Consumer Reviews and Regulation: Evidence from NYC Restaurants[†]

Preliminary and incomplete – please do not circulate.

Chiara Farronato
Harvard Business School
and NBER

Georgios Zervas
Questrom School of Business
Boston University

February 21, 2018

Abstract

We investigate how two signals of restaurant quality, health inspections and consumers reviews, jointly affect consumer choice and restaurants' incentives to comply with hygiene standards. We first examine whether consumer reviews can be used to detect hygiene violations that health inspectors look for. To do so, we use machine learning methods to isolate information contained in consumers reviews specifically pertaining to each type of violation, which we use to predict violations. We find substantial heterogeneity in prediction accuracy. Violations relating to food temperature and pests are more accurately predicted than facility maintenance violations. Next, we estimate the supply and demand effects of violation-specific information contained in consumer reviews. On the demand side, we find that the disclosure of hygiene conditions in reviews affects where consumers choose to eat. On the supply side, we find that relative to restaurants not on Yelp, restaurants reviewed on Yelp score better on hygiene dimensions that are predictable from consumer reviews compared to less predictable aspects of hygiene. Our results have implications for the design of regulation when consumers rate their service experiences online.

[†]We thank Yi Fung for outstanding research assistance. We thank the New York City Department of Health and Mental Hygiene for sharing data and insights.

1 Introduction

When consumers have limited information about service providers' underlying quality, a market for services can easily break down. Low quality providers can drive away high quality providers, and consumers might either stop purchasing or engage in risky transactions. This is the classic lemons market problem (Akerlof, 1970). Regulation has historically been the solution to protect consumers from risky transactions and prevent a market breakdown. Regulation relies on expert evaluations to inform and protect consumers. Many services are subject to specific licensing requirements, and health and safety rules. For example, restaurants are evaluated by public health inspectors for compliance with health and safety regulations.

While regulation aims to inform and protect the public, consumers have turned to review platforms as their primary source of information for various products and services. Review platforms collect and publish information from non-expert consumers who voluntarily provide information about their experiences with various providers in the form of reviews, ratings, and pictures. Popular platforms like Yelp and TripAdvisor publish millions of reviews for restaurants, hotels, and other local businesses, and are visited by tens of millions of consumers every month.[1]

In this paper, we investigate the extent to which review platforms can achieve the two main goals of regulation: informing the public about dimensions of quality relevant to regulators, and incentivizing compliance with hygiene standards. Our empirical context is New York restaurants. We combine detailed inspection records from the New York City Department of Health and Mental Hygiene (DoH), and consumer reviews from Yelp. Health inspectors periodically visit restaurants to look for different kinds of health and safety violations. We measure the informativeness of consumers reviews about health violations by asking how well we can predict the occurrence of different types of violations using Yelp data. A novel feature of our approach is that we use machine learning methods to isolate the features of Yelp reviews that are specifically predictive of each type of violation rather than relying on hard-coded lists of keywords. We find substantial heterogeneity in how accurately Yelp reviews can predict different types of violations. For example, pests and food handling violations are more accurately predicted by the text of recent Yelp reviews than violations related to facility maintenance.

Next, we estimate the impact of review content pertaining to each type of violation on consumer choice and the incentives of restaurants to comply with regulation. On the demand side, we find that hygiene information contained in Yelp reviews drives customer demand

[1] See, for example, `https://www.yelp.com/factsheet`.

towards cleaner restaurants. On the supply side, we find that restaurants on Yelp tend to violate less along hygiene dimensions that are more predictable from consumer reviews relative to restaurants not on Yelp.

Overall, our findings highlight the potential of consumer reviews to partially substitute for regulation aimed at violations that consumers can discover from consumer reviews. For policy makers, an implication of our results is that limited resources of health inspectors could be targeted to monitor violations that are not easily detectable from consumer reviews. Of course the viability of online reviews as a partial substitute to regulation depends on the review platform's incentives to provide unbiased ratings, and on its ability to collect truthful reviews while policing providers' attempts to manipulate the system. We turn to this issue in our concluding remarks.

The debate over consumers reviews and regulation has intensified, with the diffusion of online peer-to-peer markets such as Uber and Airbnb. This debate focuses on whether these new marketplaces should be subject to existing regulation, just like traditional service providers such as taxis and hotels. On one hand, new online marketplaces argue that their screening mechanisms and reputation systems are effective at ensuring service quality and monitoring service providers. On the other hand, policymakers are skeptical about the efficacy of consumers reviews, and instead favor stricter regulation.[2] Our results in this paper suggest the answer to this debate may be somewhere in the middle: consumer reviews can inform the public about certain types of hazards better than others, and thus regulation could be tailored to hazards that consumers cannot easily detect.

An extensive literature documents the value of regulation in ensuring health and safety standards. Within the restaurant industry, existing work has shown that inspections are an effective way to incentivize restaurants to be clean, and that consumers are sensitive to the information disclosed by health inspections (Jin and Leslie, 2003). At the same time, research has also shown that regulation can be costly to consumers because it harms competition. For instance, occupational licensing can limit entry of manicurists who do not meet licensing standards but whose services are still valuable to consumers (Federman et al., 2006).

The need for regulation is typically justified if other ways to ensure consumer protection are impractical or too costly. However, recent technological developments have made it cheaper to collect and aggregate information about service providers. In particular, review platforms facilitate the collection and dissemination of information on providers' quality, and have the potential to offer an alternative to regulatory efforts. Moreover, if they are accurate,

[2]Last year both Uber and Lyft stopped operating in Austin to protest stricter background checks for their service providers. See: `http://www.nytimes.com/2016/05/10/technology/uber-and-lyft-stop-rides-in-austin-to-protest-fingerprint-background-checks.html`.

consumer reviews offer some practical advantages over regulation. 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 restaurant health inspections can be costly.[3] In addition, inspectors have been shown to have discretionary power over their evaluations (Ibanez and Toffel, 2017). If a large number of consumers contribute reviews about restaurant hygiene, they have the potential to be more accurate than inspections.

We organize the remaining of this paper as follows. In § 2 we discuss the relevant literature. In § 3 we describe our empirical context and the data we use for our analysis. In § 4 we present our approach to predict health violations from the text of Yelp reviews. In § 5 we estimate the effect of Yelp hygiene information on demand and supply incentives, and in § 6 we conclude by discussing the limitations and implications of our work.

2 Related work

Our work is most closely related to three strands of literature. The first set of related papers study the role of online reviews in ensuring trust and affecting consumer choice. Lewis and Zervas (2017) look at the effect of hotel ratings on travelers' demand. They find that an additional star on TripAdvisor, Expedia, and Hotels.com increases hotel demand by 25%, and that hotels are able to charge 9% higher prices. Luca (2011) studies the effect of Yelp star ratings on restaurant revenues, finding that an additional Yelp stars increases restaurant revenues by 5-9%. Other papers, including Resnick and Zeckhauser (2002) and Cabral and Hortacsu (2010), have looked at reputation mechanisms in e-commerce.

There is also a set of papers on the biases of online reviews, which is important to keep in mind when evaluating whether online reviews are an accurate measure of providers' quality. Mayzlin et al. (2014) show evidence that hotels have incentives to fake reviews on TripAdvisor, by submitting positive reviews to themselves and negative reviews to their competitors. Luca and Zervas (2016) confirm that incentives to manipulate reviews are also present on Yelp.com.

The second set of related work focuses on the effects of regulation on service quality. Jin and Leslie (2003) focus on a policy change in Los Angeles county that required restaurants to start displaying their health grades at their door. They find that consumer demand became sensitive to changes in restaurant hygiene quality, restaurants improved their hygienic conditions, and the number of hospitalizations due to foodborne illnesses decreased. Sep-

[3]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`

arate work by Simon et al. (2005) confirms their results. Jin and Leslie (2009) find that the improvements in restaurant hygiene occur despite the fact that restaurants have different reputational incentives depending on the degree of repeat-customers, and franchised restaurants tend to free-ride on chain reputation. Online, Hui et al. (2016) investigate the addition of a buyer protection program on eBay. They find a reduction in moral hazard after eBay started mandating that sellers refund buyers when items are not received as described. This reduction in moral hazard increases surplus, above and beyond the benefits from the reputational system already present on eBay.

In our work, we combine the study of the effect of online reviews and health inspection on customer demand. A joint analysis is important at this point because as Kang et al. (2013) and Harrison et al. (2014) have shown, online ratings and health inspection scores are correlated: when online reviews are positive, a restaurant is also more likely to be found clean by inspectors. By separately studying the effect of one measure of restaurant quality (e.g. health inspection scores) on customer demand, the estimates can actually confound the true effect of health grade with the separate effect of online reviews.

The third and final set of related papers use machine learning techniques to predict expert decisions. For example Kleinberg et al. (2017) use bail decisions to show that judges' could reduce jailing rates or crime rates if they complemented their expertise with algorithms that predict crime rates for released defendants. Specific to the setting of restaurant inspections, Kang et al. (2013) and more recently Mejia et al. (2017) show that Yelp reviews are able to track hygiene conditions of restaurants.

It is important to notice one key difference between our work and existing efforts to predict restaurant hygiene from online reviews. For example, Glaeser et al. (2016) compare competing machine learning models to maximize prediction accuracy. An important goal of their exercise is to help inspectors identify which restaurants are not complying with existing regulation. Depending on its accuracy, the prediction can partially substitute random selection of restaurants to inspect, which is the status quo. In existing work, the algorithms are created with the purpose of improving experts' decision making, rather than substitute it.

Our focus is different. The goal of our exercise is to evaluate whether consumer reviews can substitute rather than inform inspectors to guarantee hygiene standards, at least on some dimensions of hygiene. The difference has important implications. A prediction problem to inform the health department should include all available sources of information, external from online reviews and internal from within the health department. A prediction problem to evaluate whether consumers can substitute for the health department should only include information available to consumers through online reviews.

By incorporating Yelp review text as signals 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. (2017), and Greenstein et al. (2016)). Research on online reviews has mostly focused on the aggregate numeric rating that consumers assign to service providers. 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. So the extent to which Yelp stars capture restaurant hygiene will depend on the weight consumers place on hygiene compared to other quality dimensions. For example, Lehman et al. (2014) show that ratings are less susceptible to unsanitary conditions for restaurants that are perceived as being more "authentic". Using the text of reviews allows us to break down a consumer overall assessment of a restaurant and separate the hygienic dimensions from everything else.

Our work can inform the debate over the regulation of peer-to-peer markets. Recent papers have focused on the welfare benefits of flexible labor for Uber drivers (Chen et al. (2017)) and passengers (Cohen et al. (2016)). Farronato and Fradkin (2016) study the welfare implications from Airbnb for travelers, Airbnb hosts, and hotels jointly. But still 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)). To our knowledge, we are among the first to shed light on potential substitutabilities between online reviews and regulation.

3 Data

This section decribes the data on health grades, online reviews, and demand that we use in our empirical analysis.

First, data on health inspections come from the New York City Department of Health and Mental Hygiene (DoH). As the DoH describes on its website,[4] *"[it] conducts unannounced inspections of restaurants at least once a year. Inspectors check for compliance in food handling, food temperature, personal hygiene and vermin control. Each violation of a regulation gets a certain number of points. At the end of the inspection, the inspector totals the points, and this number is the restaurant's inspection score"*. More points imply more violations, so a higher *health score* implies worse restaurant health quality. We obtained inspection level data directly from the DoH between July 2007 and September 2016.

The DoH performs inspections of restaurants, coffee shops, bars, nightclubs, most cafeterias, and fixed food stands in the five boroughs of New York. The three stated objectives of

[4] `http://www1.nyc.gov/site/doh/services/restaurant-grades.page`

the inspection program are: 1) give consumers better information about sanitary conditions of restaurants 2) improve restaurants' food preparation practices 3) reduce foodborne illness attributable to restaurants.

We have data at the level of each violation code. There are about 50 violations codes that evaluate restaurants on multiple dimensions: vermin (e.g. evidence of mice), food temperature (hot food item not held at or above 140F), facilities (improper sewage disposal system), food handling (raw food not properly washed), overall hygiene (inadequate personal cleanliness of staff), contamination (worker does not wash hands throughly), and regulatory (wash-hand sign not posted).[5] On July 27th 2010 the DoH started assigning a *health grade*, A through C, following each inspection. Since then, restaurants must display their grade card conspicuously enough for customers to see. The program uses dual inspections to help restaurants improve before being graded.

Every year a restaurant 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. A 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 card. With a score of 14 or more points on initial and 28 or more points on reinspection a restaurant receives a C grade card. The restaurant should post the grade card, or alternatively can post a "Grade Pending" card if it decides to dispute B or C grades at an administrative tribunal.[6]

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, and date of entry and exit. In total, we have records for about 448 thousand inspections covering 57 thousand restaurants.

The second dataset includes scraped online reviews from Yelp.com. Yelp contains business information, such as zip code and phone number, and a historical record of reviews. From this record we are able to construct the average *Yelp rating score* of a business at any point in time. We also scrape the text of each review, and we describe in Section 4 how to summarize the health information contained within. The complete set of reviews include 2.8 million reviews for 31 thousand of the restaurants included in the health inspection dataset.

The third and last dataset comes from OpenTable. We have scraped OpenTable data

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

[6]More details on inspection regulation and grading can be found at `http://www1.nyc.gov/assets/doh/downloads/pdf/rii/inspection-cycle-overview.pdf`.

since April 2013. For every day and restaurant on OpenTable, we have information on whether the restaurant had a table available for 2 people around 7PM. This dataset includes 2.1 million restaurant-days and covers 2.5 thousand of the restaurants in the health inspection dataset.

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.

3.1 Descriptives

Overall, of the 57,062 individual businesses present in the DoH data, 58 percent were matched to a business on Yelp, and 4.4 percent 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 Christmas season). The interval between inspection cycles depends on the sanitary condition of the restaurant during the previous inspection. Namely, 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.[7] In practice, there is a substantial amount of variability in the time interval between inspection cycles, as pictured in Fig. 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 Fig. 3. Only 25 percent of restaurants obtain an A grade during the initial inspection. The other restaurants obtain a violation score corresponding to a B (43 percent) or a C (32 percent). Those restaurants whose score would imply a B or a C-grade are reinspected within a few weeks. After reinspection, well over half of the restaurants get to display an A-card. As the bottom panel of Fig. 3 shows, 73 percent of restaurants end an inspection cycle with an A

[7]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`.

grade, 20 percent end it with a B grade, and only 7 percent 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 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. In Section 4 we use outcomes of initial inspections as our measure of restaurant hygiene.

Table 2 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 percent score between 0 and 13 points during the initial inspection of cycle $t+1$, 36 percent score between 14 and 27 points, and 22 percent score 28 or more points. By the end of the cycle most restaurants end up with a A-grade, but still 15 percent of restaurants with a A-grade in t lose it by inspection cycle $t+1$. Of those, 13 percent drop to B-grade, and 2 percent drop to C-grade.

In the next section we describe how we use the text of online reviews to predict restaurants' hygiene outcomes during initial inspections.

4 Can Online Reviews Predict Restaurant Hygiene?

A first step to evaluate when consumer reviews can substitute for inspections is understanding whether consumers have visibility into the dimensions of hygiene that inspectors care about. This visibility is a necessary but not sufficient condition for substitutability. The ability to predict the occurrence of a particular violation – say, presence of mice – from the text of Yelp reviews will be our measure of visibility.

Inspectors rate restaurants on multiple dimensions of hygiene, from the way in which staff handle raw food to the quality of the ventilation system. We focus our analysis on the 20 most frequent violation codes, which are listed in Table 3 and constitute over 80 percent of all violations recorded during initial inspections since July 2010.

In order to detect health violations from consumer reviews on Yelp, we exploit the text contained in each review. 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 vio-

lations 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. So we use the review text to predict each violation separately, and if a violation is well predicted by consumer reviews we define it as *detectable by consumers*.

To incorporate review text in our 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 a three step approach, which was recently applied to analyzing congressional speech by Gentzkow et al. (2016).

In the first step, we learn what reviewers say when hygiene violations occur. We associate each restaurant 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.[8] The second reason 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 47,726 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.

The combined text of the 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 in-

[8]This statistic is likely a lower bound 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.

spection i, recalling that the dimension of c_i is $3 \times 47,726$. We assume c_i to be drawn from a multinomial distribution

$$c_i \sim MN(q_i, m_i), \tag{1}$$

where m_i is the document length – number of words – and q_i is a vector of probabilities with length equal to the number of distinct words that consumers could use – our vocabulary. 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 $\eta_{ij} = \alpha_j + x_i\beta_j + \phi_j r_j \times v_i$. Here, x_i denote restaurant characteristics and v_i are dummies for different violation codes interacted with the rating groups r_j associated with word j. In x_i we include star-rating fixed effects and various restaurant characteristics provided by the DoH such as cuisine, zipcode, and chain affiliation.

Given the distributional assumption in Equation 1, estimating the coefficient vector (α, β, ϕ) as a multinomial logit model is prohibitively expensive because the coefficients for each word in c_i depend on the coefficients of all the other words in c_i. Fortunately, we can approximate the multinomial logit model with $3 \times 47,726$ independent Poisson regressions following the distributed multinomial regression framework of Taddy (2015). This approximation makes the estimation tractable, but we lose any correlation in the occurrence of distinct words.

To estimate the model we split inspections in a training set that contains 70% of the restaurants and a testing set that contains the remaining 30%. We estimate each Poisson lasso regression using the same 10-fold cross-validation and block sampling at the restaurant level. The estimated model yields a separate set of coefficients $(\alpha_j, \beta_j, \phi_j)$ for each word. We combine all coefficient estimates for the violation covariates into the matrix Φ. In Φ the number of rows corresponds to the number of violation code-star rating pairs, while the number of columns corresponds to the number of word-star pairs.

In the second step, we construct low-dimensional statistics for review text by projecting text onto violations. We use the estimates from the first step to compute Φf_i, where $f_i = c_i / \sum c_i$ are word frequencies in document i. This vector is a sufficient reduction (SR) in the sense that violations are independent of the word frequencies conditional on the SRs (Taddy, 2015). This step allows us to reduce the dimensionality of the text data from hundreds of thousand to 60, i.e. the number of violation code-star rating pairs, and it implies that once we have the sufficient reduction we can ignore text for the purpose of predicting violations.

Finally in the third step, we predict violations using the low-dimensional representation of review text described in the second step. Separately for every violation code, we fit a gradient boosted tree, where the dependent variable is an indicator for the occurrence of a violation during inspection i. The predictors include restaurant characteristics – cuisine, chain affiliation, age, service, and zipcode – and information from Yelp – average star rating,

review counts, and the sufficient reductions which capture the effects of review text. We test the predictions on the 30% of restaurants excluded from the training data.

4.1 Prediction Results

This section presents the results of our prediction. The output of the multinomial distributed regression is a matrix Φ of coefficients, or factor loadings, one per word-violation pair. A high factor loading for the word-violation pair ("undercook" in a 2-star review, 02A "Food not cooked to required minimum temperature") means that when an inspector finds a restaurant violating 02A customers are more likely to use the word "undercook" in the Yelp reviews submitted in the preceding 3 months.

Table 4 does not show the factor loadings, but for each violation it shows the words included in 1-star and 2-star reviews with the highest loadings. So for example, when a restaurant violates 02G – Cold food item held above 41F – customers are more likely to use words such as rancid, tourist trap, puke, and diarrhea.

For some violations, such as 02G, the words are descriptive of the actual dimension of cleanliness that inspectors evaluate. For other violations, such as 10F – Non-food contact surface improperly constructed – the words do not seem to be related to the substance of the violation. This difference across violation codes in the words with high factor loadings suggests that probably consumer reviews are better at detecting violations about food temperature handling, than violations about restaurant facilities. Our results below confirm this hypothesis.

We use the factor loadings to construct sufficient reductions Φc_i. A high value of the sufficient reduction for a specific violation implies that reviews contain words typically associated with this violation. What is important to highlight is that once we estimate Φ using inspection data, we can construct the sufficient reduction at any point in time, as long as customers submit reviews on Yelp.

Finally, in order to evaluate how well we can predict each violation using the sufficient reduction, Figure 4 shows the AUC from the third step of our procedure. The AUC is a measure of the predictive power of our model. Since we are trying to predict a binary outcome – whether a restaurant is in violation or compliance with a particular hygiene code – the worst we can do it 50%, i.e. a random guess. Values closer to 1 mean that we can predict the occurrence of a particular violation with better accuracy.

Two results are worth highlighting. First, the fact that all AUC metrics are between 55% and 70% implies that it is relatively difficult to predict the occurrence 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 is itself a noisy signal of hygienic conditions.

Second, there is a substantial amount of heterogeneity in how well online reviews can predict distinct violations. Online reviews are better predictors of violations related to vermin, food temperature, and food handling than violations related to facilities and certifications. It is reassuring to see that the violations that appear to be more predictable are the ones that relate directly to food and pests in the premises – intuitively these are the violations we would expect consumers to be most likely to notice. In addition, the words predicting these violations are descriptive of the actual infringement.

The results of this section point to one main conclusion. Consumers discuss restaurant hygiene on Yelp, but not all dimensions of hygiene are captured by consumer reviews. Some dimensions, such as pests and food handling, can be more easily detected by consumers, and consumer reviews do indeed predict the occurrence of these violations at a higher rate than the occurrence of other types of violations. We want to distinguish between "detectable" and "non-detectable" violations, but any threshold based on the AUC metric would inevitably be arbitrary. However, it is important to highlight that detectability is a necessary condition for substitutability between online reviews and regulation. If consumers do not have visibility into the same dimensions of hygiene that inspectors care about, then it is impossible for consumer reviews to take the role of health and safety inspections.

Of course, it could happen that consumers do not discuss restaurant hygiene online because they know that health and safety inspections exist for this very purpose, but that if inspections did not exist consumers would step in and start monitoring hygienic conditions. We do not have access to a context where reviews exist and inspections do not, but in this sense our detectability measure is a lower bound on consumer ability to monitor restaurant hygiene.

The heterogeneity we observe in how much consumer reviews predict different types of violations suggests that dimensions of hygiene such as pests are more likely candidates to be outsourced to consumer reviews – they are relative more "detectable" – than, for example, the conditions of restaurant facilities. 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 – two of the stated goals of health and safety inspections.

Before concluding this section, a caveat is in order. The fact the our machine learning algorithm can predict one dimension of hygiene better than another does not immediately imply that Yelp readers can similarly predict the same dimensions of hygiene. One of the results that provide some confidence that our algorithm might approximate what readers can gather from online reviews is the interpretability of our results. For example, the sufficient

reduction for V04M contains the word "roach". It is seems plausible that a Yelp reader could use that same information to infer that there are pests in a restaurant. The results in the next section, which show that those "detectable" measures of hygiene are the ones associated with restaurant demand, further provide some evidence that our algorithm is picking up hygiene information that consumers can also obtain from reading online reviews.

5 Can Online Reviews Ensure Restaurant Hygiene?

The DOH has two stated goals: to inform consumers about restaurant hygiene, and to improve restaurant hygiene practices. These goals are intimately related. If consumers change their restaurant choices in response to the disclosure of hygiene information, restaurants can either improve their hygiene conditions or face decreased demand and potentially go out businesses. In the previous section we evaluated whether consumers have visibility into the same dimensions of hygiene that inspectors care about. We found that some hygiene dimensions are more visible to consumers than other dimensions, at least when predicted by our text analysis algorithm. To confirm that our exercise from Section 4 picks up information that consumers also take into account when choosing where to eat, we want to measure the effect of Yelp hygiene information on consumer choices and restaurant incentives. We devote the first subsection to consumer choices – demand – and the second subsection to restaurant incentives – supply.

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 success (Jin and Leslie (2003) for health grades, Luca (2011) and Anderson and Magruder (2012) for online ratings). Our work is the first to look at the effect of online ratings and health grades jointly. In addition, the previous section allowed us to separate hygiene information contained in the text of Yelp reviews, an additional signal of restaurant quality.

In order to analyze the effect of information contained in online reviews and health inspections on demand, we estimate two separate models. The first uses the probability of being sold out at 7PM on OpenTable (Figure 5). The second uses a restaurant's decision to go out of business in a discrete time hazard model (Figure 7). At the end of this section we provide some evidence supporting the hypothesis that sold out probability and restaurant exits are correlated with one another, so presumably with overall restaurant demand.

When we look at the probability of being sold out on OpenTable, we estimate the following linear probability model:

$$Y_{it} = \alpha_1 doh_{it} + \alpha_2 yelp_{it} + \alpha_3 (\Phi c_{it}) + \alpha_4 X_i + \alpha_5 X_t + \epsilon_{it}, \tag{2}$$

where Y_{it} is 1 if restaurant i is sold out on day t. Here, doh_{it} is a dummy for whether a restaurant displays an A-card at its doors on day t, Φc_{it}, is the vector of sufficient reductions obtained from reviews submitted within the last 90 days, and $yelp_{it}$ is a vector of dummies corresponding to the Yelp average rating rounded to the nearest half star. So for example, a 2.73 average rating would be rounded to 2.5 stars. Controls include time-invariant restaurant characteristics such as cuisine and zipcode, and time controls such as day of the week and quarter fixed effects.

We are worried about the possibility that some characteristics, unobservable to the econometrician, but observable to both the inspector and restaurant guests, might both affect the inspection outcome and guests' decision of where to eat. These characteristics could include, for example, the presence of mice, or adulterated food, that are both visible by the inspector, and by the guests eating at the restaurant.

To separate the causal effect of health grades on demand, we use an instrumental variable strategy that relies on the random assignment of inspectors to restaurants. Existing work confirms that inspectors are randomly assigned to restaurant inspections in New York City (Meltzer et al. (2017)).In Appendix A1 we provide evidence that inspectors are randomly assigned after controlling for a small set of restaurant characteristics, which are included in our second stage. We use two instruments. The first is a dummy for whether any inspector during an inspection cycle evaluated the same restaurant in the previous inspection cycle.[9] This instrument has been shown to be a good predictor of inspection outcomes, since repeated interactions lead to more generous evaluations (Ibanez and Toffel (2017)). The second instrument is the average violation score that an inspector assigns to restaurants other than the restaurant under evaluation during initial inspections.[10] Effectively, this is a measure of inspector severity, which is uncorrelated with the underlying quality of a restaurant under random assignment (Bartik (1991)).

Another source of endogeneity comes from Yelp reviews. Yelp stars are not allocated

[9] An inspection cycle is composed of multiple visits to a restaurant by different inspectors, but they all jointly contribute to the final letter grade displayed at the door. See details in Section 3.

[10] We only use the grade at initial inspection because it is the least biased of all inspection grades, as shown in Section 3. On average, each inspector evaluates 1,372 different restaurants during an initial inspection (median is 1,340), with a standard deviation of 716. Of the 286 inspectors, 5 inspectors in our data have inspected only one restaurant during an initial inspection, so for this restaurant-inspector conbinations we cannot compute a stringency instrument. Adding a dummy for this event or simply assigning a value of 0 to these observations does not change the results.

randomly to different restaurants, so they might be correlated with unobserved underlying quality. For now we ignore this, but we will be using the same instrumental variable approach described for the hygiene letter grade. So we will instrument the current restaurant rating with the average stringency of its reviewers, as measured by the ratings that these reviewers submitted to other restaurants on Yelp.[11]

We first discuss the OLS estimates without including the sufficient reductions Φc_{it} as controls. Table 5 shows the estimates of linear probability models where the outcome is a dummy for whether a restaurant is sold out on OpenTable around 7PM on a given day.[12] The estimated coefficients are marginal effects relative to the baseline of no Yelp reviews and a hygiene card different from A. The results show that displaying an A-card is associated with a 2.3 percentage points higher likelihood of being sold out on a given night relative displaying any other letter grade. This represents almost a 17% increase in sold out probability relative to the .14 baseline. Yelp ratings are associated with larger differences in sold-out probability. Having a 4.5-star rating is associated with a 33% increase in the probability of being sold out relative to having no Yelp reviews. Having a 3.5-star average rating is associated with a 47% reduction in the sold out probability.

As we include more and more controls, the effect of hygiene grade goes down, while the effect of star ratings stay relatively constant and economically large. At the extreme, when we control for restaurant fixed effects, the effect of letter grade disappears. This result suggests that letter grade could be simply picking up hygiene characteristics that are visible by customers in ways other than the hygiene score. Intuitively, given that OpenTable allows customers to book a restaurant online before seeing the letter grade displayed at the door, we would expect that the effect of hygiene inspections would be small at best. On the other hand, given that OpenTable users are likely to make their decisions based on online reviews, it is reassuring to see that Yelp stars have a larger effect on restaurant sold-out probability.

Table 6 shows the estimates when we instrument for the letter grade with inspector characteristics.[13] The first-stage estimates are displayed in Table 7, and the F-statistics are all above conventional levels. With instruments the effect of letter grade disappears, although in none of the specifications we can reject the null hypothesis that the letter grade is exogeneous in the second stage.

When we add the sufficient reductions from our prediction model (Table 8), we find that it is the hygiene information contained on Yelp that is associated with sold-out probability,

[11] We are in the process of collecting these data.

[12] In the current draft, we only include the tables of the linear probability models because the coefficients are easily interpretable as marginal effects. Qualitatively the results do not change when we estimate logit models.

[13] We do not have enough variation in our data to estimate the model with restaurant fixed effects.

rather than the letter grade displayed at the door. The coefficient on the A-card dummy drops and becomes statistically insignificant. The coefficient estimates on the average star-rating remain large and statistically significant, at least for 4.5 and 5-star ratings.

The coefficients on the (z-scored) sufficient reductions are displayed in Figure 6. The coefficients for the sufficient reductions corresponding to more *detectable* violation codes are highlighted in blue. The sufficient reductions more negatively associated with sold-out probability are those of *detectable* violation codes, such as 04H (adulterated food) and 04M (roaches present).

When we look at restaurant exit, we use a Cox proportional hazard model specification, using observations at the level of each restaurant and two-week time interval:

$$h_{it} = h_0 exp\left[\beta_1 doh_{it} + \beta_2 yelp_{it} + \beta_3\left(\Phi c_{it}\right) + \beta_4 X_i + \beta_5 X_t\right], \tag{3}$$

where h_{it} represents the hazard rate of exiting for restaurant i at age t, and the other variables contain those defined in Equation 2, plus a few more. The additional variables include whether a restaurant does not have a Yelp page, and dummies for the most recent inspection type. The endogeneity of our explanatory variables complicates the estimation. However, the close relationship between the Cox proportional model and the Poisson model (Laird and Olivier (1981)) allow us to estimate Equation 3 as an IV Poisson regression.

The current draft reports results from linear probability models. The results found using the sold-out probability are qualitatively similar to the results using restaurant exit as an outcome, although this time the effects of hygiene grades and Yelp reviews are comparable in size. Improvements in letter grades and Yelp star ratings are associated with a reduction in the probability that a restaurant goes out of business, and the results hold broadly true after adding a large set of restaurant and time controls, even after instrumenting for hygiene grades and after adding the sufficient reductions. Results are shown in Tables 9 through 11.

One might wonder whether the two outcome variables – sold out on OpenTable and going out of business – are good measures of demand. To the extent that restaurant success is the unique common underlying factor influencing both whether a restaurant is sold out on OpenTable and whether a restaurant goes out of business, we should expect these two measures to be negatively correlated.

To verify this, we estimate three-month survival curves of two groups of restaurants: those with a low probability of being sold out on OpenTable, and those with a high probability of being sold out on OpenTable. For every quarter, we compute the cumulative probability of being sold out on OpenTable up to the beginning of that quarter. We then estimate Kaplan-Meier survival curves with two strata, above and below the median probability of being sold

out. Fig. 8 plots these curves, together with confidence intervals. The blue curve at the top shows the survival probability for restaurants that start a quarter with above median probability of being sold out. The red curve at the bottom shows the survival probability for restaurants that start a quarter with below median sold-out probability. The fact that the red curve is always below the blue curve demonstrates that booking rates on OpenTable are correlated with exit. The higher the booking rate, the lower the probability that a restaurant will go out of business within 3 months.

5.2 Restaurants' Incentives

In order to assess whether restaurants change their hygiene quality due to online reputation, we take advantage of the fact that the DoH inspects restaurants that have different degrees of visibility on Yelp. We also take advantage of the fact that the DoH evaluates restaurants on different dimensions of quality, i.e. the different violation codes. The degree to which each violation code can be detected through online reviews differs across violation codes, as we have shown in Section 4. We define a restaurant's *visibility* on Yelp with two measures. The first measure is a dummy for whether a restaurant has received any reviews recently. The second measure is a function of the total number of recent reviews. We define a violation code's *detectability* on Yelp to be the degree of accuracy with which Yelp reviews can predict the occurrence of that violation code. The measure of detectability of each violation code is the result of our text analysis from Section 4, and it is displayed in Figure 4.

We run a OLS regression of the following type:

$$violation_{vit} = \nu_v + \nu_i + \alpha X_{vit} + \beta_v * on\_yelp_i + \gamma_v * has\_reviews_{it} + \delta_v * log(nr\_reviews_{it}+1) + \epsilon_{vit} \tag{4}$$

where $violation_{vit}$ is equal to one if restaurant i was found violating code v during inspection t. The dummy on_yelp_i is equal to one if the restaurant has a Yelp page with at least one review at some point in time. The dummy $has\_reviews_{it}$ is equal to one if the restaurant has received some reviews in the 90 days prior to inspection t. Finally, $log(nr\_reviews_{it}+1)$ is the logarithm of the total number of reviews received in the 90 days prior to inspection t.

In general, restaurants on Yelp tend to violate a little less during health inspections (Figure 9). However, to measure the effect of Yelp visibility on restaurant propensity to violate along more or less detectable dimensions of hygiene, we do not think that food-serving entities which are not on Yelp are a good comparison group for restaurants on Yelp. This is because during our sample period Yelp was already relatively popular, and because the DoH inspects establishments – such as workplace cafeterias – that are unlikely to be on Yelp but are likely to have other hygiene monitoring mechanisms.

Given the discussion above, the coefficients of interest are γ_v and δ_v, one per violation. These coefficients measure how likely restaurants on Yelp are to violate on different measures of hygiene as their short-term visibility on Yelp increases. The vector γ_v measures the difference between restaurants that are on Yelp but have not received recent reviews and restaurants that have received recent reviews. For the restaurants with recent reviews, the vector δ_v measures the marginal effect of a 1 percent increase in the number of reviews. Both coefficients should be negatively correlated with *detectability*, i.e. the coefficient estimates should be lower – more negative – for more *detectable* violations. We should note that X_{vit} includes, among other controls, the total inspection score during inspection t, so γ_v and δ_v measure the restaurant propensity to violate on specific hygiene dimensions, conditional on their overall hygiene level.

Results are displayed in Figures 10 through 12. Overall there seems to be a negative correlation between the detectability of a dimension of hygiene and restaurant propensity to violate along that dimension when the resturant is *visible* on Yelp. This is suggestive evidence that restaurant visiblity on Yelp is associated with a reduction in restaurant propensity to violate along dimensions that are detectable by consumers.

6 Conclusion

In a world where consumers cannot distinguish between high- and low-quality service providers before engaging in a transaction, the market can break down or consumers may engage in transactions that are too risky from a public policy perspective. In these contexts, regulation that guarantees minimum quality standards and informs consumers about different quality levels can help solve the market failure. However with the diffusion of online reviews – a new cheap way to collect provider quality information from consumers' past transactions – online reputation provides an additional way to reduce information asymmetries.

In the context of restaurants in New York City, we have shown that there are differences in the degree to which Yelp reviews can predict various dimensions of hygiene. Specifically, Yelp reviews contain relevant information to better predict pests and food handling violations than facilities and maintenance violations. We have also shown that the hygienic information contained in Yelp reviews is associated with consumers' choices of where to eat, above and beyond the information contained in the aggregate Yelp rating and in the hygiene grade card. Finally we find suggestive evidence that detectability of some hygiene dimensions on Yelp drive restaurants to comply with minimum hygiene standards exactly in those more detectable dimensions.

We do not have the ideal experiment to test whether telling inspectors not to check for

mice in a restaurant's premises and to instead focus on broken pipes would lead to more or less mice. However, we have shown that consumers are more informed about the presence of mice than broken pipes via Yelp reviews, and that both supply and demand seems to respond to this information available online. Our results combined provide suggestive evidence that regulators could focus their resources to inspect quality dimensions that consumers cannot detect, while leaving the more detectable dimensions to the *crowd*.

Of course, relying on online reviews to guarantee hygiene quality raises a new set of relevant questions. 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 it is useful for future transactions, it incentivizes providers to continuously engage in high quality transactions, and 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 buyers are more likely to leave a review after a positive experience (Nosko and Tadelis (2015)). They have also shown that if consumers personally interact with service providers – on Airbnb for example, when guests and hosts sometimes meet in person – they are less likely to leave negative reviews (Fradkin et al. (2016)). And if they fear retaliation after a negative review, the bias can be even worse (Resnick and Zeckhauser (2002)). 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)).

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 many to read, so inevitably consumers focus on just a (potentially non-representative) subset. 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. We leave the analysis of platform incentives to provide unbiased reviews to future research.

References

G. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, 1970. 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.

T. J. Bartik. *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute for Employment Research, 1991. ISBN 9780880991148. URL http://ideas.repec.org/b/upj/ubooks/wbsle.html.

L. Cabral and A. Hortacsu. 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. Chevalier, P. Rossi, and E. Oehlsen. The Value of Flexible Work: Evidence from Uber Drivers. Technical report, National Bureau of Economic Research, Cambridge, MA, mar 2017. URL http://www.nber.org/papers/w23296.pdf.

P. Cohen, R. Hahn, J. Hall, S. Levitt, and R. Metcalfe. Using Big Data to Estimate Consumer Surplus: The Case of Uber. Technical report, National Bureau of Economic Research, Cambridge, MA, sep 2016. URL http://www.nber.org/papers/w22627.pdf.

L. Einav, C. Farronato, and J. Levin. Peer-to-Peer Markets. *Annual Review of Economics*, 8(1):615–635, oct 2016. ISSN 1941-1383. doi: 10.1146/annurev-economics-080315-015334. URL http://www.annualreviews.org/doi/10.1146/annurev-economics-080315-015334.

C. Farronato and A. Fradkin. Market Structure with the Entry of Peer-to-Peer Platforms: the Case of Hotels and Airbnb. 2016.

M. Federman, D. Harrington, and K. Krynski. The impact of state licensing regulations on low-skilled immigrants: The case of Vietnamese manicurists. *The American Economic Review*, 2006. URL http://www.jstor.org/stable/30034649.

A. Fradkin, E. Grewal, D. Holtz, and M. Pearson. The determinants of online review informativeness: Evidence from field experiments on airbnb. 2016.

M. Gentzkow, J. M. Shapiro, and M. Taddy. Measuring polarization in high-dimensional data: Method and application to congressional speech. *NBER Working Paper*, 2016.

M. Gentzkow, B. T. Kelly, and M. Taddy. Text as Data. *NBER Working Paper*, 2017. URL http://www.nber.org/papers/w23276.pdf.

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

S. Greenstein, Y. Gu, and F. Zhu. Ideological Segregation among Online Collaborators: Evidence from Wikipedians. *NBER Working Paper*, 2016. URL `http://www.nber.org/papers/w22744`.

C. Harrison, M. Jorder, H. Stern, F. Stavinsky, and V. Reddy. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013. *MMWR*, 2014. URL `http://www.cdc.gov/MMWr/preview/mmwrhtml/mm6320a1.htm`.

X. Hui, M. Saeedi, Z. Shen, and N. Sundaresan. Reputation and regulations: evidence from ebay. *Management Science*, 62(12):3604–3616, 2016.

M. Ibanez and M. W. Toffel. Assessing the Quality of Quality Assessment: The Role of Scheduling. *SSRN Electronic Journal*, 2017. ISSN 1556-5068. doi: 10.2139/ssrn.2953142. URL `http://www.ssrn.com/abstract=2953142`.

G. Jin and P. Leslie. Reputational incentives for restaurant hygiene. *American Economic Journal: Microeconomics*, 2009. URL `http://www.ingentaconnect.com/content/aea/aejmi/2009/00000001/00000001/art00011`.

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.

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.

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, 2017.

N. Laird and D. Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240, 1981.

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. 2017.

M. Luca. Reviews, Reputation, and Revenue: The Case of Yelp.Com. *SSRN Electronic Journal*, 2011. ISSN 1556-5068. doi: 10.2139/ssrn.1928601. URL http://www.ssrn.com/abstract=1928601.

M. Luca and G. Zervas. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science*, 62(12):3412–3427, dec 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, aug 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. Watch where you eat: Restaurant hygiene inspections in new york city and moral hazard. 2017.

R. Meltzer, M. W. Rothbart, A. E. Schwartz, T. Calabrese, D. Silver, T. Mijanovich, and M. Weinstein. What are the financial implications of public quality disclosure? evidence from new york city's restaurant food safety grading policy. *Public Finance Review*, page 1091142117715112, 2017.

C. Nosko and S. Tadelis. The limits of reputation in platform markets: An empirical analysis and field experiment. 2015. doi: w20830. URL http://www.nber.org/papers/w20830.pdf.

M. F. Porter. An algorithm for suffix stripping. *Program*, 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.

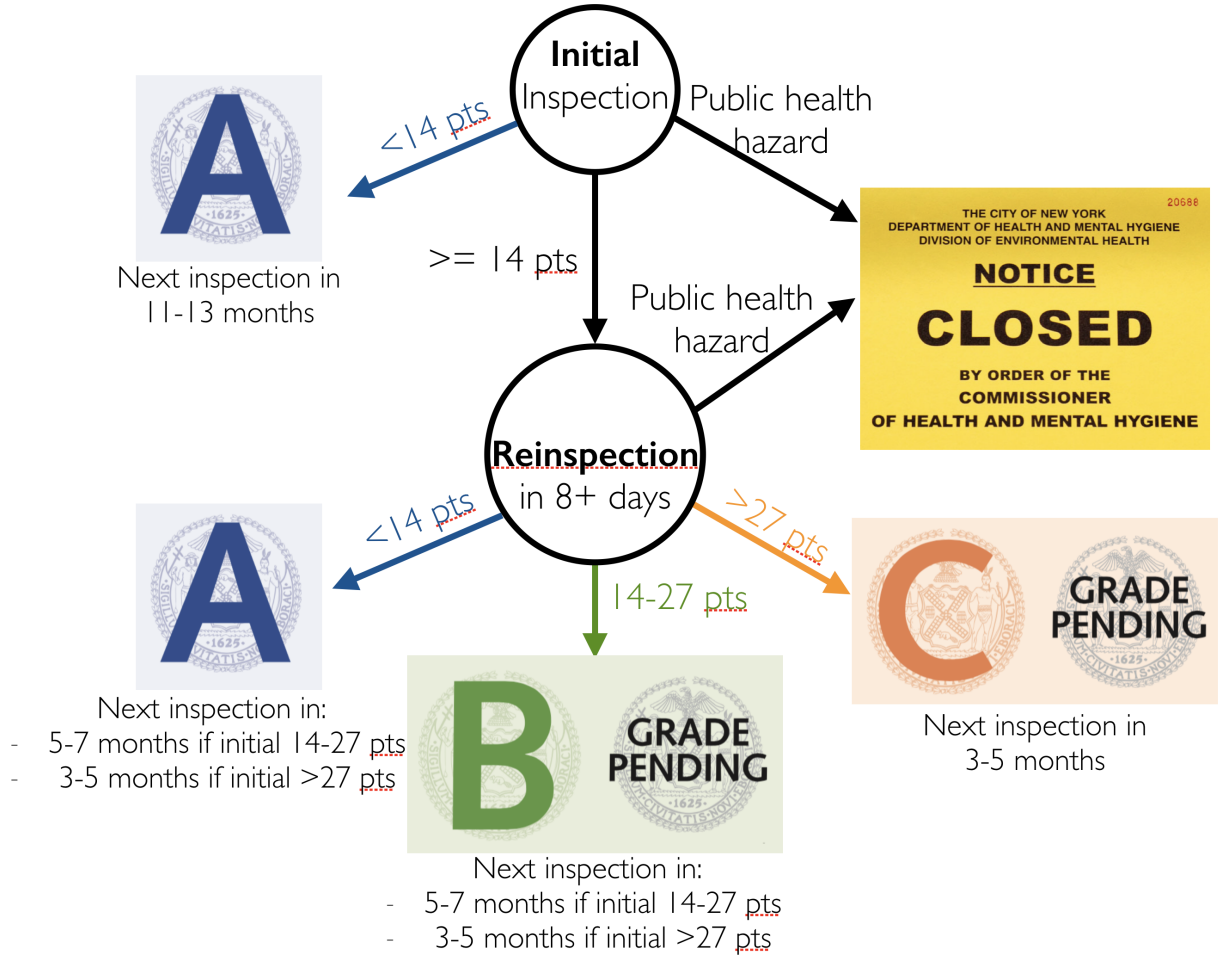
P. A. Simon, P. Leslie, G. Run, G. Z. Jin, et al. Impact of restaurant hygiene grade cards on foodborne-disease hospitalizations in los angeles county. *Journal of Environmental Health*, 67(7):32, 2005.

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=AAGBfm0AAAAAWQuYdzhTFRyG0fIMAlJFwMw6OFwvwHtd{&}scisf=4{&}ct=citation{&}cd=-1{&}hl=en.

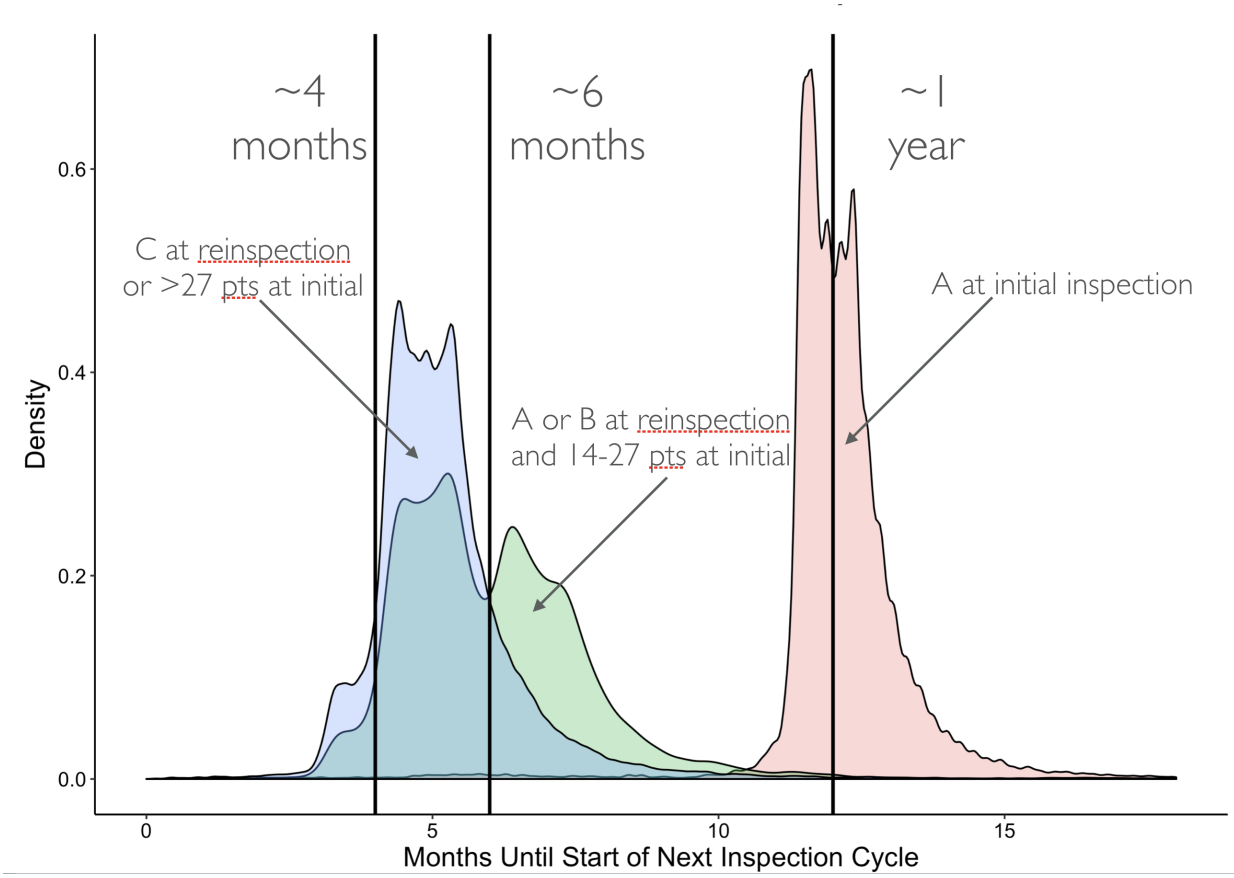
M. Taddy. Distributed Multinomial Regression. *The Annals of Applied Statistics*, 9(3):1394–1414, 2015. URL https://scholar.googleusercontent.com/scholar.bib?q=info:9CFVSMWrqbEJ:scholar.google.com/{&}output=citation{&}scisig=AAGBfm0AAAAAWQuYzwMR5EUZjgIaHlo62ql205MMzC9L{&}scisf=4{&}ct=citation{&}cd=-1{&}hl=en.

Figure 1: Inspection Cycle



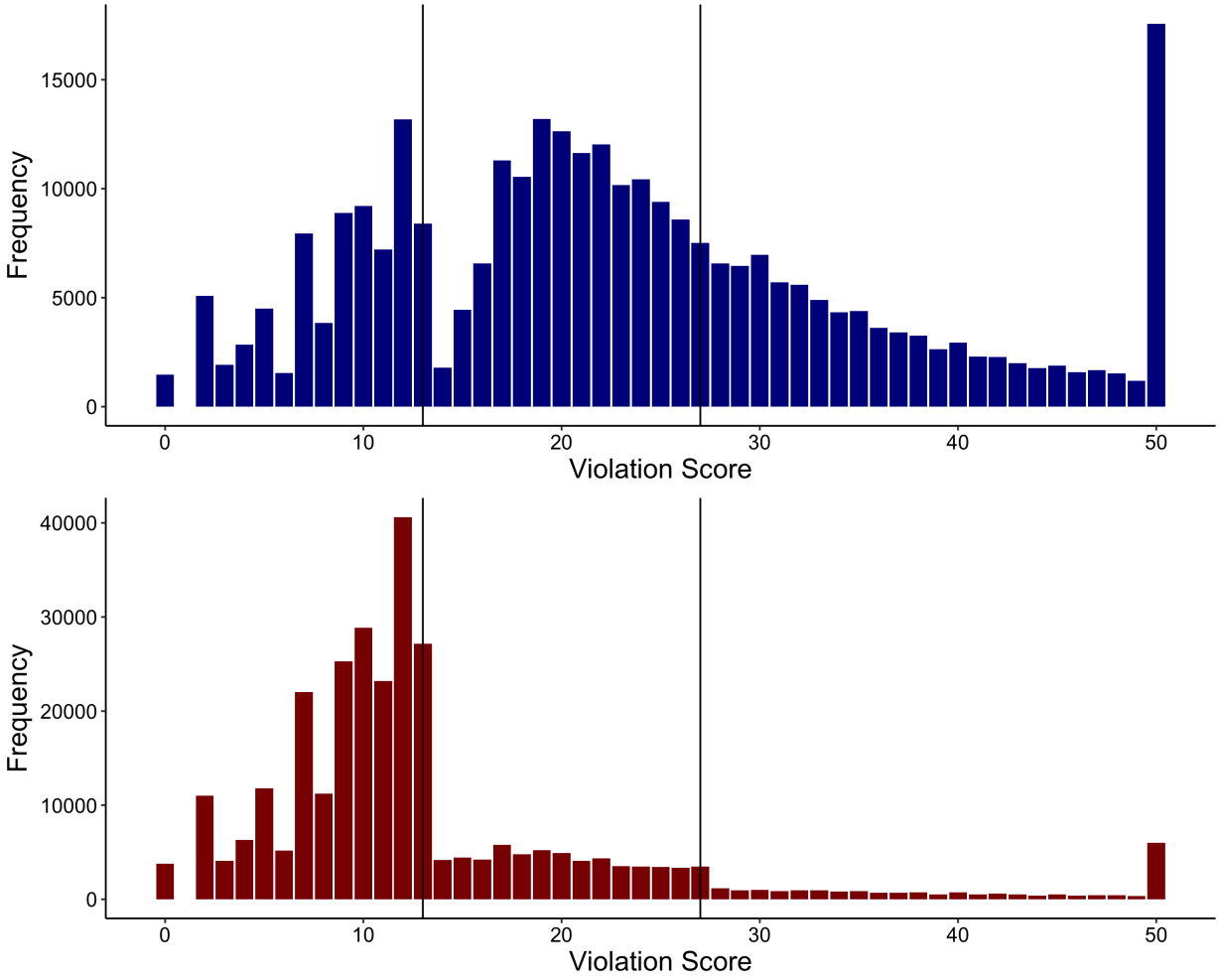
The figure shows the structure of an inspection cycle, and is adapted from `https://www1.nyc.gov/assets/doh/downloads/pdf/rii/inspection-cycle-overview.pdf` *(Accessed on February 5, 2018).*

Figure 2: Time Between Inspection Cycles



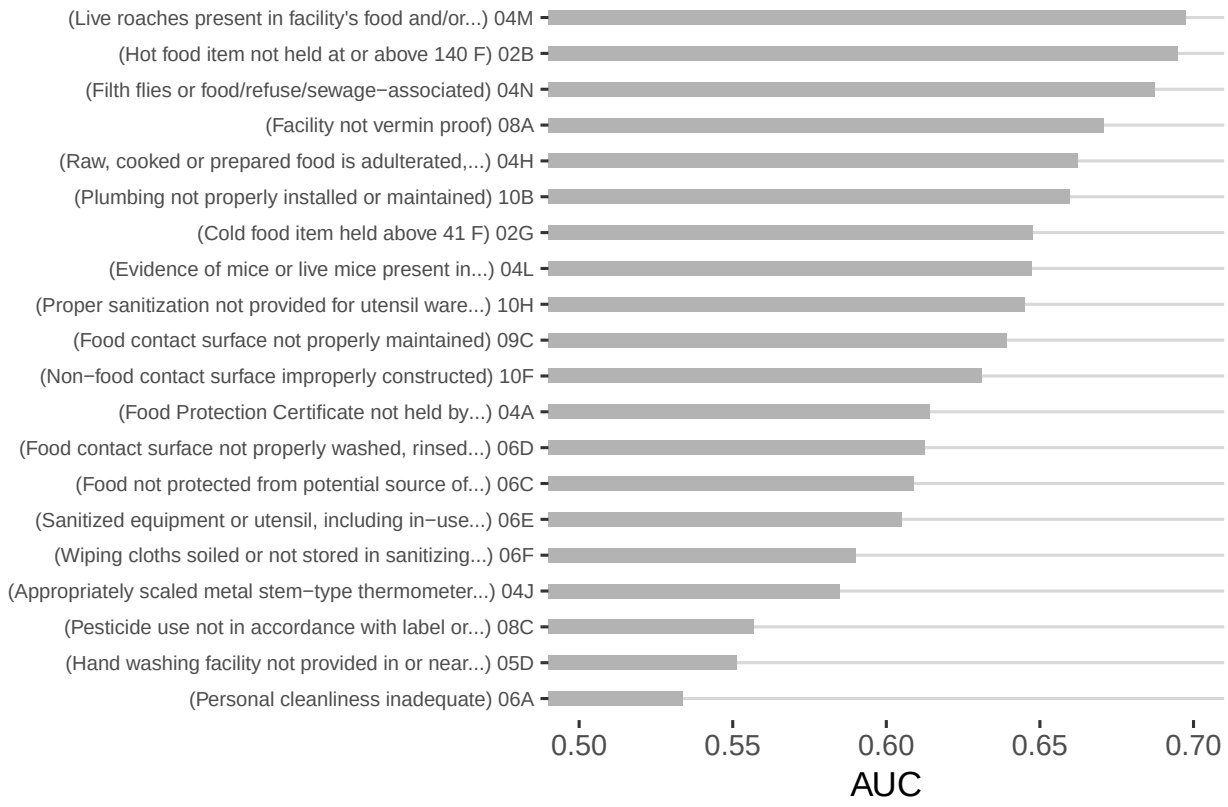
The plot shows the distribution of time between the last inspection within a cycle, and the initial inspection of the following cycle. For restaurants obtaining a A-grade at initial inspection (pink), the expected time is 12 months since the most recent inspection. For restaurants scoring 14-27 points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

Figure 3: Violation Scores after Initial Inspection and after Re-inspection



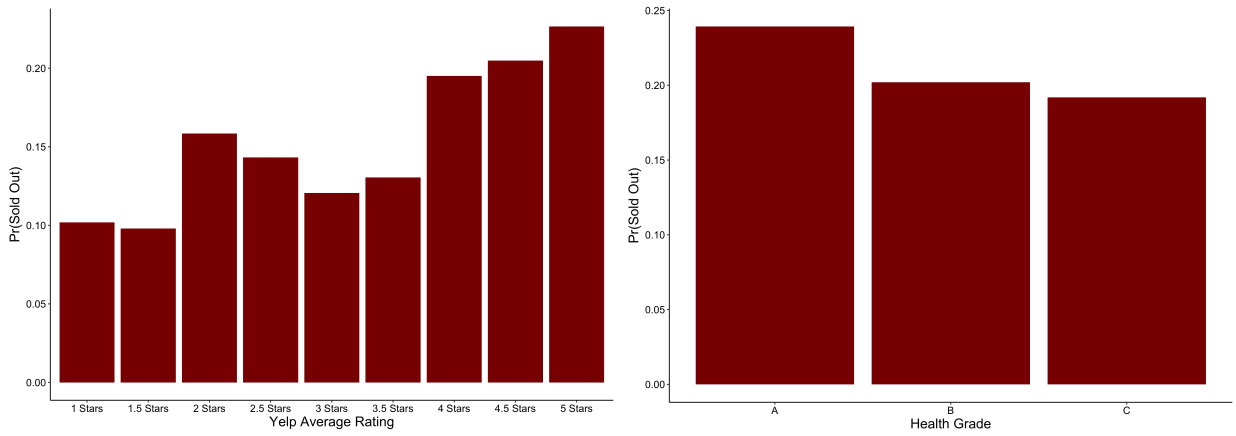
For each inspection cycle, the top panel shows the distribution of violation scores that restaurants obtain after the initial inspection. The vertical lines correspond to the score thresholds that would assign A-B-C letter grades. Scores of 13 or less automatically give an A-grade, while higher scores imply that a restaurant will be reinspected within a few weeks. The bottom panel shows the distribution of violation scores that restaurants obtain after re-inspection, and it includes those restaurants that obtained an A-grade at initial inspection.

Figure 4: Prediction Accuracy



The figure plots the AUC of the prediction of each of the 20 violations codes from restaurant characteristics and the sufficient reduction of Yelp review text. Details on the estimation algorithm are described in Section 4.

Figure 5: Quality Signals and Sold Out Probability

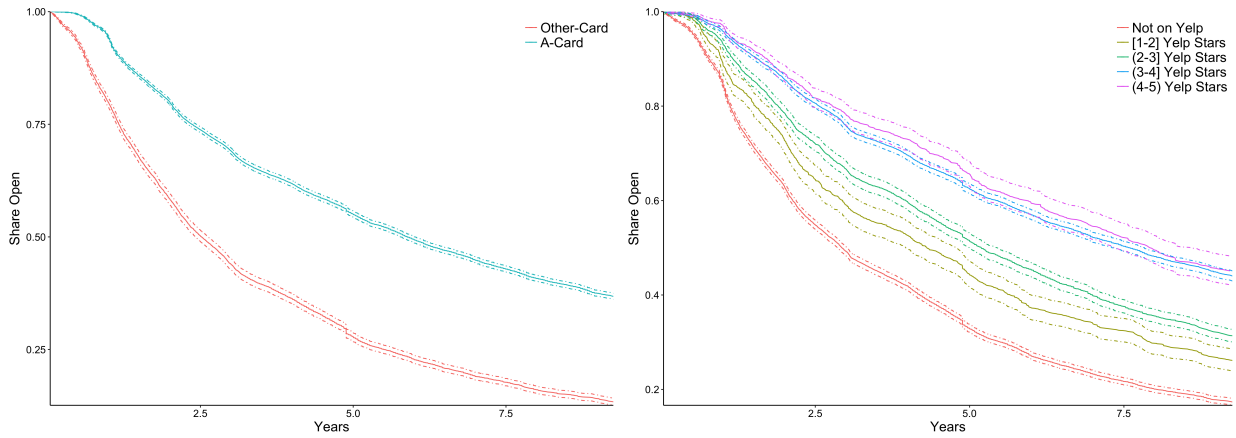


The two bar charts show the average probability of being sold out at 7PM on any given night as a function of Yelp reviews (left panel) and hygiene letter grades (right panel). 86% of restaurant-days fall into the A-card group, 9% fall into the B-card group, 2% fall into the C-card group, and 3% fall into the pending-card group. There are virtually no restaurants with average star rating below 2 on Yelp. Most restaurant-days (69%) are in the 3-4 star group, followed by the 4-5 star group (18%), and the 2-3 star or no-rating group (both at 6.5%).

Figure 6: Coefficient Estimates of Sufficient Reduction

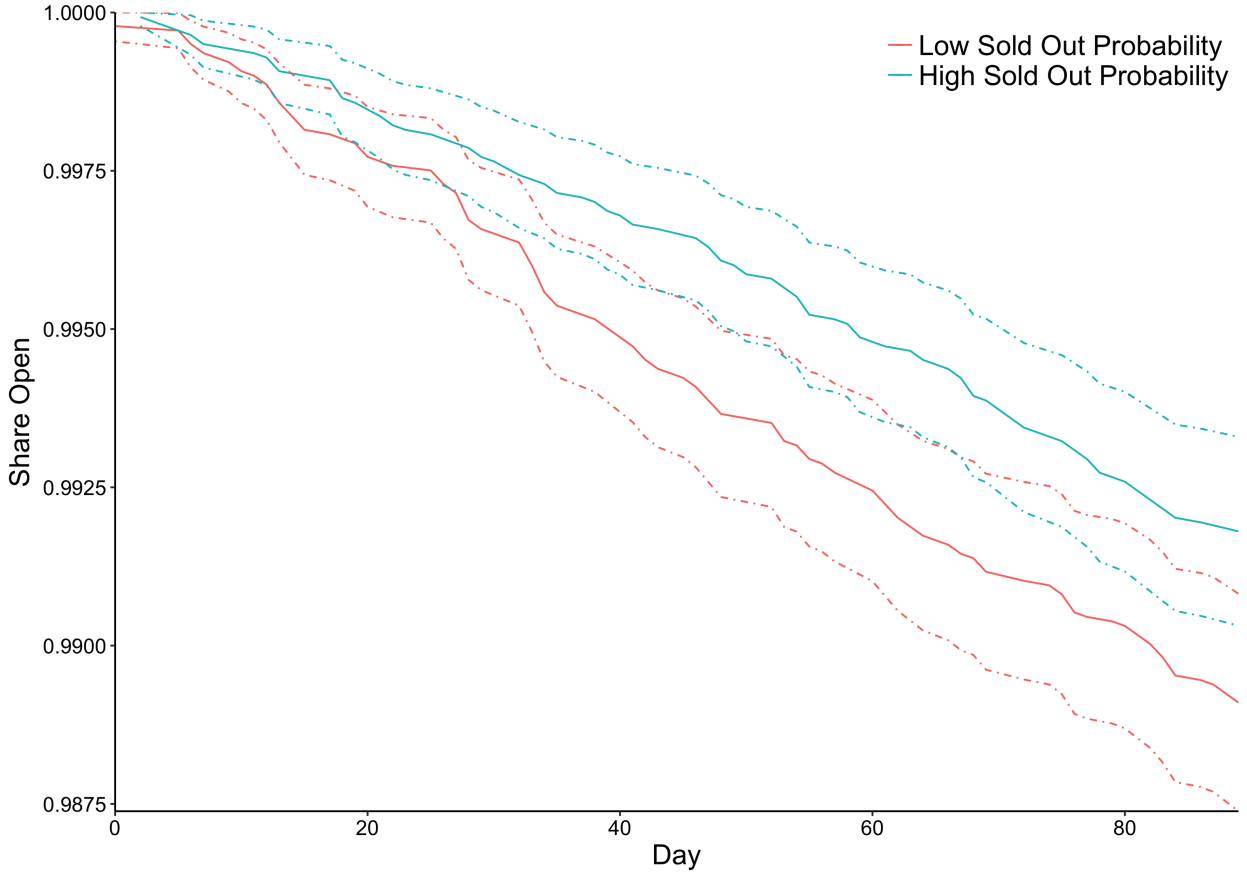
OLS coefficient estimates on the sufficient reductions, where the dependent variable is a dummy for for being sold out on OpenTable. The coefficient estimates complement the coefficients displayed in Table 8, column 3. The turquoise bars denote violation codes with high detectability (AUC > .65).

Figure 7: Quality Signals and Restaurant Exit



The figures plot restaurant survival curves for restaurants with and without an A-hygiene card (left panel), and for restaurants with different average star ratings on Yelp (right panel). The survival curves are estimated in discrete time with time-varying covariates, and the covariates are updated every 14 days. 95% confidence intervals are denoted with dotted lines.

Figure 8: Survival Curves of Restaurants on OpenTable



The figure shows three-month survival curves for restaurants. For every quarter, we compute the cumulative probability of being sold out on OpenTable up to the beginning of that quarter. We then estimate Kaplan-Meier survival curves with two strata, above and below the median probability of being sold out (median across all restaurants and quarters for which we have observations). The blue curve shows the survival probability for restaurants that start a quarter with above median probability of being sold out. The red curve shows the survival probability for restaurants that start a quarter with below median sold-out probability. The fact that the red curve is always below the blue curve demonstrates that booking rates on OpenTable are correlated with exit. The higher the booking rate, the lower the probability that a restaurant will go out of business within 3 months.

Figure 9: Inspection Score and Restaurant Visibility on Yelp

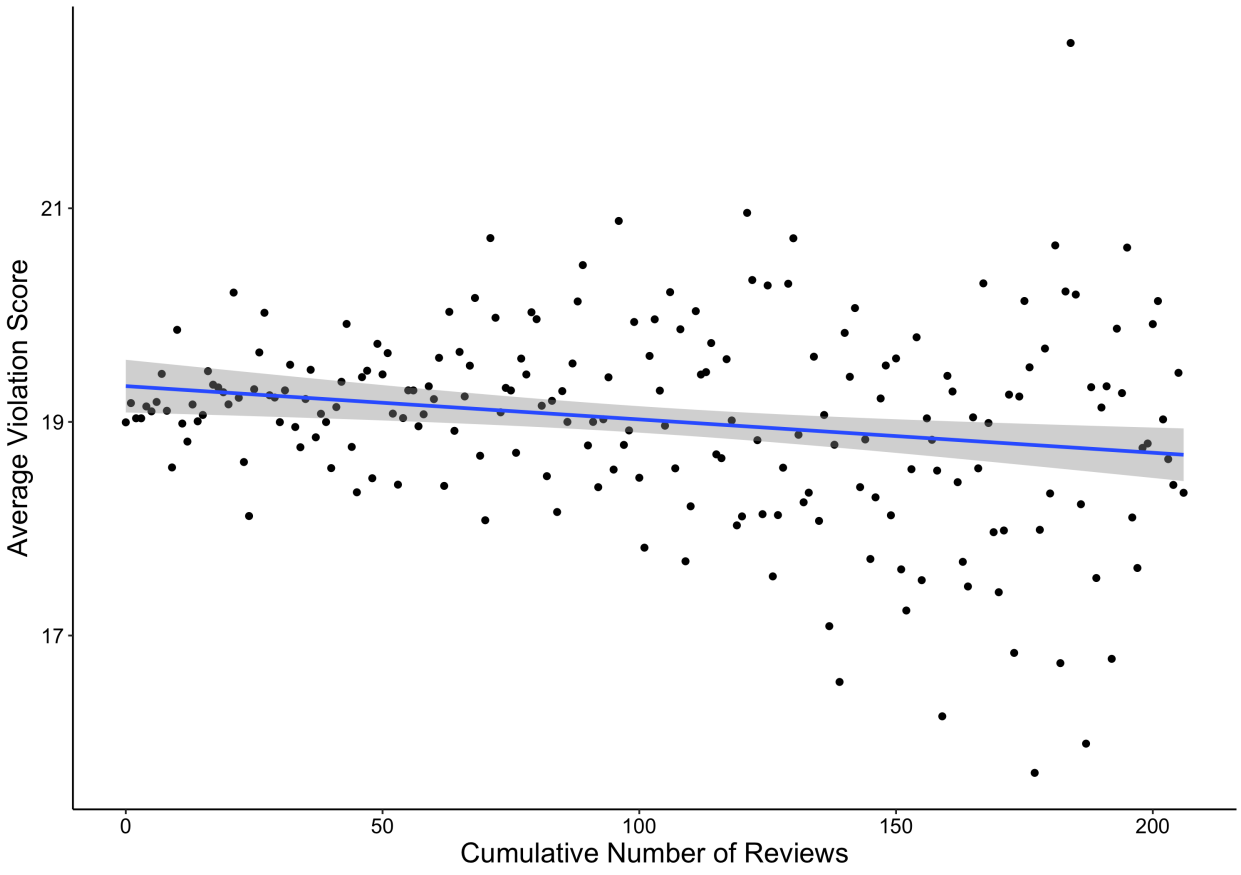
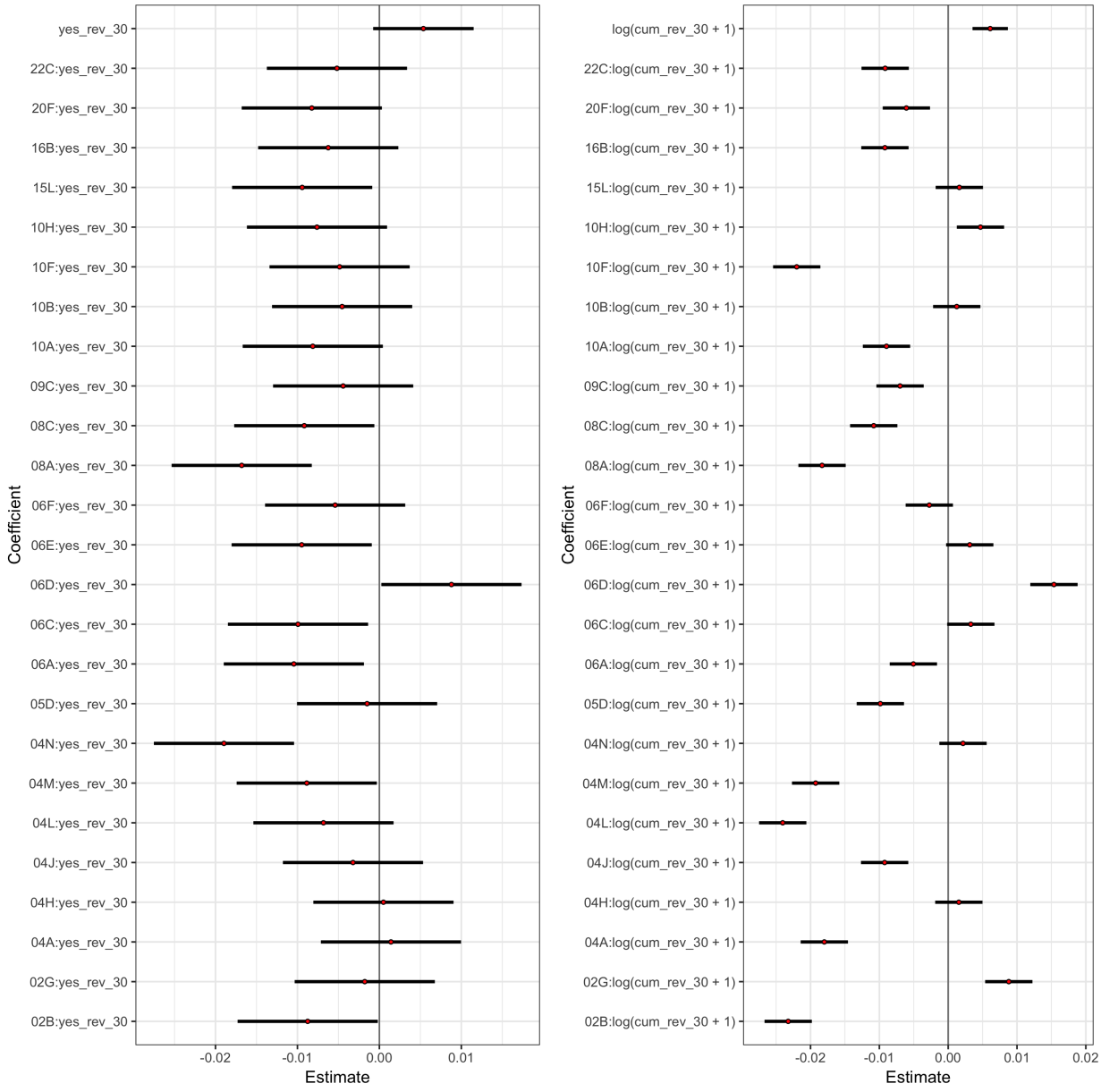
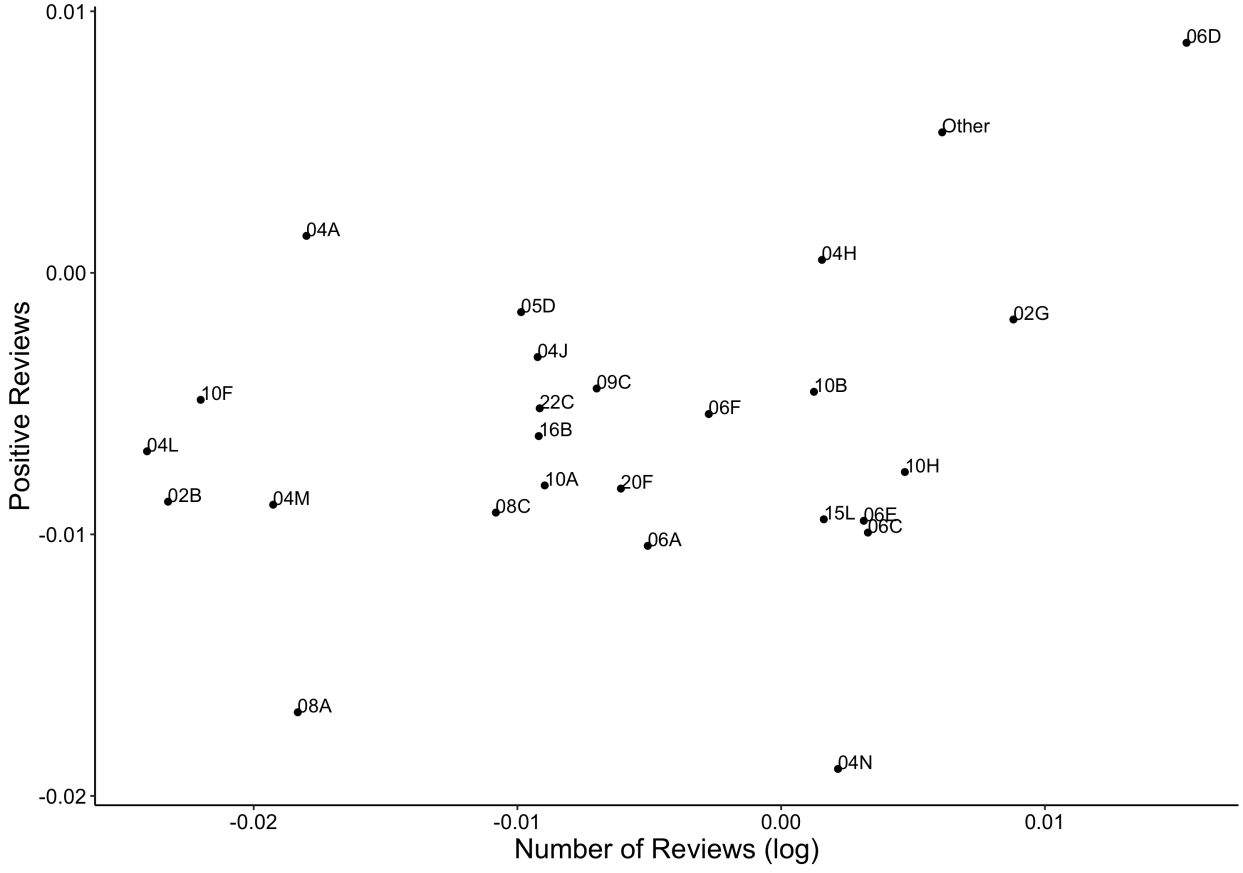


Figure 10: Diff-in-Diff Coefficient Estimates



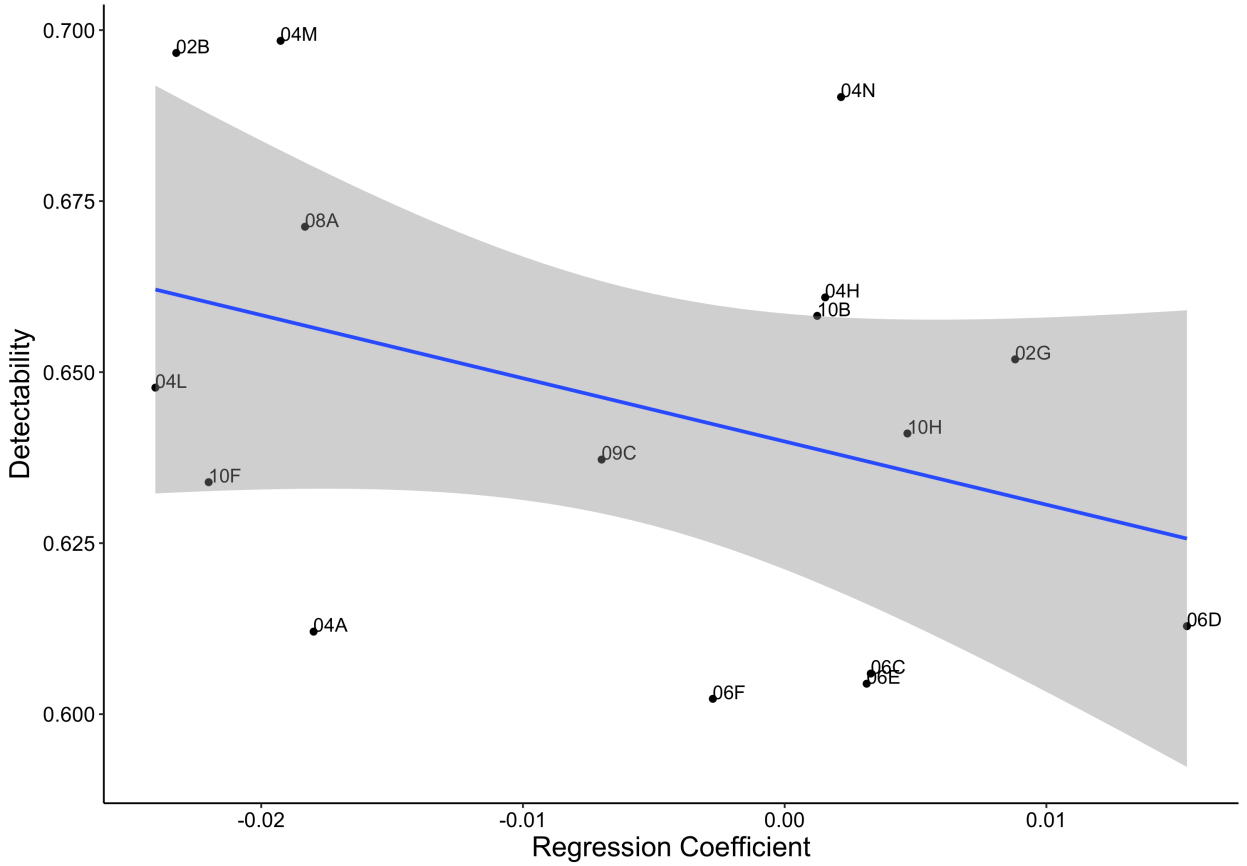
Coefficient estimates from equation 4. The left panel shows the propensity of Yelp restaurants to violate relative to restaurants without Yelp reviews in the last 90 days, γ_v. The right panel shows the propensity of Yelp restaurants to violate as the number of recent Yelp reviews increases, δ_v.

Figure 11: Diff-in-Diff Coefficient Estimates



Coefficient estimates from equation 4. The figure displays the coefficients estimates from Figure 10 as a scatterplot. On the x-axis we plot δ_v, on the y-axis we plot γ_v.

Figure 12: Propensity to violate and detectability on Yelp



The variable on the x-axis is the estimated δ_v coefficient from equation 4, one per violation. The variable on the y-axis is the R-squared from the prediction of violation v from restaurant characteristics and the text of Yelp reviews.

Tables

Table 1: Restaurant Characteristics

Characteristics	All	Yelp	OpenTable
Cuisine - American	20.6%	21.8%*	30.1%*
Cuisine - Chinese	10.2%	10.7%*	1.0%*
Cuisine - Cafe/Bakery	6.4%	7.1%*	0.5%*
Cuisine - Latin/Mexican	6.0%	5.9%	6.0%
Cuisine - Pizza	6.0%	7.3%*	2.0%*
Cuisine - Italian	3.1%	4.5%*	18.5%*
Borough - Manhattan	36.6%	44.0%*	83.5%*
Borough - Brooklyn	25.5%	25.3%	11.8%*
Borough - Queens	23.7%	20.3%*	2.9%*
Borough - Bronx	10.2%	6.5%*	1.1%*
Borough - Staten Island	3.9%	3.8%	0.8%*
Venue - Restaurant	47.5%	60.8%*	89.6%*
Venue - Fast Food	7.9%	8.8%*	0.2%*
Venue - Bar/Pub	4.7%	5.2%*	2.9%*
Share Chain Restaurants	10.2%	11.4%*	1.2%*
Share Closed	57.1%	42.8%*	22.2%*
Share Newly Opened	67.0%	67.2%	66.9%
N	57,045	31,218	2,534

$^*p < 0.01$

The 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.

Table 2: Grade Transitions

		Score at Initial Inspection		
Prior Card	N	0-13	14-27	28+
A	126,540	0.41	0.36	0.22
B	27,353	0.21	0.43	0.36
C	6,359	0.18	0.42	0.40

		Card Posted at End of Inspection Cycle		
Prior Card	N	A	B	C
A	126,540	0.85	0.13	0.02
B	27,353	0.64	0.30	0.06
C	6,359	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. So for example, of the 126,540 restaurant-inspections obtaining an A-grade during the previous inspection cycle, 41 percent scored between 0 and 13 points during the initial inspection, 36 percent scored between 14 and 27 points, and 22 percent scored 28 or more points. The right panel shows the card displayed before the cycle starts and the card displayed when the cycle ends. So of the 126,540 restaurant-inspections starting a new inspection cycle with an A-grade, 85 percent kept it, 13 percent dropped to B-grade, and 2 percent dropped to C-grade.

Table 3: Top Violation Codes

Code	Description	Share of Inspections
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.	45.9%
08A	Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.	42.1%
02G	Cold food item held above 41° F (smoked fish and reduced oxygen packaged foods above 38 °F) except during necessary preparation.	33.1%
06D	Food contact surface not properly washed, rinsed and sanitized after each use and following any activity when contamination may have occurred.	27.4%
04L	Evidence of mice or live mice present in facility's food and/or non-food areas.	26.2%
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.	24.5%
06C	Food not protected from potential source of contamination during storage, preparation, transportation, display or service.	23.7%
02B	Hot food item not held at or above 140° F.	20%
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.	13.3%
04H	Raw, cooked or prepared food is adulterated, contaminated, cross-contaminated, or not discarded in accordance with HACCP plan.	11.8%
06E	Sanitized equipment or utensil, including in-use food dispensing utensil, improperly used or stored.	11.4%
04A	Food Protection Certificate not held by supervisor of food operations.	9.9%
06F	Wiping cloths soiled or not stored in sanitizing solution.	8.7%
10H	Proper sanitization not provided for utensil ware washing operation.	8.1%
06A	Personal cleanliness inadequate. Outer garment soiled with possible contaminant. Effective hair restraint not worn in an area where food is prepared.	8.1%
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.	7.7%
04M	Live roaches present in facility's food and/or non-food areas.	7.6%
09C	Food contact surface not properly maintained.	7.6%
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.	6.8%
08C	Pesticide use not in accordance with label or applicable laws. Prohibited chemical used/stored. Open bait station used.	5.4%

The table provides a 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.

Table 4: Topics from the Multinomial Distributed Regression

Violation	Words
02B	vile, mice, driver, rush, minimum, front_desk, health, demand, incorrect, sick tourist_trap, roti, grade, counter, fusion, eel_avocado, hilton, absurd, overcook, ice
02G	rancid, applebe, rip_off, tourist_trap, mother, puke, horrend, redeem, diarrhea, wave straw, club, fall_apart, fishi, omelett, front_desk, hamburg, ticket, frozen, limp
04A	erica, promot, seafood_tower, lair, buzzer, sport, stair, samosa, smoke, threaten carrot_halwa, risotto, lentil, drunk, shoe, pot_pie, groupon, avocado, dough, smoke
04H	automat_gratuiti, trashi, douch_bag, aisl, lame, interrupt, atroci, egg_benedict, piss, attend pool_player, cafe_henri, bro, pitcher, pour, empti, suck, she, shell, stair
04J	porchetta, bouncer, turkey, honey_mustard, bland, bu_boy, overpr, garlic, raw, ined jaeger_schnitzel, rosti, chicharon_bulaklak, vindaloo, hash_brown, hudson, reuben, cod, bone, basket
04L	certif, 3d, ipad, voucher, medium_rare, ketchup, oil, music, spit, review lahmacun, flavorless, penn, frozen, chef, oliv_oil, nasti, hype, serv, entre
04M	abita, marti, fork, roach, mice, host, shove, turkey, tap, kitchen kunjip, omlett, bu_boy, onion_ring, bubbl_tea, yellowtail, eggplant, deep, frozen_margarita, smoke_salmon
04N	ghetto, toilet, voucher, straw, drunk, strike, hard_earn, disinterest, owner, tourist croissant, smoke, concept, raw, cart, pork_chop, orang_juic, pour, eventu, ice
05D	ticket_holder, brie, seafood_tower, doormen, coconut, bed_bug, deliveri, ticket, glass, appar il_bambino, chicharon_bulaklak, ticket_holder, crawfish, squid, shisha, stair, sesam, asian, skewer
06A	cl, paul, nan, bother, uncal, hill, baguett, kitchen, omelett, hung paella, bamboo_steamer, kimche, omelet, frost, nacho, soho, dough, salti, pour
06C	wife, grade, black_bean, silverwar, year, third, done, absurd, chef, threw indiffer, blech, stew, fat, tourist_trap, accept, mash_potato, either, potato, meh
06D	nightmar, inappropri, desk, tab, sarcast, shirt, remov, busboy, cash, shoe tourist_trap, bro, imposs, bummer, chipotl, smoke, chili_oil, bought, push, bone
06E	dandan_noodl, spa_castl, ticket, caraf, nacho, discount, atroci, bouncer, sit, upstair undersp, beef_randang, seva, pot_pie, cevich, oili, snapper, danc, ketchup, medium_rare
06F	pongsri, tourist_trap, guy, hair, black, argu, show, him, loung, waitress matsuhisa, bouncer, mac_n, raw, teriyaki, szechuan, mash_potato, michelin_star, music, sirloin
08A	mice, redeem, discount, downhil, voucher, social, gratuiti, grubhub, groupon, crappi reheat, toilet, gratuiti, runni, groupon, blah, forgett, hollandais_sauc, badli, anytim_soon
08C	dp, beard, dieci, coupon, groupon, cop, samosa, spinach, spanish, risotto croqueta_de, floyd, gra, do_camino, ihop, ghetto, rabbit, pancit, pork_chop, guac
09C	gil, snow_crab, maitr_d, iranian, forgotten, float, needless, entre, fee, parti pigal, matsuhisa, hostest, macaroni, schnitzel, mani_pedi, mash_potato, yucca, groupon, loud
10B	risotto, rush, ipad, bu_boy, remov, grade, oxtail, dick, scream, chef ketchup, tourist_trap, bleh, rigatoni, poach_egg, magnolia, hash_brown, ined, eggplant, dri
10F	certif, jacket, feet, coupon, fight, stolen, credit_card, tourist_trap, ticket, minimum ticket_holder, greas, asian, passabl, bonchon, ehh, veal, glass, bottl, medium
10H	insult, doorman, acknowledg, wear, credit_card, internet, honor, later, attend, brown porch_swing, sum, shabu, scallop, veal, sea_bass, fish, potenti, barista, plate
rating1	refund, unprofession, disgust, poison, worst, refus, zero, disrespect, insult, appal
rating2	meh, tasteless, below_averag, flavorless, downhil, unprofession, underwhelm, worst, disgust, unaccept
rating3	write_home, meh, alright, averag, eh, wow_factor, unmemor, fell_short, underwhelm, solid
rating4	highli_recommend, gem, hidden_gem, perfect, yum, solid, delici, fantast, excel, divin
rating5	highli_recommend, gem, hidden_gem, heaven, impecc, amaz, everi_penni, phenomen, perfect, fantast

Words with the highest loadings in 1- and 2-star reviews.

Table 5: Effect of Yelp Reviews and Health Grades on Sold Out Probability – OLS

	Sold Out at 7PM (OLS)			
	(1)	(2)	(3)	(4)
1.5 Yelp Stars	0.093	0.103	0.114	-0.002
	(0.124)	(0.104)	(0.098)	(0.052)
2 Yelp Stars	0.017	0.013	-0.005	0.039
	(0.028)	(0.036)	(0.037)	(0.030)
2.5 Yelp Stars	0.200	0.168	0.160	0.012
	(0.207)	(0.194)	(0.192)	(0.022)
3 Yelp Stars	-0.015	-0.029	-0.032	0.009
	(0.029)	(0.028)	(0.028)	(0.020)
3.5 Yelp Stars	-0.065***	-0.052***	-0.056***	0.011
	(0.017)	(0.017)	(0.017)	(0.017)
4 Yelp Stars	-0.038**	-0.034**	-0.039**	0.015
	(0.016)	(0.016)	(0.016)	(0.016)
4.5 Yelp Stars	0.046**	0.055***	0.056***	0.026*
	(0.018)	(0.017)	(0.017)	(0.015)
5 Yelp Stars	0.125***	0.128***	0.135***	0.032**
	(0.035)	(0.032)	(0.032)	(0.014)
Grade A	0.023***	0.013*	0.013*	-0.004
	(0.008)	(0.007)	(0.007)	(0.003)
Constant	0.138***	0.027	-0.001	0.145***
	(0.016)	(0.099)	(0.100)	(0.015)
Observations	1,635,231	1,635,231	1,635,231	1,635,231
R-squared	0.019	0.072	0.085	0.406
Geo and Restaurant Controls	No	Yes	Yes	No
Time Controls	No	No	Yes	Yes
Camis Fixed Effects	No	No	No	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Linear probability model of the probability of being sold out on health outcomes and Yelp reviews. One observation is a restaurant-day. The outcome variable is equal to 1 if the restaurant has no availability on OpenTable between 6:30PM and 7:30PM. Controls include time-invariant restaurant characteristics such as cuisine and zipcode, and time controls such as day of the week and quarter fixed effects. Standard errors are clustered at the restaurant level.

Table 6: Effect of Yelp Reviews and Health Grades on Sold Out Probability – IV

	Sold Out at 7PM (IV)		
	(1)	(2)	(3)
1.5 Yelp Stars	0.086	0.093	0.107
	(0.128)	(0.109)	(0.103)
2 Yelp Stars	0.024	0.021	0.003
	(0.028)	(0.036)	(0.037)
2.5 Yelp Stars	0.198	0.165	0.157
	(0.206)	(0.192)	(0.191)
3 Yelp Stars	-0.021	-0.036	-0.038
	(0.029)	(0.029)	(0.029)
3.5 Yelp Stars	-0.066***	-0.054***	-0.058***
	(0.017)	(0.017)	(0.017)
4 Yelp Stars	-0.039**	-0.034**	-0.039**
	(0.016)	(0.016)	(0.016)
4.5 Yelp Stars	0.048***	0.057***	0.057***
	(0.018)	(0.017)	(0.017)
5 Yelp Stars	0.127***	0.131***	0.137***
	(0.035)	(0.032)	(0.032)
Grade A	-0.035	-0.051	-0.041
	(0.043)	(0.042)	(0.046)
Constant	0.188***	0.080	0.039
	(0.040)	(0.105)	(0.106)
Observations	1,635,231	1,635,231	1,635,231
R-squared	0.017	0.069	0.082
Geo and Restaurant Controls	No	Yes	Yes
Time Controls	No	No	Yes
Camis Fixed Effects	No	No	No

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Linear probability model of the probability of being sold out on health outcomes and Yelp reviews. One observation is a restaurant-day. The outcome variable is equal to 1 if the restaurant has no availability on OpenTable between 6:30PM and 7:30PM. Same regression as in Table 5, except that the letter grade posted on the restaurant door – dummy for grade A – is instrumented with two variables that shift the inspection violation score, but are unlikely to be correlated with other quality characteristics observable by restaurant guests. The first instrument is whether the inspector of the current inspection was the same as the inspector of the previous restaurant inspection. The second instrument is an average of the inspector's violation scores assigned to restaurants other than the one being inspected. Controls include time-invariant restaurant characteristics such as cuisine and zipcode, and time controls such as day of the week and quarter fixed effects. Standard errors are clustered at the restaurant level. In all specifications, we reject the null hypothesis of exogeneity of hygiene letter grades (p-values range between 0.10 and 0.20.

Table 7: Effect of Yelp Reviews and Health Grades on Sold Out Probability – IV First Stage

	Grade A (IV First Stage)		
	(1)	(2)	(3)
1.5 Yelp Stars	-0.149	-0.187	-0.153
	(0.123)	(0.124)	(0.125)
2 Yelp Stars	0.120***	0.136***	0.157***
	(0.035)	(0.052)	(0.054)
2.5 Yelp Stars	-0.041	-0.057**	-0.058*
	(0.029)	(0.028)	(0.030)
3 Yelp Stars	-0.099***	-0.106***	-0.102***
	(0.035)	(0.037)	(0.037)
3.5 Yelp Stars	-0.035**	-0.038**	-0.041**
	(0.016)	(0.016)	(0.016)
4 Yelp Stars	-0.012	-0.009	-0.016
	(0.012)	(0.012)	(0.012)
4.5 Yelp Stars	0.028**	0.036***	0.032**
	(0.012)	(0.012)	(0.012)
5 Yelp Stars	0.048***	0.048***	0.049***
	(0.016)	(0.017)	(0.017)
Inspector Stringency	-0.015***	-0.015***	-0.014***
	(0.001)	(0.001)	(0.001)
Same Inspector	-0.002	0.002	0.005
	(0.021)	(0.021)	(0.021)
Constant	1.192***	1.148***	1.046***
	(0.019)	(0.066)	(0.069)
Observations	1,635,231	1,635,231	1,635,231
R-squared	0.045	0.079	0.087
Geo and Restaurant Controls	No	No	No
Time Controls	No	No	No
Camis Fixed Effects	No	No	No

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

First stage of Table 6. The F-statistics of the first stage regressions are all above conventional levels.

Table 8: Effect of Yelp Reviews and Health Grades on Sold Out Probability – SR

	Sold Out at 7PM (OLS)			
	(1)	(2)	(3)	(4)
1.5 Yelp Stars	0.115	0.132	0.135	0.011
	(0.132)	(0.108)	(0.104)	(0.054)
2 Yelp Stars	0.039	0.034	0.012	0.043
	(0.031)	(0.041)	(0.041)	(0.033)
2.5 Yelp Stars	0.256	0.222	0.210	0.021
	(0.209)	(0.194)	(0.194)	(0.025)
3 Yelp Stars	0.029	0.023	0.014	0.019
	(0.037)	(0.032)	(0.033)	(0.022)
3.5 Yelp Stars	-0.015	0.001	-0.009	0.020
	(0.029)	(0.026)	(0.026)	(0.020)
4 Yelp Stars	0.007	0.014	0.003	0.024
	(0.029)	(0.025)	(0.026)	(0.019)
4.5 Yelp Stars	0.073**	0.087***	0.082***	0.035**
	(0.030)	(0.026)	(0.026)	(0.018)
5 Yelp Stars	0.129***	0.138***	0.138***	0.039**
	(0.040)	(0.034)	(0.034)	(0.016)
Grade A	0.018**	0.010	0.010	-0.004
	(0.007)	(0.007)	(0.007)	(0.003)
Constant	0.356***	0.246**	0.230**	0.144***
	(0.045)	(0.110)	(0.111)	(0.020)
Observations	1,635,231	1,635,231	1,635,231	1,635,231
R-squared	0.041	0.089	0.101	0.406
Geo and Restaurant Controls	No	Yes	Yes	No
Time Controls	No	No	Yes	Yes
Camis Fixed Effects	No	No	No	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The table displays estimation results from a specification identical to Table 5, except that now we include the sufficient reductions as additional explanatory variables. The estimates of the coefficients for the sufficient reductions are shown in Figure 6.

Table 9: Effect of Yelp Reviews and Health Grades on Restaurant Exit – OLS

	Out of Business (OLS)			
	(1)	(2)	(3)	(4)
1.5 Yelp Stars	-0.001	0.001	0.001	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)
2 Yelp Stars	0.001	0.002***	0.002***	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
2.5 Yelp Stars	0.002***	0.004***	0.003***	0.001
	(0.000)	(0.000)	(0.000)	(0.001)
3 Yelp Stars	0.001***	0.002***	0.002***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
3.5 Yelp Stars	0.001***	0.002***	0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
4 Yelp Stars	-0.000***	0.001***	0.000	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)
4.5 Yelp Stars	-0.001***	-0.000***	-0.001***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)
5 Yelp Stars	-0.001***	-0.001***	-0.001***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)
Not on Yelp	0.004***	0.004***	0.004***	
	(0.000)	(0.000)	(0.000)	
Grade A	-0.004***	-0.003***	-0.004***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.007***	0.009	-0.001	-0.020***
	(0.000)	(0.008)	(0.009)	(0.001)
Observations	3,682,413	3,682,413	3,682,413	3,682,413
R-squared	0.002	0.004	0.005	0.040
Geo and Restaurant Controls	No	Yes	Yes	No
Time Controls	No	No	Yes	Yes
Camis Fixed Effects	No	No	No	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Linear probability model of the probability of being out of business. For all inspection types A and B, the outcome variable measures whether a restaurant closes within 90 days since the inspection. Yelp ratings are computed as cumulative average of all the reviews received up to the date of the inspection. In column (1) we only include inspection type fixed effects as controls. Column (2) includes inspection type fixed effects, as well as fixed effects for zipcode, cuisine type, service type, venue, and whether it's a chain restaurant. Column (3) includes all the controls in column (2) plus time controls, i.e. day of the week FE, year-quarter FE, restaurant age group FE. Column (4) has inspection type FE, restaurant FE, and time controls from column (3). Standard errors are clusterd at the restaurant level. Because the DoH can actually decide to close a restaurant during an inspection if hygiene violations are deemed too severe, we run the regressions both with and without those restaurants that went out of business immediately following a closure by the DoH. The table presents the results exluding those inspections, but the results are virtually identical.

Table 10: Effect of Yelp Reviews and Health Grades on Restaurant Exit – IV

	Out of Business (IV)		
	(1)	(2)	(3)
1.5 Yelp Stars	-0.001	0.001	0.000
	(0.002)	(0.002)	(0.002)
2 Yelp Stars	0.001	0.002***	0.002**
	(0.001)	(0.001)	(0.001)
2.5 Yelp Stars	0.002***	0.004***	0.003***
	(0.000)	(0.000)	(0.000)
3 Yelp Stars	0.001***	0.002***	0.002***
	(0.000)	(0.000)	(0.000)
3.5 Yelp Stars	0.001***	0.002***	0.001***
	(0.000)	(0.000)	(0.000)
4 Yelp Stars	-0.000***	0.001***	0.000
	(0.000)	(0.000)	(0.000)
4.5 Yelp Stars	-0.001***	-0.000*	-0.001***
	(0.000)	(0.000)	(0.000)
5 Yelp Stars	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)
Not on Yelp	0.004***	0.004***	0.004***
	(0.000)	(0.000)	(0.000)
Grade A	-0.005***	-0.006***	-0.009***
	(0.001)	(0.001)	(0.001)
Constant	0.008***	0.010	-0.001
	(0.001)	(0.008)	(0.008)
Observations	3,682,413	3,682,413	3,682,413
R-squared	0.002	0.004	0.004
Geo and Restaurant Controls	No	Yes	Yes
Time Controls	No	No	Yes
Camis Fixed Effects	No	No	No

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Linear probability model of the probability of being out of business. The specifications are the same as in Table 9, but the health grade is instrumented with inspector-specific characteristics: same inspector as in the previous inspection, and inspector's average violation points assigned to other restaurants. In all specifications we reject the null hypothesis of exogeneity of hygiene letter grades.

Table 11: Effect of Yelp Reviews and Health Grades on Restaurant Exit – IV First Stage

	Grade A (IV First Stage)		
	(1)	(2)	(3)
1.5 Yelp Stars	-0.004	-0.069	-0.117***
	(0.044)	(0.044)	(0.044)
2 Yelp Stars	0.020	-0.016	-0.057***
	(0.016)	(0.015)	(0.016)
2.5 Yelp Stars	0.019**	0.002	-0.036***
	(0.009)	(0.008)	(0.008)
3 Yelp Stars	0.012**	0.012**	-0.022***
	(0.005)	(0.005)	(0.005)
3.5 Yelp Stars	0.025***	0.033***	0.001
	(0.003)	(0.003)	(0.003)
4 Yelp Stars	0.025***	0.040***	0.008**
	(0.003)	(0.003)	(0.003)
4.5 Yelp Stars	0.055***	0.068***	0.035***
	(0.003)	(0.003)	(0.003)
5 Yelp Stars	0.085***	0.089***	0.059***
	(0.003)	(0.004)	(0.004)
Inspector Stringency	-0.011***	-0.012***	-0.010***
	(0.000)	(0.000)	(0.000)
Same Inspector	0.016***	0.020***	-0.003
	(0.005)	(0.005)	(0.005)
No other score	-0.676***	-0.630***	-0.508***
	(0.158)	(0.157)	(0.124)
Constant	1.133***	1.210***	0.623**
	(0.003)	(0.285)	(0.261)
Observations	3,682,413	3,682,413	3,682,413
R-squared	0.201	0.220	0.250
Geo and Restaurant Controls	No	No	No
Time Controls	No	No	No
Camis Fixed Effects	No	No	No

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

First stage of the IV regressions in Table 10. The smallest F-statistic of these first stages is 2,321.

Table 12: Effect of Yelp Reviews and Health Grades on Restaurant Exit – SR

	Out of Business (OLS)			
	(1)	(2)	(3)	(4)
1.5 Yelp Stars	-0.001	0.001	0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)
2 Yelp Stars	0.001	0.002***	0.002***	0.000
	(0.001)	(0.001)	(0.001)	(0.001)
2.5 Yelp Stars	0.002***	0.004***	0.003***	0.002**
	(0.000)	(0.000)	(0.000)	(0.001)
3 Yelp Stars	0.002***	0.003***	0.003***	0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
3.5 Yelp Stars	0.002***	0.002***	0.002***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
4 Yelp Stars	0.001***	0.001***	0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
4.5 Yelp Stars	0.000***	0.001***	0.001***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)
5 Yelp Stars	0.000**	0.000*	0.000	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Not on Yelp	0.009***	0.008***	0.008***	
	(0.000)	(0.000)	(0.000)	
Grade A	-0.004***	-0.003***	-0.004***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.008***	0.012	0.003	-0.019***
	(0.000)	(0.008)	(0.008)	(0.001)
Observations	3,682,413	3,682,413	3,682,413	3,682,413
R-squared	0.003	0.005	0.006	0.041
Geo and Restaurant Controls	No	Yes	Yes	No
Time Controls	No	No	Yes	Yes
Camis Fixed Effects	No	No	No	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The table displays estimation results from a specification identical to Table 5, except that now we include the sufficient reductions as additional explanatory variables. The estimates of the coefficients for the sufficient reductions are shown in Figure 6.

Table 13: Restaurant visibility on Yelp and hygiene performance

	(log) Inspection Score		
	(1)	(2)	(3)
On Yelp	0.069***	0.001	−0.044***
	(0.007)	(0.008)	(0.013)
Has Recent Yelp Reviews	0.019**	0.022**	−0.020
	(0.010)	(0.009)	(0.012)
(log) Number of Recent Yelp Reviews	−0.007*	−0.017***	0.025***
	(0.004)	(0.004)	(0.008)
(log) Restaurant Age		0.023***	0.001
		(0.001)	(0.002)
Constant	2.431***		
	(0.004)		
Geo and Restaurant Controls	No	Yes	No
Restaurant FE	No	No	Yes
Observations	244,121	244,117	244,117
R^2	0.001	0.028	0.180
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 14: Effect of restaurant visibility on Yelp on restaurant incentives to improve hygiene performance

	Violation Found
Inspection Score (log)	0.109^{***}
	(0.0004)
Restaurant Age (log)	0.001^{***}
	(0.0001)
On Yelp	-0.010^{***}
	(0.001)
Has Recent Yelp Reviews	0.0003
	(0.001)
(log) Number of Recent Yelp Reviews	0.004^{***}
	(0.001)
On Yelp * Detectable Violation	0.013^{***}
	(0.002)
Has Recent Yelp Reviews * Detectable	-0.004^{*}
	(0.002)
(log) Number of Recent Yelp Reviews * Detectable	-0.006^{***}
	(0.001)
Violation Code FE	Yes
Restaurant FE	Yes
Observations	3,326,192
R^2	0.140
Note:	*p<0.1; **p<0.05; ***p<0.01

Coefficient estimates of equation 4. Here, instead of computing one set of (β, γ, δ) per violation, we compute $(\beta_d, \gamma_d, \delta_d)$ for the 8 most detectable violations and $(\beta_{nd}, \gamma_{nd}, \delta_{nd})$ for the 8 least detectable violations, as measured by our text analysis.

A1 How are inspectors assigned to restaurants?

The instrumental variable approach described in Section 5.1 relies on random assignment of inspectors to restaurants. Here we verify that conditional on observables, we cannot reject the hypothesis that inspectors are randomly assigned to evaluate restaurants. To do that, we compute two probability distributions. First, we compute the unconditional distribution of a particular restaurant characteristic, denoted $P(X)$. Second, we compute the distribution of X conditional on a particular inspector Z, denoted $P(X|Z)$. If inspectors are assigned randomly conditional on the observable characteristic X, then the conditional distribution of restaurant characteristic X should be the same as its distribution conditional on a particular inspector. So for every inspector we compute the difference $P(X) - P(X|Z)$ across all possible values of X.

Figures A1 through A7 display the distribution of $P(X) - P(X|Z)$ across all inspectors and for different observable characteristics. If inspectors were randomly assigned to restaurants conditional on observable X, the dotted and solid density functions could not be distinguished from one another. The figures show that inspectors tend to specialize by geography, inspecting restaurants in one New York City borough more than in other boroughs (Figure A1), and clustering inspections within a few zipcodes (Figure A5). Beyond geography, there does not seem to be specialization of inspectors across other observable restaurant characteristics.

Figure A1: Independence of inspectors and boro's.

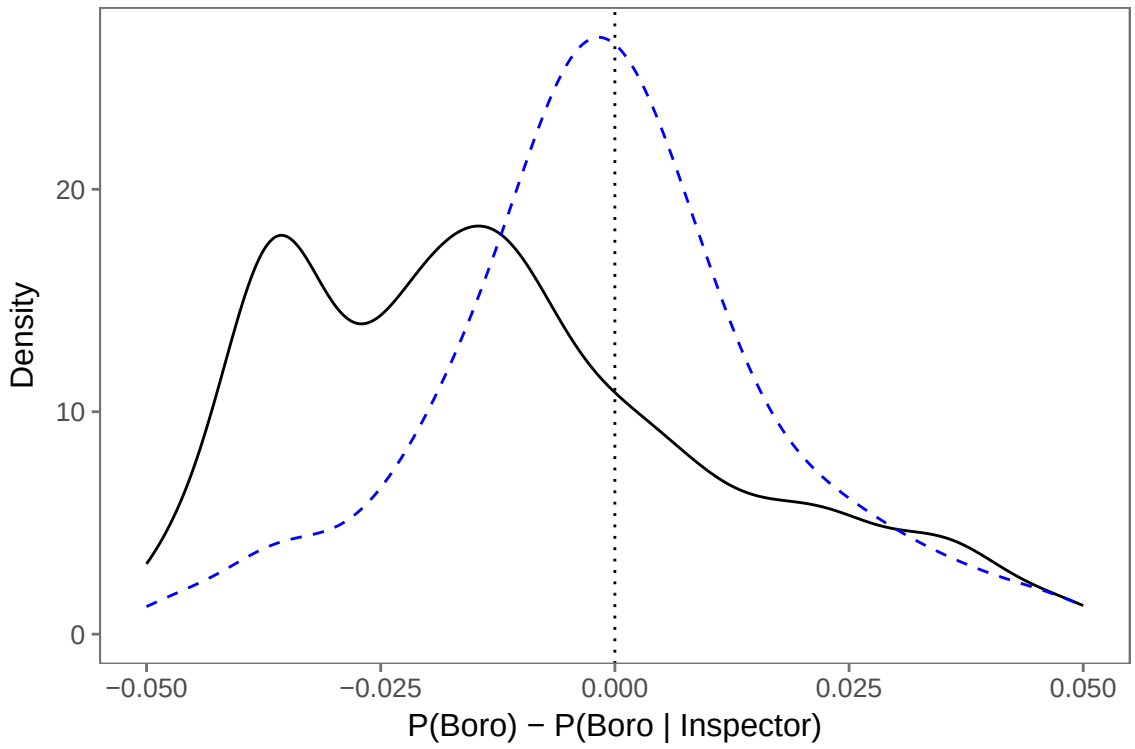


Figure A2: Independence of inspectors and venue.

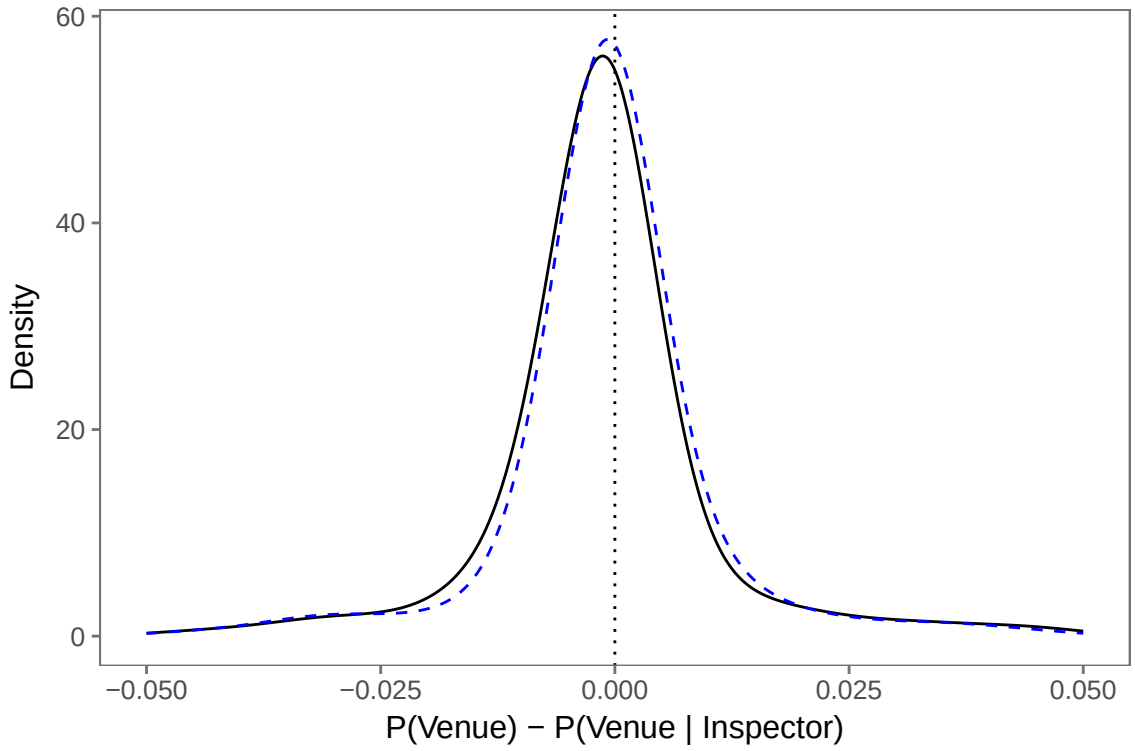


Figure A3: Independence of inspectors and service.

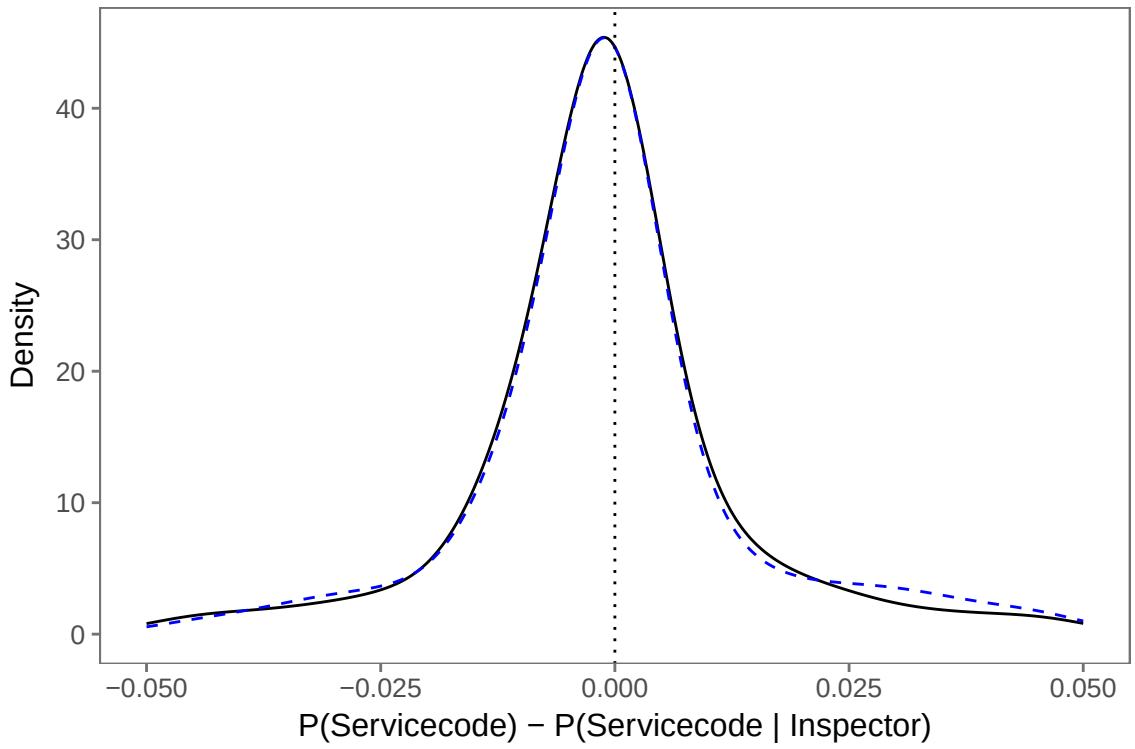


Figure A4: Independence of inspectors and cuisine.

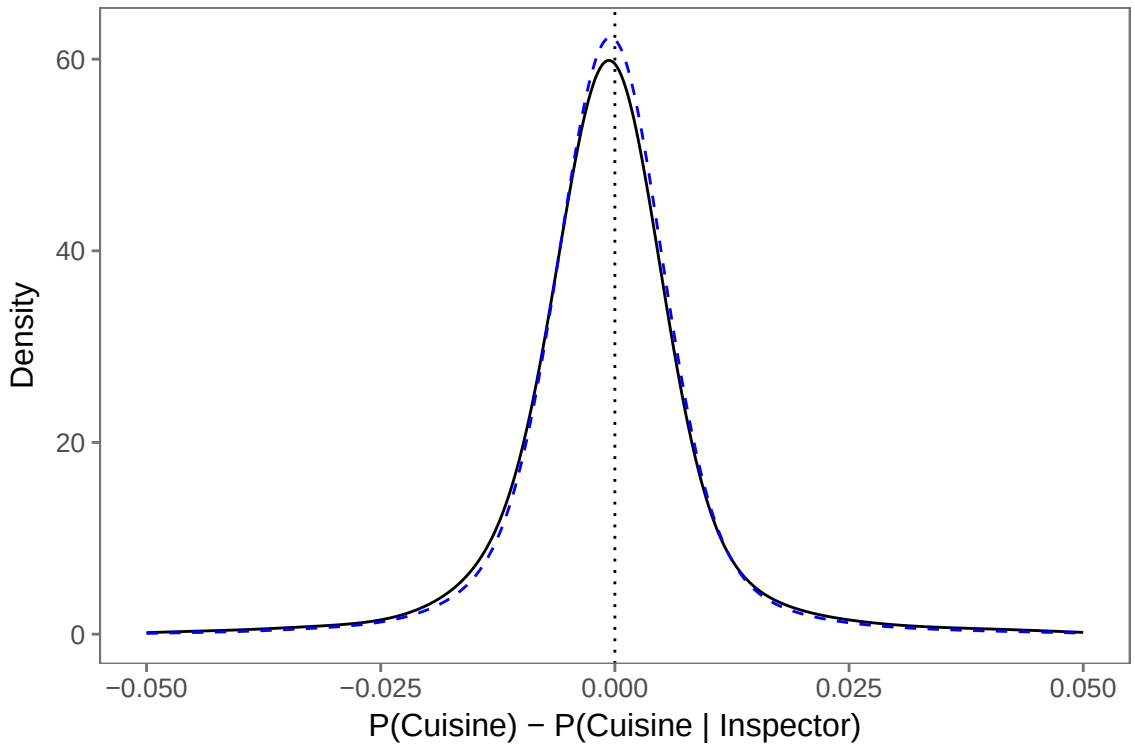


Figure A5: Independence of inspectors and zipcodes.

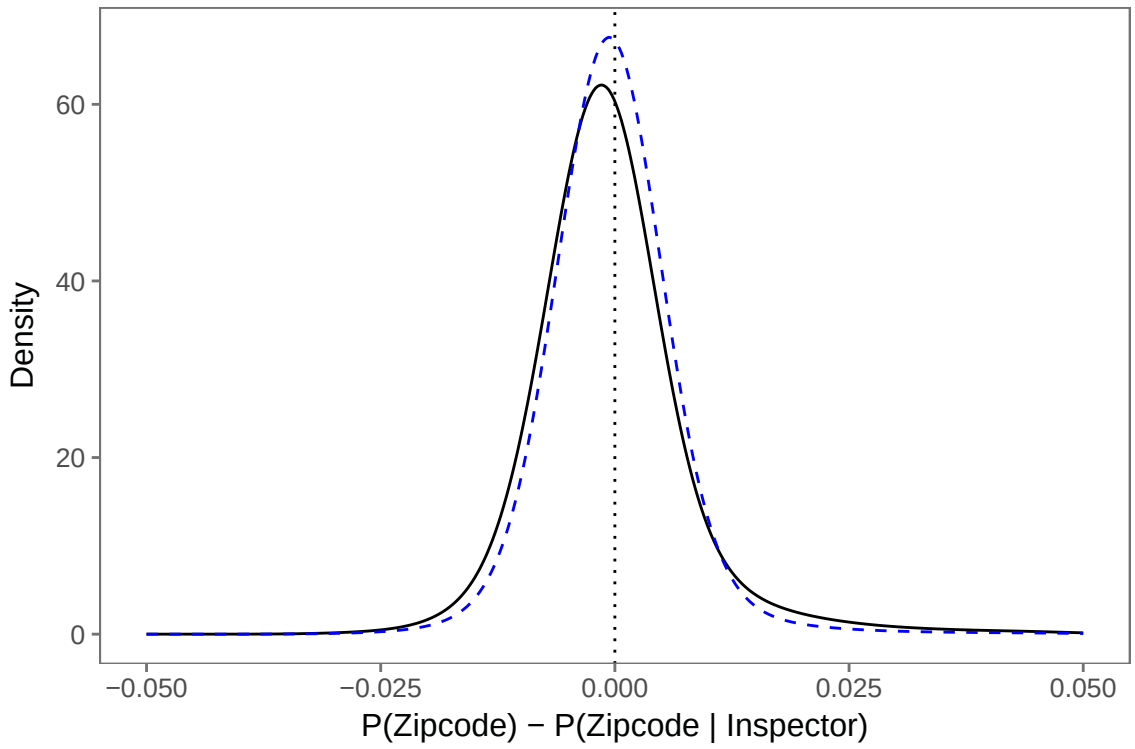


Figure A6: Independence of inspectors and zipcode-cuisines.

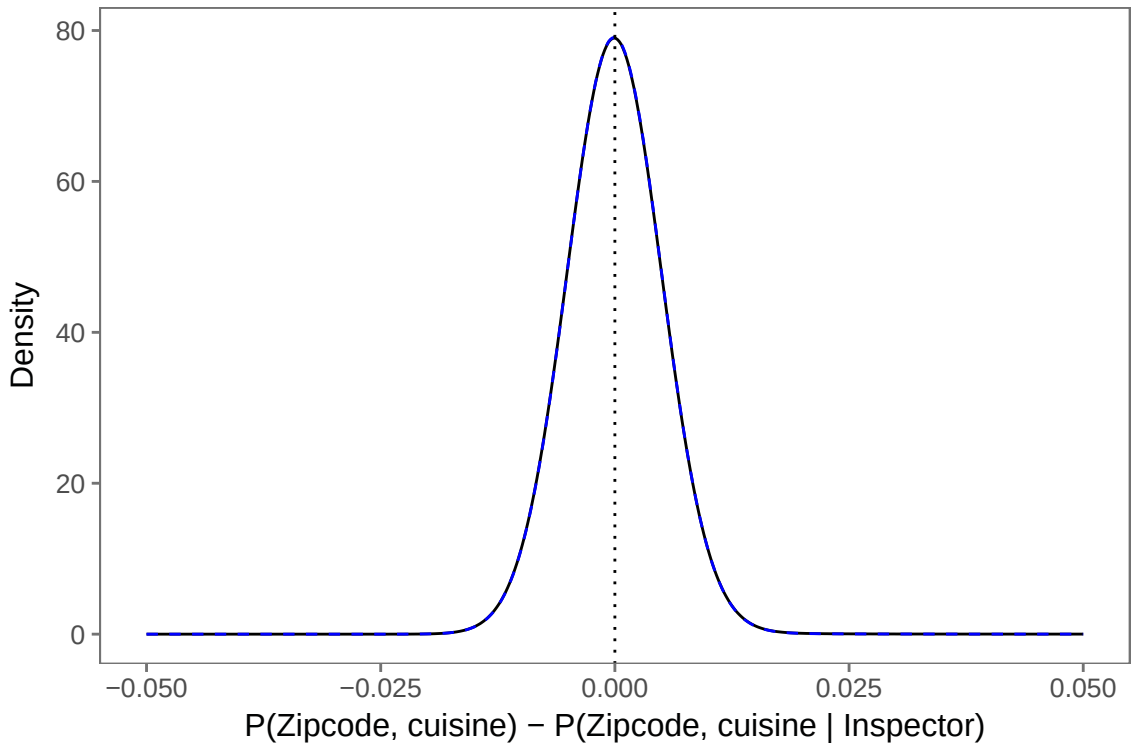


Figure A7: Independence of inspectors and camis.