Ashwood, Z.C., Elias, B., & Ho, D.E. (2021). Improving the reliability of food safety disclosure:

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

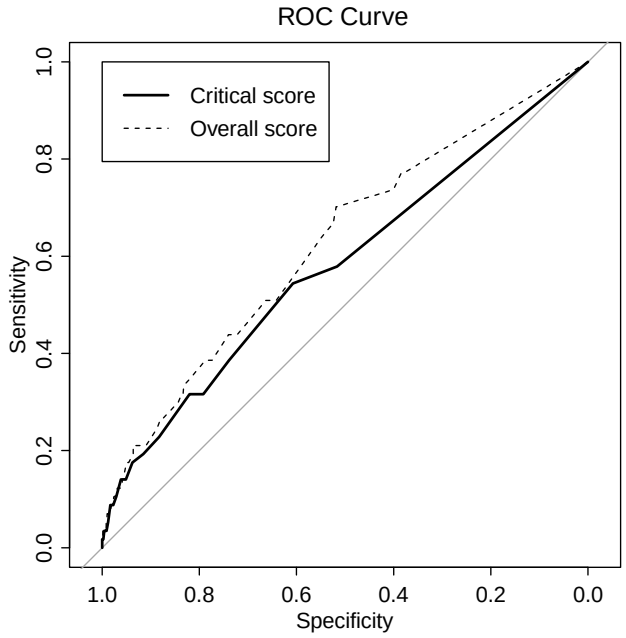


Figure 1: Receiver operating characteristic (ROC) curve of logistic regression model predicting probable or lab-confirmed foodborne illness outbreaks. The solid line represents the ROC curve for a model with critical points as the explanatory variable. The dashed line represents the ROC curve for a model 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

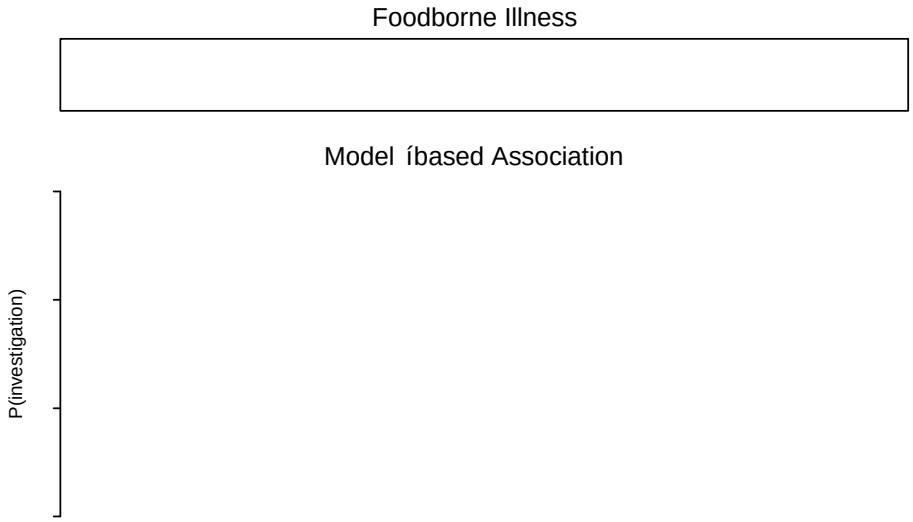


Figure 2: Correlation between critical violation points and probable or lab-confirmed cases of foodborne illness based on full investigations. The bottom panel plots the histogram of critical violation points of all establishments. The top panel plots the critical violation points in the routine inspection 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		

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 6 by 0_5 (omit 6 indices on all variables to simplify notation, although each variable is also dependent on year), the number in stratum F by $0_{5\,\acute{y}}$ the number of control businesses in stratum F by $0_{4\,\acute{y}}$ and the estimators for the mean

[redacted]

$\$_{5\,\acute{y}}$ and $\$_{4\,\acute{y}}$ for the treatment and control groups

bin,[1] we calculate estimators for the mean

ons, $\$_5$ and $\$_4$ for the treatment and control

ì¸ L ;5; F ;4 L Í l 05Ý / 05 p ìÝ,   (2)

where ìÝ L $;5Ý F $;4Ý

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 i

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

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

I. Persistence of Inspector Differences across Different Areas

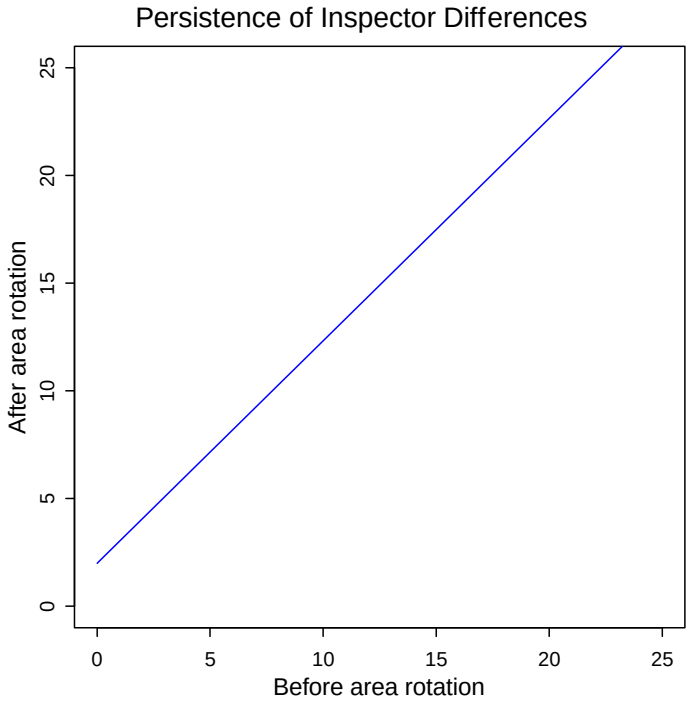


Figure 6: Correlation of inspector average critical score before an area rotation (2012-13) and after area rotation (2014-15). Each dot represents one inspector, weighted by the average number of inspections conducted across both periods to account for sampling variability. The line indicates prediction 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 „ and vector , with matrix „ being of dimensions J H 4 and vector being of length J, where Jis the number of restaurants to be graded (in our case, the number of high risk restaurants in King County). Entry „ ü ¡is the inspection score for restaurant Ein the Eth most recent inspection, while üjis the ZIP code for restaurant E(although in principle, üjcould 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 „ would read:

$$\text{„} = \begin{bmatrix} 5 & 5 & 1 & 2 \\ 3 & 4 & 5 & 10 \end{bmatrix}$$

and vector :

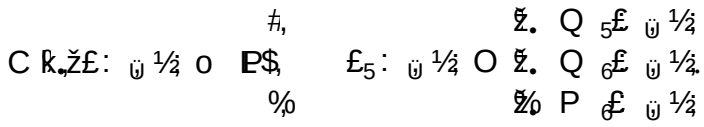
$$= \begin{bmatrix} 10001 \\ 10002 \end{bmatrix}$$

Let us also store grade cutoff scores in vector £, in which £5is the A/B cutoff score745 TD24.44 0 TD0 T9.400

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 $„_{ü\,s}$ with $\%_{ü}$ in (1), where $ž_{ü}$ is the Eth row vector in matrix $„$ and $\%_{ü}$ is the mean inspection score for restaurant E over its four most recent inspections (or fewer if it has not yet been subject to as many). Furthermore, $£$ is no longer independent of ZIP code. In particular, let $½$ be a vector of percentiles of length 2 with $½_{s} = 0$ $½_{6} = 1$. Let $›_{:ü;}$ be the vector of unique mean (critical) inspection scores for ZIP code $ü$ of length $J_{í_{ü}}$ and without loss of generality, let us assume that scores are ordered from smallest to largest. Let vector $•_{:›_{:ü;};}$ contain the weights associated with each mean score in ZIP code $ü$ i.e., let $•_{ý:›_{:ü;};}$ the Eth element of $•_{:›_{:ü;};}$ be the proportion of restaurants in ZIP code $ü$ with score $›_{ý:ü;}$. Then the grade cutoffs for ZIP code $Í@$



(4)

The vector of percentiles, ½is independent of ZIP code: the core idea of our adjusted grading scheme is to differentiate as close to the top ⅛% of restaurants in ZIP code ü from the middle

businesses have a mean inspection score, $\bar{x}$, of 2.5. The problem with the percentile method is demonstrated if the desired proportion of restaurants to gain 'A' grades, p, falls between 0.52 and 0.595. In this instance, the returned A/B cutoff for Tukwila, $c_{Tukwila,p}$, 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 $0.52 < p < 0.595$). If $p > 0.9$, 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 p, 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 p, returns 1.25 as $c_{Tukwila,p}$, and selects the B/C cutoff so that as close as is possible to $1 - p = 10\%$ 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 '`ibqgFxwriv`' 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

in level 3. This is valuable because the levels, an

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 £:ڽ½, the vector of grade cutoffs, for each ZIP code ڽ as well as adjusted grades, C:%, £:g½; and unadjusted grades, C:„ڽ,5£;. To integrate easily with King County's EnvisionConnect system, we anticipate that the package will

requirement for „ is that it is an J H Lnumerical matrix, where Jis the number of entities to be graded and Lis the number of scores that should be averaged to calculate 5% in the adjusted system. Similarly, need only be a character vector of length J. Although we have designed the package with King County in mind, the package can also be readily used in jurisdictions where

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.