appendix is available at http://www.nber.org/data-appendix/w29715

1 Introduction

When consumers have limited information about service providers' quality, a market can easily break down or put consumers at risk, either because low quality providers drive away high quality providers (Akerlof (1970)) or because providers do not have enough incentives to invest in quality (Arrow (1963)). Government regulation, such as occupational licensing and regulatory inspections, has historically been the solution to increase quality and inform consumers. More recently however, consumers have turned to online reviews as a primary source of information for various products and services. Popular platforms like Yelp and TripAdvisor publish millions of consumer reviews for restaurants, hotels, and other local businesses, and are visited by tens of millions of consumers every month.[1] It is unclear whether online reviews actually capture information on those dimensions of quality that government regulation has identified as important to monitor. Similarly, it is unclear whether the choices of consumers and providers are affected by the availability of such information in consumer reviews.

In this paper, we explore the value of online reviews as a signal of hygiene. Our empirical context is New York City restaurants. We implement machine learning methods to obtain hygiene signals from the text of Yelp reviews and evaluate how informative reviews are about different hygiene dimensions that regulators monitor. We find that reviews are more informative about hygiene dimensions that consumers directly experience – pests and food handling practices – than other dimensions – workers' hygiene and facilities maintenance. We then use our newly constructed hygiene signal to estimate its effects on restaurant demand. Using an event study approach around the time of review submission, we find that restaurants are significantly less likely to sell out in the weeks following a review with a poor hygiene signal. Finally, we find suggestive evidence that restaurants may take into account the signaling role of online reviews when choosing their hygiene level.

Our paper is motivated by the recent diffusion of consumer reviews for providers that are often subject to regulatory screening and monitoring. The need for regulation is typically justified if other ways to ensure consumer protection are impractical or too costly (Shapiro (1986) and Friedman (1962)). In fact, an extensive literature documents the value of regulation in ensuring health and safety standards (e.g., Jin and Leslie (2003)). At the same time, regulation can increase entry barriers and reduce competition (Federman et al. (2006)).

Recent technological developments have made it cheaper to collect and aggregate information about service providers from past customers. If informative about quality dimensions that matter for consumer protection, online reviews offer some practical advantages over reg-

[1] See, for example, `https://www.yelp.com/factsheet` (accessed on January 15, 2021).

ulation. They are cheaper to collect and more frequent than regulatory inspections, at least for certain businesses. Major US cities are currently financially constrained, and conducting health inspections of local businesses can be costly.[2] In addition, inspectors have been shown to have discretionary power over their evaluations (Ibanez and Toffel (2019)). Online reviews have also been shown to be biased (Mayzlin et al. (2014)), but some review platforms have recently implemented systems to detect and remove fake reviews (Luca and Zervas (2016)).

To make progress on understanding the role of online reviews in informing consumers and increasing providers' quality along dimensions that matter to regulators, we combine detailed inspection records from the New York City Department of Health and Mental Hygiene (DoH), consumer reviews from Yelp, and reservation data from OpenTable. Health inspectors periodically visit restaurants to look for different kinds of hygiene violations, from the presence of mice to workers' hygiene. In Section 4 we predict the occurrence of each type of violation during a regulatory inspection as a function of the text contained in recent consumer reviews (Gentzkow et al. (2019b)). We find substantial heterogeneity in how accurately Yelp reviews can predict different types of violations. Such heterogeneity is consistent with the degree to which customers can directly experience those hygiene dimensions. Additionally, we can verify that the words identified by our algorithm as predictive of a particular violation are actually semantically related to that violation. For example, food handling violations are predicted by words like *sick* and *nauseous*, while pests are predicted by words like *cockroaches* and *filthy*. The semantic connection gives us some reassurance that consumers reading those reviews could actually infer hygiene information.

Next, we use the output from our prediction model to estimate the impact of review-based hygiene signals on consumer demand and on restaurants' incentives to comply with hygiene regulation (Section 5). On the demand side, our machine learning procedure allows us to construct signals of hygiene from the text of Yelp reviews. We use those signals to identify Yelp reviews that discuss poor hygiene conditions. We then compare the probability that the restaurant is sold out in the two weeks before and after the submission of the focal review. We find that a restaurant is between 0.4 and 0.7 percentage points less likely to be sold out on OpenTable after the focal review, a 1.8%-3.2% reduction in sold-out probability. The effect of poor hygiene captures about half of the effect of a low-star review.

On the supply side, our machine learning procedure allows us to separate the violations for which Yelp is most informative from other violations. We use a difference-in-differences approach to compare hygiene compliance across violations for which Yelp is more versus less

[2]Glassdoor reports base salaries for NYC heath inspectors of about $40,000. See: `https://www.glassdoor.com/Salary/New-York-City-Department-of-Health-and-Mental-Hygiene-Health-Inspector-Salaries-E212691_D_KO54,70.htm` (accessed on January 15, 2021).

informative by restaurants that are more versus less visible on Yelp. Since review recency is predictive of restaurants' rank in search results, we proxy for visibility with whether a restaurants has received recent reviews and use reviewers' history across all Yelp as an instrument. Our findings suggest that restaurants that are more visible on Yelp tend to violate less along hygiene dimensions for which reviews provide more informative signals.

Overall, our findings suggest that consumer reviews can inform the public only about certain types of hazards that are monitored through regulation. For policy makers, an implication of our results is that limited resources of health inspectors could be targeted to monitor aspects of hygiene for which online reviews are least informative. Of course, the viability of online reviews as a substitute to certain tasks that regulation is responsible for depends on the review platforms' incentives to provide truthful and frequent reviews while policing providers' attempts to manipulate them. We turn to these issues in the conclusion.

Our study contributes to the broad literature on the role of government regulation and online reviews in reducing asymmetric information and moral hazard. While some papers find quality improvements following the disclosure of regulatory inspection outcomes to consumers (Jin and Leslie (2003) and Jin and Leslie (2009)),[3] there is also ample evidence that other forms of regulation designed to increase quality, such as occupational licensing, do not achieve similar quality improvements (e.g., Kugler and Sauer (2005), Farronato et al. (2020), and Barrios (2021)) or only benefit high income geographies (Larsen et al. (2020)).

In our work, we take the government's decision to monitor certain quality dimensions as given, and evaluate the extent to which online reputation mechanisms capture information related to those quality dimensions. We already know that online ratings and health inspection scores are correlated (Kang et al. (2013) and Harrison et al. (2014)), but less is known about whether they capture similar quality dimensions.

Our empirical approach closely relates to papers that use machine learning techniques to predict expert decisions (Kleinberg et al. (2018)). Specific to the setting of restaurant inspections, Kang et al. (2013) and more recently Mejia et al. (2019) show that Yelp reviews are able to track restaurants' inspection outcomes. It is important to notice one key difference between our work and existing efforts to predict restaurant hygiene from online reviews. In existing work, the algorithms are created with the purpose of improving experts' decision making (Glaeser et al. (2016) and Glaeser et al. (2019)). Our focus is different. The goal of our exercise is twofold. First, we evaluate for which dimensions of restaurant hygiene online reviews can offer informative signals. Second, we measure the effect of hygiene signals extracted from review text on customer demand and restaurant hygiene choices. The

[3]Separate work by Simon et al. (2005) confirms the results of Jin and Leslie (2003), although more recently Ho et al. (2019) have challenged their conclusions.

difference has important implications. A prediction problem to inform the health department should include all available sources of information, from online reviews and within the health department. A prediction problem to evaluate whether consumers can obtain hygiene information when reading restaurant reviews, which is the goal of our exercise, should only include information available to consumers through online reviews.

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), Taddy (2015), Gentzkow et al. (2019a), and Greenstein et al. (2016)). Research on online reviews has mostly focused on the aggregate numeric rating that consumers assign to service providers (e.g., Lewis and Zervas (2019), Luca (2019)). But a Yelp star rating typically reflects a consumer overall satisfaction with the provider, and it is a function of several quality dimensions, such as food taste, service, price, and hygiene conditions. The extent to which Yelp stars capture restaurant hygiene will depend on the weight consumers place on hygiene compared to other quality dimensions.[4] Using the text of reviews allows us to break down a consumer overall assessment of a restaurant and separate the effect of hygiene information from everything else, an approach that can be extended to other quality dimensions.

Finally, our work can inform the debate over the regulation of online marketplaces, which heavily rely on consumer reviews to screen and monitor providers. Recent papers have focused on the welfare benefits of flexible labor for Uber drivers (Chen et al. (2019)) and passengers (Cohen et al. (2016)). Farronato and Fradkin (2018) study the welfare implications from Airbnb for travelers, Airbnb hosts, and hotels jointly. Farronato et al. (2020) study the role of occupational licensing regulation when choosing service providers online. Still, very little is known about the role that online reviews have in providing adequate information about providers' quality as an alternative to regulatory screening (Einav et al. (2016)). Our work sheds some initial light on the signaling value of online reviews as an alternative to regulatory disclosures.

We organize the paper as follows. In Section 2 we present a simple Bayesian framework to understand the role of hygiene signals in affecting restaurant demand and quality decisions. In Section 3 we describe our empirical context and the data we use for our analysis. In Section 4 we present our approach to predict health violations from the text of Yelp reviews. In Section 5 we estimate the effect of review-based hygiene signals on demand and supply incentives, and in Section 6 we conclude by discussing the limitations and implications of our work.

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

2 Theoretical Framework

In this section we present a simple theoretical framework of a market with asymmetric information to motivate our empirical section. We assume that consumers are uncertain about providers' quality, but online reviews serve as an informative signal. Consumers update their beliefs in a Bayesian fashion, and choose providers to maximize their expected utility. In turn, providers invest in quality as a function of related costs and benefits, where the benefits are induced by quality signals driving demand.

Let θ denote the true underlying single-dimensional quality of a restaurant. Consumers believe that θ is distributed according to a normal prior distribution with mean θ_0 and unit variance. We let s denote a signal of restaurant quality, and assume that its distribution conditional on the true quality level is normal with mean θ and variance σ_s^2. Restaurant demand is linear in price and expected quality conditional on the signal realization:

$$D(p, s) = -\alpha p + \beta E(\theta|s). \tag{1}$$

Given our distributional assumptions, we have a closed form solution for the expected quality: $E(\theta|s) = \theta_0 \frac{1}{1+1/\sigma_s^2} + s\frac{1/\sigma_s^2}{1+1/\sigma_s^2}$. Expected quality is a weighted average of the prior mean and the signal realization where the weights are a function of the informativeness of the prior (normalized to 1) and the signal ($1/\sigma_s^2$).

The restaurant chooses its quality level, which is costly, to maximize its profits:

$$\max_{\tilde{\theta}} E\left[D(p, s|\tilde{\theta})\right] p - C(\tilde{\theta}).$$

Prices are assumed to be fixed, and $C(\tilde{\theta}) = c\tilde{\theta}^2 + \eta\tilde{\theta}$ is an increasing and convex cost function, where $\eta \sim N(0, c^2)$. The assumptions on the cost function simplify the closed-form solutions below.

The timeline of decisions is as follows: the restaurant chooses its quality level, the signal realizes, then the representative consumer chooses where to eat. The perfect Bayesian equilibrium is defined by a quality level θ^* ($1/\sigma_s^2$) that solves the restaurant's profit maximization problem, and by a prior distribution that is consistent with the distribution of cost shocks. The distributional assumptions imply that in the restaurant maximization problem we have $E\left[D(p, s|\tilde{\theta})\right] = -\alpha p + \beta\left[\theta_0 + (\theta - \theta_0)\left(\frac{1}{1+\sigma_s^2}\right)\right]$. The restaurant's first order condition determines the optimal quality investment as a function of the signal informativeness as well

as customers' sensitivity to the signal:

$$\theta^* \left({}^1\!/\!_{\sigma_s^2}, \beta \right) = \frac{\beta}{c} \left(\frac{{}^1\!/\!_{\sigma_s^2}}{1 + {}^1\!/\!_{\sigma_s^2}} \right) p - \frac{\eta}{c}. \tag{2}$$

Given that η is normally distributed, the prior distribution of quality levels is normal with mean $\theta_0 = \frac{\beta}{c} \left(\frac{{}^1\!/\!_{\sigma_s^2}}{1+{}^1\!/\!_{\sigma_s^2}} \right) p$ and unit variance.

Note that in our setup restaurant prices are fixed and not a function of restaurant quality. This is not an innocuous assumption, but one that we make to avoid multiplicity of equilibria. It is reasonable in a context where restaurants make short-run quality decisions that are unlikely to affect menu prices – e.g., whether to install mouse-traps or how frequently to clean kitchen counters. Indeed, the National Restaurant Association does not include hygiene maintenance as an important factor when considering whether to raise prices.[5] 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 prices on their menus.[6]

The theory can be extended to multi-dimensional quality ($\boldsymbol{\theta}$ is a now vector) as long as demand and restaurant costs are additively separable in the various dimensions of quality. Under this assumption, the restaurant's demand and optimal quality choices are independent across the elements of $\boldsymbol{\theta}$.

This simple model provides useful comparative statics. First, Equation 1 implies that a bad signal realization should decrease demand as long as customers are sensitive to the signal ($\beta > 0$) and the signal is informative (${}^1\!/\!_{\sigma_s^2} > 0$):

$$D(p, s) < D(p, s') \text{ for any } s < s'. \tag{3}$$

Second, Equation 2 implies that the chosen quality level is increasing in the informativeness of the signal (${}^1\!/\!_{\sigma_s^2}$) and in the sensitivity of demand to the signal (β). We can thus compare quality choices across two quality dimensions, θ_1 and θ_2 such that ${}^1\!/\!_{\sigma_{s_1}^2} > {}^1\!/\!_{\sigma_{s_2}^2}$, and across two restaurants, i and j such that $\beta_i > \beta_j$. When the signal for quality dimension θ_1 is more informative than the signal for θ_2 and demand for restaurant i is more responsive to signal realizations than demand for restaurant j, Equation 2 implies that

$$\theta_1^* \left({}^1\!/\!_{\sigma_{s_1}^2}, \beta_i \right) - \theta_2^* \left({}^1\!/\!_{\sigma_{s_2}^2}, \beta_i \right) > \theta_1^* \left({}^1\!/\!_{\sigma_{s_1}^2}, \beta_j \right) - \theta_2^* \left({}^1\!/\!_{\sigma_{s_2}^2}, \beta_j \right) \tag{4}$$

for ${}^1\!/\!_{\sigma_{s_1}^2} > {}^1\!/\!_{\sigma_{s_2}^2}$ and $\beta_i > \beta_j$. In words, this means that the quality level chosen by restaurant

[5] See `http://www.restaurant.org/Manage-My-Restaurant/Marketing-Sales/Food/Is-it-time-to-raise-your-prices` (accessed on January 15, 2021).

[6] See `http://smallbusiness.chron.com/restaurant-food-pricing-strategies-14229.html` (accessed on January 15, 2021).

i for quality dimension θ_1 is higher relative to the quality level chosen for quality dimension θ_2 and relative to the choices of restaurant j.

We will test Equations 3 and 4. We take the underlying quality of restaurants, $\boldsymbol{\theta}$, to be the outcome of DoH's health and safety inspections. In Section 3 we justify our choice based on the design of inspection cycles in New York. We let $\boldsymbol{\theta}$ be a 20-element vector, corresponding to the most frequent violation codes inspected by the DoH. For each element of $\boldsymbol{\theta}$, in Section 4 we extract the corresponding signal from the text of Yelp reviews with machine learning methods. The prediction accuracy with which each element of $\boldsymbol{\theta}$ is predicted by the text of Yelp reviews is our measure of signal informativeness.

3 Data

This section describes the data on health inspections, online reviews, and online reservations that we use in our empirical analysis.

First, data on health inspections come from the New York City Department of Health and Mental Hygiene through a FOIA (Freedom of Information Act) request. The DoH conducts unannounced inspections of food serving entities in the five boroughs of New York City at least once a year.[7] Inspectors check for compliance across many practices, including food handling, personal hygiene, and vermin control. Inspectors separately assign points for each violation code (more points imply more severe violations), and then tally the points to assign a final inspection score.[8]

We have data at the level of each violation code for all inspections conducted between July July 2010 (when the most recent overhaul of restaurant inspections was implemented) and September 2016. There are over 100 violation codes that evaluate restaurants on multiple dimensions: vermin (e.g., evidence of mice), food temperature (e.g., hot food item not held at or above 140F), facilities (e.g., improper sewage disposal system), food handling (e.g., raw food not properly washed), overall hygiene (e.g., inadequate personal cleanliness of staff), contamination (e.g., worker does not wash hands throughly), and regulatory (e.g., wash-hand sign not posted).[9]

The DoH program uses dual inspections to help restaurants improve before being assigned a letter grade, A through C, to post at the restaurant door. This feature is useful for our prediction exercise, as explained in Section 3.1. In particular, every year a restaurant

[7] The DoH performs inspections of restaurants, coffee shops, bars, nightclubs, most cafeteria, and fixed food stands.

[8] `http://www1.nyc.gov/site/doh/services/restaurant-grades.page`, accessed on January 15, 2021.

[9] For more details, see `https://fivethirtyeight.com/features/how-data-made-me-a-believer-in-new-york-citys-restaurant-grades/`, accessed on January 15, 2021.

undergoes an inspection cycle. An inspection cycle is a series of inspections consisting of an initial inspection, possibly followed by a reinspection and compliance inspections that lead to a letter grade update (Figure 1). An initial inspection is followed by a reinspection several weeks later for restaurants that do not receive an A grade on initial inspection. An inspection score of less than 14 points on either initial or reinspection results in an A grade. With a score of 14 or more points on initial and 14-27 points on reinspection a restaurant receives a B grade. With a score of 14 or more points on initial and 28 or more points on reinspection a restaurant receives a C grade. The restaurant must post the letter grade at its entrance for patrons to see, or alternatively it can post a "Grade Pending" card if it decides to dispute B or C grades at an administrative tribunal and until the tribunal makes a final decision.[10] From these records, we can re-construct the card posted at the door at any point in time.

In addition to inspection outcomes, the DoH inspection data provide us with restaurant level characteristics, such as the type of cuisine, the type of restaurant (chain or independent), date of entry (and exit if it went out of business by September 2016), and anonymized inspector identifiers for each inspection. We confirmed with the DoH that the assignment of inspectors to restaurant inspections is random.

The second dataset includes online reviews from Yelp.com, which are publicly available. Yelp contains business information, such as zip code and phone number, and a historical record of reviews, inclusive of text, time of review submission, and an identifier of the reviewer.[11] From this record we are able to construct the average Yelp star rating of a business at any point in time. In order to construct instruments for number of reviews of our focal restaurants, we also collect the entire Yelp activity of every user who ever submitted a review for a restaurant in New York City. This allows us to construct user level measures of propensity to rate restaurants online.

The third and last dataset comes from OpenTable. Since April 2013, for every day and restaurant on OpenTable, we have information on whether the restaurant had a table available for 2 people between 6:30 and 7:30PM. We match businesses from the DoH with businesses on Yelp and OpenTable using Yelp search algorithm, and matching on restaurant name, address, and phone number.

[10] More details on inspection regulation and grading can be found at `http://www1.nyc.gov/assets/doh/downloads/pdf/rii/inspection-cycle-overview.pdf`, accessed on January 15, 2021. In principle consumers could access the list of violations found during an inspection by visiting the DoH website at `http://www1.nyc.gov/site/doh/services/restaurant-grades.page`, but anecdotally this rarely happens.

[11] Reviews that Yelp deems "fake" are not displayed online nor count towards the average rating of a restaurant (see Luca and Zervas (2016)).

3.1 Descriptives

Overall, of the 49,034 individual businesses present in the DoH data since July 2010, 61.3% were matched to a business on Yelp,[12] and 4.5% were matched on OpenTable. Table 1 presents descriptive statistics at the restaurant level for the three samples: all restaurants inspected by the DoH, restaurants with Yelp reviews, and restaurants available for booking on OpenTable. 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.

Initial inspections occur throughout the year, with a lower number of inspections during vacation periods (summer and winter holiday season). The interval between inspection cycles depends on the sanitary condition of the restaurant during the previous inspection. If a restaurant has an A-grade at initial inspection, it will be inspected again after approximately one year. If a restaurant scores 14-27 points during the initial inspection and gets a A- or B-grade at reinspection, it will be inspected after 5-7 months since the most recent reinspection. If the restaurant scores 28 points or more at initial inspection or it gets a C-grade at reinspection, it will be inspected 3-5 months since the most recent compliance inspection.[13] In practice, there is a substantial amount of variability in the time interval between inspection cycles, as pictured in Figure 2. Such variability is intentional, given that inspectors show up unannounced to evaluate restaurants' sanitary conditions.

The distribution of violation scores at initial inspection is depicted in the top panel of Figure 3. 36% of restaurants obtain an A-grade during the initial inspection. The other restaurants obtain a violation score corresponding to a B (38%) or a C (26%). Those restaurants whose score would imply a B- or a C-grade are reinspected within a few weeks, and the inspector (likely not the one who conducted the initial inspection) again shows up unannounced.[14] After reinspection, the vast majority of restaurants get to display an A-grade. As the bottom panel of Figure 3 shows, 77% of restaurants end an inspection cycle with an A-grade, 17% end it with a B-grade, and only 6% end it with a C-grade.

Figure 3 shows that compared to the final grade, at initial inspection there is much less bunching at the threshold between A and B grades, and no bunching at all beween B and C

[12] This is not because it is difficult to match businesses, but rather because food serving places that the DoH inspects include entities that are unlikely to be on Yelp, such as workplace cafeterias.

[13] Compliance inspections are follow-on inspections conducted to check that restaurants have resolved specific critical violations. A list of critical violations is at `http://www1.nyc.gov/assets/doh/downloads/pdf/rii/blue-book.pdf`, accessed on January 15, 2021.

[14] See Appendix Figure A1 for a distribution of the time lag between a reinspection and the initial inspection.

grades. This results from the specific structure of the inspection cycle, which gives restaurants a second chance to obtain an A grade after initial inspection. This structure works in our favor because it makes initial inspections a more truthful assessment of restaurant hygiene.

Table 2 shows that even if observable restaurant characteristics such as price group or cuisine are somewhat correlated with the outcome of initial inspections, with the exception of chain affiliation the effects of observables are relatively small in size, and altogether explain about 8% of the variation in inspection scores. Inspector fixed effects alone explain more of the variation in inspection scores than restaurant characteristics. Indeed, the R-squared of the regression more than doubles when including inspector fixed effects (columns 2 and 4 in Table 2). This result emphasizes the discretionary power that inspectors have over restaurants' hygiene grades, although the random allocation of inspectors to restaurants does not lead to systematic differences in hygiene violations across restaurants.

Table 3 shows that there is substantial variation in the scores that restaurants obtain during consecutive inspection cycles. For example, of the 126,540 inspection cycles obtaining an A-grade during cycle t, 41% score between 0 and 13 points during the initial inspection of cycle $t+1$, 36% score between 14 and 27 points, and 22% score 28 or more points. By the end of the cycle most restaurants end up with a A-grade, but still 15% of restaurants with a A-grade in t lose it by inspection cycle $t+1$. Of those, 13% drop to B-grade, and 2% drop to C-grade.

Initial descriptives highlight sizable variability in restaurant's hygiene conditions, both across restaurants and within restaurants over time. In addition, the low incentives for initial inspections to be manipulated and the random allocation of inspectors to restaurants allow us to leverage the outcome of inspections to construct hygiene signals from Yelp reviews, as described in the next section.

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 the true underlying quality of a restaurant along the various dimensions of hygiene. In the previous section we have argued that initial inspections are a relatively truthful measure of restaurant hygiene, so we use the outcome of initial inspections as a restaurant's true hygiene level. We use machine learning methods to predict the occurrence of specific violations during initial inspections – say, presence of cockroaches – from the text of online reviews. The prediction accuracy of our model, which will differ by violation code, will be our measure of

signal informativeness. The better Yelp can predict the occurrence of 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 Table 4 and constitute over 80 percent of all violations cited during initial inspections.

Unlike Yelp stars, the text of Yelp reviews provides a breakdown of a consumer overall assessment of restaurant quality. For example, a consumer might reveal that they gave a 3-star rating to a restaurant with excellent food but rude staff. Or a consumer might describe that the food arrived lukewarm and that they experienced stomach cramps shortly after. To the extent that consumer reviews are a function of restaurant hygiene, different violations discovered during an inspection will result in different words used in online reviews. For example, if an inspector finds evidence of roaches in the restaurant's premises, it is more likely that in the weeks preceding the inspection consumers mention cockroaches when reviewing that restaurant.

To incorporate review text in our subsequent analysis we need to reduce the dimensionality of our text data. Yelp reviews contain hundreds of thousands of unique words, and using each one of these as a covariate is impossible as we would end up with more covariates than observations. We solve this problem with an approach that was recently applied to analyze congressional speech by Gentzkow et al. (2019b), whose results were also used to evaluate political slant in Wikipedia articles by Greenstein et al. (2016). We first extract signals of hygiene from text, and then we measure the informativeness of those signals. The two subsections below describe each of the two steps.

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, and we 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 3 months prior to the inspection. There are two reasons for choosing the preceding 3 months. The first reason is that online reviews are not extremely frequent, so a longer time interval can capture more heterogeneity across restaurants. Indeed, the median restaurant in our sample receives one review every 28 days.[15] Appendix Figures A3 and A4 show the distribution of the number of recent reviews across restaurant-inspections, and how the frequency of reviews

[15]This is likely a lower bound on the time between reviews because we compute it by dividing the number of a restaurant's reviews received by the number of days between its last and first reviews. 28 days is the median of the distribution of this metric across restaurants, and the interquartile range is one review every 10-77 days. The distribution is highly skewed, with some restaurants being frequently reviewed – 20% of the restaurants have at least one review per week – and other restaurants being almost never reviewed.

has significantly increased over time. The second reason for choosing a 3-month time window is that the minimum time between inspection cycles is about 3 months, which allows us to allocate each review to at most one inspection.

We do not keep every word as it appears on Yelp. To construct our vocabulary of words, we take the raw text of Yelp reviews and eliminate all elements other than words, such as punctuation and numbers. We then replace each word with their root using Porter stemming algorithm (Porter, 1980). Finally, to exclude both common and rare words, we exclude (stemmed) words that appear in fewer than 5 reviews, or in more than 50% of reviews. We end up with a vocabulary containing 12,176 words.

In the construction of our vocabulary, we perform a final pre-processing step that is best illustrated with an example: the word "clean" has a different meaning depending on the rating of the review it appears in ("clean" likely implies "dirty" in the context of a 1-star review.) To deal with this issue, we separately count word frequencies in three rating groups: 1- and 2-star reviews; 3-star reviews; and 4- and 5-star reviews. This effectively multiplies the size of our vocabulary by 3, the number of rating groups we consider. In the rest of the paper, unless otherwise noted, we denote each word-star rating combination as a separate *word* in our vocabulary.

The combined text of reviews submitted in the 3 months preceding each initial inspection constitutes a document, which is simply a collection of word counts in no particular order. We let c_i denote the observed vector of word counts in reviews associated with inspection i. We assume that c_i is drawn from a multinomial distribution

$$c_i \sim \text{MN}(q_i, m_i), \tag{5}$$

where m_i is the document length – total number of words in reviews linked to inspection i – and 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, $q_{ij} = \frac{e^{\eta_{ij}}}{\sum_k e^{\eta_{ik}}}$, where

$$\eta_{ij} = \mu_j + \alpha_j r_j + \beta_j x_i + \phi_j (r_j \cdot v_i) + \epsilon_{ij}. \tag{6}$$

In the above equation, the coefficients of interest are contained in the vectors ϕ_j, and they tell us by how much the frequency of word j changes when each violation in v_i occurs. In particular v_i is a vector of dummies, one per violation code, that are equal to 1 in the presence of a violation. The intercept μ_j captures the overall frequency of word j. The terms r_j are rating-group dummies that allow word probabilities to vary depending on whether a review is positive (4-5 stars), neutral (3 stars), or negative (1-2 stars). For instance, we

expect the word "great" to be more frequent in positive than in negative reviews. The vector v_i contains dummies for each violation code. We interact v_i with the rating-groups dummies r_j to flexibly capture changes in word frequencies by rating-group and violation code. For example, if violation code 04M (which pertains to roaches) occurs, we expect the frequency of the word "roach" to increase in the restaurant's negative reviews but not its positive reviews. The vector x_i includes various controls: year and month of inspection fixed effects, cuisine fixed effects, ZIP code 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 ϕ_j can pick up correlations between a restaurant's propensity to commit a specific violation and the restaurant's observable characteristics. For example, consider violation `02G`, which pertains to food not being held at a cool enough temperature and which sushi restaurants are more likely to violate 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, in and by itself, the word "sushi" does not suggest that during a specific inspection the restaurant in question was more likely to be cited for violation `02G`. Including restaurant controls helps avoid spurious correlations between words and violations like this one.

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. Fortunately, we can approximate the multinomial logit model with as many independent Poisson regressions as we have words following the distributed multinomial regression framework of Taddy (2015). This approximation makes estimation tractable at the cost of ignoring correlations in word occurrence. For example, the model treats the frequencies of the words "pad" and "thai" as independent.

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

$$\min_{\beta_j,\phi_j} \frac{1}{N} \sum_{i=1}^{N} l(c_{ij}, \eta_{ij}; \mu_j) + \lambda_j(\|\alpha_j\|_0 + \|\beta_j\|_0 + \|\phi_j\|_0), \tag{7}$$

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.[16] Lasso is natural in our setting as we expect many coefficients for our rich set of controls to be zero. For example, we expect the dummy for Japanese cuisine to have a zero coefficient for the word "pizza."

[16] We apply a tiny penalty to the intercept term to aid convergence.

We tune the word-specific penalty parameters λ_j using 5-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 end 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.

Results

The matrix of estimated coefficients $\hat{\Phi}$ with entries $\hat{\phi}_{jv}$ tells us by how much the frequency of word j changes when violation v occurs. While $\hat{\Phi}$ 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 $\hat{\Phi}$. These remaining entries correspond to the strongest predictors of word frequencies.

One complication we have to deal with is that each word is estimated using a different lasso path. Thus, the same increase in penalty will induce different amount 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 CV 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 $\mathrm{SE}_j(\lambda) = \sqrt{\mathrm{Var}(\mathrm{CV}_i(\lambda), \ldots, \mathrm{CV}_5(\lambda))/5}$ is the cross-validation standard error. Then, for each violation code v, we sort words in descending order γ_{jv}, which provides us with a ranking of the most predictable changes in word frequency when violation v occurs.

Table 5 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 the presence of roaches, we see that Yelp reviewers

tend to increase their use of words like *roach* and *filth* in the 3 months leading to the violation being uncovered by a heath inspector. A similar pattern appears for violations `02B` and `02G`, which are predicted by words like *sick* and *nauseous* and pertain to keeping food at an appropriate temperature. If a consumer were to read reviews containing these words, we might expect them to correctly predict that a restaurants has roaches, or that food is not kept at the right temperature. By contrast, consider violation code `10F`, which pertains to inappropriate construction of non-food contact surfaces. A priori we might not expect the average Yelp reviewer to know what materials or methods are permitted for the construction of non-food contact 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 specifically alert a consumer that 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`, for which the text is descriptive of the actual violation, compared to `10F`, for which text is much less specific. We quantify these differences in informativeness in Section 4.2.

Constructing Low-Dimensional Signals of Hygiene

We use the predictive model to construct interpretable low-dimensional signals of hygiene from review text. For each inspection and violation code, we map the text contained in all negative reviews occurring up to 90 days prior to the inspection to a single dimensional score. This score will be higher when the reviews preceding the inspection contain many words (such as those in Table 5) that are typically associated with the violation in question.

To compute these scores for a single inspection i, we multiply the estimated matrix of coefficients $\hat{\Phi}$ by the vector of word frequencies c_i to obtain

$$z_i = \hat{\Phi} c_i,$$

which is a vector with 20 entries, one for each violation code. These scores 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. A key property of these SR projections is that they are sufficient statistics for the violation codes: $v_i \perp c_i | z_i$. In words, given the low-dimensional SR projections z_i, the high-dimensional vector of text c_i is orthogonal to the violation code dummies. This property of the SR projections allows us to reduce the dimensionality of the text data from tens of thousands of words down to a single score for each violation code that captures variation in review text specifically pertaining to the occurrence of that violation.

4.2 Evaluating the Informativeness of Yelp Hygiene Signals

We can now evaluate the informativeness of Yelp hygiene signals constructed as SR projections. Our basic approach is to compare the predictive power of two classifiers predicting the occurrence of 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 the restaurant hygiene from different sources of information they have access to.[17] Algorithm 1 describes the steps we take to build and evaluate these two classifiers in detail. Next, we discuss a few key components of our algorithm.

A key decision we have to make is what features to include in each classifier. This decision requires assumptions regarding the information sets of consumers. For the baseline classifier, we use the letter grade posted on the restaurant door at the time of the inspection. This is the letter grade that a consumer would see if they were to walk into the restaurant.[18] 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 constructed from reviews submitted in the preceding 3 months.

When computing SR projections for each inspection, we exclude all inspections of the restaurant whose violations we are trying to predict to avoid data leakage.[19] To do that, we make careful use of cross-validation. We divide our data into 5 different folds, with each restaurant's entire set of inspections assigned to the same fold. Given that a restaurant is in, say, the first fold, we use the other four folds to estimate $\hat{\Phi}$, and we use the first fold to compute the SR projections and evaluate their predictive power.

We train gradient boosted tree classifiers (Ke et al., 2017) as described in Algorithm 1.[20] We evaluate the performance of the baseline and review-augmented classifiers using the AUC (area under the curve) metric.[21] An AUC of 0.5 means that our classifier performs no better

[17] Our objective is not to maximize prediction accuracy, in which case we would include every available feature for prediction.

[18] 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, from other consumers, or they might look up a restaurant's entire history of inspections in the DoH database. We assume that most consumers do not engage in this costly search behavior.

[19] To see how data leakage can arise recall that SR projections associate violations with changes in word frequencies via the learned projection matrix $\hat{\Phi}$. If we included the focal inspection when learning the projection matrix $\hat{\Phi}$, we would be peeking at the outcome we are trying to predict, resulting in data leakage and overstated classifier accuracy.

[20] We also carry out the analysis with a penalized logistic regression, and obtain similar results.

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

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

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 words counts

Output: $\text{AUC}_0, \text{AUC}_1$: 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 $\hat{f}_1(\mathcal{D}_V^{-k_2}, \mathcal{D}_H^{-k_2}, \mathcal{D}_R^{-k_2}, Z^{-k_2}); h)$;

$\text{AUC}_{0,k_1} \leftarrow$ AUC of $\hat{f}_0$ for test fold k_1;

$\text{AUC}_{1,k_1} \leftarrow$ AUC of $\hat{f}_1$ for test fold k_1;

`/* Compute average CV AUC of each classifiers */`

$\text{AUC}_0 \leftarrow \frac{1}{5}\sum_{k_1=1}^{5} \text{AUC}_{0,k_1}$; `/* AUC without Yelp hygiene signals */`

$\text{AUC}_1 \leftarrow \frac{1}{5}\sum_{k_1=1}^{5} \text{AUC}_{1,k_1}$; `/* AUC with Yelp hygiene signals */`

than a random guess.

Figure 4a displays AUCs for each violation code and each of the two classifiers separately. A few interesting patterns emerge. First, all AUC metrics range between 0.51 and 0.68, suggesting that it is relatively difficult for consumers to predict the incidence of 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. After all, the inspector evaluates multiple dimensions of hygiene on a random day during just a few hours.

Second, although the letter grade is in general a poor predictor of each individual violation – there is no AUC above 0.55 in the baseline classifier – we observe more sizable variation in the performance of the review-augmented classifier, with some violations being easier to predict than others. Figure 4b displays the incremental improvement in AUC from the review-augmented classifier relative to the baseline classifier. We use this incremental improvement as the degree of informativeness of Yelp reviews for each violation code. Note that the vast majority of the improvement comes from review text, rather than star rating (see Appendix Figure A6).

Comparing the violation codes ranked higher and lower in Figure 4b, it becomes apparent that Yelp reviews tend to be better predictors of violations such as vermin, food temperature and food handling than 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. However, the fact the our machine learning algorithm can predict one dimension of hygiene better than another does not immediately imply that Yelp readers can infer hygiene information from Yelp reviews. One of the results that provide some reassurance that our algorithm might approximate what readers can gather from online reviews is the interpretability of our results. As Table 5 shows, the words predicting the violations ranked highest are descriptive of the actual infringement. For example, *roach* is highly predictive of `04M` (Live roaches present in facility's food) and *nauseous* is highly predictive of `02B` (Hot food item not held above 140^o F.)

The results of this section point to one main conclusion: consumers discuss restaurant hygiene on Yelp, but not all dimensions of hygiene are equally captured by consumer 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 consumer choice of where to eat, and restaurant incentives to be clean.

5 Effects of Hygiene Signals on Demand and Supply

In order to confirm that our exercise from Section 4 picks up information that consumers also take into account when choosing where to eat, we provide causal evidence that Yelp hygiene information affects consumer choices and some suggestive evidence that it may even affect restaurant incentives. We devote the first subsection to consumer choices and the second subsection to restaurant incentives.

5.1 Consumer Demand

Demand is a function of signals that customers receive about restaurant quality. These signals include online ratings and health grades posted at a restaurant door, and have been found to affect restaurant demand (Jin and Leslie (2003) for health grades, Luca (2019) and Anderson and Magruder (2012) for online ratings). In this section we use the hygiene information contained in the text of Yelp reviews to estimate its effect on consumer demand.

In order to analyze how the information contained in online reviews is related to demand, we use the probability of being sold out on OpenTable as our demand proxy (see Appendix Figure A7 for descriptives on this outcome). The advantage of using sold-out probability is that it is a measure of restaurant success that changes on a daily basis and thus allows us to look at changes in demand immediately following the submission of a particular review. The drawback is that we only have it for a small subset of restaurants (see Table 1 for selection on observables).

To test for a causal relationship between the hygiene signal on Yelp and the reduction in the sold out probability we take advantage of the submission time of Yelp reviews. Assuming that the *timing* of review submission 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 the submission of a review that discusses the restaurant's poor hygiene.

How do we identify reviews discussing poor hygiene? As the previous section highlighted, the text of Yelp reviews is more informative for some violation codes than others, so we rank violations according to the informativeness of Yelp reviews as described in the previous section (Figure 4b). We then restrict attention to the 5 violations for which Yelp is most informative: 02B (Hot food item not held at or above 140^o F), 04H (Raw, cooked or prepared food is adulterated, contaminated, cross-contaminated, or not discarded in accordance with HACCP plan), 04M (Live roaches present in facility's food and/or non-food areas), 04A (Food Protection Certificate not held by supervisor of food operations), and 02G (Cold food item held above 41^oF except during necessary preparation). In truth, the choice to focus on the top 5 violations is somewhat arbitrary, but the results do not depend on the specific

threshold.[22]

Restricting attention to the 5 violation codes for which Yelp is most informative, we construct a hygiene signal for restaurant i on a given day by summing the corresponding sufficient reductions contained in 1-2-3 star reviews that were submitted for restaurant i on that same day. Recall that each sufficient reduction approximates the probability that the corresponding violation occurs. The sum of multiple sufficient reductions informs consumers about the corresponding violations jointly. A higher hygiene signal constructed this way means worse conditions, and it can originate from one review discussing one hygiene dimension in a very negative way, one review discussing multiple hygiene dimensions, or even multiple reviews submitted on the same day discussing one or more hygiene dimensions.

Among all days when restaurants receive 1-2-3 star reviews, we then identify the focal event as a day when a restaurant receives a hygiene signal among the 20% most negative. We consider a month around the focal event to estimate the following regression,[23] motivated by Equation 3 in the theory section:

$$sold\_out_{ijt} = \sum_{t=-15}^{15} \beta_t + \alpha X_{ijt} + \nu_{ij} + \epsilon_{ijt}. \tag{8}$$

The subscript i denotes a restaurant, j denotes the set of (possibly multiple) reviews submitted on the event day, and t denotes the days since the event. The outcome, $sold\_out_{ijt}$, is equal to 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. Since we look at a month around the focal event, t goes from -15 days to +15 days. The coefficient on the day of the submission is normalized to zero. Controls include restaurant-event fixed effects ν_{ij}, day of the week fixed effects, Yelp average star ratings, and the hygiene card posted at the door. We cluster standard errors at the restaurant level.

The results are presented in the left plot of Figure 5 and show a decrease in the probability of selling out that is gradual at first and stabilizes at its minimum around a week later. We can aggregate the days until the review submission and those after the review submission to estimate a single "post review submission" coefficient. The results are presented in column 1 of Table 6 and imply a reduction in the sold out probability of 0.7 percentage points. This represents a 3.2% reduction in the average sold out probability of 0.22.

In order to rule out the possibility that this is simply given by the low-star review, we can perform the event study analysis as a difference-in-differences, comparing the probability

[22]Results with other aggregations are presented in Appendix Table A2.

[23]In practice low-star reviews are relatively rare events, so time windows around focal events do not tend to overlap.

of selling out of restaurants receiving a 1-2-3 star review with the probability of selling out of restaurants receiving a 1-2-3 star review containing a negative hygiene signal among the worst 20%:

$$sold\_out_{ijt} = \sum_{t=-15}^{15} \beta_t + \sum_{t=-15}^{15} \gamma_t * bad\_hygiene\_signal_{ij} + \alpha X_{ijt} + \nu_{ij} + \epsilon_{ijt}. \quad (9)$$

Relative to Equation 8, the new specification includes all days when restaurants receive 1-2-3 star reviews as focal events, but we also interact the day-since-event fixed effects with a dummy for whether the reviews on the focal day contain a hygiene signal among the 20% worst 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 a negative hygiene signal. The γ_t coefficients are plotted in the right panel of Figure 5. Albeit noisier, the estimates confirm an additional decrease in the sold out probability compared to restaurants that simply receive low-star reviews. The vast majority of the daily coefficients after the event are negative and significantly different from zero. Aggregating all days before the event and all days after the event leads to a single difference-in-differences coefficient as presented in column 2 of Table 6. This estimate confirms that restaurants receiving a negative hygiene signal in a low-star review are 0.4 percentage points less likely to be sold out compared to restaurants receiving a low-star review in the two weeks following the review submission. The difference-in-differences coefficient estimate is about 57% of the coefficient estimate from column 1 (0.004/0.007), thus implying that more than half of the reduction in sold out probability following bad reviews is attributable to the hygiene information contained in the text of the reviews.

The estimated effects are robust to a number of checks. First, we verify that the results do not change if instead of using all 1-2-3 star reviews as the control group, we just focus on 1-2-3 star reviews with the 20% least negative hygiene signals. Second, because receiving a review is more likely to happen when a restaurant has high demand, we remove 5 days around the review date (from 2 days prior to 2 days following) to avoid capturing reverse causality. 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 A1 shows that the difference-in-differences coefficients do not change across these different specifications. Finally, because the decision to sum the sufficient reduction of the 5 violations 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 the sufficient reduction of a single violation, and end with the sum of the sufficient reductions across all 20 most frequent violation codes. Every time we re-estimate the difference-in-differences

specification. The coefficients of interest are presented in Appendix Table A2. The first coefficient is small and statistically indistinguishable from 0, while all other coefficients are around 0.004 as in the baseline.

5.2 Restaurants' Incentives

In the previous subsection we showed that negative hygiene signals on Yelp decrease restaurant demand. In this subsection we focus on supply. Since restaurant quality choices are increasing in the informativeness of a signal and in the responsiveness of demand to that signal, Equation 4 implies that restaurants with higher exposure to Yelp should violate less along dimensions of hygiene for which Yelp is more informative.

In order to test this hypothesis, we take advantage of our finding that the informativeness of Yelp signals differs across violation codes (Section 4). We also take advantage of the fact that Yelp's search algorithm effectively controls how visible a restaurant is to consumers, directly affecting the sensitivity of consumers to information contained in Yelp reviews. Consumers typically search for restaurants in a given location, and Yelp identifies which restaurants are displayed and the order in which they are displayed. Effectively, even if a restaurant is on Yelp, it is 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 compared to lower ranked restaurants.

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 the search results. (We provide evidence of this in Appendix A3.) For this reason we define a restaurant's exposure to Yelp with a dummy for whether a restaurant has received any reviews in the last 90 days. We want to test the hypothesis that 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 run OLS regressions of the following type:

$$\begin{aligned} violation_{vit} = \alpha * has\_recent\_reviews_{it} + \beta * yelp\_informative_{v} + \\ \gamma * has\_recent\_reviews_{it} * yelp\_informative_{v} + \delta X_{vit} + \epsilon_{vit} \end{aligned} \tag{10}$$

where $violation_{vit}$ is equal to one if restaurant i was found violating code v during inspection t. The dummy $has\_recent\_reviews_{it}$ is equal to one if the restaurant has received any Yelp reviews in the 90 days prior to inspection t. Finally, $yelp\_informative_{v}$ is equal to 1 for the

top 5 violation codes for which Yelp is most informative, as defined in Section 5.1.[24] In the most restrictive specification, the vector of controls X_{vit} includes inspection fixed effects and violation code fixed effects.

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 restaurants that were recently reviewed on Yelp compared to other restaurants and other dimensions of hygiene. In the specification with the most stringent set of controls, γ measures the restaurant propensity to violate on specific hygiene dimensions, conditional on their overall hygiene level and the inspector's effort on that particular day, and conditional on time invariant factors that make it easier or harder to comply with a specific hygiene code. We expect γ to be negative.

Despite the inclusion of stringent controls, unobservable characteristics may affect both whether a restaurant receives reviews on Yelp and the restaurant's hygiene level during an inspection. For example, cleanliness and Yelp reviews may be the result of the restaurant's effort to increase its appeal to customers. Note that our specification would suffer from omitted variables only if this effort impacted certain hygiene dimensions – those for which Yelp is more informative – more than others. To address this possibility, we take advantage of reviewers' behavior across the entire Yelp platform. Specifically, we define a reviewer's propensity to rate businesses online as the average number of reviews submitted on Yelp during their tenure on the site. We instrument for whether a restaurant at time t has received any reviews in the last 90 days with the average rating propensity of *all* reviewers to the focal restaurant who have rated it up until time t. Note that this instrument is valid if consumers' propensities to review businesses online is uncorrelated with restaurant hygiene decisions except through the effect it has on a restaurant's exposure to Yelp. This assumption must hold conditional on the stringent controls included in Equation 10, so we allow for the fact that more frequent reviewers may select more popular restaurants or specific cuisines when dining out. Because we interact restaurant's exposure with whether Yelp is informative for a particular violation code, we effectively have two endogenous variables, $has\_recent\_reviews_{it}$ and $has\_recent\_reviews_{it} * yelp\_informative_v$. We use the main instrument and its interaction with $yelp\_informative_v$.

We present both OLS and IV results in Table 7. The first stage regression results for our IV estimates are presented in Table 8, and the Kleiberger-Paap rk Wald F statistic, which

[24]Like in Section 5.1, the results are not sensitive to the precise number of violation codes for which we consider Yelp to be informative, although since the difference-in-differences coefficient effectively compares the the propensity to violate across two groups of violations, as we expand the violations in one group we reduce the violations in the other, so the coefficient flips sign when the treated group includes more than 6 violation codes. See results in Appendix Table A3.

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 first column displays results with no controls, while the last column displays results with inspection fixed effects and violation code fixed effects. In all columns the OLS coefficient implies that a restaurant with recent reviews is 0.6 percentage points less likely to violate along dimensions of hygiene for which Yelp is more informative. The IV coefficient is bigger in absolute value, although we cannot statistically distinguish it from the OLS estimate. The IV estimate implies a 1.1 percentage point reduction in the propensity to violate, or a 6% decrease off the baseline probability to violate, which is 17%. This result provides some support for our hypothesis that exposure on Yelp makes restaurants clean up along hygiene 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 and respond to this information by cleaning up more when their hygiene quality is exposed online.

6 Conclusion

So much of regulation, especially for consumer protection, relies on justifications that have asymmetric information and moral hazard at their core. But the same is true for arguments behind online reviews. Our key insight is that reviews and ratings are limited in their ability to spot certain quality dimensions. On one hand, reviews miss some features that inspectors are better trained to identify and legally able to see (i.e., because they can go in the back kitchen). On the other hand, reviews' information about other features affect consumer choices. In this paper we can tease these dimensions apart and evaluate the consequences for consumers.

In the context of restaurants in New York City, we have shown that there are differences in the degree to which Yelp reviews can be an informative signal of various dimensions of hygiene monitored through city inspections. Yelp reviews contain relevant information on dimensions of hygiene that consumers directly experience, such as pests and food handling violations, compared to other violations, such as facilities maintenance. We have also shown that the hygiene signals contained in Yelp reviews affects 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 some suggestive evidence that Yelp's hygiene signals may drive restaurants that are more exposed to Yelp to comply with those hygiene standards for which Yelp is most informative.

We do not have the ideal experiment to test whether telling inspectors not to check for roaches in a restaurant's premises and to instead focus on broken pipes would lead to more or less pests. However, we have shown that consumers are more informed about the presence of roaches than broken pipes via Yelp reviews, and that both supply and demand seem to respond to this information available online. Our combined results open the possibility for regulators to consider focusing their resources relatively more on inspecting quality dimensions that consumers cannot see and write about online, while complementing their effort on the other quality dimensions with consumer reviews. While our results show that government monitoring cannot be replaced by online reviews, they also imply that government inspections could be made more efficient with more reliance on consumer reviews.

Of course, relying on online reviews to ensure hygiene quality raises a new set of relevant questions, all candidates for future research. In particular, the online platform collecting reviews has two important roles to play: it must convince customers to share their past experiences, and it must be able to aggregate and summarize this information so that 1. it is useful for future transactions, 2. it incentivizes providers to continuously invest in high quality, and 3. it does not create excessive barriers to entry for new providers.

At the information gathering stage, at least two issues are worth considering. The first is that reviews are not necessarily representative of all transaction experiences. Research has shown that reviews can be strategically submitted (Resnick and Zeckhauser (2002) and Cabral and Hortaçsu (2010)) or simply positively selected (Nosko and Tadelis (2015)), especially when consumers and providers interact on a personal basis (Fradkin et al. (2019)). The second issue is that providers have incentives to manipulate the reviews, by offering discounts on future purchases, or by directly writing fake reviews about their business or their competitors' (Mayzlin et al. (2014) and Luca and Zervas (2016)).

At the information aggregation stage, platforms need to consider how to display reviews, and how to rank service providers on the basis of those reviews. Aggregate ratings can be too coarse a measure of quality, and individual reviews can be too idiosyncratic. In addition, if quality changes over time or if certain reviews, or non-reviews, are more useful than others, aggregation needs to incorporate those features. The analysis of platform incentives to provide current and unbiased signals of quality is a valuable avenue for future research.

Finally, not every business is reviewed online – at all or frequently enough for reviews to be useful – which further highlights the importance of regulation to protect consumers from restaurants with poor hygiene. However, our research has shed some light into the benefits of complementing regulation with information from consumers, which could ultimately allow regulators to reduce the level of monitoring while holding constant the benefits that mandated monitoring in a world without online reviews was intended to achieve.

References

G. A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/1879431.

M. Anderson and J. Magruder. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563):957–989, 2012.

K. J. Arrow. Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973, 1963. ISSN 00028282. URL http://www.jstor.org/stable/1812044.

J. M. Barrios. Occupational licensing and accountant quality: Evidence from the 150-hour rule. *NBER Working Paper No. 29318*, (2018-32), 2021.

L. Cabral and A. Hortaçsu. The dynamics of seller reputation: Evidence from eBay. *The Journal of Industrial Economics*, 58(1):54–78, 2010. URL https://scholar.googleusercontent.com/scholar.bib?q=info:yCFzWkIArGcJ:scholar.google.com/{&}output=citation{&}scisig=AAGBfm0AAAAAWQuVlWmNPQ4IG33SbsRAwmUj2CsH3Sv9{&}scisf=4{&}ct=citation{&}cd=-1{&}hl=en.

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.

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*, September 2016. doi: 10.3386/w22627. URL http://www.nber.org/papers/w22627.pdf.

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

C. Farronato and A. Fradkin. The Welfare Effects of Peer Entry in the Accommodation Market: The Case of Airbnb. *Working Paper*, 2018.

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.

M. N. Federman, D. E. Harrington, and K. J. Krynski. The impact of state licensing regulations on low-skilled immigrants: The case of vietnamese manicurists. *The American Economic Review*, 96(2):237–241, 2006. ISSN 00028282. URL http://www.jstor.org/stable/30034649.

A. Fradkin, E. Grewal, and D. Holtz. The determinants of online review informativeness: Evidence from field experiments on airbnb. *Working Paper*, 2019.

M. Friedman. *Capitalism and Freedom*. University of Chicago Press Chicago, 1962.

M. Gentzkow, B. Kelly, and M. Taddy. Text as data. *Journal of Economic Literature*, 57 (3):535–74, September 2019a. doi: 10.1257/jel.20181020. URL http://www.aeaweb.org/articles?id=10.1257/jel.20181020.

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. doi: 10.3982/ECTA16566.

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.

E. L. Glaeser, A. Hillis, H. Kim, S. D. Kominers, and M. Luca. How does compliance affect the returns to algorithms? evidence from boston's restaurant inspectors. *Working Paper*, 2019.

S. Greenstein, Y. Gu, and F. Zhu. Ideological Segregation among Online Collaborators: Evidence from Wikipedians. *NBER Working Paper No. 22744*, 2016. URL http://www.nber.org/papers/w22744.

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. *MMWR*, 63(20):441–5, 2014. URL http://www.cdc.gov/MMWr/preview/mmwrhtml/mm6320a1.htm.

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

D. E. Ho, Z. C. Ashwood, and C. Handan-Nader. New evidence on information disclosure through restaurant hygiene grading. *American Economic Journal: Economic Policy*, 11 (4):404–28, 2019.

M. Ibanez and M. W. Toffel. Assessing the Quality of Quality Assessment: The Role of Scheduling. *Working Paper*, 2019.

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. ISSN 19457669, 19457685. URL `http://www.jstor.org/stable/25760354`.

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.

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

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.

B. Larsen, Z. Ju, A. Kapor, and C. Yu. The effect of occupational licensing stringency on the teacher quality distribution. Technical report, National Bureau of Economic Research, 2020.

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*, 2019.

M. Luca. Reviews, Reputation, and Revenue: The Case of Yelp.Com. *Working Paper*, 2019.

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. ISSN 0025-1909. doi: 10.1287/mnsc.2015.2304. URL `http://pubsonline.informs.org/doi/10.1287/mnsc.2015.2304`.

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. ISSN 0002-8282. doi: 10.1257/aer.104.8.2421. URL http://pubs.aeaweb.org/doi/10.1257/aer.104.8.2421.

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. doi: w20830. URL http://www.nber.org/papers/w20830.pdf.

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.

P. A. Simon, P. Leslie, G. Run, G. Z. Jin, R. Reporter, A. Aguirre, and J. E. Fielding. Impact of restaurant hygiene grade cards on foodborne-disease hospitalizations in los angeles county. *Journal of Environmental Health*, 67(7):32–36, 2005.

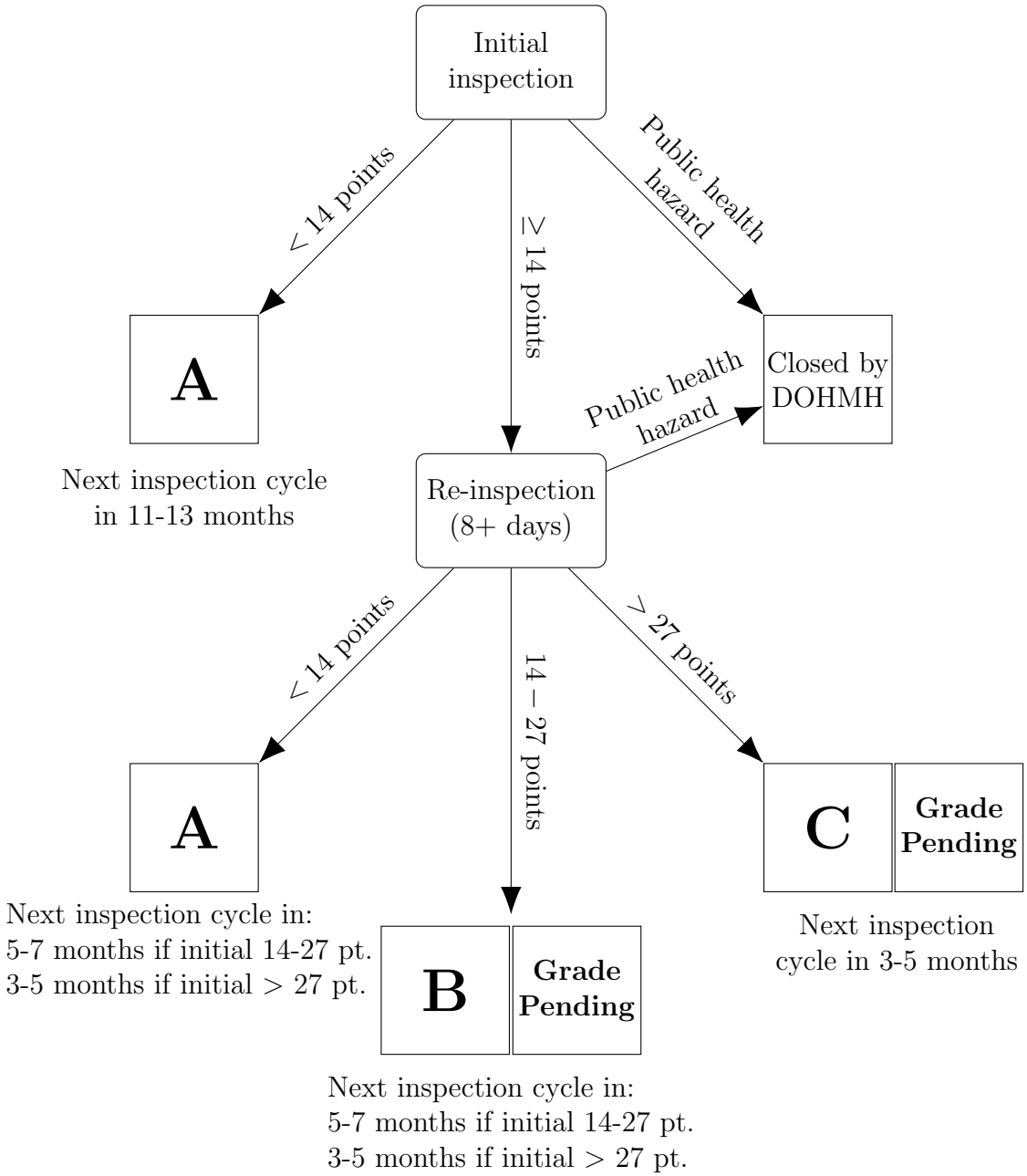
J. Stock and M. Yogo. Testing for Weak Instruments in Linear IV Regression. In *Identification and Inference for Econometric Models*. Cambridge University Press, 2005. URL http://www.economics.harvard.edu/faculty/stock/files/TestingWeakInstr{_}Stock{%}2BYogo.pdf.

M. Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013. URL https://scholar.googleusercontent.com/scholar.bib?q=info:THyJGbWP2nEJ:scholar.google.com/{&}output=citation{&}scisig=AAGBfm0AAAAAWQuYdzhTFRyG0fIMAlJFwMw60FwvwHtd{&}scisf=4{&}ct=citation{&}cd=-1{&}hl=en.

M. Taddy. Distributed Multinomial Regression. *The Annals of Applied Statistics*, 9(3):1394–1414, 2015. URL https://scholar.googleusercontent.

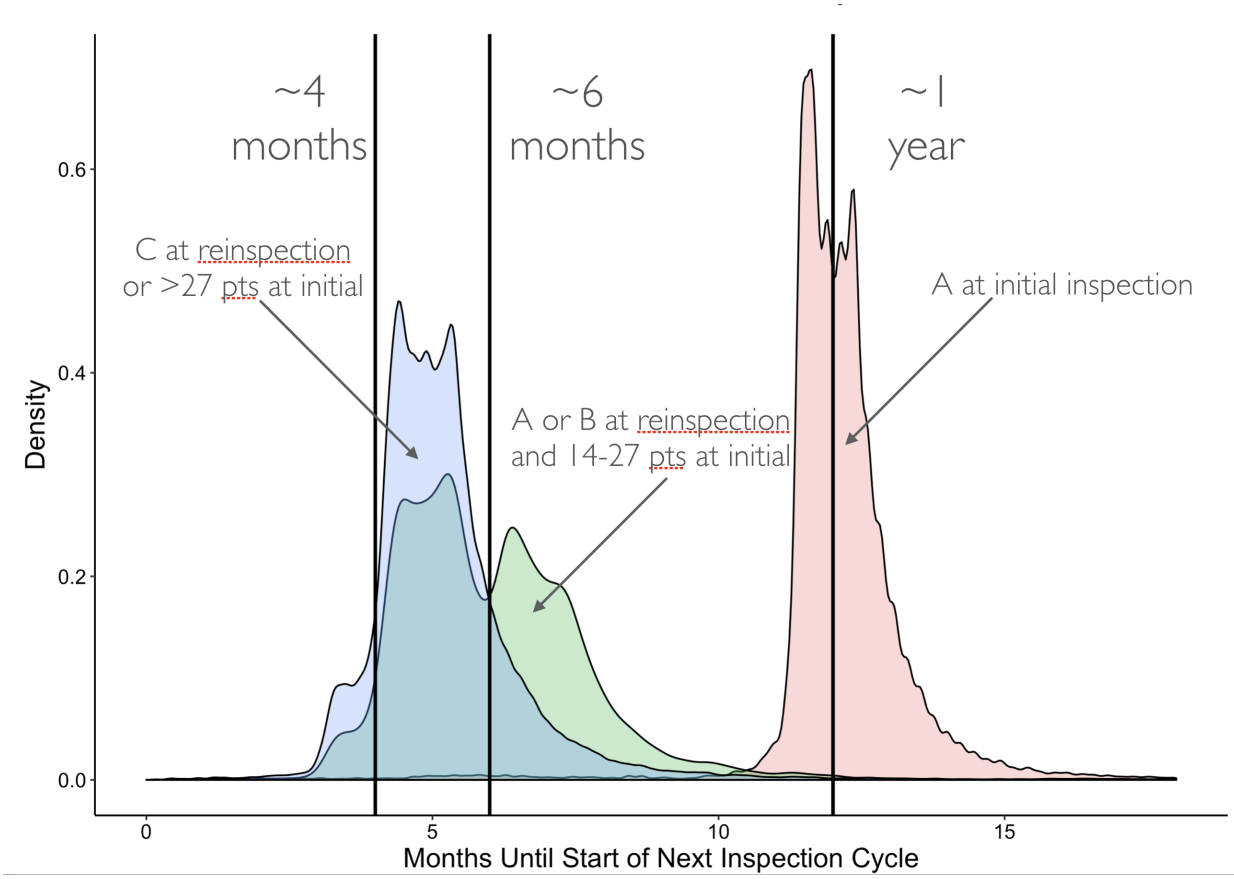
com/scholar.bib?q=info:9CFVSMWrqbEJ:scholar.google.com/{&}output=citation{&}scisig=AAGBfm0AAAAAWQuYzwMR5EUZjgIaHlo62ql205MMzC9L{&}scisf=4{&}ct=citation{&}cd=-1{&}hl=en.

Figure 1: Inspection Cycle in New York City



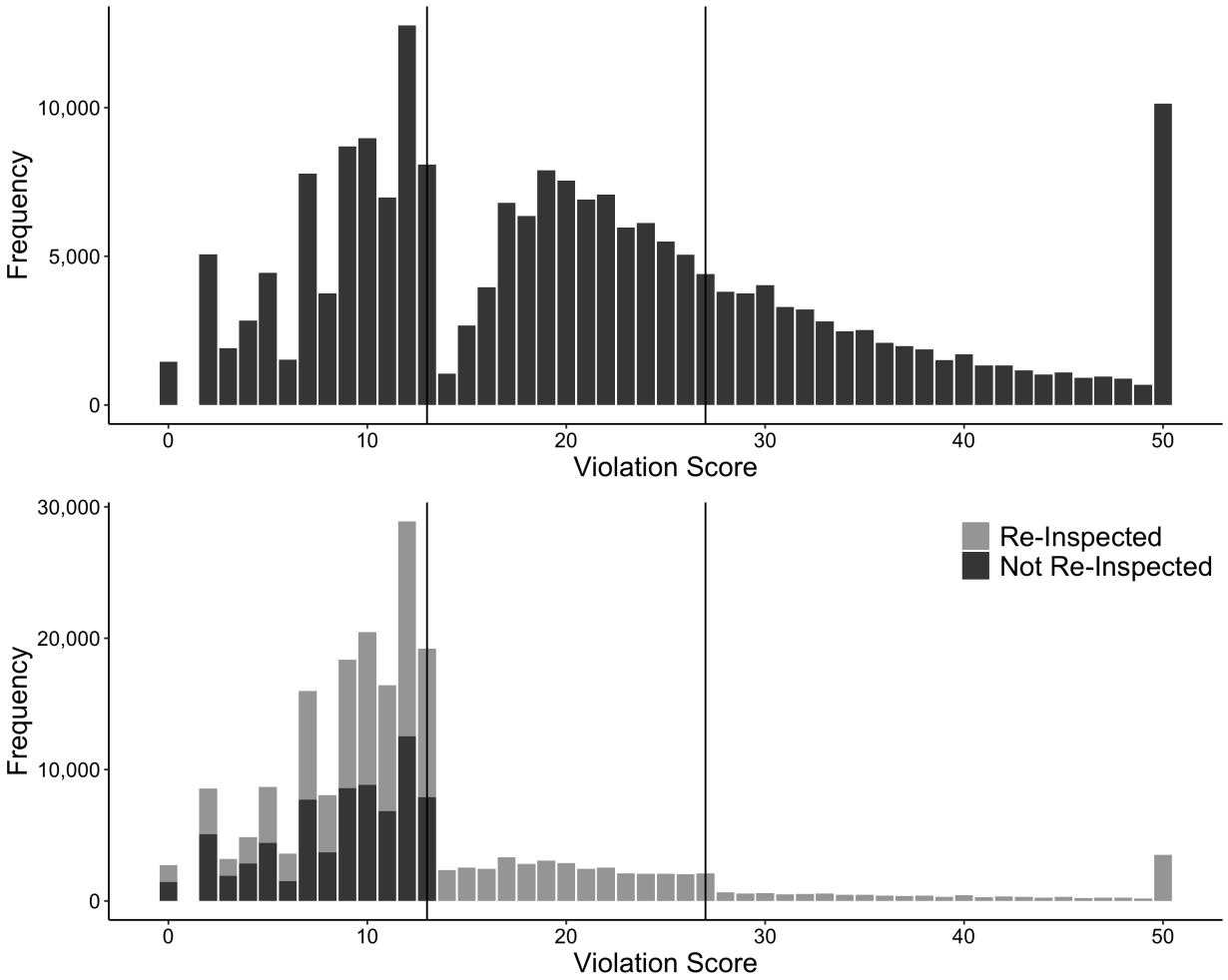
This figure plots the restaurant inspection cycle conducted by the New York City Department of Health and Mental Hygiene (adapted from `https://www1.nyc.gov/assets/doh/downloads/pdf/rii/inspection-cycle-overview.pdf`*, accessed on January 15, 2021).*

Figure 2: Time Between Inspection Cycles



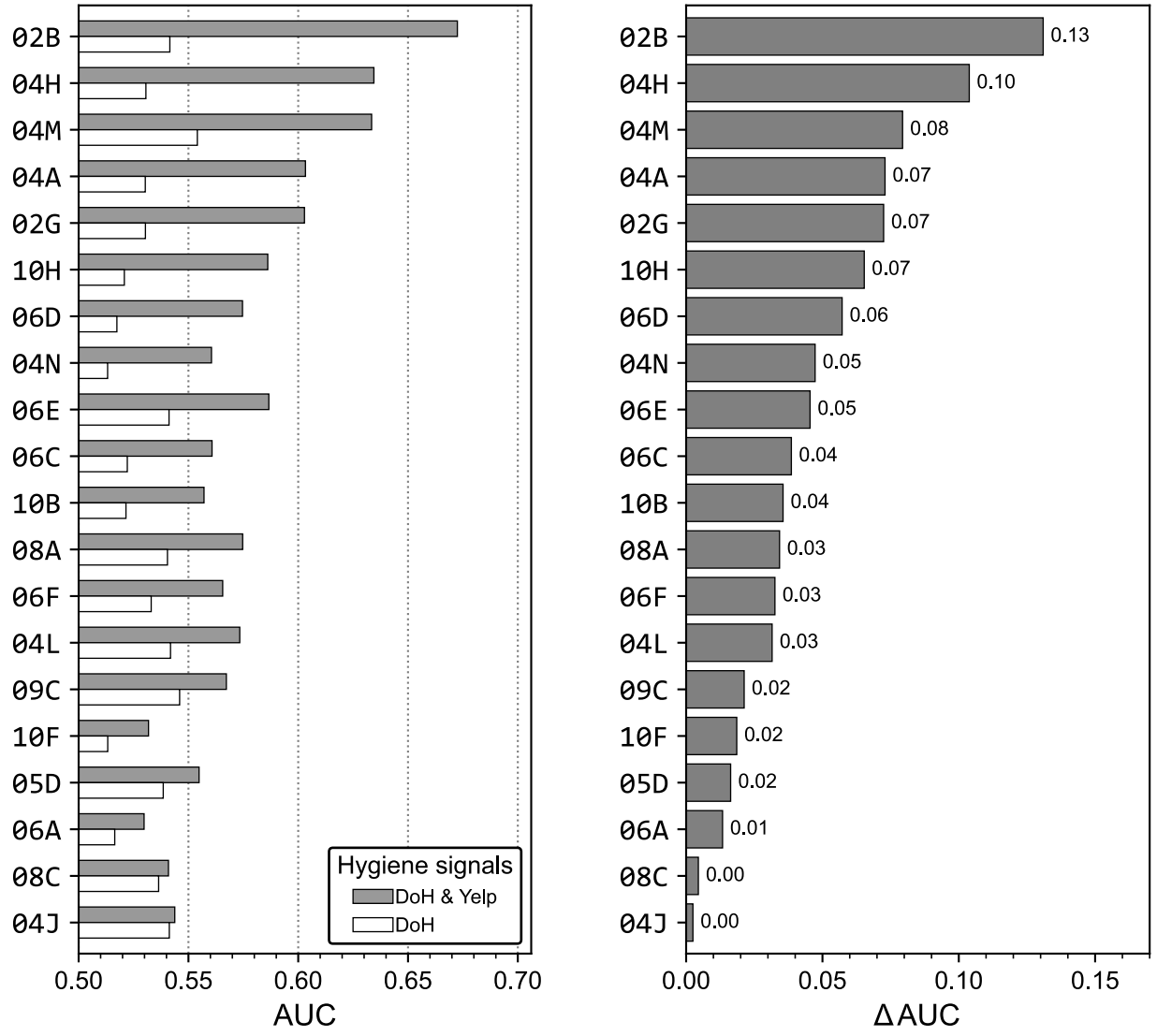
This plot shows the distribution of time between the last inspection of the current inspection cycle and the first inspection of the next cycle. For restaurants obtaining a A-grade at initial inspection during the current inspection cycle (pink), the expected time is 12 months since the last inspection.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

Figure 3: Violation Scores at Initial Inspection and Re-Inspection



For each inspection cycle, the top panel 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. The vertical lines correspond to the score thresholds that would assign A-B-C letter grades. Scores of 13 or less automatically give an A-grade, while higher scores imply that a restaurant will be reinspected within a few weeks (see Appendix Figure A1 for the distribution of the time lag between initial and reinspection). The bottom panel shows the distribution of violation scores that restaurants obtain during a re-inspection, and it also includes those restaurants that obtained an A-grade at initial inspection and thus were not re-inspected (dark gray). For histograms by price group, see Appendix Figure A2.

Figure 4: Prediction accuracy

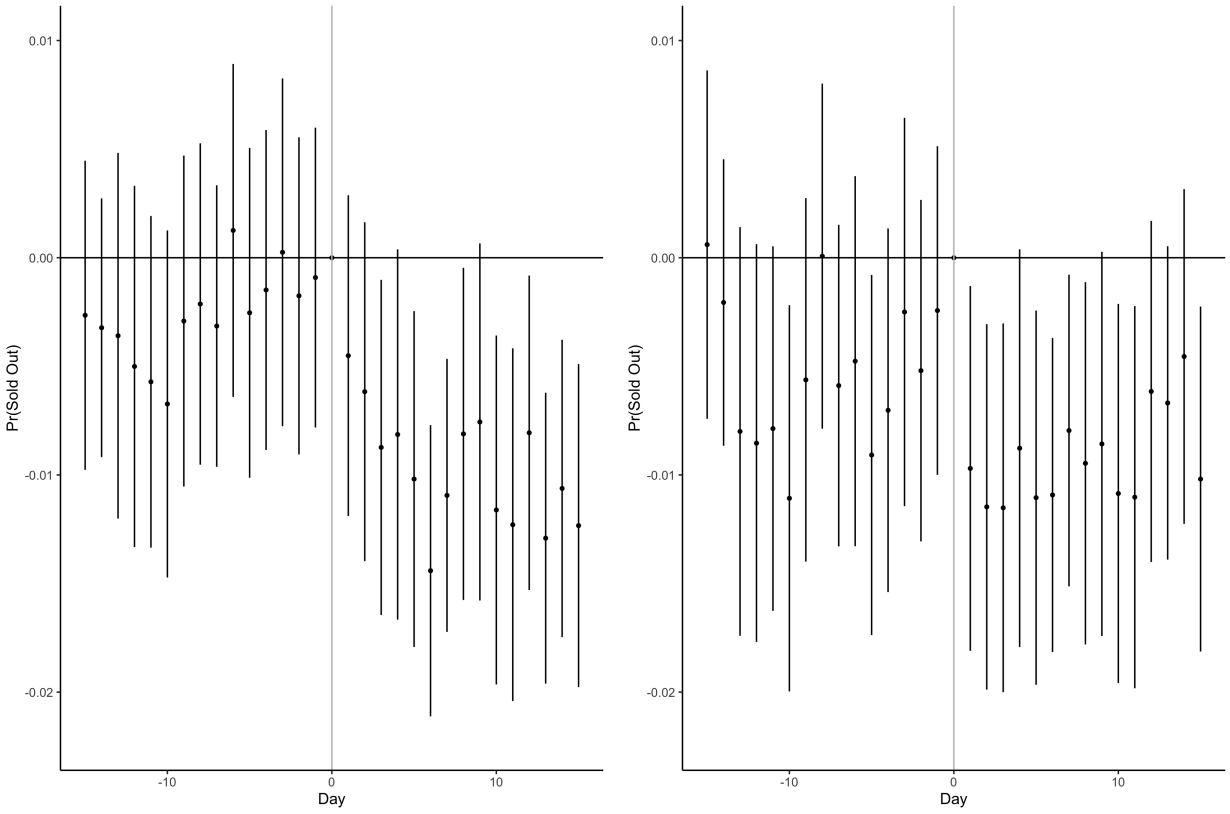


(a) AUC by violation code.

(b) Difference in AUC between the classifiers.

This figure plots the area under the curve (AUC) of the prediction of the 20 most frequent violation codes, separately for the baseline and review-augmented classifiers (panel a). The violation codes are ordered according to the increment in AUC obtained between the review-augmented and baseline classifier (panel b). For a full description of the violation codes, see Table 4. Details on the classifiers are in Section 4. For a comparison with two other classifiers, which independently use star ratings and review text, see Appendix Figure A6.

Figure 5: Yelp Hygiene Signal and Sold Out Probability – Event Study



Results from event study regressions (Equation 8). The left plot only includes 1-2-3 star reviews with a negative hygiene signal, denoted as a sufficient reduction for the top 8 violations codes for which Yelp is informative that is above 5 (23% of 1-2-3 star reviews fall in this category). The right plot includes all 1-2-3 star reviews and the coefficients are the relative difference in sold out probability between restaurants receiving a 1-2-3 review with a negative signal and restaurants receiving a 1-2-3 star review without a negative hygiene signal. Table 6 provides regression results where day fixed effects are replaced by a dummy variable for whether the day is after the submission of the focal review.

Tables

Table 1: Restaurant Characteristics

Characteristics	All	Yelp	OpenTable
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%	6.5%*	1%*
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%	8.9%	0.1%*
Venue - Restaurant	54.5%	64%*	94.8%*
Share Chain	10.8%	11.4%*	1%*
Share Closed	47.6%	37.7%*	12.9%*
Share Newly Opened	49.5%	48.9%*	53%*
N	49,647	30,447	2,215

This table presents a 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 star denotes statistical significance at 5% confidence level for the difference in means between the Yelp and non-Yelp subsamples, and between the OpenTable and non-OpenTable subsamples.

Table 2: 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, where higher scores denote worse hygiene. In columns (3)-(4) the outcome is whether the score is 14 points or more, which is the threshold past which the restaurant is not assigned a A-grade and is re-inspected 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 are clustered at the restaurant level. To improve readability, standard errors are excluded and the symbol ‡ denotes a coefficient that is not statistically significant at 5% confidence level.

Table 3: Grade Transitions

Prior Card	N	Score at Initial Inspection 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

Prior Card	N	Card Posted at End of Inspection Cycle 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 and the score obtained during the initial inspection. 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 and the card displayed when the cycle ends. 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.

Table 4: Top 20 Violation Codes

Code	Description	Share of Inspections
02B	Hot food item not held at or above 140° 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°F (smoked fish and reduced oxygen packaged foods above 38°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%

This table provides the list of the 20 violation codes that most frequently occur during 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 4b based on the informativeness of Yelp reviews towards that specific violation (highest informativeness at the top).

Table 5: Top-10 Associations Between Violation and Word Occurrences.

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

This table provides a list of the words from low-star reviews that are most predictive of each violation. It is the list of words with the highest coefficients in the matrix $\hat{\Phi}$, whose estimation is described in Section 4. Violation codes are ordered as in Figure 4b based on the informativeness of Yelp reviews towards that specific violation (highest informativeness at the top).

Table 6: Yelp Hygiene Signal and Sold Out Probability – Event Study

	Sold Out on OpenTable	
	(1)	(2)
After Review	−0.007	−0.003
	(0.001)	(0.001)
Bad Yelp Hygiene Signal		0.001
		(0.007)
Bad Yelp Hygiene Signal*After Review		−0.004
		(0.001)
Day of Week FE	Yes	Yes
Restaurant-Review FE	Yes	Yes
Observations	694,132	3,430,377
Adjusted R^2	0.517	0.510

Results from event study regressions (Equation 8) where day fixed effects are replaced by a dummy variable for whether the day is after the submission of the focal review. Column 1 only includes 1-2-3 star reviews with a negative hygiene signal, denoted as a sufficient reduction for the top 8 violations codes for which Yelp is informative that is above 5 (23% of 1-2-3 star reviews fall in this category). Column 2 includes all 1-2-3 star reviews and the coefficient of interest is the difference-in-differences coefficient estimated from the interaction of the dummy for whether the day is after the submission of the review and whether the review has a negative hygiene signal. Standard errors (in parentheses) are clustered at the restaurant level. Figure 5 provides event study plots.

Table 7: Yelp Signal and Restaurants' Hygiene Compliance – OLS and IV

	Panel A: Violation Found – OLS			
	(1)	(2)	(3)	(4)
Has Recent Reviews	0.003	0.003	0.001	
	(0.001)	(0.0004)	(0.001)	
Informative	−0.009			
	(0.001)			
Has Recent Reviews*Informative	−0.006	−0.006	−0.006	−0.006
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.175			
	(0.001)			
Adjusted R^2	0.0002	0.147	0.142	0.120
	Panel B: Violation Found - IV			
Has Recent Reviews	0.004	0.009	0.0003	
	(0.001)	(0.001)	(0.002)	
Informative	−0.006			
	(0.001)			
Has Recent Reviews*Informative	−0.011	−0.011	−0.011	−0.011
	(0.002)	(0.002)	(0.002)	(0.002)
Constant	0.174			
	(0.001)			
Adjusted R^2	0.0002	0.147	0.142	0.120
Violation Code Fixed Effects	No	Yes	Yes	Yes
Inspection Controls	No	Yes	Yes	No
Restaurant Fixed Effects	No	No	Yes	No
Time Controls	No	No	Yes	No
Inspection Fixed Effects	No	No	No	Yes
Observations	2,904,680	2,904,680	2,904,680	2,904,680

This table presents coefficient estimates of Equation 10. An observation is a restaurant-initial inspection-violation code triplet for restaurants on Yelp. The outcome variable is equal to 1 if during an initial inspection, the inspector finds that the restaurant is infringing on a particular violation. The outcome is 0 if the restaurant is compliant. The variable "has recent reviews" is equal to 1 if the restaurant has received Yelp reviews in the 90 days preceding the initial inspection. The variable "informative" is equal to 1 if the violation code is one of the top 8 codes for which Yelp is most informative – the top 8 codes in Figure 4a. The diff-in-diff coefficient of interest is the coefficient on the interaction between "has recent reviews" and "informative." Inspection controls include the aggregate inspectio score and inspector fixed effects. Time controls include include day of the week fixed effects, year-quarter fixed effects, and restaurant age. The last column includes the most stringent set of controls, inspection fixed effects and violation code fixed effects. Standard errors (in parentheses) are clustered at the restaurant level. Panel B is the same regression, except that we instrument for "has recent reviews" with the average number of reviews that past reviewers to the focal restaurant have submitted to restaurants other than the focal restaurant. First stage regression results are presented in Table 8. Results where Yelp is considered informative for an increasingly larger set of violation codes are presented in Table A3.

Table 8: Yelp Signal and Restaurants' Hygiene Compliance – IV First Stage

	Panel A: Has Recent Reviews – First-Stage			
	(1)	(2)	(3)	(4)
Log(Reviewers' Reviews + 1)	0.131	0.128	0.091	
	(0.001)	(0.001)	(0.001)	
Adjusted R-Squared	0.287	0.304	0.658	
	Panel B: Has Recent Reviews*Informative – First-Stage			
Log(Reviewers' Reviews + 1)*Informative	0.131	0.131	0.131	0.131
	(0.001)	(0.001)	(0.001)	(0.001)
Adjusted R-Squared	0.703	0.705	0.739	0.765
Violation Code Fixed Effects	No	Yes	Yes	Yes
Inspection Controls	No	Yes	Yes	No
Restaurant Fixed Effects	No	No	Yes	No
Time Controls	No	No	Yes	No
Inspection Fixed Effects	No	No	No	Yes
Observations	2,904,680	2,904,680	2,904,680	2,904,680
Note:				$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

This table presents the first stage estimates of Panel B in Table 7.