Ashwood, Z.C., Elias, B., & Ho, D.E. (2021). Improving the reliability of food safety disclosure: Restaurant grading in Seattle and King County, Washington. *Journal of Environmental Health*, *84*(2), 30–37.

Corresponding Author: Daniel E. Ho, Stanford University, Crown Quadrangle, 559 Nathan Abbott Way, Stanford, CA 94305-8610.
Email: dho@law.stanford.edu.

Note. This supplemental file was submitted by the authors along with the respective peer-reviewed article and has been posted online due to space limitations at https://www.neha.org/jeh/supplemental. The *Journal of Environmental Health* did not copy edit this file, nor was the file peer reviewed. The authors have provided this supplemental file as an extra resource should the reader want more information.

SUPPLEMENTAL APPENDIX: IMPROVING THE RELIABILITY OF FOOD SAFETY DISCLOSURE: RESTAURANT GRADING IN SEATTLE AND KING COUNTY, WASHINGTON

We here present supplemental results and figures to document our findings in support of the new grading system:

- Sections A-C presents results on the association between critical violation points and foodborne illnesses.
  - Table 1 provides a contingency table between critical and non-critical violation points and probable or lab-confirmed foodborne outbreaks.
  - Figure 1 presents receiver operating characteristic curve from a logistic regression of predicting foodborne illness outbreaks, showing that while violations are statistically significant predictors, the marginal predictive power is relatively low.
  - Figure 2 visualizes the correlation between violation points and foodborne illness outbreaks.
- Sections D-E demonstrate that reliability is much higher with critical violation points.
  - Figure 3 presents data from 378 peer review inspections showing that when two inspectors observed identical conditions, their agreement rate was much higher in the citation of critical than non-critical violations.
  - Figure 4 presents formal tests from logistic regressions of agreement for each violation across 378 peer review inspections, confirming that critical violations were statistically significantly more likely to result in agreement.

- Section F provides methodological details and results of the matched sample analysis of repeat violations.
    - Table 2 presents effects from 2006-14, demonstrating that repeat violations do not systematically predict worse outcomes.
- Section G presents results from the matched sample analysis of order effects.
    - Table 3 shows that the time trend, conditional on the average inspection score, does not systematically predict outcomes.
- Section H displays results from the investigation into predictive power going back multiple rounds of inspections.
    - Figure 5 plots the magnitude and 95% confidence interval of regression coefficients, showing that there is a sharp break in marginal predictive power around 4-5 prior inspections.
- Section I shows that area rotations do not substantially affect the average critical score of an inspector, meaning that inter-inspector differences dwarf area differences.
    - Figure 6 shows that inspector differences persist across area rotations. These findings justify particular attention to account for inter-inspector variability rather than inter-area variability.
- Section J provides a formal description of unadjusted and adjusted grading systems.
- Section K describes the easy-to-use software we make available in the R language to implement adjusted grading.
- Section L calculates the grade distribution for 60 establishments subject to full investigations with probable or confirmed instances of foodborne illness under both unadjusted and adjusted grading.

A. Lab-Confirmed Foodborne Illness and Violations

	Lab-Confirmed Foodborne Illness		
	Yes	No	Difference
Critical point score	18.42	9.95	8.47^{**}
	(3.67)	(0.07)	(3.67)
Non-critical point score	6.40	2.98	3.42^{***}
	(1.09)	(0.02)	(1.09)
N	57	51,757	

Table 1: Correlation between number of critical and non-critical violations and probable or lab-confirmed cases of foodborne illness based on full investigations. Each cell presents the conditional mean with standard errors in parentheses below. The "Difference" column indicates the difference in points between establishments with lab-confirmed foodborne illnesses and those without. $^{**}/^{***}$ indicate statistical significance at 0.05 and 0.01-levels, respectively, using a difference-in-means *t*-test.

B. Predictive Power of Critical Score

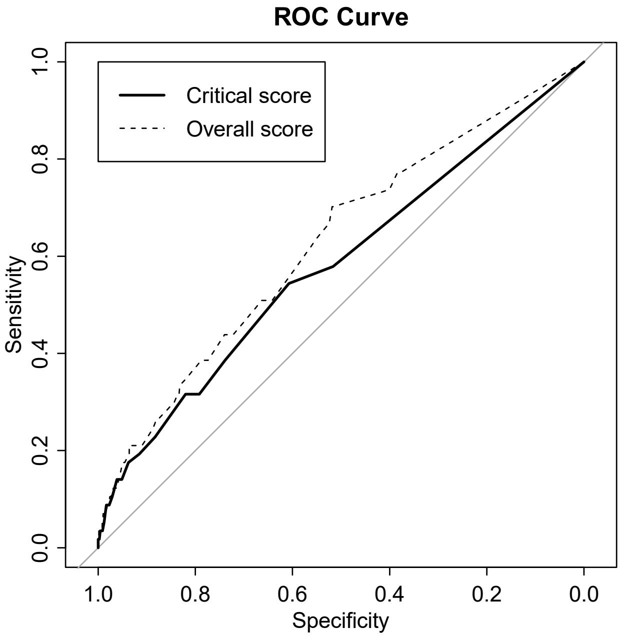

Figure 1: Receiver operating characteristic (ROC) curve of logistic regression model predicting probable or lab-confirmed foodborne illness outbreaks. The solid line represents the ROC curve for a model with critical points as the explanatory variable. The dashed line represents the ROC curve for a model with total points (the sum of critical and non-critical points) as the explanatory variable. While both predictors are statistically significant (*p*-value < 0.01), the substantive predictive power is low. For instance, sensitivity (the true positive rate) at 50% has a specificity (true negative rate) of only 61-67%.

C. Predicted Probability of Investigation

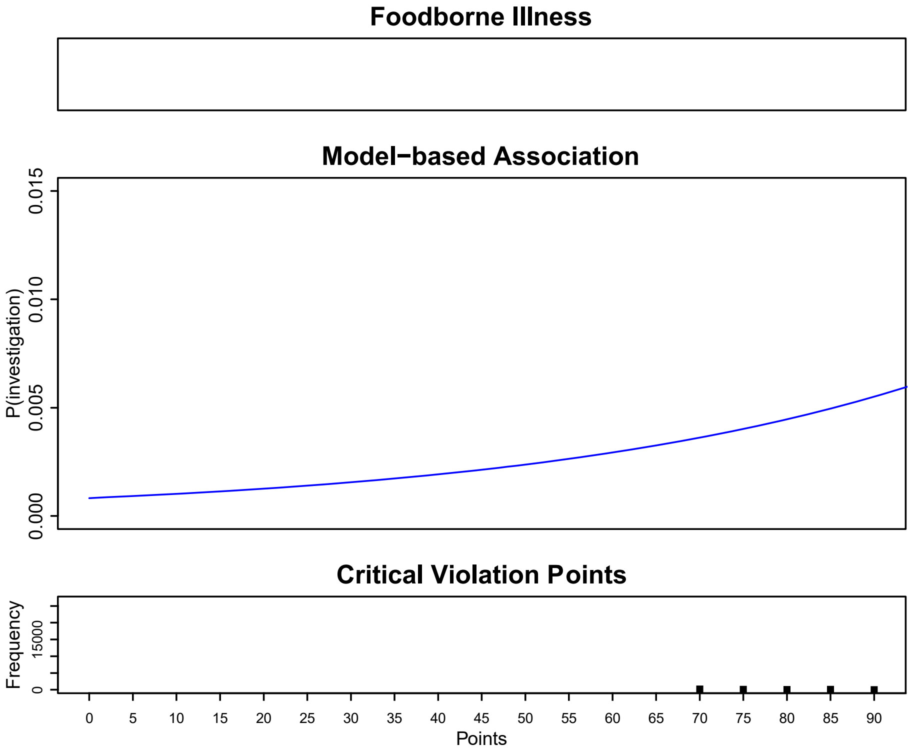

Figure 2: Correlation between critical violation points and probable or lab-confirmed cases of foodborne illness based on full investigations. The bottom panel plots the histogram of critical violation points of all establishments. The top panel plots the critical violation points in the routine inspection immediately preceding the case of foodborne illness. The middle panel plots the model-based association, using a logistic regression with foodborne illness as the outcome and critical violation points as the explanatory variable. The curve plots the predicted probability, with 95% confidence intervals. The coefficient is statistically significant (p-value < 0.001), but because the baseline rate of foodborne illnesses traced back to an establishment is so low, the substantive predictive power is low.

D. Reliability of Critical vs. Non-Critical Violations

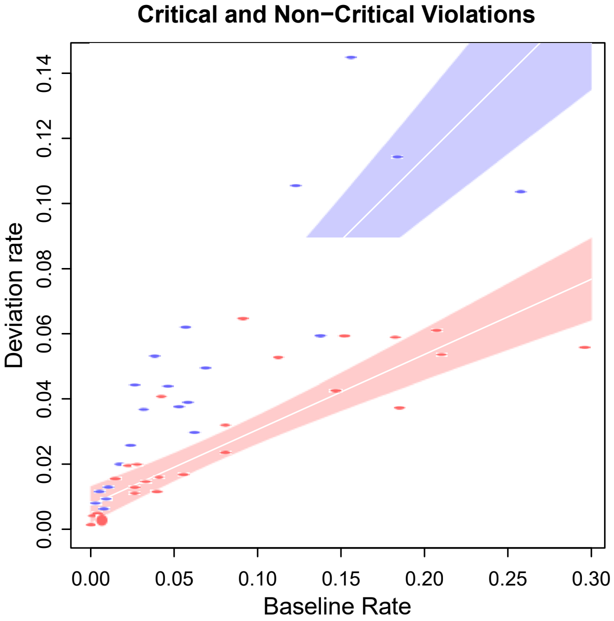

Figure 3: Results from 378 peer review inspections. The *x*-axis plots the baseline rate at which each violation was cited and the *y*-axis plots the rate at which two inspectors observing the same conditions deviated on whether or not to cite the violation. Red (blue) corresponds to critical (non-critical) violations, and the bands present correlation from a simple linear fit separate to critical and non-critical violations, with 95% confidence intervals. Critical violations exhibit much lower deviation rates, so that basing a grade on critical violations has a better public health rationale and improves reliability of grades.

E. Regression Tests of Critical Violation Reliability

	Model 1	Model 2	Model 3
Critical violation	0.43^{***}	0.59^{***}	0.65^{***}
	(0.08)	(0.08)	(0.08)
Baseline citation rate		-7.83^{***}	-8.64^{***}
		(0.43)	(0.46)
Business FE	No	No	Yes
No. businesses	378	378	378
No. violations	19,656	19,656	19,565
AIC	5912	5613	5275

Figure 4: Logistic regression estimates of whether critical violations are more reliably inspected, based on 378 peer review inspections. Each observation represents one of 52 violations during 378 peer review inspections, with a binary outcome of whether two inspectors observing the same conditions agreed on citation of the violation. Coefficients are presented with standard errors in parentheses. *** indicates statistical significance at the 0.01-level. In each model, critical violations exhibit statistically significantly higher rates of agreement. Business FE indicates business fixed effects and AIC indicates the Akaike Information Criterion.

F. Matched Samples Analysis of Repeat Violations

Methods

We analyzed inspection data for King County businesses with “risk level 3” permits (highest risk category) with at least one inspection score in the year of interest (specified in the “Year” column in **Error! Reference source not found.**), and with at least two subsequent inspections (between January 1 in the year of interest and July 2016). We matched businesses with the same inspection scores in the first and second rounds of inspections (with each unique set of first and second round scores corresponding to one stratum), and identified, as members of a treatment group, those businesses in each stratum that were cited for the same violation in the first and second inspections.

Denote the total number of treatment businesses in year T by N_1 (omit T indices on all variables to simplify notation, although each variable is also dependent on year), the number in stratum j by N_{1j}, the number of control businesses in stratum j by N_{0j}, and the estimators for the mean third round inspection scores in stratum j by $\bar{Y}_{1j}$ and $\bar{Y}_{0j}$, for the treatment and control groups respectively. As described in Imbens and Rubin,[1] we calculate estimators for the mean inspection scores in the third round of inspections, $\bar{Y}_1$ and $\bar{Y}_0$, for the treatment and control groups, as follows:

$$\bar{Y}_1 = \sum_j \left(\frac{N_{1j}}{N_1}\right)\bar{Y}_{1j} = \sum_j \left(\frac{N_{1j}}{N_1}\right)\left(\frac{1}{N_{1j}}\right)\sum_{k=1}^{N_{1j}} Y_{1jk} \tag{1}$$

$$\bar{Y}_0 = \sum_j \left(\frac{N_{1j}}{N_1}\right)\bar{Y}_{0j} = \sum_j \left(\frac{N_{1j}}{N_1}\right)\left(\frac{1}{N_{0j}}\right)\sum_{k=1}^{N_{0j}} Y_{0jk}$$

where Y_{1jk} is the observed third round inspection score for the kth business in the treatment group of stratum j, and similarly Y_{0jk} is the third round inspection score for the kth business in the control group of stratum j. $\bar{Y}_1$ and $\bar{Y}_0$ are recorded in **Error! Reference source not found.** for years T between 2006 and 2014. The average treatment effect on the treated is reported in **Error! Reference source not found.**:

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0 = \sum_j \left(\frac{N_{1j}}{N_1}\right) \hat{\tau}_j \,, \tag{2}$$

where $\hat{\tau}_j = \bar{Y}_{1j} - \bar{Y}_{0j}$, and is the treatment effect within each stratum.

To test the null hypothesis that the average treatment effect, $\hat{\tau}$, is zero, we calculate the generalized Neyman sampling variance for each stratum[1]:

$$\widehat{\mathbb{V}}(\hat{\tau}_j) = \frac{1}{N_{0j}(N_{0j}-1)} \sum_{k=1}^{N_{0j}} \left(Y_{0jk} - \bar{Y}_{0j}\right)^2 + \frac{1}{N_{1j}(N_{1j}-1)} \sum_{k=1}^{N_{1j}} \left(Y_{1jk} - \bar{Y}_{1j}\right)^2,$$

and we calculate the sampling variance of $\hat{\tau}$ by summing the within-stratum variances when each is weighted by the square of the proportion of treatment businesses within the stratum:

$$\widehat{\mathbb{V}}(\hat{\tau}) = \sum_j \left(\frac{N_{1j}}{N_1}\right)^2 \widehat{\mathbb{V}}(\hat{\tau}_j) \,. \tag{3}$$

We weight each stratum's contribution to the variance (as well as the treatment effect, and mean inspection scores) by the proportion of treatment restaurants.

Finally, the resulting t-statistic we use is[1]:

$$t = \frac{\hat{\tau}}{\sqrt{\widehat{\mathbb{V}}(\hat{\tau})}} \,. \tag{4}$$

Results

Table 1 presents results for matched sample analyses from 2006-2014. Results are quite mixed for the period from 2006-2012, where the 2008 data might even suggest that repeat violators perform better than the control group. In 2013 and 2014, there is some evidence that repeat violators perform worse. We interviewed county officials to understand why such a difference might exist. The major difference is that around 2013, the county adopted an electronic tablet system, as well as an online dashboard, which facilitated looking up violation history, making the presence of repeat violations more salient. It is highly plausible that such salience might affect inspection conduct in the third inspection: if inspectors believed repeat violators to be "bad apples," they may cite more violations in the third inspection, even conditional on the same risk factors present. As a result, we believe that the 2006-12 period provides a cleaner test of the repeat violation hypothesis. In any case, because the results are mixed, it does not provide a strong evidence base for using repeat violations as an input to the grading system.

	Treated		**Control**		Avg. treatment effect
Year	Avg. score	N_1	Avg. score	N_0	
2006	13.03	743	13.59	1439	-0.55
2007	13.36	758	12.62	910	0.75
2008	13.06	575	14.68	1131	-1.62*
2009	13.52	730	12.30	873	1.22
2010	12.71	770	12.20	967	0.52
2011	13.29	691	14.40	872	-1.11
2012	14.33	705	14.17	974	0.16
2013	17.30	688	15.05	910	2.25**
2014	19.21	851	17.12	954	2.10**

Table 2: Treatment effects from separate matched sample analyses of repeat violations form 2006-14. The first two columns present (weighted) average scores and sample sizes for the matched treated units and the next two columns present (weighted) average scores and sample sizes for the matched control units. The last column indicates the estimate of the average treatment effect on the treated. */** indicate statistical significance at the 0.1 and 0.05 levels, respectively.

G. Matched Samples Analysis of Order Effects

Year	No. businesses, Treatment	No. businesses, Control	Score, Treatment	Score, Control	Score Difference (Treatment – Control)
2006	1818	1884	11.19	11.40	-0.21
2007	1949	1863	11.53	10.81	0.72
2008	1886	1892	10.42	10.55	-0.13
2009	1947	1803	11.49	10.05	1.44^{***}
2010	1730	2072	9.89	8.98	0.91^{**}
2011	1789	1832	10.49	10.07	0.42
2012	1938	1854	11.04	10.55	0.48
2013	1890	1800	13.76	13.19	0.57
2014	1769	1934	14.38	14.07	0.31

Table 3: Matched Samples and Inspection Score Trends. We analyzed inspection data for level 3 permit businesses with at least one inspection score in the year of interest (specified in the "Year" column above), and with at least two subsequent inspections (between January 1 in the year of interest and July 2016). We matched businesses with the same inspection scores in the first and second rounds of inspections, and sorted businesses into treatment and control groups based on whether scores across the first and second rounds were increasing (treatment) or decreasing (control). We then calculated the mean inspection score in the third round of inspections for both treatment and control groups, as in Equation (1) of Section F, as well as the treatment effect as in Equation (2). Finally, we calculated the variance for the score difference in order to test the null hypothesis, as in Equations (3) and (4) of Section F. **/*** indicate statistical significance at the 0.05 and 0.01 level, respectively. Inspection score trend is generally not a good predictor of future performance.

H. Analysis of Time Periods

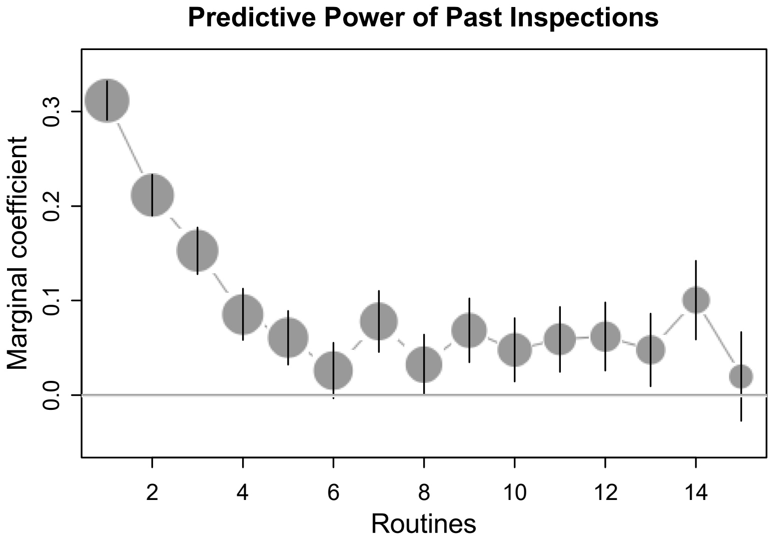

Figure 5: Analysis of historical predictive power. The figure presents the marginal coefficient estimates from least squares models regressing the most recent inspection score on prior inspection scores. Each model sequentially adds an additional round of inspections, and the dots represent the coefficient point estimate with 95% confidence interval in vertical lines, weighted by the number of establishments with the requisite number of inspections. We can see that the marginal predictive power decreases, and levels off sharply around 4-5 routine inspections.

I. Persistence of Inspector Differences across Different Areas

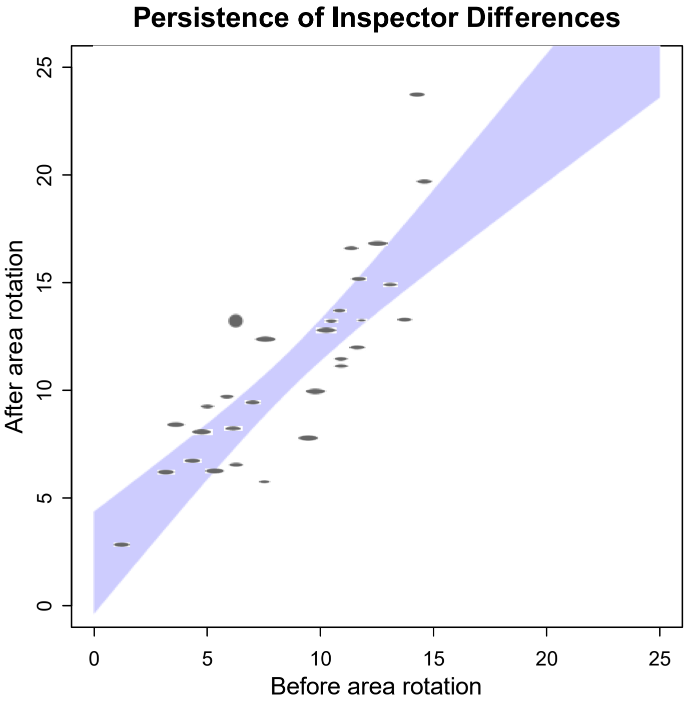

Figure 6: Correlation of inspector average critical score before an area rotation (2012-13) and after area rotation (2014-15). Each dot represents one inspector, weighted by the average number of inspections conducted across both periods to account for sampling variability. The line indicates prediction from a least squares fit, with 95% confidence interval. This figure demonstrates that differences in food safety quality across areas are dwarfed by inter-inspector differences. Regardless of the rotation, inspectors center their scores around the pre-rotation mean.

J. Formal Description of Grading Systems

Let us encode restaurant information within matrix $\boldsymbol{X}$ and vector $\boldsymbol{z}$, with matrix $\boldsymbol{X}$ being of dimensions $n \times 4$ and vector $\boldsymbol{z}$ being of length n, where n is the number of restaurants to be graded (in our case, the number of high risk restaurants in King County). Entry $\boldsymbol{X}_{ij}$ is the inspection score for restaurant i in the jth most recent inspection, while $\boldsymbol{z}_i$ is the ZIP code for restaurant i (although in principle, $\boldsymbol{z}_i$ could represent any unit of aggregation that is meaningful within the grading system, e.g., inspector assignment areas, census tracts, municipalities, or district offices). For example, imagine that restaurant *A* in ZIP code 10001 scored 5, 5, 1 and 2 points in its most recent, second most recent, third most recent and fourth most recent inspections respectively; and that restaurant *B* in ZIP code 10002 scored 3, 4, 5, and 10 points in its most recent, second most recent, third most recent and fourth most recent inspections respectively (these are artificial scores and should not be associated with real restaurants in these ZIP codes). Then matrix $\boldsymbol{X}$ would read:

$$\boldsymbol{X} = \begin{bmatrix} 5 & 5 & 1 & 2 \\ 3 & 4 & 5 & 10 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

and vector $\boldsymbol{z}$:

$$\boldsymbol{z} = \begin{bmatrix} 10001 \\ 10002 \\ \vdots \end{bmatrix}.$$

Let us also store grade cutoff scores in vector $\boldsymbol{\Gamma}$, in which $\boldsymbol{\Gamma}_1$ is the A/B cutoff score and $\boldsymbol{\Gamma}_2$ is the B/C cutoff score. An example set of cutoffs could be:

$$\boldsymbol{\Gamma} = \begin{pmatrix} 10 \\ 20 \end{pmatrix}.$$

In conventional restaurant grading systems (unadjusted grading), if we let $g(\boldsymbol{X}_{i1}, \boldsymbol{\Gamma})$ represent the grade awarded to restaurant i, then the grades for restaurants *A* and *B* are calculated according to the following rule:

$$g(\boldsymbol{X}_{i1}, \boldsymbol{\Gamma}) = \begin{cases} A, & \boldsymbol{X}_{i1} \leq \boldsymbol{\Gamma}_1 \\ B, & \boldsymbol{\Gamma}_1 < \boldsymbol{X}_{i1} \leq \boldsymbol{\Gamma}_2. \\ C, & \boldsymbol{X}_{i1} > \boldsymbol{\Gamma}_2 \end{cases} \quad (1)$$

From Equation (1), unadjusted grading only uses the most recent inspection score for grade assignment and grade cutoff vectors are independent of ZIP code. (Restaurants *A* and *B* would both receive ‘A’ grades in this scheme, since 5 and 3 are both less than or equal to 13).

Quantile Adjustment (with “Ties Resolution”)

In our proposed grading system (adjusted grading), we replace $\boldsymbol{X}_{i1}$ with $\overline{\boldsymbol{x}}_i$ in (1), where $\boldsymbol{x}_i$ is the ith row vector in matrix $\boldsymbol{X}$ and $\overline{\boldsymbol{x}}_i$ is the mean inspection score for restaurant i over its four most recent inspections (or fewer if it has not yet been subject to as many). Furthermore, $\boldsymbol{\Gamma}$ is no longer independent of ZIP code. In particular, let $\boldsymbol{\gamma}$ be a vector of percentiles of length 2 with $\boldsymbol{\gamma}_1 < \boldsymbol{\gamma}_2$. Let $\boldsymbol{u}(\boldsymbol{z}_i)$ be the vector of unique mean (critical) inspection scores for ZIP code $\boldsymbol{z}_i$ of length n_{z_i}, and without loss of generality, let us assume that scores are ordered from smallest to largest. Let vector $\boldsymbol{w}(\boldsymbol{u}(\boldsymbol{z}_i))$ contain the weights associated with each mean score in ZIP code $\boldsymbol{z}_i$, i.e., let $\boldsymbol{w}_j(\boldsymbol{u}(\boldsymbol{z}_i))$, the jth element of $\boldsymbol{w}(\boldsymbol{u}(\boldsymbol{z}_i))$, be the proportion of restaurants in ZIP code $\boldsymbol{z}_i$ with score $\boldsymbol{u}_j(\boldsymbol{z}_i)$. Then the grade cutoffs for ZIP code $\boldsymbol{z}_i$ in our adjusted grading system are:

$$\boldsymbol{\Gamma}(\boldsymbol{z}_i, \boldsymbol{\gamma}) = \begin{pmatrix} \boldsymbol{u}_{f(1)}(\boldsymbol{z}_i) \\ \boldsymbol{u}_{f(2)}(\boldsymbol{z}_i) \end{pmatrix}, \tag{2}$$

where

$$f(x) = \begin{cases} \min\limits_{l} \left| \boldsymbol{\gamma}_1 - \sum\limits_{j=1}^{l} \boldsymbol{w}_j\big(\boldsymbol{u}(\boldsymbol{z}_i)\big) \right| & \boldsymbol{x} = 1 \\ \min\limits_{l} \left| (\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) - \sum\limits_{j=f(1)+1}^{l} \boldsymbol{w}_j\big(\boldsymbol{u}(\boldsymbol{z}_i)\big) \right| & \boldsymbol{x} = 2 \end{cases} \tag{3}$$

In words, $f(1)$ is the index that minimizes the absolute difference between desired proportion $\boldsymbol{\gamma}_1$ for an “A” grade and the proportion of restaurants in $\boldsymbol{z}_i$ with scores less than or equal to $\boldsymbol{u}_{f(1)}(\boldsymbol{z}_i)$. $f(2)$ returns the index that minimizes the absolute difference between desired proportion $(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1)$ for a “B” grade and the proportion of restaurants in $\boldsymbol{z}_i$ with scores between $\boldsymbol{u}_{f(1)}(\boldsymbol{z}_i)$ and $\boldsymbol{u}_{f(2)}(\boldsymbol{z}_i)$.

The grade awarded to restaurant i is then:

$$g\left(\bar{\boldsymbol{x}}_i, \boldsymbol{\Gamma}(\boldsymbol{z}_i, \boldsymbol{\gamma})\right) = \begin{cases} A, & \bar{\boldsymbol{x}}_i \leq \boldsymbol{\Gamma}_1(\boldsymbol{z}_i, \boldsymbol{\gamma}) \\ B, & \boldsymbol{\Gamma}_1(\boldsymbol{z}_i, \boldsymbol{\gamma}) < \bar{\boldsymbol{x}}_i \leq \boldsymbol{\Gamma}_2(\boldsymbol{z}_i, \boldsymbol{\gamma}). \\ C, & \bar{\boldsymbol{x}}_i > \boldsymbol{\Gamma}_2(\boldsymbol{z}_i, \boldsymbol{\gamma}) \end{cases} \tag{4}$$

The vector of percentiles, $\boldsymbol{\gamma}$, is independent of ZIP code: the core idea of our adjusted grading scheme is to differentiate as close to the top $\boldsymbol{\gamma}_1$% of restaurants in ZIP code $\boldsymbol{z}_i$ from the middle $(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1)$ % and the bottom $(\mathbf{1} - (\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_1))$%, which cannot be done with unadjusted grading due to inspector differences. The specific $\boldsymbol{\gamma}$ values are chosen so that overall proportions of A/B/C grades across the entire county are the same (or within a certain tolerance) as those for unadjusted grading (at the time of grading), when the cutoff scores for unadjusted grading are selected to be meaningful values within the county's inspection system (e.g., a closure threshold).

Quantile Adjustment – Alternative Implementation ("Percentile Method")

In our software package, we provide an alternative mapping between percentiles $\boldsymbol{\gamma}$ and ZIP code cutoffs, $\boldsymbol{\Gamma}(\boldsymbol{z}_i, \boldsymbol{\gamma})$**,** to the one outlined in equations (2) and (3). In particular, let $\boldsymbol{\gamma}$ continue to represent a vector of percentiles of length 2 with $\boldsymbol{\gamma}_1 < \boldsymbol{\gamma}_2$; but now let $\mathcal{M}(\boldsymbol{z}_i)$ represent the set of all mean inspection scores for ZIP code $\boldsymbol{z}_i$, ordered from smallest to largest ($\mathcal{M}(\boldsymbol{z}_i)$ is different to $\boldsymbol{u}(\boldsymbol{z}_i)$, which is the vector of all *unique* mean scores in ZIP code $\boldsymbol{z}_i$). Let N_{z_i} be the number of restaurants to be graded in ZIP code $\boldsymbol{z}_i$. Then the grade cutoffs for ZIP code $\boldsymbol{z}_i$ in our adjusted grading system are:

$$\boldsymbol{\Gamma}(\boldsymbol{z}_i, \boldsymbol{\gamma}) = \begin{pmatrix} \lceil \boldsymbol{\gamma}_1 \times (N_{z_i}) \rceil \text{th element of } \mathcal{M}(\boldsymbol{z}_i) \\ \lceil \boldsymbol{\gamma}_2 \times (N_{z_i}) \rceil \text{th element of } \mathcal{M}(\boldsymbol{z}_i) \end{pmatrix}, \tag{5}$$

where $\lceil \ldots \rceil$ represents the ceiling function. The grade awarded to restaurant i is then calculated using equation (4).

A motivation for adopting the first "Ties Resolution" method over the second, arguably simpler, "Percentile" method is illustrated in Figure 7. In particular, we display part of the cumulative distribution function for the Tukwila ZIP code when it was inspected by a lenient inspector in the pre-rotation period. Tukwila has a total of 176 restaurants to be graded, and 27 of these businesses have a mean inspection score, $\bar{\boldsymbol{x}}_i$, of 2.5. The problem with the percentile method is demonstrated if the desired proportion of restaurants to gain 'A' grades, $\boldsymbol{\gamma}_1$, falls between 0.52 and 0.595. In this instance, the returned A/B cutoff for Tukwila, $\boldsymbol{\Gamma}(\text{Tukwila}, \boldsymbol{\gamma_1})$, calculated by the percentile method, is 2.5; and 67% of restaurants in Tukwila gain an 'A' grade. This is despite the fact that choosing 1.25 as the A/B cutoff results in 52% of restaurants scoring 'A's, which is closer to the percentage of restaurants gaining 'A' grades in other ZIP codes (most other ZIP codes do not have such large ties problems, so have proportions closer to the desired $0.52 < \boldsymbol{\gamma}_1 < 0.595$). If $\boldsymbol{\gamma}_2 = 0.9$, 23% of restaurants in Tukwila gain a "B" grade with the percentile method (the ties problem is not an issue for the upper end of the Tukwila score distribution), while this is closer to, depending on the choice of $\boldsymbol{\gamma}_1$, 31% - 38% of restaurants in other ZIP codes. With such a large difference in the proportion of 'B' grades between Tukwila and other ZIP codes, the B/C cutoff in Tukwila seems an arbitrary choice. In comparison, the "Ties Resolution" method, for the same $\boldsymbol{\gamma}_1$, returns 1.25 as $\boldsymbol{\Gamma}(\text{Tukwila}, \boldsymbol{\gamma_1})$, and selects the B/C cutoff so that as close as is possible to $(0.9 - \boldsymbol{\gamma}_1)$% of restaurants gain "B" grades. In order to minimize geographic differences in the presence of ties in ZIP code score distributions, we prefer quantile adjustment with ties resolution. This is the default method applied inside the '`ibqgFxwriiv`' function of our software package.

Additional Implementation Details for the Quantile Grading System

While the majority of establishments are graded according to the protocol described above, there are some edge cases that we discuss here. Firstly, in the case that a ZIP code has fewer than 10 establishments, we aggregate inspection scores for establishments in neighboring ZIP codes before calculating cutoffs. This ensures that cutoffs (and, thus, grades) are less likely to dramatically change as a result of a single inspection for a single establishment.

Secondly, while the analyses presented in text focus on the riskiest establishments ("level 3" permit holders), Seattle-King County rolled out grading on all food establishments within the county, with the exception of food trucks and grocery stores. Cutoffs are calculated separately for level 3 establishments compared to for level 1 and 2 establishments to compare businesses of a similar food complexity to one another (establishments in levels 1 and 2 compared to other establishments in levels 1 and 2, and establishments in level 3 compared to other establishments in level 3. This is valuable because the levels, and food complexity that they represent, result in some violation types being applicable to level 3 establishments that would not be for levels 1 and 2. This then enables greater differentiation of performance between business that are like one another. Because of the relatively small number of level 1 and level 2 establishments, cutoffs were calculated by pooling the two levels together.

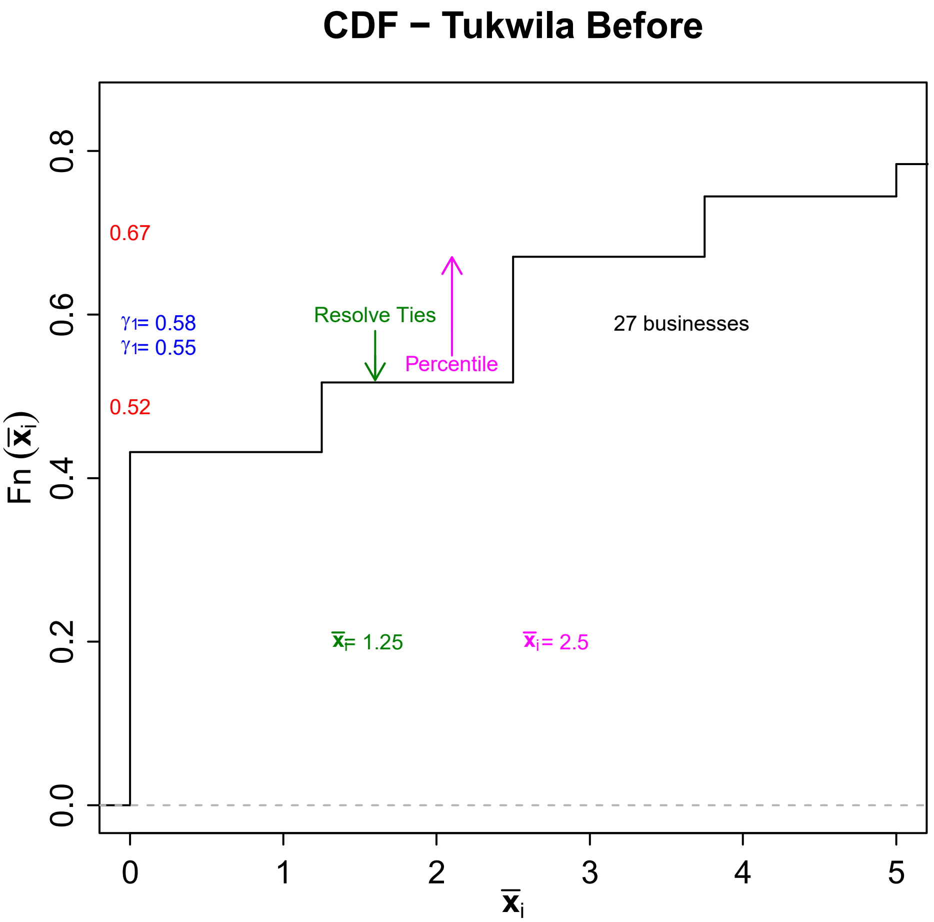

Figure 7: Illustration of the “Tukwila Ties Problem” and both methods of Quantile Adjustment (Quantile Adjustment with Ties Resolution, and the Percentile Method). We plot the empirical cumulative distribution function for mean inspection scores, $\bar{x}_i$, in the Tukwila ZIP code. If the desired global proportion of restaurants to gain an ‘A’ grade is $0.52 < \gamma_1 < 0.595$, the percentile method returns $\Gamma(\text{Tukwila}, \gamma_1) = 2.5$, which results in 67% of restaurants gaining ‘A’ grades. In comparison, $\Gamma(\text{Tukwila}, \gamma_1) = 1.25$ for the ‘Resolve Ties’ method and 52% of restaurants gain an ‘A’ grade. The ‘Resolve Ties’ method is better at reducing geographic differences in grade proportions than the percentile method, and correspondingly is the default method within our software.

K. Software

To easily implement the grading system in any jurisdiction, we have designed an open source statistical software package called "QuantileGradeR" written in the R language.[2] The package is available at https://cran.r-project.org/web/packages/QuantileGradeR/index.html. This package enables the calculation of $\boldsymbol{\Gamma}(\boldsymbol{z}_i, \boldsymbol{\gamma})$, the vector of grade cutoffs, for each ZIP code $\boldsymbol{z}_i$, as well as adjusted grades, $g(\overline{\boldsymbol{x}}_{\boldsymbol{i}}, \boldsymbol{\Gamma}(\boldsymbol{z}_{\mathrm{i}}, \boldsymbol{\gamma}))$, and unadjusted grades, $g(\boldsymbol{X}_{i1}, \boldsymbol{\Gamma})$. To integrate easily with King County's EnvisionConnect system, we anticipate that the package will be used to calculate grade cutoffs for ZIP codes, with the table imported into the database. However, QuantileGradeR is a standalone package that can be used to grade restaurants entirely on its own: as input, all that is required is matrix $\boldsymbol{X}$, vector $\boldsymbol{z}$ and the vector of natural grade cutoff values for the inspection system, $\boldsymbol{\Gamma}$.

The central functions contained in the package are 'ilqgFxwriiv' and 'judghDooExv'. Within the package we also provide two anonymized King County data samples: '[1nf' is a toy matrix $\boldsymbol{X}$ and '}lsv1nf' is a toy vector $\boldsymbol{z}$. To call the central functions and perform adjusted grading on the example datasets is easy. The user first locates grade cutoffs in each ZIP code:

```
zip.cutoffs <- findCutoffs(X.kc, zips.kc, c(0, 30)),
```

and grades are then calculated using the cutoff scores,

```
grades <- gradeAllBus(rowMeans(X.kc, na.rm = TRUE),
                      zips.kc, zip.cutoffs).
```

QuantileGradeR is a versatile package: the number of grade classifications is not limited to three (simply increasing or decreasing the number of inspection system relevant values within the $\boldsymbol{\Gamma}$ vector will alter the number of grade classifications); and both methods of quantile adjustment (the 'Resolve Ties' and 'Percentile' methods outlined in Appendix J) can be readily implemented ('Resolve Ties' is the default option, but the Percentile method is invoked by setting 'uhvroyh1wlhv' to IDOVH when calling 'ilqgFxwriiv'). Furthermore, our package is not limited to grading restaurants, nor to performing the percentile adjustment on the ZIP code level – the only requirement for $\boldsymbol{X}$ is that it is an $n \times p$ numerical matrix, where n is the number of entities to be graded and p is the number of scores that should be averaged to calculate $\overline{\boldsymbol{x}}_{\boldsymbol{i}}$ in the adjusted system. Similarly, $\boldsymbol{z}$ need only be a character vector of length n. Although we have designed the package with King County in mind, the package can also be readily used in jurisdictions where higher scores correspond to reduced risk: inspection scores should be transformed before ‘ilqgFxwriiv’ or ‘judghDoExv’ are called with a simple transformation function like $f(score) = -score$. The resulting cutoff values can be transformed back, if desired, using (in this case) the same transformation function.

L. Grades and Foodborne Illness

	A	B	C
Unadjusted	24	24	12
Adjusted	22	25	13

Table 4: Incidence of probable or confirmed foodborne illness from 2012-May 2016 across establishments by grading system. Each row indicates the distribution of grades existing at the time of the illness under the unadjusted or adjusted grading system. The adjustment moves two establishments from the ‘A’ to the ‘B’ category, and one from the ‘B’ to the ‘C’ category.

SUPPLEMENTAL REFERENCES

1. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press; 2015.

2. Team RC. *R: A Language and Environment for Statistical Computing [Computer Software]. Vienna: R Foundation for Statistical Computing*.; 2016.