A for Effort? Using the Crowd to Identify Moral Hazard in NYC Restaurant Hygiene Inspections

Jorge Mejia
Indiana University
Bloomington IN
jmmejia@iu.edu

Shawn Mankad
Cornell University
Ithaca NY
smankad@cornell.edu

Anandasivam Gopal
University of Maryland
College Park MD
agopal@rhsmith.umd.edu

Abstract

From an upset stomach to a life-threatening foodborne illness, getting sick is all too common after eating in restaurants. While health inspection programs are designed to protect consumers, such inspections typically occur at wide intervals of time, allowing restaurant hygiene to remain unmonitored in the interim periods. Information provided in online reviews may be effectively used in these interim periods to gauge restaurant hygiene. In this paper, we provide evidence for how information from online reviews of restaurants can be effectively used to identify cases of hygiene violations in restaurants, even after the restaurant has been inspected and certified. We use data from restaurant hygiene inspections in New York City from the launch of an inspection program from 2010 to 2016, and combine this data with online reviews for the same set of restaurants. Using supervised machine learning techniques, we then create a hygiene dictionary specifically crafted to identify hygiene-related concerns, and use it to identify systematic instances of moral hazard, wherein restaurants with positive hygiene inspection scores are seen to regress in their hygiene maintenance within 90 days of receiving the inspection scores. To the extent that social media provides some visibility into the hygiene practices of restaurants, we argue that the effects of information asymmetry that lead to moral hazard may be partially mitigated in this context. Based on our work, we also provide strategies for how cities and policy-makers may design effective restaurant inspection programs, through a combination of traditional inspections and the appropriate use of social media.

Key Words: Hygiene Inspections, Moral Hazard, Online Reviews, Machine Learning, Restaurants, Information Systems, Public Health,

Introduction

The Center for Disease Control estimates that one in six Americans gets sick each year as a result of foodborne illness (Centers for Disease Control 2016a). Approximately 60% of such cases are estimated to be the result of inadequately managed hygiene-related practices at restaurants (Gould et al. 2013, Hedberg et al. 2006). Understandably, strictly adhering to hygiene standards requires significant and constant investment by restaurants in terms of training, additional staff, and equipment (Jin and Leslie 2009). Beyond providing standards for food safety and hygiene, regulators and public policy makers also work on hygiene inspection programs to ensure and certify the hygiene practices followed by restaurants. The potential benefits of such programs are clear and have been empirically shown to decrease foodborne illness outbreaks (Irwin et al. 1989), particularly in cities where the hygiene scores are made public (Jin and Lee 2014, Jin and Leslie 2009). However, inspecting restaurants is a costly and time-consuming process, and real-time changes in hygiene quality are difficult to observe through infrequent inspections. Continuous monitoring is infeasible; New York City (NYC), for example, has over 20,000 restaurants, and inspections often take hours while incurring significant costs for the city. Regulators invest considerable effort, therefore, in designing inspection programs that balance out the costs of inspections and the risk of hygiene-related illness outbreaks. One such program, widely viewed as being highly successful, is the NYC Restaurant Inspection Program (RIP) program, spearheaded by the Bloomberg Administration in 2010 (Ho 2012, New York City Department of Health and Mental Hygiene 2016e).

The RIP was started in 2010 during Mayor's Bloomberg administration with three goals: give consumers easy access to information about the quality of hygiene of restaurants; improve restaurants' hygiene practices; and reduce the amount of restaurant-related foodborne illness. The innovative aspect of the RIP program was a two-step inspection process (Ho 2012, New York City Department of Health and Mental Hygiene 2016e), wherein each food service establishment, a category that includes restaurants, coffee shops, bars, nightclubs, and most cafeterias (New York City Department of Health and Mental Hygiene 2016e), was inspected twice in an inspection cycle and given the chance to correct any hygiene-related problems identified in the first inspection. As will be described in the next section, this approach was designed to quickly identify problems and fix them without imposing a heavy regulatory footprint on

the services industry (Ho 2012). Data released by the NYC Health Department indicates that the RIP has been highly effective. In terms of addressing restaurant hygiene practices, the number of restaurants that receive an A grade (the highest grade) at the end of an inspection cycle has increased significantly, from less than 30% in 2010, the first year of the program, to 42% in 2012 (Farley 2011), and 80% in 2014 (Farley 2016). More importantly, the rates of foodborne illness in NYC have declined significantly; according to the CDC, the city had over 2.1 million foodborne illness episodes in 2009 and approximately 1 million in 2014 (Centers for Disease Control 2016b). There has also been a 14% decrease in reported cases of Salmonella, one of the most dangerous foodborne bacteria, between 2010 and 2013 (New York City Department of Health and Mental Hygiene 2016c).

In this paper, we identify a classic moral hazard problem that if addressed can make the inspection program even more effective and create sustained improvements in hygiene practice. With any certification scheme, there exists the possibility that the certified entity will shirk on the required diligence needed to retain certification, once the certification process has been completed (Shapiro 1986). This is consistent with moral hazard, wherein firms or individuals change their behavior after acquiring risk protection or insurance (Hölmstrom 1979). In this particular context, it is therefore conceivable for restaurants who have acquired the highest hygiene certification to gradually discontinue their efforts to maintain their hygiene standards, given the knowledge that the next inspection cycle is more than a year away (New York City Department of Health and Mental Hygiene 2016b). The problem facing the regulating agency therefore is one of information asymmetry, i.e. the regulator does not know which restaurants are likely to shirk on their hygiene practices after receiving certification, since maintaining hygiene requires costly effort (Starbird 2005). If restaurants appear to bypass these efforts to maintain hygiene (whether intentionally or otherwise), the resulting risk of disease outbreaks in the city increase significantly. One solution to addressing moral hazard within the literature is to invest in information systems that provide continuous visibility and observability (Hölmstrom 1979). However, this would necessitate a system of flexible and continuous hygiene inspections, an option that is clearly economically infeasible. To identify instances of moral hazard, and to truly evaluate the effectiveness of the RIP program, we build on recent work that has argued that social media may help alleviate the problems of information asymmetry (Kang et al. 2013, Schomberg et al. 2016). Using a crowd-based "information system" created through the appropriate extraction of

information from online reviews, we show that incidents of moral hazard can be identified, thereby helping policy makers and regulators in their decision making.

Online reviews, provided by restaurant patrons and collated by firms like Yelp and TripAdvisor, have been used in previous research to capture service quality and feedback and link these to economic outcomes such as sales and purchasing behavior (Y. Lu et al. 2013, Park and Allen 2013). We view these as potential information sources about the restaurant's hygiene-related status that, when aggregated across multiple reviewers, can provide insight into whether the restaurant is shirking on its hygiene status. Effectively, we allow consumers to act as hygiene inspectors, albeit unintentionally and in the aggregate, by sharing their content online without any forward-looking expectations of restaurant efforts or influence from restaurants that may potentially bias their responses. The advantage here is that online reviews of restaurants, when utilized through appropriate statistical techniques, can help provide city regulators with the functionality of a "real-time" information system that can identify, first, those restaurants that are likely to be at risk of important hygiene violations even after receiving high hygiene grades (moral hazard), and second, identify those that are consistently diligent about their hygiene practices.

Specifically, we develop, test, and externally validate a social media-based hygiene dictionary that captures the observed counts of related words within the online reviews received by restaurants on an ongoing basis, using supervised machine learning techniques. Subsequently, we estimate statistical models that correlate the actual hygiene scores from inspections to word counts generated through the hygiene dictionary, to establish that the word counts are an accurate and reasonable proxy for hygiene scores. Having established this correlation, it is then possible to use word counts to gain visibility into restaurant behavior with respect to hygiene. We then use word counts to identify restaurants that display behavior consistent with moral hazard, as well as those that retain their hygiene status even after inspections. Our work extends recent work (Kang et al. 2013, Schomberg et al. 2016) in three distinct ways. First, rather than manually specifying the semantic terms that should be associated with disease outbreaks, we create our dictionary using a rigorous methodology that is based on the text present in online reviews. Second, rather than predict inspection grades per se, we focus on identifying shirking behavior. Hygiene inspections are complex activities requiring experience and training – we do not believe that actual inspections can be fully substituted with data from online reviews. Rather, we argue that social media data, when appropriately

used, can help improve the efficiency of such inspection programs by identifying non-compliance. Finally, we account for several sources of heterogeneity, such as restaurant characteristics, economic indicators, restaurant characteristics, and the specific format of the RIP, that help provide a more complete picture of hygiene inspections and moral hazard in this context.

Our analysis provides several insights relevant to the RIP program as well as the use of social media. First, as a baseline, we find that restaurants do indeed respond to the threat of regulatory action by significantly improving their hygiene scores; restaurants that score badly on their initial hygiene inspections are able to generate much higher scores on re-inspections within the same inspection cycle. Such a finding is important since it shows that hygiene problems are rarely structural and can be corrected by restaurants, given the right incentives. Equally importantly, it also suggests that maintaining hygiene is costly in terms of restaurant effort; on exerting such effort, hygiene improves. Second, based on the hygiene word counts, we find that roughly 30% of all restaurants in NYC deteriorate in terms of their hygiene into what would be considered appropriate for a lower grade (i.e. B or C grade) *within 90 days of certification*, i.e. behavior consistent with moral hazard. Effectively, we find that the effectiveness of the NYC RIP program is likely overstated, and that there continue to be significant public health risks in the city arising from hygiene violations. From a policy perspective, augmenting the hygiene inspections regime with information from online reviews would enhance the effectiveness of the RIP program, while also allowing it to be generalizable to other urban areas.

Our research contributes to the literature in several ways. First, we add to the small but influential literature in the restaurant sector that provides evidence for how social media can be used to tackle policy-oriented or socially relevant problems, thus responding to the call by Aral et al. (2013) to look beyond the for-profit sector in the utilization of online reviews. Second, our work adds to the small but growing literature that combines machine learning and economics (Athey 2015, Athey and Imbens 2006) to address social problems. Vast amounts of unstructured data exist in online repositories, including social media platforms, social networking sites, and review platforms – machine learning techniques can help extract suitable insights from these so that they can be used in economic modeling. We use this information to develop a dictionary that can help identify moral hazard with some accuracy. Third, we contribute to ongoing work in the public health domain, where tackling the risks of foodborne illnesses remains a hard

problem within large urban areas like New York City (Harrison et al. 2014). While inspection regimes and random checks are no doubt effective in general, there still remain inefficiencies in terms of the extent to which hygiene is maintained across the city. Our work provides a blueprint for how social media content can be used to identify gaps in the inspection regime and potentially improve outcomes, beyond simply providing a prediction model for hygiene inspections (Schomberg et al. 2016). As a means to this end, we also construct the first available hygiene-related dictionary based on a rigorous process of identifying related text, which can be used to identify potential offenders in the restaurant industry in general.

Since much of our work is closely linked to the structure of the NYC RIP program, we first describe this setting in some detail next. Subsequently, we briefly review the relevant literature in moral hazard, online reviews, and public health that we draw from, before describing the analysis conducted.

The NYC Restaurant Inspection Program

To operate a food service establishment in NYC, owners must have their facilities inspected and graded by the NYC Department of Health and Mental Hygiene (DoHMH) as part of the RIP program. Food service establishments are defined as fixed-site food vendors, a category that includes restaurants, coffee shops, bars, nightclubs, and most cafeterias (New York City Department of Health and Mental Hygiene 2016e). The program excludes mobile food vending units or temporary food service establishments, such as food trucks, correctional facilities, and charitable organizations. Every food service establishment receives an unannounced random inspection at least once every 365 days from DoHMH, which inspects over 20,000 establishments each year (New York City Department of Health and Mental Hygiene 2016b). The visit may take place anytime the establishment is open to the public or preparing food.

For each inspection, the inspector follows an established procedure to record in a mobile device the observed violations to the health code. Lower inspection scores show better adherence to the city's Health Code. Each violation is associated with a range of points, which depends on the type of violation and the risk it presents to the potential consumer. At the end of the inspection, the points are summed, and the total becomes the final inspection score, which is publicly available. Scores with 13 or fewer points, 14 to 27 points, and 28 or more points result in A, B, and C grades, respectively (see Figure A1 in Appendix A). The grade cards are required to be displayed within 5 feet of the entrance of the restaurant.

An important element that makes the RIP distinct from others in the U.S. is their two-step inspection process (Ho 2012, New York City Department of Health and Mental Hygiene 2016b). Each restaurant is inspected through an "inspection cycle", which begins with an initial inspection, but which can also include a re-inspection to ensure that any problems identified in the earlier inspection (within the same cycle) have been corrected. If the initial inspection yields an A hygiene score, the establishment posts an A sticker and is not subject to a re-inspection until the next inspection cycle (roughly 12 months later). However, if it receives a B or C score, the establishment is scheduled for re-inspection within the same "inspection cycle", to be completed within 3 months approximately. Meanwhile, the restaurant posts a "grade pending" or P sticker. The score generated from the re-inspection, which may be an A, B or C grade, must be posted immediately unless the restaurant requests a hearing at the NYC Health Administrative Tribunal. The underlying rationale is that restaurants can improve their hygiene processes by exerting the appropriate effort, and having achieved the required level of hygiene, such practices can be institutionalized within the restaurant so that there is no return to the earlier state of lower hygiene. Implicit in this model is the recognition that ensuring hygiene requires ongoing effort, which restaurants may be unwilling to invest in unless needed (Jin and Leslie 2003). Apart from these two-step inspection design, the program in NYC classifies hygiene violations as critical and non-critical, grades different conditions depending on the severity of violation, and is tasked with posting the public grades on the restaurant (for more details on the program details, please see Appendix A).

The NYC RIP is generally perceived as a success by the public (Farley 2011). Recent consumer surveys find that 89% of New Yorkers consider grades before dining out, 91% approve of publicizing grades, and 77% feel more confident dining in an A grade restaurant (Farley 2012). Furthermore, the program is credited with increasing the average level of hygiene within NYC restaurants. The number of A grade restaurants in the city has steadily increased, based on data released by the city, during the period of the program, as shown in Figure A2 in Appendix A. By 2015 almost 90% of all restaurants in NYC received A grades, ranging from 87.7% in Queens to 90.2% in Manhattan. These improvements are unrelated to local poverty rates, based on NYC Open Data (OpenData 2016), showing that the program is successful across the NYC area. However, while many restaurants eventually receive an A grade each inspection cycle, approximately 54% of such restaurants actually achieve initial inspection scores within an inspection cycle

that would have resulted in a B or C grade. Thus, at any point in time, there were roughly 4000 restaurants operating with a P grade (New York City Department of Health and Mental Hygiene 2016c).

In spite of the apparent success of the program, media reports indicate that the inspection program may not be as successful as the high compliance statistics would suggest.[1] Indeed, the high frequencies of A grades and the two-inspection regime allows restaurants with low hygiene practices to temporarily improve their hygiene levels until they are re-inspected, after which they can choose to not continue exerting this effort. Thus, given the protection offered by an A hygiene grade, such restaurants effectively undercut hygiene requirements and display risky behavior – by definition, this is behavior reflecting moral hazard (Holmstrom 1979). The negative effects of moral hazard in the context of hygiene inspections of food, produce, and service establishments has been studied in some detail in past work (Buchholz et al. 2002, Irwin et al. 1989). However, the availability of crowd-sourced data, in the form of online reviews, as well as text analysis methods provide the opportunity to test for and identify instances where such ex post shirking behavior can be isolated. Since this research is located at the intersection of information asymmetry, public health surveillance, and online reviews research in information systems, we briefly review these streams of research next before moving on to the research methodology.

Theoretical Background

Information Asymmetry and Moral Hazard in Restaurant Services

Markets for services are prone to significant information asymmetry, which result in adverse selection and moral hazard, i.e. it is hard to gauge the true quality of the counter-party ex ante as well as difficult to ensure compliant behavior ex post (Akerlof 1970, Hölmstrom 1979). These market limitations are particularly pressing in certain industries where vital information is needed for transactions to be conducted, verifiable information is hard to procure, and the costs of ex post renegotiation and opportunism are very high (Shavell 1979, Stiglitz 2002). Certain industry contexts are particularly prone to problems that arise from moral hazard in particular, such as insurance (Rubinstein and Yaari 1983) and healthcare (Gaynor et al. 2000). One particular context where moral hazard remains a critical problem, and is relevant to our work here, is in the context of public health surveillance of food safety (Starbird 2005). As food products travel through the supply chain from the source of production, there are various points where problems of food quality

[1] http://nypost.com/2014/04/13/city-restaurant-health-inspection-grades-a-shame-expert/

and hygiene can manifest; the traditional responses to these issues have been based on inspections, certifications, and labeling practices (Crespi and Marette 2001, Starbird and Amanor-Boadu 2007).

The final step in this supply chain is at the food service establishment (Angulo and Jones 2006), i.e. the restaurant. It is notable that significant information asymmetry exists at this point too, in terms of hygiene and food quality, with non-compliance potentially leading to the spread of food-borne illness. The CDC estimates that, as a result of foodborne illness, 3,000 Americans die each year (Centers for Disease Control 2016a), with approximately 60% of cases estimated to be a result of food prepared at restaurants (Gould et al. 2013, Hedberg et al. 2006). Public policy makers have reacted by launching programs to inspect and certify the hygiene of restaurants. The potential benefits of such programs are clear and have been empirically shown to decrease foodborne illness outbreaks, particularly in cities where the hygiene scores are publicized such as the NYC RIP program (Jin and Leslie 2003). However, inspecting restaurants is a costly and time-consuming process, and "real-time" changes in hygiene quality is difficult to observe through infrequent inspections. Thus, with any certification scheme with imperfect information on post-certification compliance (Shapiro 1986), there always remains the possibility of moral hazard and shirking.

Prior research has studied the issue of information asymmetry and moral hazard associated with restaurant hygiene inspections and letter grades. Jin and Leslie (2009) studied hygiene grades in Los Angeles County restaurants and found that franchises tended to free-ride on the hygiene reputations of the brand owner. However, reputational incentives, i.e. the need to retain reputation within local markets, served as motivators to maintain hygiene, especially since these scores were publicly available. In earlier work, the authors show that the number of foodborne illness hospitalizations decreased after the Los Angeles restaurant hygiene inspection program posted the hygiene letter grades to the public, again driven by local reputational incentives (Jin and Leslie 2003). Hygiene inspection programs pose a two-sided moral hazard problem – while restaurants may violate due process with respect to hygiene, it is also possible that inspectors may not report or detect all violations. Recent work indeed shows that inspectors that are new to the particular restaurant find between 12% and 17% more violations than repeat inspectors, potentially since new inspectors may have "fresh eyes" (Jin and Lee 2014). Thus, while previous research has established inspector heterogeneity can cause behavior that is consistent with moral hazard on behalf of the inspector, we study moral hazard and strategic behavior from the perspective of the restaurant. Finally, cultural and

social norms related to food preparation and quality may conflict with regulatory norms about hygiene; in such contexts, the importance of maintaining hygiene at the cost of authenticity may be too high for restaurants to bear (Lehman et al. 2014). While the mechanism here is not based on opportunism or shirking, the challenge of maintaining hygiene remains a problem for policy. It is here that recent developments in supervised machine learning, text analytics and big data technologies present opportunities for enhancing socially desirable outcomes, as described next.

Social Media and Machine Learning in Policy

Recent work in addressing policy problems has highlighted the coming together of two important trends – the access to large volumes of data on economic and social transactions through technology-enabled platforms ("big data") *and* the concurrent development of techniques for extracting insight from these data sources (Athey 2017). In particular, the use of statistical and machine learning techniques to study economic and policy problems has become increasingly effective, since these methods address key technical challenges that arise with big data (Einav and Levin 2014, Kleinberg et al. 2015, Varian 2014). For instance, it is common with big data to encounter high-dimensional (i.e. having many more variables than observations) or unstructured data, which challenges traditional methodology. However, statistical and machine learning methods are often developed from the perspective of prediction instead of causal inference, leading to the observation that "economists have not immediately shifted to new statistical approaches, despite changes in data availability" (Einav and Levin 2014).

These problems notwithstanding, recent work has harnessed social media data and machine learning techniques in trying to address the issue of public health surveillance as a way to reduce food-borne illness more broadly (Chan et al. 2010), and restaurant-related hygiene issues, more narrowly. Because social media often contains information not captured through traditional channels, they are often useful to public health agencies for surveillance activities (Brownstein et al. 2009). HealthMap, for example, is an openly available public health surveillance system that includes data from disparate sources, such as public health officials, clinicians, and travelers, to produce a global view of continuing disease threats (HealthMap 2017). The CDC has a similar program (*Foodborne Outbreak Online Database*) to make disease outbreak surveillance system data available to the public. In the specific case of restaurant hygiene, Harrison et al. (2014) conducted a study where foodborne disease epidemiologists analyzed Yelp

restaurant reviews. They found that many of the reviews posted online were consistent with a foodborne illness episode. However, only 3% of these episodes were reported to city authorities, thus demonstrating the promise of social media data, particularly online reviews, in successfully identifying foodborne public health outbreaks where the official reporting mechanism to public health agencies is ineffective.

The closest analogs to our work here are Kang et al. (2013) and Schomberg et al. (2016). Kang et al. (2013) use a set of data from Yelp reviews of restaurants in Seattle to predict the hygiene scores received by the same restaurants. They use a diverse set of input variables, including review volume and rating, restaurant meta-data, as well as lexical cues (selected set of unigrams and bigrams) and show good accuracy in the prediction task using these variables. In a similar exercise, Schomberg et al. (2016) develop prediction models for hygiene scores of the most obvious offenders in San Francisco, based on previous history, using a combination of restaurant meta-data (referred to as tags) and keywords in their model. In contrast to Kang et al. (2013), the specific keywords in Schomberg et al. (2016) are generated from a manual perusal of the major health code violations associated with inspection process. Interestingly, Schomberg et al. (2016) used only the top-ranked reviews for each restaurant in the model and focused on "high risk populations" of restaurants (p.6). In order to ease interpretation, the keywords were further condensed into three dimensions using principal components analysis – *Reviewer Sentiment, Physical Environment and Vermin, and Foodborne Illness Related Symptoms*. The overall objective in both papers is to predict hygiene scores such that regulators use the models to help identify restaurants where hygiene is problematic, allowing further remedial action.

Our work here differs from these papers, both methodologically and conceptually, in important ways. First, we do not develop our own keywords to be used in extracting information manually – rather, we allow text data on reviews over multiple years, and a classifier for identifying hygiene-related issues, to generate the appropriate dictionary. Furthermore, we use all reviews on all restaurants to do so, thereby mitigating any potential biases that may emerge from the high-risk populations. To the extent that we consider a census of restaurants in NYC, our approach is more applicable for regulators who bear responsibility for the hygiene of all restaurants. Second, we contend that the prediction of hygiene scores is only the first step in using social media data. The inspection of restaurants is a complex task performed by trained inspectors, and extends beyond what may be visible to the consumer, such as the state of the

kitchen and food preparation processes. Therefore, social media is better used to not predict hygiene scores per se but to identify non-compliant behavior, such as with moral hazard, that can then be remedied fully through an inspection. Such non-compliant behavior can be exhibited not only by repeat offenders but also by capable restaurateurs who may experience one-off hygiene problems. Therefore, online review data may be used to stand alone as an independent "information system" that generates information to be used in conjunction with actual inspections. This rationale implies that all restaurants, including those with no history of violations, should be evaluated using a dictionary-based approach. Thus, conceptually, rather than better prediction power, we argue that social media data should help lead to a more *effective* inspection process by filling the gaps that are left behind by the process – our approach here is based on this rationale.

The approach we take to studying moral hazard builds on the large body of research in IS that has proposed econometric models linking online review data to economic outcomes, such as sales and revenues. We use supervised learning techniques to transform the high dimensional and unstructured data from the text in online reviews into an interpretable independent variable for input into a classical regression setting where we model the inspection results. We retain a level of parsimoniousness in our models since inspection score prediction is not the sole objective here, as much as addressing a pressing economic problem. Furthermore, the relative simplicity of our approach ensures that the process can be easily automated as new data from reviews and hygiene inspections flow in. As we build our models on prior IS research studying online reviews in retail settings, we briefly review this literature next.

Online Reviews as Sources of Text Data

Online reviews have long since been accepted as a strategic resource for firms in the services sector, capturing elements of quality, reputation, consumer feedback, and market power (Dellarocas 2003, Duan et al. 2008, Li and Hitt 2008). In more recent research, online reviews have also been associated with a host of offline business outcomes such as offline revenues (Duan et al. 2008), firm survival (Mejia et al. 2015) and even critical decisions such as the choice of physicians (Gao et al. 2012). Viewed as a data source, online reviews typically contain multiple pieces of information that can be used by consumers. First, the aggregate or individual rating provided by reviewers, typically on a 1-5 star(s) scale, is available. Researchers have also used measures built on the flow of reviews, representing traffic or interest in the product or service being reviewed, such as the number of reviews provided in each time-period (Dellarocas

and Narayan 2006). Extant literature has used these two sources of variation extensively in evaluating the firm-level impacts of online reviews. While a comprehensive review of the online reviews literature is outside of the scope of this study but is available (Cheung and Thadani 2012), we briefly review the more recent (and more relevant to our study) work extracting information from the actual text of online reviews.

Despite representing a collective corpus of opinions, judgments, evaluations and suggestions from consumers that are of considerable utility, the text of online reviews has often been overlooked, with some notable exceptions. The first efforts to study text were based on explaining firm performance. This stream of research has focused on understanding the effect of the review text on firm performance outcomes (i.e. sales, product rankings, and revenues) using the number of repeated words on a review (Berger et al. 2010) or content and sentiment (Dellarocas et al. 2007, Ghose et al. 2012, Godes and Mayzlin 2004, Ludwig et al. 2013). Alternatively, others have focused on inferring product, brand, and market-level attributes from the corpus of reviews. Decker and Trusov (2010), for example, suggest a methodology to enable the estimation of the relative effect of product attributes and brand names on the overall evaluation of products by using text analysis to classify the pros and cons of each product while Netzer et al. (2012) propose a combination of text mining and semantic network analysis to understand consumers' associative network of products and understand the implied market structure. Archak et al. (2011) are at the intersection of these streams of research and attempt to infer the economic value of online reviews by identifying the weight that consumers put on individual evaluations and product features, as well as estimating the overall impact of the product features discussed in the review text on sales. Moreover, a related research stream in this literature relates to understanding the features of the review text that increase the helpfulness or maximize the impact of reading a review on consumers (Cao et al. 2011, Ghose and Ipeirotis 2011, Lee and BradLow 2011, Zhao et al. 2012). Research focused on extracting insight from the text of online reviews is still relatively new to the literature but it is increasingly accepted that the text represents a vital and credible source of information for the service provider (Cao et al. 2011) that may even reduce bias in online reviews platforms (Lee et al. 2015). Building on this work, we first use the review text to create a hygiene dictionary, and second, use this dictionary in order to identify cases of moral hazard.

Research Methodology

Dataset and Exploratory Analysis

We start by constructing a composite dataset from two sources. First, data on NYC RIP was gathered through the NYC Open Data program (NYCOD), which makes public data generated by various NYC agencies available for public use (OpenData 2016). The dataset contains information on all NYC restaurant inspections since the beginning of the program in 2010 and is updated monthly. Our data sample includes restaurant-specific information, such as the restaurant's address, phone number, and cuisine; inspection-specific information, such as the inspection date, type, resulting grade, and inspector id[2]; and violation-specific information, such as the violation code, type, severity, and points.

We supplement the NYC Open Data information on each restaurant with online review data from Yelp.com, the leading online review platform for restaurants. Restaurants in the two datasets were matched using the restaurant address and telephone number, with over 95% matching between the two datasets. The data in Yelp contains more detailed restaurant characteristics, such as price point, average consumer rating, hours, and payment options. In addition, the reviews are time-stamped, thereby allowing us to partition the review data by time over the period of analysis. While the hygiene inspection data starts in 2010, when the new inspection regime was implemented, we collect review data for NYC restaurants from their first occurrence on the platform, dating back to 2004. The variable descriptions for the composite dataset are shown in Table 1. The dataset includes 24,625 restaurants in total across the five NYC boroughs, while the reviews dataset contains approximately 1.3 million individual reviews. We exclude reviews that are flagged as fraudulent by Yelp.com. Yelp has its own algorithm for identifying fake reviews, which we implicitly use, and which has also been used to conduct analysis on fraudulent reviews (Luca and Zervas 2016).

We observe, as DOHM and several media outlets have asserted (Barron 2015, Park 2015), that approximately 90% of restaurants in the city achieve an A grade in hygiene inspections. Recall that the two-step inspection approach allows restaurants scoring in the B or C range to first post a P grade, then be assigned a final grade after re-inspection within the same cycle. Therefore, some "A" restaurants achieve an A grade on the initial inspection, while others first receive a P and then an A on re-inspection. Table 2

[2] The de-identified inspector id is not available in the current public dataset on NYC Open Data (as of 1/8/2017). The data also only goes back to 2012 instead of 2010. Attempts to reinstate inspector id and 2010-2012 data were denied by NYC Open Data.

presents a distribution of grades from all inspections cycles in our dataset, and shows that these two subgroups, which we call "A" and "PA", represent 41.0% and 48.7% of the data, respectively. Since they constitute the majority of the data, and are used as an indicator that the NYC RIP is working successfully, they are the focus of our analysis.

We first examine the differences in initial hygiene inspection scores received by these two groups. This baseline analysis is useful in establishing the magnitude of hygiene improvement that is within the capacity of "PA" restaurants. Such improvements demonstrate that hygiene violations are not structural to the restaurant, but can be rectified through effort and investment. This is an important assumption in establishing moral hazard, since it implies the existence of agency on the part of the restaurateur (Jin and Leslie 2003). As expected, we see a substantial difference in initial inspection scores between the "A" and the "PA" subgroups across all inspection cycles ($Mean_A = 8.71$; $Mean_{PA} = 21.51$; $p<0.001$; lower scores represent better hygiene). Notably, the median score for "PA" restaurants is 26.5, near the threshold for C grade (28 points). Yet, upon re-inspection these restaurants achieve scores that are statistically indistinguishable from "A" subgroup restaurants ($Mean_A = 8.71$; $Mean_{PA} = 9.18$; $p=0.76$). We also estimate the changes in hygiene scores before and after receiving a P grade using a longitudinal differences-in-differences model (Bertrand et al. 2002). This model is presented in Appendix B and shows that the opportunity for re-inspection results in an estimated reduction of 6.23 points in hygiene scores, after controlling for restaurant and inspection characteristics. These restaurants are therefore able to achieve high levels of hygiene quality despite significant violations in their initial inspections. This suggests that hygiene inspection grades are important enough to incentivize remedial action on the part of restaurants, and that such action is indeed feasible (Jin and Leslie 2009).

After such improvements are made by "PA" restaurants between initial and re-inspection, does their hygiene quality remain high over time, or do they revert to their previous practices? The RIP dataset provides longitudinal information on NYC restaurants over multiple inspection cycles, allowing us to empirically examine this question. Indeed, we observe a clear dichotomy between "A" and "PA" patterns over *multiple* inspection cycles. To illustrate this, we focus on restaurants in the "A" and "PA" subgroups in the second inspection cycle and observe their performance in their third inspection cycle. We avoid using the first inspection cycle, since for many restaurants it occurred during the early, evolving days of the

program and may also represent learning on the part of the restaurant. Looking at later cycles would exclude many newer restaurants that have undergone fewer inspections. Among those 70.3% of restaurants that have undergone at least three inspection cycles, 85.5% of restaurants that achieve an "A" on initial inspection in the second cycle do so again in the third cycle; we refer to this subgroup as the "AA" subset. In contrast, 90.4% of restaurants that exhibit the "PA" pattern in the second cycle do so again in next cycle; we refer to this group as the "PAPA" subset. Therefore, for the majority of "A" restaurants, the focus on maintaining hygiene remains in place even between inspection cycles, while the majority of "PA" restaurants appear to revert in terms of their hygiene to a lower level at some point in the period between inspection cycles. The distribution of third inspection cycle grades for those restaurants achieving an "A" in the second inspection cycle is displayed in Appendix Table A1.

Figure 1 displays the hygiene inspection scores for all restaurants in the "AA" and "PAPA" subgroups across the second and third inspection cycles. Figure 2 shows their mean initial and re-inspection scores at each cycle. The behavior of restaurants in the "PAPA" subgroup illustrated in these figures is indicative of *moral hazard*, i.e. their *ex post* behavior is not compliant with hygiene expectations from the awarded A grade. The dramatic difference between re-inspection scores and initial inspection scores of the next cycle for the "PAPA" subgroup shows the potential social cost associated with moral hazard (Stiglitz 2010). While these restaurants post a grade of "A" for approximately one year after re-inspection, their hygiene practices are not consistent with that grade for at least some of that time. We observe similar patterns between future sequential inspection cycles. While the high cost of monitoring for hygiene (Akerlof 1970, Starbird 2005) prevents continuous monitoring through inspections, an "information system" providing visibility into *ex post* behavior between inspection cycles can help improve efficiency (Hölmstrom 1979). As described in the next section, we propose utilizing text from online reviews to construct a hygiene-related dictionary that can provide such visibility.

Creating a Social Media Sourced Hygiene Dictionary

Online data sources include many different kinds of information pertaining to service quality and firm performance (Abbasi and Chen 2008, Dellarocas 2003, Tetlock et al. 2008). Issues that relate to hygiene may be included in online reviews in various forms, but extracting and validating such information is not trivial (Schomberg et al. 2016). One option here is to manually use a small set of lexical cues that may

signal when a hygiene-related issue is present at a restaurant, similar to the approach taken by Schomberg et al. (2016) and Kang et al. (2013). However, as the volume of reviews increases and more types of food-borne diseases emerge from restaurants, access to larger data volumes allows the use of relatively unsupervised techniques for extracting lexical cues and key words of relevance. We adopt such an approach to construct a Social Media Sourced Hygiene (hereafter *SMASH*) Dictionary from online review text, consistent with best practice in dictionary construction for use in policy making (Liu 2015, Loughran and Mcdonald 2011, Young and Soroka 2012).

The primary process behind creating the SMASH dictionary is to use a thesaurus to repeatedly augment an initial seed list of hygiene words based on synonym relationships. We perform this in a series of steps. First, we identify an initial list of negative and hygiene-related words using the Naïve Bayes Classifier on the text of online reviews.[3] In order to reduce the possibility of bias and allay any concerns about endogeneity of word usage in the reviews, we only use online reviews that were provided *before* the onset of the inspections program that provided publicly visible hygiene grades to create the seed list, i.e. prior to 2010. Secondly, we recruit subjects in Amazon Mechanical Turk outside of the state of New York (again to minimize potential confounders) to perform the labeling of the reviews. Finally, the initial seed list of hygiene words is repeatedly augmented based on synonym relationships in WordNet (Hornik et al. 2016, Miller 1995), a popular online dictionary and thesaurus. The process is repeated until no newer synonyms can be found. The dictionary is then manually curated to remove clear errors due to homonyms, ensuring no errors propagate in the validation process. For example, *roach* is short for *cockroach* (relevant) and synonymous with *Mexican valium* (not relevant). We describe these steps in detail.

Building the List of Seed Words with Naïve Bayes Classifier

Our procedure for creating the SMASH dictionary is consistent with best practices in the sentiment and opinion mining literature (Feldman 2013, Liu 2015, Tsai et al. 2013) with one important extension. In sentiment mining, the initial set of seed words are often manually specified ((Hu and Liu 2004, Valitutti et al. 2004). However, manually specifying words for the hygiene context is much less obvious compared to tonal sentiment, given the wide range of issues that may show up in restaurants, and may be unduly

[3] In principle, there are many supervised learning methods to identify the initial list of negative words. A benchmarking study described in Appendix C shows the Naïve Bayes classifier to be the most accurate method.

influenced by the researcher's own vocabulary or lack thereof. Indeed, poorly created seed lists lead to less accurate dictionaries (Tang et al. 2009, Wallach et al. 2009). Therefore, to mitigate this potential bias, we generate the initial word list through the Naïve Bayes classifier, a simple yet effective machine learning technique to identify the initial word list in a data-driven manner (Li 2010, Liu 2015).

We first introduce some notation that will help facilitate discussion of the Naïve Bayes classifier (Hand et al. 2001, Tang et al. 2009). Suppose we are given a training dataset with n documents that are labeled by their hygiene polarity, or d_j, defined as a binary variable denoting whether a review (document) $j = 1, \dots, n$ is discussing hygiene negatively. Let w_{jk} denote the number of times word $k = 1, \dots, p$ occurs in review j, where p is the number of unique words appearing in all reviews. The Naïve Bayes classifier estimates the probability that each document discusses hygiene negatively based on the word occurrences, i.e., $P(d_j = 1|w_{j1}, w_{j2}, \dots, w_{jp})$. Through an application of Bayes Rule,

$$P\left(d_j = 1|w_{j1}, w_{j2}, \dots, w_{jp}\right) = \frac{P\left(w_{j1}, w_{j2}, \dots, w_{jp}\middle|d_j = 1\right)P\left(d_j = 1\right)}{P\left(w_{j1}, w_{j2}, \dots, w_{jp}\right)} \quad (1)$$

However, in practice calculating the joint distributions requires an unrealistically large amount of data (Hand et al. 2001). To overcome this issue, the joint distribution is simplified under the assumption of conditional independence,

$$P\left(w_{j1}, w_{j2}, \dots, w_{jp}\middle|d_j = 1\right) = \prod_{k=1}^{p} P\left(w_{jk}\middle|d_j = 1\right) \quad (2)$$

By the law of total probability (Ross 1996), the probability that a document discusses hygiene negatively based on the word occurrences can be expressed as

$$P\left(d_j = 1|w_{j1}, w_{j2}, \dots, w_{jp}\right) = \frac{\prod_{k=1}^{p} P\left(w_{jk}\middle|d_i = 1\right)P\left(d_j = 1\right)}{\prod_{k=1}^{p} P\left(w_{jk}\middle|d_i = 1\right)P\left(d_j = 1\right) + \prod_{k=1}^{p} P\left(w_{jk}\middle|d_i = 0\right)P\left(d_j = 0\right)}, \quad (3)$$

where $P(w_{jk}|d_j = 1)$ and $P(w_{jk}|d_j = 0)$ can easily be calculated by inspecting how often the kth word appears in documents that are labeled as discussing hygiene negatively ($d_j = 1$) or those that are not ($d_j = 0$) (Hand et al. 2001). Thus, the conditional independence assumption is a fundamental one that defines the Naïve Bayes classifier. Though from a probabilistic perspective the conditional independence assumption is usually not strictly valid, the performance of Naïve Bayes in various machine learning contexts, particularly with text data, has been widely demonstrated (Go et al. 2009, McCallum et al. 1998,

Sebastiani 2002). We explore this issue in detail with a benchmarking study comparing the Naïve Bayes classifier to other machine learning methods, described in Appendix C. We find that the Naïve Bayes classifier is the most accurate on our dataset among competing machine learning methods for text classification. After estimating the Naïve Bayes classifier (described in the next section), we sort all words by their estimated $P(w_{jk}|d_j = 1)$ and keep the *top 5%* as the initial seed list to ensure that seed words chosen are strongly associated with negative discussion of hygiene in the online reviews. The selection of the 5% threshold is based on best practice and is recommended in large datasets to maintain tractability (Loughran and Mcdonald 2011). Words outside of the 5% level are typically phrases composed of root words that make up the seed list.

Obtaining Training Data and Implementation

We begin with a dataset containing the text of online reviews for NYC restaurants from Yelp, matched to those in the inspections dataset based on name, address and phone number, yielding a composite dataset with inspections data, restaurant meta-data, and reviews. We perform standard preprocessing of all review text, such as moving to lower case, removing stopwords and stemming (Hornik et al. 2016). In order to identify documents in which hygiene is discussed negatively (i.e. lower hygiene ratings from consumers), we first randomly sample 1,200 restaurants with high inspection scores, which are most likely to have hygiene problems, and then randomly select one document (review) for each restaurant. The next task is to manually assign a label regarding the hygiene relation of each document, for which we recruited 1,200 subjects from Amazon Mechanical Turk (MTurk) for pay. MTurk subjects have been shown to represent the broader population (Buhrmester et al. 2011), be reliable for experimental research (Goodman et al. 2013), and to generate high quality data (Ipeirotis 2010). Each subject was asked, "*Given a restaurant review, answer questions about whether a review indicates problems related to hygiene*" and was then presented a single document (review). After reading the review, the subject indicated whether it was related to the hygiene of the restaurant using a 7-point scale adapted from Egan et al. (Egan et al. 2007). Subjects were also asked to select the type(s) of hygiene problem described (such as food preparation or cleanliness). Finally, to ensure high quality responses, we discarded responses from subjects that took a large amount of time to finish to the task (more than 15 minutes), which has been shown to decrease subject's attention span

and the quality of responses (Buhrmester et al. 2011). Of 1,200 documents sampled, we received clean responses for 1,191 documents, of which 15% were labeled as relating to restaurant hygiene ($d_j = 1$).

These 1,191 training documents, their MTurk-generated labels d_j, and the word occurrences w_{jk} allow us to calculate $P(w_{jk}|d_j = 1)$, which forms the basis of the seed word list for SMASH. By expanding the seed list using the synonym approach described above, we created a hygiene dictionary based on social media. The entire process was performed using single words, two-word phrases, and three-word phrases, i.e., *n-grams of up to order three* (Lodhi et al. 2002), so that the final dictionary contains single words along with two- and three-word phrases. For example, as a result of using n-grams, phrases like "barely edible" were retained in our dictionary even though the individual words "barely" and "edible" were not included. This strategy is recommended in recent work in this literature, advising authors to move beyond word-level analysis, since this tends to oversimplify language (Cambria et al. 2013). Therefore, we extend the bulk of existing work that relies only on dictionaries based on individual words (see (Liu 2015) and references therein). Further, compared to the set of keywords in Schomberg et al. (2016), our dictionary generated a richer and more varied set of terms associated with hygiene. While certain words are common, such as "vomit", "diarrhea", and "nausea", our dictionary also included several phrases that, on first glance, may not appear relevant if manual curating of words associated with food-borne diseases were utilized, such as "pungency" and "wiping nose". For simplicity, we refer to all phrases as "words".

A subset of words in the final SMASH dictionary is shown in Table 3. We choose to not display all the words from our dictionary due to space limitations and to avoid the strategic use of these words in reviews to create false or fraudulent impressions of hygiene concerns (Luca and Zervas 2016). Indeed, the utilization of the dictionary to identify instances of hygiene violations relies on keeping the specific words and phrases within the dictionary private so that restaurants or individuals can rely on the unbiased use of these words and phrases in reviews. Further refinements to the dictionary as well as systematically re-validating it at regular time intervals would be desirable to ensure that it is not used in fraudulent ways.[4]

Once the directory has been assembled, we can use it to assign measures to each restaurant based on its reviews. For a given document (review), the SMASH score is defined as the total number of times

[4] While the full dictionary is not shown here, we are happy to provide it to facilitate journal review if requested. We stress the need to retain confidentiality of the dictionary for its effectiveness.

words from the SMASH dictionary appears in the document. Letting $\delta_k = 1$ if word k is included in the SMASH dictionary, 0 otherwise, the SMASH score for document j is $\text{WC}_j = \sum_k w_{jk}\delta_k$, where w_{jk} is the number of times word k appears in document j. The SMASH scores for a particular restaurant and time period by summing over the reviews published in that time period for that restaurant for the purposes of longitudinal econometric models, which are described next.

Econometric Analysis

We present our econometric analyses in two stages. First, we validate the dictionary by estimating the correlation between the word counts and actual observed inspection scores; this external validation of the dictionary is needed to ensure that it can indeed be used as a reasonable proxy for hygiene quality. We also perform several robustness tests to ensure that the relationship between word counts and hygiene scores is reliable. Second, we use the word counts to detect instances of moral hazard by identifying restaurants that appear to have regressed in their hygiene quality after posting a grade.

Validating the SMASH Dictionary

To validate the SMASH dictionary, we use a longitudinal mixed effects regression model (Liang and Zeger 1986) to estimate the association between SMASH word counts and hygiene inspection scores. Since most restaurants in our sample are inspected multiple times over the study period, our unit of analysis is restaurant-inspection. For each restaurant and inspection, we use the SMASH score for that restaurant in the time period immediately *preceding* the inspection to predict the ultimate inspection score, using month as the unit of time.[5] We lag the SMASH scores in the model to guarantee that all reviews contributing to the SMASH score occurred prior to inspection. This is done to minimize the risk of reverse causality, which may occur when reviewers are influenced the new inspection letter grade posted at the restaurant.

Let t index month-year time periods, and let $INS_{i,t}$ be the numerical hygiene inspection score for restaurant i, if that restaurant was inspected during time period t. If such an inspection is a re-inspection within an inspection cycle, $ReIns_{i,t}$ equals 1 and equals 0 otherwise. Let $WC_{i,t-1}$ be the SMASH score for restaurant i based on reviews published on Yelp during time period $t-1$, the primary independent variable of interest. We estimate the following model:

[5] This avoids the possibility of inspections occurring for the same restaurant in subsequent time periods, since for initial inspections resulting in a "P" grade, re-inspections occur on average 54 days (standard deviation 11 days) later, and in our sample always occur at least one month later.

$$INS_{i,t} = \beta_0 + \beta_1 WC_{i,t} + \beta_2 Rating_{i,t-1} + \beta_3 Reviews_{i,t-1} + \beta_4 ReIns_{i,t} + \boldsymbol{\gamma R}_i + \boldsymbol{\delta I}_{i,t} + \alpha t + \varepsilon_{i,t} \quad (4)$$

$Rating_{i,t-1}$ and $Reviews_{i,t-1}$ are the average numerical rating and number of reviews published on Yelp for restaurant i during time period $t-1$. $\boldsymbol{R}_i$ and $\boldsymbol{\gamma}$ are restaurant-level indicator variables and fixed effects, respectively, which we include to capture heterogeneity among restaurants[6]; similarly, $\boldsymbol{I}_{i,t}$ and $\boldsymbol{\delta}$ are inspector-level indicator variables and fixed effects, respectively, included to account for the influence of individual inspectors on inspection scores (Jin and Lee 2014). Finally, we account for linear time trends.[7] For model estimation, we exclude restaurants that have undergone only one inspection, as well as inspections with less than five Yelp reviews for the same restaurant in the previous time period. The final dataset contains 21,488 unique restaurants with 136,503 inspections over a 5-year period.

We first estimate a base model (Table 4: Model 1), assuming a linear relationship between hygiene word counts and inspection results, controlling for a series of variables from prior research (Kang et al. 2013). We observe that the coefficient for $Rating_{it}$ is negative and significant (-0.36, $p<0.01$), indicating that negative online reviews are associated with higher (worse) inspection scores, as found previously by Schomberg et al. (2016). More pertinent to our analysis, the coefficient for SMASH word count is positive and significant (0.59, $p<0.01$), indicating that more mentions of hygiene-related words in recently published reviews is associated with higher inspection scores, even after controlling for the average rating and number of reviews, as well as restaurant-level fixed effects, inspector-level fixed effects, and linear time trends. This establishes that the proposed hygiene dictionary is significantly associated with the "ground truth" of hygiene inspections, in spite of the limited access that consumers have into many aspects of a restaurant's operations relative to inspectors.

Since hygiene inspection scores range from 0 and 90, we also estimate a Tobit regression of the model in equation (4) to account for truncation in the inspection scores. The results (Table 4: Model 2) show the linear association between SMASH word counts and hygiene inspections to be slightly higher (0.65, $p<0.05$) than estimated by OLS. These results provide clear evidence of a positive relationship between SMASH word counts and hygiene scores.

[6] We do not include time-invariant restaurant-level characteristics, since restaurant-level fixed effects are included in the model. To account for heteroscedasticity, we report robust Huber-White standard errors.
[7] We checked for seasonality and non-linear trends across the years of the study and found no evidence for either.

Note that we do not postulate that hygiene inspection scores are causally linked to online reviews in our model, but instead that they are driven by the same latent variable, true restaurant hygiene quality, and therefore should be correlated. Our interest lies in estimating this correlation accurately. One potential source of bias in the estimation could be reverse causality due to inspectors choosing restaurants based on information gleaned from online reviews. If so, the observed correlation would be in part due to online reviews generating inspections for at-risk restaurants, rather than because both review text and inspection scores independently reflect the true underlying level of hygiene. However, we believe this is unlikely due to the way inspections are scheduled. Scheduling is not performed by the actual inspectors, but by administrative staff at the NYC RIP, who are responsible for scheduling thousands of inspections every year. Further, inspections are meant to occur randomly at approximately one-year intervals. Thus, the correlation between SMASH word counts and hygiene scores, shown in Table 4, can be assumed to be estimated with low risk of reverse causality.

Robustness Tests

The results presented above show that consumers' hygiene-related feedback through online reviews is highly correlated with inspection scores. Here, we explore three alternative models to examine the robustness of this correlation. First, as Schomberg et al. (2016) points out, consumers may not be able to spot all types of hygiene violations. For example, consumers would not be expected to accurately report on the cleanliness of the kitchen, which could only be observed through full access to the restaurant premises. Generally, critical violations related to issues like meat handling, kitchen cleanliness, and food preparation processes would not be directly observable by consumers, unlike non-critical violations, which include unclean bathrooms and dining areas. Each inspection has a score for critical violations and a score for non-critical violations, with the total score being the sum of both. This distinction provides us with a plausible test for falsifiability, i.e. we should not observe a strong correlation between SMASH word counts and critical violation scores. We therefore re-fit model (4) using critical violation scores as the dependent variable, and separately using non-critical violation scores. As before, we estimate linear regressions for both dependent variables. The results are presented in Table 5, with Model 1A based on critical violations and Model 2A based on non-critical violations. As expected, we find that SMASH word counts are not significantly associated with critical inspection scores (0.94, p=0.22) but are significantly associated with

non-critical violations (0.43, p<0.01). Tobit regression for the same models shows similar results as OLS. These results, beyond providing a falsifiability test, also emphasize the argument that while crowd-sourced reviews are useful as complements to traditional inspections, they cannot and should not be viewed as substitutes. Further, while prediction models of inspections based on such reviews may be useful, they have limited value in predicting *critical* violations within restaurants, which are more likely to be the drivers of food-borne illnesses. We depart from prior research in this domain (Kang et al. 2013, Schomberg et al. 2016) on this subtle but important point.

Second, thus far we have only considered a linear relationship between word counts and inspection scores. We now consider a possible non-linear relationship between word counts and aggregate inspection scores (critical + non-critical), by including a squared term as follows:

$$INS_{i,t} = \beta_0 + \beta_1 WC_{i,t-1} + \beta_2 WC_{i,t-1}^2 + \beta_3 Rating_{i,t-1} + \beta_4 Reviews_{i,t-1} + \beta_5 ReIns_{it} + \boldsymbol{\gamma R}_i + \boldsymbol{\delta I}_{it} + \alpha t + \varepsilon_{it} \quad (5)$$

The results from this analysis, shown in Table 4, are similar to the results from above. The coefficient of the word counts is positive and significant (0.43, p<0.01) for the base OLS model (Table 4: Model 3) and the Tobit (Table 4: Model 4). However, the squared term is not significant in either model, supporting the assumption of a linear relationship between SMASH word counts and inspection scores.

Finally, review sentiment may be associated with inspection scores, as observed by Kang et al. (2013); if sentiment is also associated with SMASH word counts, this could help explain the observed relationship between SMASH scores and inspection scores. We therefore examine whether including average sentiment for reviews received in time period $t-1$ alters the estimated correlation between SMASH word counts and inspection scores.[8] We find that review sentiment is highly correlated with the average rating, as one might expect, raising the potential issue of multicollinearity; therefore, we first estimate our model using sentiment but not numerical rating (Table 6: Model 1) and then both sentiment and numerical rating (Table 6: Model 3). We find that the relationship between SMASH word counts and inspection scores remains consistent even after including sentiment. Given the high correlation between

[8] The sentiment measure is calculated using the NRC Yelp Restaurant Sentiment Lexicons, which were automatically generated using online reviews of restaurants within the Yelp Phoenix Academic Dataset (Cambria et al. 2013, DrivenData 2016). Note that the sentiment dictionary was created from text generated in a similar domain and context, thus ensuring that sentiment scores are meaningful and appropriate for our context (Liu 2015).

sentiment and rating, it is not surprising that the coefficients for these variables are not significant when both are included in the model.

To summarize, we find a robust and significant positive linear association between SMASH word counts and hygiene scores, after controlling for restaurant-specific, inspector-specific, and time trends. We can therefore now utilize these word counts to try to identify shirking ex-post on the part of restaurants, and the timing associated with such behavior.

Using SMASH Dictionary to Identify Moral Hazard

We use restaurant-specific SMASH word counts, aggregated by day, to track hygiene of NYC restaurants in the time *between* hygiene inspection cycles, a period of approximately 12 to 15 months. Recall that, as illustrated in Figures 1 and 2, our exploratory analysis suggests that "PAPA" restaurants shirk on hygiene at some point during this period. The data used to generate Figures 1 and 2 were based on connecting inspection results for the same restaurant (or group of restaurants) across the inspection dataset; in reality, the average customer or inspector is blind to the actual hygiene-related practices within restaurants during the period between inspections. Here, we propose to "connect the dots" between inspections by using SMASH word counts as a proxy for hygiene practices of restaurants between inspections.

Figure 3 shows the smoothed daily SMASH word counts for 90 days[9] after each inspection in cycles 2 and 3 for "PAPA" restaurants; Figure 4 shows the same thing for "AA" restaurants. Note that the "AA" restaurants only have two inspections across the two cycles, since they did not require re-inspection, while the "PAPA" restaurants have four inspections, since they were inspected twice during each cycle. The top-left panel of Figure 3 shows a downward trend in word counts following the initial inspection of cycle 2, consistent with the discrepancy in inspection scores between initial and re-inspection for these restaurants. More interestingly, the top-right panel of Figure 3 shows an upward trend in the days following re-inspection of cycle 2, showing that the hygiene practices of "PAPA" restaurants tend to worsen after posting an A grade. The bottom row shows a similar pattern over the third inspection cycle. In contrast, Figure 4 shows that the hygiene practices of "AA" restaurants tend to be relatively steady after achieving an A grade.

[9] For "PAPA" restaurants between initial and re-inspection, we only used days prior to re-inspection for each restaurant. On average, re-inspections occur 54 days after initial inspection. Therefore, different numbers of restaurants contribute to the average across the x-axis of Figure 3. However, re-inspections are scheduled randomly within 3 months of initial inspection, so this does not introduce bias, and the trends are similar considering only 30 and 60 days after initial inspection.

Notably, while we found in our exploratory analysis that "PAPA" and "AA" restaurants achieve similar final hygiene inspection scores, the SMASH word counts of "PAPA" restaurants immediately *after* their final inspection are substantially higher on average than those of "AA" restaurants immediately after their final inspection at each cycle. Thus, "PAPA" restaurants may strive to comply with the "letter of the law" regarding hygiene inspections, but continue to exhibit behavior that consumers perceive as unhygienic.

We employ a longitudinal linear regression model to compare SMASH word count scores for the "AA" and "PAPA" subgroups over the 90 days following the final inspection of each cycle (i.e. initial inspection for "AA" restaurants and re-inspection for "PAPA" restaurants). Our objective is to quantify any negative effect on hygiene that manifests immediately post-inspection for both subgroups after an A grade is granted. Letting $0 \leq D_{it} \leq 90$ be the number of days since final inspection for restaurant i, we fit the following model:

$$WC_{it} = \beta_0 + \beta_1 D_{it} + \beta_2 D_{it} AA_i + \beta_3 D_{it} PAPA_i + \beta_4 Rating_{it} + \beta_5 Reviews_{it} + \boldsymbol{\gamma} \boldsymbol{R}_i + \varepsilon_{it}, \qquad (6)$$

where WC_{it} is the total SMASH word count of restaurant i at day t. Since ε_{it} may be correlated across subsequent observations for the same restaurant, we use weighted least squares estimation with an AR(1) autocorrelation structure on the residuals. AA_i and $PAPA_i$ are indicators of restaurant i's membership in groups "AA" or "PAPA", respectively. As in model (4), $\boldsymbol{\gamma}$ contains restaurant-level fixed effects to account for unobserved heterogeneity between restaurants. This model does not include effects for AA_i and $PAPA_i$, which are time-invariant and therefore incompatible with restaurant-level fixed effects, $\boldsymbol{\gamma}$. Based on our observations in Figures 3 and 4, we expect no significant interaction between D_{it} and AA_i, since "AA" restaurants tend to maintain stable hygiene practices over time, while we would expect a positive interaction coefficient between D_{it} and $PAPA_i$, since "PAPA" restaurants tend to have higher SMASH word counts after re-inspection.

The results from this analysis are displayed in Table 7: Model 1. The main effect for D_{it} is close to zero, but the interaction coefficient between D_{it} and $PAPA_i$ is significant and positive (0.73, $p<0.001$), indicating an increase in SMASH word counts for the "PAPA" group immediately following re-inspection (i.e. after receiving an A sticker). In contrast, the coefficient for the interaction coefficient between D_{it} and AA_i is near zero (0.01, $p<0.05$), indicating that the "AA" group displays fairly consistent hygiene practices

even after the conclusion of the inspection cycle. We find consistent results when fixed effects for Yelp reviewers are included in the model (Table 7: Model 2).

Combining this model with model (4), we can also predict the average inspection score that *would have been received* if restaurants were inspected again 90 days after the final inspection, when an A grade is assigned. For the "PAPA" group, we estimate an average inspection score of 32.85 90 days after receiving an A grade. This is in the range for a C grade (28 or higher) and is consistent with moral hazard. Using model (4) and its estimated coefficients reported in Table 7, we predict the hygiene score for restaurants that are in the "AA" and "PAPA" subgroups for 90 days after that restaurant receives an A grade. We perform this analysis over 2012-2015. Based on the predicted scores, we assign the letter grade (A, B or C) that each restaurant would have received, had it been inspected. The results, displayed by year in Figure 5, are striking and show clearly the presence of moral hazard within the NYC RIP. In 2012, for example, of the "PAPA" subgroup restaurants, roughly 57.9% are predicted to receive a B grade 90 days after receiving an A grade, while 18.7% are predicted to receive a C grade. Therefore, approximately 76.6% of the "PAPA" subgroup regresses to lower hygiene scores shortly after achieving an A grade. Since the number of A grade restaurants (including restaurants who earned a P grade first) account for over 90% of restaurants, the issue of moral hazard may be viewed as a significant issue within the NYC RIP. Interestingly, 15.6% of "AA" restaurants in 2012 also display some regression in their hygiene scores, indicating that the tendency to not fully comply with hygiene requirements between inspections is not only limited to the "PAPA" restaurants. These trends are consistent over the four years of our study.

We also investigate which restaurant characteristics are associated with higher SMASH word counts within 90 days of obtaining an A for "PAPA" restaurants. We use the restaurant-level fixed effect coefficients from model (6) for each "PAPA" restaurant in a linear regression model with restaurant characteristics as independent variables. Note that model (6) controlled for the number of reviews received by each restaurant over time. We find that several variables related to location, cuisine and price point are related with higher SMASH word counts. For example, restaurants located in Manhattan, Brooklyn, and the Bronx tend to have higher SMASH word counts than restaurants in Queens and Staten Island. Regarding cuisine, Lehman et al. (2014) identified differences in hygiene inspection scores across cuisines. Our analysis complements that work, since we identify cuisines that tend to regress in their hygiene as measured

by SMASH word counts. We find that South Asian (Afghan, Bangladeshi, Pakistani, and Indian) restaurants are among the worst regressors, along with restaurants in the Pizza and Caribbean categories. We also find that lower priced-restaurants ($-$$) tend to regress more relative to higher-priced ones ($$$-$$$$). Interestingly, restaurants in some categories show little to no increase in SMASH word counts after receiving an A. These include those belonging to the top 100 national chains (franchises) in the U.S., arguably due to the tighter franchise requirements and more controlled operational processes, as well as restaurants in the Bakery, Breakfast & Brunch, Vegetarian, and Cafes categories.

Finally, we test whether certain restaurant characteristics are more associated with higher SMASH word counts in the "PAPA" group within a shorter time period. We estimate that within 30 days of receiving an A, restaurants in Brooklyn tend to regress more than those in the other boroughs, and that Pizza restaurants and very low-priced restaurants ($) tend to regress much faster than those in other cuisines and price points. We find that the greater set of location, cuisine and price variables associated with SMASH word counts 90 days post-inspection are similar to those associated with 30 days post-inspection. While these statistics are based on word counts associated with worse hygiene inspection scores and should not be established as causal arguments, they provide some initial indications for policy makers in terms of which categories of restaurants are likely to exhibit moral hazard in hygiene inspections. Unlike previous work in this area focused on detection and prediction using social media data (Kang et al. 2013, Schomberg et al. 2016), the objective of our work is to help inform planners by revealing and localizing blind spots in the current NYC restaurant inspection program.

Discussion and Conclusion

Moral hazard continues to be a pressing problem in the services context (Nayyar 1990), with particular relevance to the food preparation industry where hygiene problems can lead to localized incidences of foodborne illnesses (Gould et al. 2013, Hedberg et al. 2006). However, newer techniques of text analysis as well as access to crowd-sourced data from online review platforms like Yelp provide fresh opportunities to tackle such public policy problems (Athey 2017). This paper presents an analysis of the effects of moral hazard in a restaurant hygiene inspection program in New York City, which provides enough incentives for restaurants to improve their hygiene performance prior to receiving a hygiene sticker but appears limited in its ability to provide ongoing monitoring. To surmount these limitations, we use a combination of

supervised machine learning and econometrics to devise a proxy "information system" for hygiene using online reviews provided by restaurant consumers over time, thereby allowing for a more "continuous" form of monitoring of restaurant hygiene. Since "foodborne illness remains one of the top public health challenges for the city, the state, and the entire country" (Centers for Disease Control 2016b), our approach using machine learning to inform inspections program can help improve their design and efficacy.

Our results strongly suggest that NYC currently may be underperforming in the manner in which it communicates inspection results to consumers, and thereby ensuring a socially optimal level of hygiene within the city, in several ways. First, by publicly posting a "P" grade without translating the current grade to a "B" or "C" that consumers can understand, the program allows for the presence of opportunism, whether intentional or otherwise. Short-term compliance is easy but allows relatively costless shirking behavior soon thereafter. Without a penalty for this initial lower performance, the incentives for moral hazard remain unchecked. Second, by only showing the most recent inspection cycle result instead of the trajectory (or previous history), the program reinforces a short-term perspective, as shown in our analysis. Indeed, certain municipal authorities are changing this approach; Louisville, KY, for instance, requires restaurants to show the last three scores (regardless of the hygiene grade) in addition to the current one, thereby creating incentives to remain compliant over time and arguably translating inspection results more effectively to consumers.[10] Finally, by scheduling re-inspection times for regular and critical offenders using the same system again reduces the incentive to comply. Even in 90-days post-inspection, we observe variations in the segments of restaurants that appear to reverse their hygiene standards. Thus, using prior scores in scheduling repeat inspections could be useful in incentivizing compliance. Furthermore, in the presence of critical violations, restaurants face an additional random check, beyond the regular schedule. To the extent that these additional steps are able to reduce the observed moral hazard, inspection effort and resources are likely to be more effectively utilized.

The increasing prevalence of social media data actually enhances the practical implications of this work for restaurants and for-profit firms. Investors and restaurant franchises may benefit by being able to design franchising contracts that potentially incorporate clauses based on social media discourse and online review data, so as to avoid the typical ex-post contracting problems that emerge in such settings (Lafontaine

[10] https://louisvilleky.gov/government/health-wellness/food-safety

1992). Finally, the restaurant itself would benefit from monitoring social media discourse; while restaurants are likely to be intimately aware of their own hygiene issues, it is still useful for them to realize the extent to which such issues are discussed in the public domain. Interestingly, although our dictionary was crafted based on reviews in NYC, in robustness tests reported earlier, we show how our dictionary is successful in explaining hygiene inspections in Boston (see Appendix C), where a different inspection program is used, showing the generalizability of the approach.

Beyond the policy and managerial implications discussed above, our work adds to the current literature in IS research on the use of social media, in part by responding to the sentiment expressed by Aral et al. (2013) that social media and user-generated content can, and should, be utilized in addressing pressing social problems. Whether it is the use of online reviews and social media data in medicine (Gao et al. 2012), the study of elections and politics (Wattal et al. 2010), or disaster recovery (Oh et al. 2010), we believe that the potential for crowd-sourced data from social media platforms within the non-profit and policy sector is enormous. Public health is one such context where social media can be particularly useful for two reasons. First, the use of online reviews and social media within the important restaurant sector is already the default, with most consumers not only consuming this information but also contributing (X. Lu et al. 2013). Second, restaurants and other stake-holders have recognized the influence that social media discourse exerts, thereby creating a positive virtuous cycle that can be used for the greater good. Though our work here is focused on the New York City market, the larger contribution of our work is to provide a set of econometric models that can be replicated elsewhere, as well as provide the dictionary of hygiene, which may be utilized in other policy and regulatory recommendations (Bichler et al. 2010). Although we focus on Yelp here, the described process can also easily be extended to data from other sources, such as Facebook and Twitter, as well as other review sites such as TripAdvisor and OpenTable, speaking to generalizability of our work.

Our work also responds to the call issued by Athey (2017), calling for more work at the intersection of machine learning and economics. Traditional economic modeling has postulated the effects of many factors that affect markets, such as moral hazard, adverse selection, and shirking (Akerlof 1970, Hölmstrom 1979) but empirical approaches to identify these have often been limited by the unavailability of data. Interestingly, social media platforms have contributed to the enhanced efficiency of market by explicitly addressing some of these traditional pathologies, for instance by providing decentralized institutions that

measure reputation (Dellarocas 2003, Moreno and Terwiesch 2014), improve transparency (Luca and Zervas 2016) and enhance trust (Duan et al. 2008). We see our work has continuing on this path.

Finally, we note that our approach, blending modern statistical methods with classical regression, contributes to ongoing work fostering IT innovations in the public sector (Kankanhalli et al. 2017), particularly in the public health domain, where tackling the risks of foodborne illnesses remains a hard problem within large urban areas like New York City (Harrison et al. 2014). Inasmuch as such hybrid approaches provide better outcomes for regulatory agencies and city authorities, we also believe that caution must be exercised in understanding how such hybrid methods may inform policy, and the potential for unintended and negative consequences. We do argue that the use of hybrid methods can help in making the inspection process more efficient above, as do Harrison et al. (2014) who propose incorporating online reviews directly into the scheduling of inspectors. However, such an approach would make online review platforms endogenous to the inspection process and likely incentivize a new breed of fraudulent reviews, effectively reducing the efficacy of online reviews as a monitoring tool. Our objective here is to show how such hybrid methods can be viable complements to traditional inspection methods, but a detailed analysis of the tradeoffs inherent in including such methods into official processes is needed. Platform owners like Yelp and TripAdvisor are important parts of this ecosystem and should be included in all such evaluations, since they control the data and are accepted by consumers and merchants alike.

In summary, in this paper, we use a combination of supervised machine learning and econometrics to devise a proxy "information system" for hygiene using online reviews provided by restaurant consumers over time. Through the creation of a hygiene-related dictionary, we show how a more "continuous" form of monitoring of restaurant hygiene may be achieved, thereby identifying, and hopefully reducing, behavior associated with moral hazard. We also show the efficacy of this approach in the context of the NYC RIP, providing evidence for how the success of the program may be overstated. Our work provides a blueprint for how social media content and machine learning techniques can be used in hybrid approaches to address persistent social issues in an effective yet scalable manner (Athey 2017, Kang et al. 2013).

References

Abbasi A, Chen H (2008) CyberGate: a design framework and system for text analysis of computer-mediated communication. *Mis Q.*:811–837.

Akerlof GA (1970) The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Q. J. Econ.* 84(3):488–500.
Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: an empiricist's companion* (Princeton university press Princeton).
Archak N, Ghose A, Ipeirotis PG (2011) Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Manag. Sci.* 57(8):1485–1509.
Athey S (2015) Machine Learning and Causal Inference for Policy Evaluation. (ACM), 5–6.
Athey S (2017) Beyond prediction: Using big data for policy problems. *Science* 355(6324):483–485.
Athey S, Imbens GW (2006) Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74(2):431–497.
Autor DH (2003) Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *J. Labor Econ. January*.
Barron J (2015) Restaurants Follow Consultants' Advice to the Letter for an A Grade. *The New York Times* (October 4) https://www.nytimes.com/2015/10/05/nyregion/health-exam-help-for-restaurants-to-avoid-rodents-or-worse-a-c.html.
Berger J, Sorensen AT, Rasmussen SJ (2010) Positive Effects of Negative Publicity: When Negative Reviews Increase Sales. *Mark. Sci.* 29(5):815–827.
Bertrand M, Duflo E, Mullainathan S (2002) *How much should we trust differences-in-differences estimates?* (National Bureau of Economic Research).
Bichler M, Gupta A, Ketter W (2010) Research Commentary—Designing Smart Markets. *Inf. Syst. Res.* 21(4):688–699.
Brownstein JS, Freifeld CC, Madoff LC (2009) Digital Disease Detection — Harnessing the Web for Public Health Surveillance. *N. Engl. J. Med.* 360(21):2153–2157.
Buchholz U, Run G, Kool J, Fielding J, Mascola L (2002) A risk-based restaurant inspection system in Los Angeles County. *J. Food Prot.* 65(2):367–372.
Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6(1):3–5.
Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* 28(2):15–21.
Cao Q, Duan W, Gan Q (2011) Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decis. Support Syst.* 50(2):511–521.
Centers for Disease Control (2016a) Foodborne Germs and Illnesses. Retrieved (April 10, 2016), http://www.cdc.gov/foodsafety/foodborne-germs.html.
Centers for Disease Control (2016b) *Trends in Foodborne Illness*
Chan EH, Brewer TF, Madoff LC, Pollack MP, Sonricker AL, Keller M, Freifeld CC, Blench M, Mawudeku A, Brownstein JS (2010) Global capacity for emerging infectious disease detection. *Proc. Natl. Acad. Sci.* 107(50):21701–21706.
Cheung CMK, Thadani DR (2012) The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decis. Support Syst.* 54(1):461–470.
Crespi JM, Marette S (2001) How Should Food Safety Certification be Financed? *Am. J. Agric. Econ.* 83(4):852–861.
Decker R, Trusov M (2010) Estimating aggregate consumer preferences from online product reviews. *Int. J. Res. Mark.* 27(4):293–307.
Dellarocas C (2003) The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Manag. Sci.* 49(10):1407–1424.
Dellarocas C, Narayan R (2006) A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth. *Stat. Sci.* 21(2):277–285.
Dellarocas C, Zhang X (Michael), Awad NF (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. Interact. Mark.* 21(4):23–45.
DrivenData (2016) Competition: Keeping it Fresh: Predict Restaurant Inspections. Retrieved https://www.drivendata.org/competitions/5/.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter? — An empirical investigation of panel data. *Decis. Support Syst.* 45(4):1007–1016.
Egan M, Raats M, Grubb S, Eves A, Lumbers M, Dean M, Adams M (2007) A review of food safety and food hygiene training studies in the commercial sector. *Food Control* 18(10):1180–1190.
Einav L, Levin J (2014) Economics in the age of big data. *Science* 346(6210):1243089.
Farley T (2011) *Restaurant Letter Grading: The First Year* (New York City Department of Health and Mental Hygiene and The City University of New York).
Farley T (2012) *Restaurant Grading in New York City at 18 Months* (New York City Department of Health and Mental Hygiene and The City University of New York).
Farley T (2016) *Enforcement Guidelines for Common Sanitary Violations* (New York City Department of Health and Mental Hygiene).
Feldman R (2013) Techniques and applications for sentiment analysis. *Commun. ACM* 56(4):82–89.
Gao GG, McCullough JS, Agarwal R, Jha AK (2012) A Changing Landscape of Physician Quality Reporting: Analysis of Patients' Online Ratings of Their Physicians Over a 5-Year Period. *J. Med. Internet Res.* 14(1):e38.
Gaynor M, Haas-Wilson D, Vogt WB (2000) Are Invisible Hands Good Hands? Moral Hazard, Competition, and the Second-Best in Health Care Markets. *J. Polit. Econ.* 108(5):992–1005.
Ghose A, Ipeirotis PG (2011) Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Trans. Knowl. Data Eng.* 23(10):1498–1512.
Ghose A, Ipeirotis PG, Li B (2012) Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Mark. Sci.* 31(3):493–520.
Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. *CS224N Proj. Rep. Stanf.* 1:12.
Godes D, Mayzlin D (2004) Using Online Conversations to Study Word-of-Mouth Communication. *Mark. Sci.* 23(4):545–560.
Goodman JK, Cryder CE, Cheema A (2013) Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *J. Behav. Decis. Mak.* 26(3):213–224.
Gould LH, Rosenblum I, Nicholas D, Phan Q, Jones TF (2013) Contributing factors in restaurant-associated foodborne disease outbreaks, FoodNet sites, 2006 and 2007. *J. Food Prot.* 76(11):1824–1828.
Hand DJ, Mannila H, Smyth P (2001) *Principles of data mining* (MIT press).
Harrison C, Jorder M, Stern H, Stavinsky F, Reddy V, Hanson H, Waechter H, Lowe L, Gravano L, Balter S (2014) Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013. *MMWR* 63(20):441–445.
HealthMap (2017) About | HealthMap. Retrieved (March 24, 2017), http://www.healthmap.org/site/about.
Hedberg CW, Smith SJ, Kirkland E, Radke V, Jones TF, Selman CA, Group ENW (2006) Systematic environmental evaluations to identify food safety differences between outbreak and nonoutbreak restaurants. *J. Food Prot.* 69(11):2697–2702.
Ho DE (2012) Fudging the nudge: information disclosure and restaurant grading. *Yale LJ* 122:574.
Hölmstrom B (1979) Moral hazard and observability. *Bell J. Econ.*:74–91.
Hornik K, Rauch J, Buchta C, Feinerer I, Hornik MK (2016) *Package 'textcat'*
Hu M, Liu B (2004) Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM), 168–177.
Imbens G, Wooldridge J (2007) Difference-in-differences estimation. *Natl. Bur. Econ. Res. Work. Pap.*
Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* 47(1):5–86.
Ipeirotis PG (2010) Analyzing the Amazon Mechanical Turk Marketplace. *XRDS* 17(2):16–21.
Irwin K, Ballard J, Grendon J, Kobayashi J (1989) Results of routine restaurant inspections can predict outbreaks of foodborne illness: the Seattle-King County experience. *Am. J. Public Health* 79(5):586–590.

Jin GZ, Lee J (2014) Inspection technology, detection, and compliance: evidence from Florida restaurant inspections. *RAND J. Econ.* 45(4):885–917.
Jin GZ, Leslie P (2003) The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards. *Q. J. Econ.* 118(2):409–451.
Jin GZ, Leslie P (2009) Reputational incentives for restaurant hygiene. *Am. Econ. J. Microecon.* 1(1):237–267.
Kang JS, Kuznetsova P, Luca M, Choi Y (2013) Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews. *EMNLP*:1443–1448.
Kankanhalli A, Zuiderwijk A, Tayi GK (2017) Open innovation in the public sector: A research agenda. *Gov. Inf. Q.* 34(1):84–89.
Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z (2015) Prediction Policy Problems. *Am. Econ. Rev.* 105(5):491–495.
Lechner M (2011) *The estimation of causal effects by difference-in-difference methods* (Now).
Lee TY, BradLow ET (2011) Automated Marketing Research Using Online Customer Reviews. *J. Mark. Res.* 48(5):881–894.
Lee YJ, Hosanagar K, Tan Y (2015) Do I follow my friends or the crowd? Information cascades in online movie ratings. *Manag. Sci.* 61(9):2241–2258.
Lehman DW, Kovács B, Carroll GR (2014) Conflicting Social Codes and Organizations: Hygiene and Authenticity in Consumer Evaluations of Restaurants. *Manag. Sci.* 60(10):2602–2617.
Li F (2010) The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *J. Account. Res.* 48(5):1049–1102.
Li X, Hitt LM (2008) Self-Selection and Information Role of Online Product Reviews. *Inf. Syst. Res.* 19(4):456–474.
Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73(1):13–22.
Liu B (2015) *Sentiment analysis: Mining opinions, sentiments, and emotions* (Cambridge University Press).
Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C (2002) Text classification using string kernels. *J. Mach. Learn. Res.* 2(Feb):419–444.
Loughran T, Mcdonald B (2011) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *J. Finance* 66(1):35–65.
Lu X, Ba S, Huang L, Feng Y (2013) Promotional Marketing or Word-of-Mouth? Evidence from Online Restaurant Reviews. *Inf. Syst. Res.* 24(3):596–612.
Lu Y, Jerath K, Singh PV (2013) The emergence of opinion leaders in a networked online community: A dyadic model with time dynamics and a heuristic for fast estimation. *Manag. Sci.* 59(8):1783–1799.
Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and Yelp review fraud. *Manag. Sci.*
Ludwig S, de Ruyter K, Friedman M, Brüggen EC, Wetzels M, Pfann G (2013) More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates. *J. Mark.* 77(1):87–103.
McCallum A, Nigam K, Others (1998) A comparison of event models for naive bayes text classification. (Citeseer), 41–48.
Mejia J, Mankad S, Gopal A (2015) More Than Just Words: Latent Semantic Analysis, Online Reviews and Restaurant Closure. *Acad. Manag. Proc.* 2015(1).
Miller GA (1995) WordNet: a lexical database for English. *Commun. ACM* 38(11):39–41.
Moreno A, Terwiesch C (2014) Doing business with strangers: Reputation in online service marketplaces. *Inf. Syst. Res.* 25(4):865–886.
Nayyar PR (1990) Information asymmetries: A source of competitive advantage for diversified service firms. *Strateg. Manag. J.* 11(7):513–519.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Mark. Sci.* 31(3):521–543.
New York City Department of Health and Mental Hygiene (2016a) *Food Service Establishment Inspection Scoring Parameters: A Guide to Conditions*
New York City Department of Health and Mental Hygiene (2016b) *How We Score and Grade*
New York City Department of Health and Mental Hygiene (2016c) *Letter Grading Workshop Materials*
New York City Department of Health and Mental Hygiene (2016d) *Self-Inspection Worksheet for Food Service Establishments*
New York City Department of Health and Mental Hygiene (2016e) *What to Expect When You're Inspected: A Guide for Food Service Operators*
New York City Department of Health and Mental Hygiene The Impact of Letter Grading in NYC.
Oh O, Kwon K, Rao H (2010) An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010. *ICIS 2010 Proc.*
OpenData (2016) NYC Open Data: Restaurant Inspections. https://nycopendata.socrata.com/.
Park A (2015) New York City Restaurants Are Cleaner Than Ever. *Time Mag.* Retrieved (October 2, 2017), http://time.com/3940482/new-york-city-restaurants-clean/.
Park SY, Allen JP (2013) Responding to online reviews problem solving and engagement in hotels. *Cornell Hosp. Q.* 54(1):64–73.
Ross SM (1996) *Stochastic processes* (John Wiley & Sons New York).
Rubinstein A, Yaari ME (1983) Repeated insurance contracts and moral hazard. *J. Econ. Theory* 30(1):74–97.
Schomberg JP, Haimson OL, Hayes GR, Anton-Culver H (2016) Supplementing public health inspection via social media. *PloS One* 11(3):e0152117.
Sebastiani F (2002) Machine learning in automated text categorization. *ACM Comput. Surv. CSUR* 34(1):1–47.
Shapiro C (1986) Investment, moral hazard, and occupational licensing. *Rev. Econ. Stud.* 53(5):843–862.
Shavell S (1979) On Moral Hazard and Insurance. *Q. J. Econ.* 93(4):541–562.
Starbird SA (2005) Moral hazard, inspection policy, and food safety. *Am. J. Agric. Econ.* 87(1):15–27.
Starbird SA, Amanor-Boadu V (2007) Contract selectivity, food safety, and traceability. *J. Agric. Food Ind. Organ.* 5(1):1542–0485.
Stiglitz J (2002) Information and the Change in the Paradigm in Economics. *Am. Econ. Rev.* 92(3):460–501.
Stiglitz J (2010) Regulation and failure. *Rev. Econ. Inst.* 12(23):13–28.
Tang H, Tan S, Cheng X (2009) A survey on sentiment detection of reviews. *Expert Syst. Appl.* 36(7):10760–10773.
Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: Quantifying language to measure firms' fundamentals. *J. Finance* 63(3):1437–1467.
Tsai ACR, Wu CE, Tsai RTH, Hsu JY jen (2013) Building a concept-level sentiment dictionary based on commonsense knowledge. *IEEE Intell. Syst.* 28(2):22–30.
Valitutti A, Strapparava C, Stock O (2004) Developing affective lexical resources. *PsychNology J.* 2(1):61–83.
Varian HR (2014) Big Data: New Tricks for Econometrics. *J. Econ. Perspect.* 28(2):3–27.
Wallach HM, Mimno DM, McCallum A (2009) Rethinking LDA: Why Priors Matter. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Adv. Neural Inf. Process. Syst. 22*. (Curran Associates, Inc.), 1973–1981.
Wattal S, Schuff D, Mandviwalla M, Williams CB (2010) Web 2.0 and Politics: The 2008 U.S. Presidential Election and an E-Politics Research Agenda. *MIS Q.* 34(4):669–688.
Young L, Soroka S (2012) Affective News: The Automated Coding of Sentiment in Political Texts. *Polit. Commun.* 29(2):205–231.
Zhao Y, Yang S, Narayan V, Zhao Y (2012) Modeling Consumer Learning from Online Product Reviews. *Mark. Sci.* 32(1):153–169.

Figure 1. Inspection Scores of Restaurants that Consistently Score A (AA Group) Versus Those that Post a Grade P in their Initial Inspection and then an A (PAPA Group)

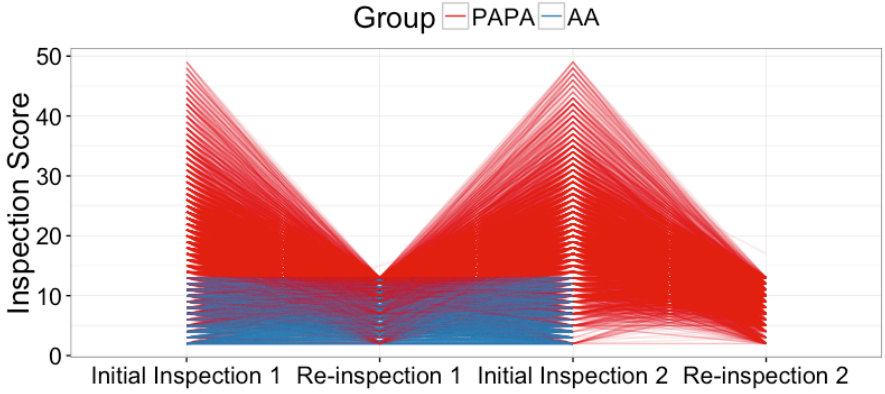

Figure 2. Mean Inspection Scores of Restaurants in AA and PAPA Groups

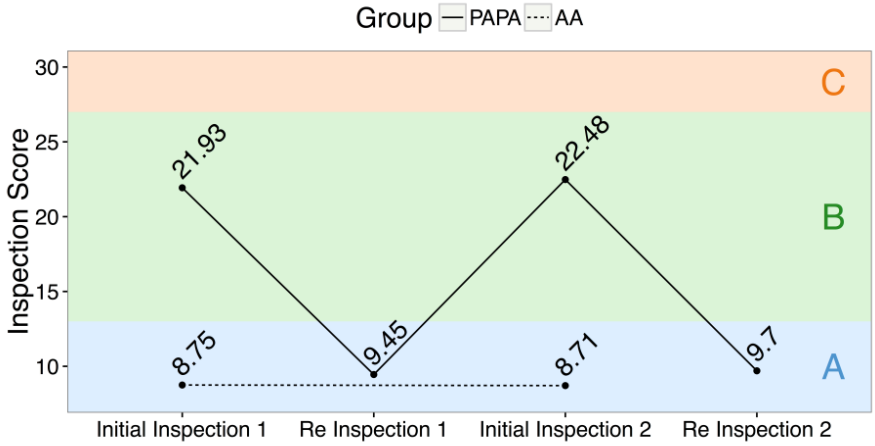

Figure 3. SMASH Score Trends After Initial and Re-Inspection of Two Cycles for "PAPA" Sample

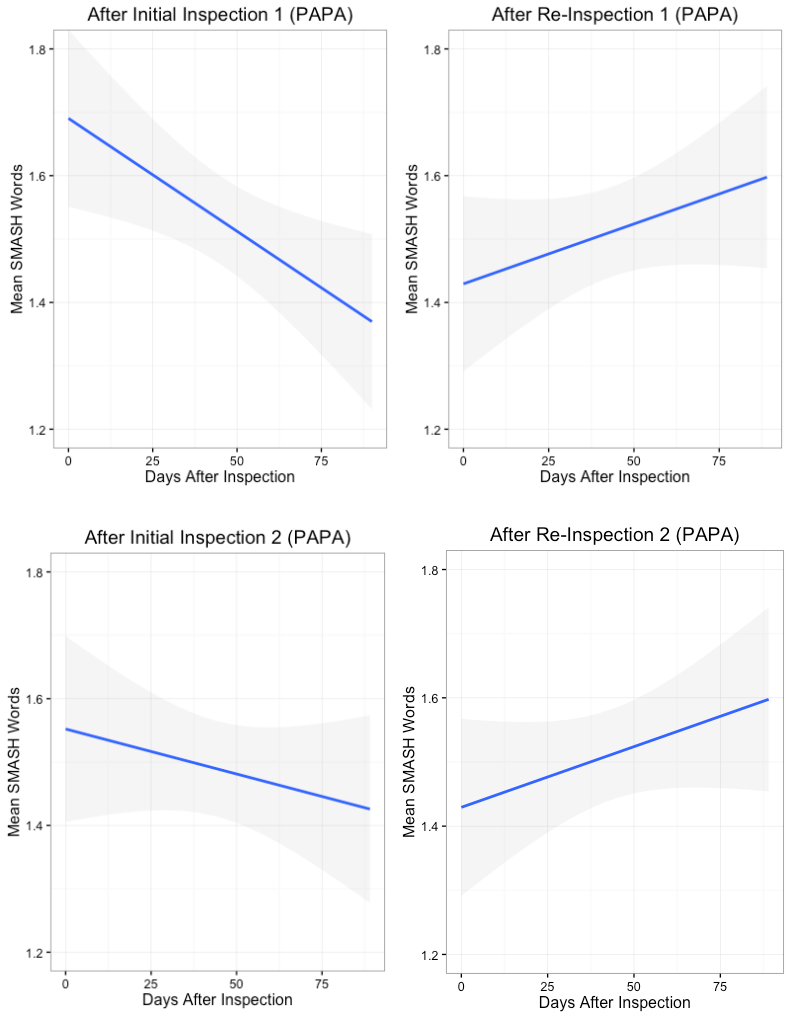

Figure 4. Trends in SMASH Scores After Initial Inspections for "AA" Sample

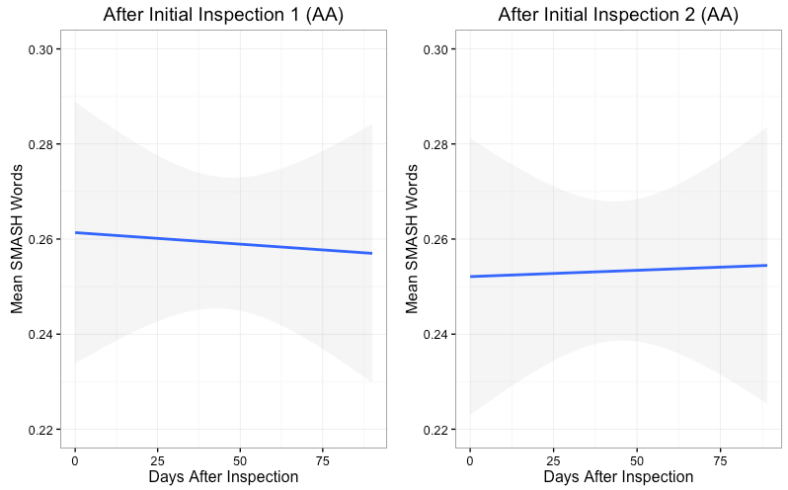

Figure 5. Predicted Inspection Scores for AA and PAPA Restaurants 90 Days after Inspection Based on Model (4) using SMASH Dictionary Word Counts

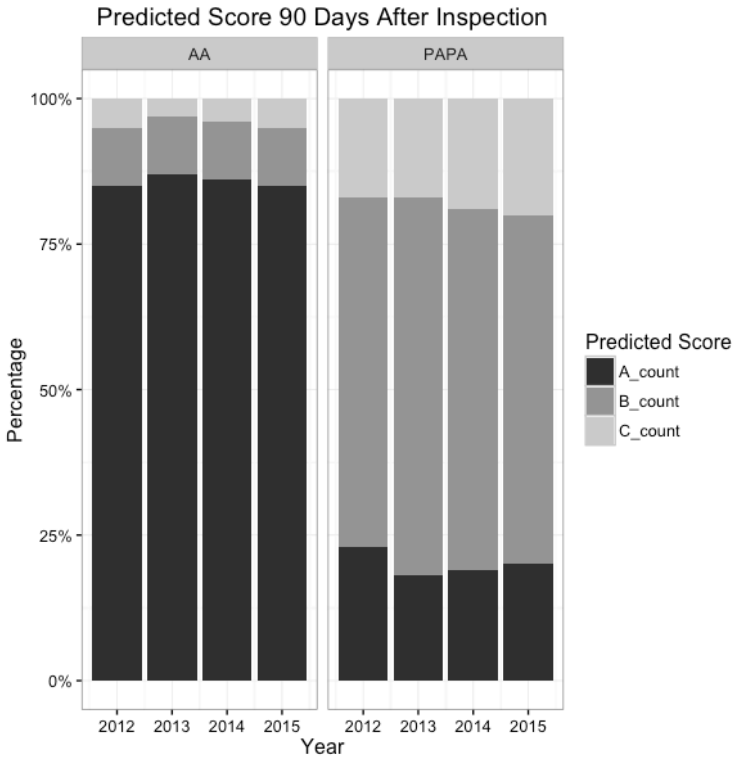

Table 1. Variable Descriptions

Source	Variable(s)	Description
Open Data	Boro	NYC Borough
Open Data	Address	Physical address
Open Data	Phone	Phone number
Open Data	Cuisine	Main cuisine declared
Open Data	Inspection Date	The date the inspection took place
Open Data	Inspection Type	They type of inspection (i.e. initial, re-inspection)
Open Data	Inspection Grade	The grade received as a result of the inspection (i.e. A, B, C, or P)
Open Data	Violation Code	The specific health code violation
Open Data	Violation Description	Description of the health code violation
Open Data	Violation Critical	1 if the violation is a critical violation
Open Data	Violation Score	The points added for the violation
Yelp.com	Rating	The overall rating
Yelp.com	Number of Reviews	Total number of reviews
Yelp.com	Price	Price range of the restaurant (1-4 dollar signs)
Yelp.com	Chars	Hours, Takes Reservations, Delivery, Take-out, Accepts Credit Cards, Accepts Apple Pay, Good For, Parking Options, Bike Parking, Good for Kids, Good for Groups, Attire, Ambience, Noise Level, Alcohol Beer & Wine, Outdoor Seating, Wi-Fi, Has TV, Waiter Service, Caters

Table 2. Inspection Grades from Inspection Cycles: 90% of Restaurants Achieve an A through Initial Inspection (A) or Re-Inspection (PA)

Grade (s)	Percentage
A	41.01%
PA	48.69%
PB	5.02%
PC	5.28%
Total	100.00%

Table 3. 25 Representative "Words" in the SMASH Dictionary

bugs	dirty	food disgusting	horrendous	insects
cockroaches	disgusting food	food horrible	ill-scented	insects food
contaminate	eaten saw cockroaches	gross	inedible	like insects
contaminating	eggs gross	hair baked eggs	inhuman	moth-eaten
defecate	filthy	health department	insect bite	wiping nose

Table 4. Relationship Between SMASH Word Counts and Hygiene Inspection Results for Restaurants, DV = Inspection Scores

	Model 1: OLS	**Model 2: Tobit**	**Model 3: OLS W/ Squared WC**	**Model 4: Tobit W/ Squared WC**
Variables	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
Intercept	13.17 (3.44) ***	14.84 (4.01) ***	9.91 (3.90) **	14.31 (5.11) **
WC	0.59 (0.21) **	0.65 (0.24) **	0.43 (0.15) **	0.58 (0.21) **
WC^2			0.01 (0.10)	0.20 (0.26)
Rating	-0.36 (0.14) **	-0.67 (0.20) ***	-0.36 (0.06) ***	-0.42 (0.19) *
Reviews	0.00 (0.00) ψ	0.00 (0.00) ψ	0.00 (0.00) ψ	0.00 (0.00) ψ
ReIns	-0.11 (0.05) *	-0.10 (0.05) *	-0.31 (0.15) *	-0.37 (0.18) *
Restaurant Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Inspector Fixed Effects	Yes	Yes	Yes	Yes
Groups	21,488	21,488	21,488	21,488
Observations	136,503	136,503	136,503	136,503
AIC		710.11		953.1
R^2	0.74		0.70	
Robust standard errors in parentheses Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 'ψ' 0.1				

Table 5. Examining Differences in Critical Versus Non-Critical Inspections, DV=Inspection Scores

	Model 1A: OLS	Model 1A: Tobit	Model 2A: OLS	Model 2B: Tobit
Type of inspections	**Only Critical**		**Only Non-Critical**	
Variables	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
Intercept	7.93 (4.01) *	9.62 (4.05) *	9.44 (3.01) ***	11.49 (4.45) **
WC	0.94 (1.22)	0.28 (0.77)	0.43 (0.0.17) **	0.75 (0.29) **
Rating	-0.54 (0.85)	-0.47 (1.63)	-0.49 (0.21) *	-0.57 (0.24) *
Reviews	0.00 (0.00) ψ	0.01 (0.00) ψ	0.00 (0.00) ψ	0.00 (0.00) ψ
Inspection-2nd	-0.80 (0.21) ***	-0.95 (0.35) **	-0.36 (0.11) ***	-0.52 (0.15) ***
Restaurant Fixed Effects	included	included	included	included
Inspector Fixed Effects	included	included	included	included
Time Fixed Effects	included	included	included	included
Groups	18,640	18,640	21,488	21,488
Observations	75,076	75,076	61,426	61,426
AIC		2886.9		931.4
R^2	0.41		0.65	
Robust standard errors in parentheses Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 'ψ' 0.1				

Table 6. Results Including Sentiment and Numerical Rating, DV=Inspection Scores

	Model 1: OLS W/ Sentiment	Model 2: Tobit W/ Sentiment	Model 3: OLS W/ Sentiment and Rating	Model 4: Tobit W/ Sentiment and Rating
Variables	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
Intercept	8.78 (3.00) **	9.94 (2.99) ***	11.11 (3.90) **	13.74 (4.10) ***
WC	0.45 (0.15) **	0.68 (0.24) **	0.58 (0.19) **	0.71 (0.21) ***
Rating			-0.10 (0.58)	-0.32 (0.73)
Sentiment	-0.15 (0.05) **	-0.21 (0.08) **	-0.26 (0.19)	-0.35 (1.12)
Reviews	0.00 (0.00) ψ	0.00 (0.00) *	0.00 (0.00) ψ	0.01 (0.00) ψ
ReIns	-0.11 (0.05) *	-0.11 (0.05) *	-0.11 (0.05) *	-0.11 (0.05) *
Restaurant Fixed Effects	included	included	included	included
Inspector Fixed Effects	included	included	included	included
Time Fixed Effects	included	included	included	included
Groups	21,488	21,488	21,488	21,488
Observations	136,503	136,503	136,503	136,503
AIC		1193		1211.8
R^2	0.70		0.70	
Robust standard errors in parentheses Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 'ψ' 0.1				

Table 7. The Effect of Belonging to the AA Versus PAPA Groups on Hygiene Word Counts, DV = SMASH WC (up to 90 days after inspection)

	Model 1: AR-1	**Model 2: AR-1 W/ Reviewer FE**
Variables	**Estimate (SE)**	**Estimate (SE)**
Intercept	0.98 (0.21) ***	0.66 (0.25) **
Offset	0.00 (0.00) ψ	0.01 (0.00) ψ
Offset*AA	0.01 (0.00) *	0.01 (0.00) **
Offset*PAPA	0.73 (0.22) ***	0.79 (0.30) **
Rating	0.55 (1.47)	0.72 (1.51)
Reviews	-0.19 (0.47)	-0.65 (0.83)
Restaurant Fixed Effects	Yes	Yes
Reviewer Fixed Effects	No	Yes
Groups	21,488	21,488
Observations	1,579,001	1,579,001
R^2	0.51	0.55
Robust standard errors in parentheses Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 'ψ' 0.1		