FEASIBLE POLICY EVALUATION BY DESIGN: A RANDOMIZED SYNTHETIC STEPPED-WEDGE TRIAL IN KING COUNTY

Cassandra Handan-Nader
Department of Political Science, Stanford University

Daniel E. Ho
Senior Fellow, SIEPR

Becky Elias
Public Health, Seattle & King County

November, 2018
Working Paper No. 18-040
Feasible Policy Evaluation by Design:
A Randomized Synthetic Stepped-Wedge Trial in King County

Cassandra Handan-Nader†  Daniel E. Ho‡  Becky Elias§

November 16, 2018

Abstract

Evidence-based policy is limited by the perception that randomized controlled trials (RCTs) are expensive and infeasible. We argue that carefully tailored research design can overcome these challenges and enable randomized evaluations of policy implementation. We demonstrate how a stepped-wedge (randomized roll-out) design that adapts synthetic control methods overcame substantial practical, administrative, political, and statistical constraints to evaluating King County’s new food safety rating system. The core RCT component of the evaluation came at little financial cost to the government, allowed the entire county to be treated, and resulted in no functional implementation delay. The case of restaurant sanitation grading has played a critical role in the scholarship on information disclosure, and our study provides the first evidence from a randomized trial of the causal effects of grading on health outcomes. We find that the grading system had no appreciable effects on foodborne illness, hospitalization, or food handling practices, but that the system may have marginally increased public engagement by encouraging higher reporting.

†Research Fellow, Stanford Law School, 559 Nathan Abbott Way, Stanford, CA 94305; Ph.D. student, Department of Political Science, Stanford University
‡Corresponding Author; William Benjamin Scott and Luna M. Scott Professor of Law, Professor (by courtesy) of Political Science; & Senior Fellow at Stanford Institute for Economic Policy Research; Stanford University, 559 Nathan Abbott Way, Stanford, CA 94305; Tel: 650-723-9560; Fax: 650-725-0253; Email: dho@law.stanford.edu
§Former Manager, Food Protection Program, Environmental Health Services, Public Health – Seattle & King County

*We thank Zoe Ashwood for research assistance, Jenny Lloyd (King County) and Karen Wong (CDC) for help in validating hospitalization discharge codes, David Engelskirchen (FDA) and Katey Kennedy (FDA) for help in training our independent inspection team, Phil Wyman and Eyob Mazengia for designing the streamlined risk factor checklist, Ki Straughn, Jennifer Jessen, Hasina Wong, and Dalila Zelkanovic for carrying out the risk factor study, Meagan Kay and Sargis Pogosjans for help in assembling notifiable conditions and complaints data, and Carrie Cihak, Eyob Mazengia, and Sargis Pogosjans for helpful comments. This evaluation was supported exclusively by internal funding from Stanford University, with no compensation from King County or any other entity.
1 Introduction

While evidence-based policy has made substantial strides over the generation (see, e.g., Gueron and Rolston 2013; Haskins and Margolis 2014), vast parts of policymaking remain uninformed by a rigorous evidence base. Haskins and Margolis (2014, p. 4) whimsically suggest that research constitutes only 1% of factors driving the adoption of social legislation. Bridgeland and Orszag (2013) estimate that only $1 of every $100 of federal spending is evidence-based. Rather than “evidence-based policy making,” the default may be “policy-based evidence making” (House of Commons Science and Technology Committee 2006).

A core challenge is that randomized controlled trials (RCTs) – the gold standard of policy evaluation (Imbens 2010; Rubin 2008) – are perceived of as limited in applicability when it comes to substantial questions of law and policy (Abramowicz et al. 2011; Ho 2015; Levitt and List 2009; Nathan 2008). Some perceive RCTs as infeasible (Easterly 2009; Ravallion 2012; Rodrik 2008), when policymakers seek to implement large legal and policy changes wholesale. Many perceive RCTs as slow and expensive (Bothwell et al. 2016). And others charge RCTs as unethical, as, for instance, the (perceived) benefit of treatment should not be withheld (Desposato 2015; Greiner and Matthews 2016). These concerns can be particularly acute from the perspective of public management and administration (Margetts 2011; Perry 2012).

We argue that these critiques stem in part from an unduly restrictive notion of RCTs. Practical research design can overcome many of these challenges to public administration, while retaining the benefits of a randomized design. Implementation typically cannot be carried out instantaneously across a large jurisdiction. Vaccinations cannot be administered to all children simultaneously (Hemming et al. 2015); cash transfers cannot be granted in one fell swoop (Fernald et al. 2008); and universal health care cannot be rolled out instantaneously across a country (King et al. 2007).

As a result, practical constraints often mandate that implementation be sequenced; and sequencing

\footnote{RAND’s landmark 15-year health care experiment, for instance, cost upward of $640M in present dollars (Greenberg and Shroder 1998).}
facilitates rigorous evaluation of the policy during implementation. Our argument is similar to that of King et al. (2007), namely to identify areas of discretion in policy implementation and use research design to facilitate randomized evaluation of implementation.

We illustrate with an RCT we designed to evaluate the food safety rating system adopted by the public health department in Seattle & King County (Public Health – Seattle & King County). In that context, political constraints required that the grading system for restaurants be implemented across the entire county in 2017. Treatment could not be withheld. Randomization at the restaurant level was infeasible, primarily because cases of foodborne illness cannot be attributed to specific establishments at scale. Operational concerns about training inspectors and onboarding restaurants, as well as equity concerns about the impact of grading, loomed large (Ho, 2017b). The initial perception was hence that an RCT would be too expensive, operationally unmanageable, and potentially unfair. These perceptions explain why not a single randomized evaluation of grading exists to date, despite the fact that many jurisdictions have adopted grading policies.

The research design we developed in collaboration with Public Health was aimed to solve these problems. To address the fact that the entire county needed to be treated, we adapt a stepped-wedge (randomized rollout) design, which randomizes the order in which grading is rolled out in four subregions of King County. While prominent in medicine, the design remains underutilized for evaluations of law and policy (Mdege et al. 2011, pp. 938-39). The design also allowed the department to improve operational logistics based on the early implementing regions. Second, we develop an algorithm to construct synthetic regions – which are contiguous and respect municipal boundaries – that are indistinguishable in preexisting foodborne illness trends. We show that adapting insights from synthetic control methods (Abadie et al. 2010) can improve statistical power, addressing a key limitation of cluster-based randomization. Third, to control the false

---

2This kind of implementation evaluation is distinct from the idea of “implementation analyses,” whose principal goal is to assure faithful implementation of a policy initiative (Hollister 2009). Instead, we argue that sequencing inherent in implementation facilitates rigorous randomized evaluation. We do not argue that our research design addresses the full panoply of criticisms of RCTs (see, e.g., Cartwright 2007, Deaton and Cartwright 2017). Our design addressed central constraints that otherwise would likely have led King county to compare only outcomes before and after the intervention.
positive rate, which is particularly problematic in this setting due to regional outbreaks (Ho et al., 2017), we use randomization inference that directly accounts for the randomization scheme (Imbens and Rubin, 2015; Ho and Imai, 2006; Rosenbaum, 2002a,b). We show that in our setting, this approach outperforms conventional (clustering) methods for controlling size and power. Last, we leverage administrative data to minimize the costs of outcome data collection (Buck and McGee, 2015; Coalition for Evidence-Based Policy, 2015).

Our substantive results resolve a major question in the scholarship on mandated information disclosure, for which restaurant grading has proven a pivotal case (Bubb, 2015; Ben-Shahar and Schneider, 2014; Fung et al., 2007; Ho, 2012; Loewenstein et al., 2014). We find that the grading system had no statistically or substantively significant effects on foodborne illnesses or hospitalizations. We corroborate these results by examining the effects of the grading system on risk factors, as measured by an independent inspection team created specifically for this evaluation. The team was trained by the Food and Drug Administration (FDA) and for 18 months observed restaurants (without grading them or issuing any inspection report), including periods before and after grading implementation in three subregions. While this inspection component increased the budget of the evaluation, the evaluation could have been conducted without that line item. We find that the grading system may have reduced the propensity by inspectors to cite critical violations even with comparable sanitation practices, but find suggestive evidence that the grading system sparked more public submissions of food poisoning complaints. The grading system might hence facilitate greater public engagement with food safety and investigation of outbreaks, although the durability of this effect is much less certain.

The paper proceeds as follows. Section 2 provides background on food safety and our collaboration with King County. Section 3 spells out some of the key practical constraints on the evaluation. Section 4 discusses the research design to overcome these constraints. Section 5 presents results and Section 6 concludes.
2 Background

Foodborne illness accounts for some 46 million illnesses, 128,000 hospitalizations, and 3,000 deaths annually in the United States (CDC, 2011). In the past twenty years, an increasingly popular reform has been the disclosure of restaurant inspection results to consumers (Ho, 2012; Kovacs et al., 2018; Seiver and Hatfield, 2000). By simplifying the results of an inspection visit and posting placards in entryways of restaurants, “restaurant grading” promises to inform consumers about food risk and provide market incentives for restaurants to adopt safer sanitation practices (Fung et al., 2007; Weil et al., 2006). Beginning in 2014, the public health department of Seattle & King County (Public Health — Seattle & King County) explored the possibility of a public grading system, with the express goals of reducing foodborne illness and improving risk communication with the public. To provide context for the evaluation and because the process consumed considerable departmental resources, we detail background on the grading system and planning process here.

A comprehensive program review in 2014 recommended that the Public Health “implement a restaurant reporting system that is transparent, credible and intuitive” by ”[c]onducting further research into window placard system” (Public Health – Seattle & King County, 2014, p. 16). At the same time, the program review identified the challenge of maintaining the accuracy and consistency of front-line inspections. The report recommended “an ongoing quality control program …to increase operational consistency and quality in inspections” (id., p. 2).

In response to that program review, Public Health began a process of extensive stakeholder engagement. It constituted a “Food Program Stakeholder Advisory Committee,” comprised of health officials, restaurant representatives, academics, and members of the public. A subcommittee reviewed eight existing restaurant reporting systems. The committee articulated the view that the rating system, in contrast to existing ones, should consider using multiple inspections, high risk violations, relative performance of establishments, and account for effects on the diverse community in the county. The committee also considered the question of the quality and consistency of restaun-
rant inspections, as inconsistencies among front-line inspectors would undermine the reliability of disclosure (Boehnke and Graham 2000; Seiver and Hatfield 2000). Senior food program managers expressed the concern that grading might reduce the propensity by inspectors to cite code violations, as evidence of “grade inflation” exists in many other jurisdictions (Ho 2012; Kovacs et al. 2018).

The department also convened a series of all-staff meetings around restaurant grading and the consistency of the inspection process. One meeting focused on the value of consistency / reliability of the inspection process, expressly connecting the issue to the credibility of the grading system.

To address the problem of the accuracy and consistency of inspections, Public Health developed a “peer review” initiative, whereby inspectors were randomly paired up to conduct joint inspection visits, independently scored health code violations, and developed training materials based on these results. The peer review initiative was evaluated with a randomized controlled trial in 2016, showing gains in accuracy and consistency (Elias and Ho 2016; Ho 2017a).

The design of the grading system involved two parts. The first part focused on usability. Roughly 35% of King County’s population is minority, 20% is foreign-born, and over 10% has limited English proficiency. The county used a community-based participatory approach to design the placard, convening a series of (multilingual) focus groups with community partners. A graphic designer then developed placards that would communicate the notion of food risk effectively to a diverse population. Appendix A presents the placard that emerged out of this usability process.

The second part focused the reliability of grading. One of the basic criticisms of extant grading systems is that a single visit presents merely a “snapshot in time” (Boehnke and Graham 2000; Seiver and Hatfield 2000; Wiant 1999). In contrast to other jurisdictions, the county decided to avoid introducing re-inspections solely for re-grading establishments.

---


base grades on (a) multiple routine (unannounced) inspections, (b) “critical” (red) code citations (as peer review showed that non-critical code items led to greater inconsistency), and (c) an adjustment for differences in stringency across inspection areas (Ashwood et al., 2016). The grading categories were as follows:

**Excellent**
No or Few red critical violations over the last four inspections

**Good**
Some red critical violations over the last four inspections

**Okay**
Many red critical violations over the last four inspections

**Needs to Improve**
The restaurant was either closed by Public Health – Seattle & King County within the last year or the restaurant needed multiple return inspections to fix food safety practices

In January 2017, the Board of Health adopted the food safety rating system.

As part of the County Executive’s campaign for creating the “best-run government” that prioritizes evidence to improve government and employee performance, the county was open to an evaluation. Such an evaluation would be of tremendous public value, given the popularity of grading and resource costs of implementing a grading system. New York City, for instance, allocated some $3.2M for the implementation of its grading system, constituting a 19% increase in its program budget.

The only systematic evidence to date about the effect of grading on health outcomes stems from an observational (non-randomized) study in Los Angeles, which found a 20% reduction in foodborne illness hospitalizations manifesting itself in the first quarter after adoption. A rigorous understanding of the effect of a grading system on health outcomes has implications beyond public health, as restaurant grading has played in a pivotal role in the broader scholarship on mandated information disclosure (Bubb, 2015; Ben-Shahar and Schneider, 2011; 2014; Fung et al., 2007; Ho, 2012; Loewenstein et al., 2014; Weil et al., 2006; Winston, 2008).

Notwithstanding the county’s openness to an evaluation, there were substantial reservations about the feasibility, expense, and ethics of a randomized evaluation, which we detail below. The

---

6. King County Board of Health, Rule & Regulation 17-01 (Jan. 19, 2017).
8. Jin and Leslie (2003) compared municipal adoptions of grading in Los Angeles county (LA) in 1998 to the rest of California before and after LA adopting grading. Stemming from a power analysis conducted for King County’s evaluation, Ho et al. (2017) found that the reduction in Los Angeles stemmed from the largest state recorded salmonella outbreak, which affected Southern California right before LA began grading.
county had originally planned a series of surveys and community-based focus groups, but those components would not assess the effect on health outcomes. Beginning in 2014, we worked closely with Public Health to design an evaluation that could overcome the considerable practical challenges. Our goal was to facilitate an agreement on a rigorous design by defraying any perceived costs to an evaluation. None of our work was compensated, providing a unique model for academic-agency collaboration to promote evidence-based policy.

3 Practical Barriers to Unit-Level Randomization

A priori, one appealing way to identify the effect of restaurant grading on foodborne illness might be at the restaurant level. For restaurant \( i \in \{1, \ldots, N\} \) in randomly assigned condition \( T_i = 1 \) for restaurant grading and \( T_i = 0 \) otherwise, we would estimate the average treatment effect as 
\[
\tau = \frac{1}{N} \sum_{i=1}^{N} [(Y_i|T_i = 1) - (Y_i|T_i = 0)],
\]
where \( Y_i \) represents the foodborne illness cases arising from restaurant \( i \) \ ([Neyman, 1923; Rubin, 1974]). But numerous practical constraints prohibit this ideal experiment.

**Illness attribution.** The most important limitation is that foodborne illnesses cannot be attributed to specific establishments at scale. Public Health receives complaints about establishments (second panel of Figure 1), but given the incubation period of most common foodborne pathogens, attribution requires an extensive and costly epidemiological investigation. This investigation includes completing a two-week food history for potentially affected individuals, a case-control analysis to identify likely sources, laboratory testing where available, and an on-site inspection of suspected establishments. From 2011-16, only 21 out of 3,280 consumer complaints to Public Health were attributable to King County restaurants by lab confirmation and only 69 complaints resulted in an epidemiological classification of “probable” attribution. Only 53 outbreaks (defined as two or more individuals affected) were probable or lab-confirmed from complaints (left panel of Figure 1). This constitutes a very small fraction of foodborne illnesses in the county, making restaurant-level
Figure 1: Four sources of data on foodborne illness for King County: probable or lab-confirmed foodborne illness outbreaks (first panel, purple), public complaints (second panel, pink), hospitalizations (third panel, blue), and illnesses via state-mandated report (right panel, red). Panels are organized by increasing prevalence, and each sequential panel includes the previous panel’s data (in feint colors) lines for comparison. Illnesses provide roughly 8.5 times the volume of hospitalizations and 25 times the volume of complaints. The grey box indicates years available for hospitalization data, but unavailable for other outcomes.

Outbreaks. An alternative outcome, which is not attributable to specific establishments, lies in hospitalization discharge data. Indeed, because the LA effect was based on hospitalizations, our evaluation started with this outcome in mind. Given some uncertainty about diagnostic codes (Ho et al., 2017), we consulted with epidemiologists specializing in foodborne disease in King County and the Centers for Disease Control and Prevention to compile a comprehensive set of discharge codes for foodborne illnesses likely to be affected by restaurant sanitation practices (see Appendix B). Yet we found that a small number of large outbreaks meant hospitalization trends were highly stochastic, as can be seen in the third panel of Figure 1. Between 2013 and 2015, with outbreaks stemming from Chipotle and a large meat producer, the rate increased by 60%. Even assuming a large 20% grading effect from LA, a design based on hospitalization data alone would be significantly underpowered.

The best available outcome data consists of “notifiable conditions,” which laboratories, physicians, and hospitals are required to report to state authorities after a diagnosis is made. Seven illnesses that are primarily foodborne (and not travel-related) are available: campylobacteriosis,

\[9^{9}\text{In addition, diagnosing whether an illness is foodborne is challenging. So-called “syndromic surveillance” data for each hospital visit, for instance, provide only pre-diagnostic information, hence presenting too much risk of Type I error. By and large, foodborne illnesses present with non-specific symptoms (e.g., diarrhea, vomiting).}\]
cyclosporiasis, listeriosis, salmonellosis, shiga-toxin producing E. coli, vibriosis, and yersiniosis. The right panel of Figure 1 shows that these data provide roughly 8.5 times the volume of foodborne hospitalizations, providing a more stable time series. Like hospitalizations, these illnesses cannot be attributed to specific establishments, but are aggregated at the ZIP code level based on the residential address of the patient.

Nonindependence. Even if outcomes were attributable to specific restaurants, the premise of restaurant grading is that it creates competitive pressure for sanitary practices. The rating that one restaurant receives can hence affect potential outcomes for other restaurants, violating conventional independence assumptions (Rubin, 1980). Grading must hence be implemented at a large enough scale to minimize interference (e.g., diner cross-over), but small enough to allow for actual comparisons.

Political Pressure for Adoption. Due to political pressure, the grading system needed to be implemented no later than 2017. No parts of the county could be left untreated. The pressure intensified due to high-profile, large outbreaks in 2015 (e.g., Chipotle), posing an inferential threat to any simple comparison over time: foodborne illness may decrease in subsequent years by “regression-to-the-mean” alone. Reforms may appear effective solely because large outbreak have died down. This form of endogenous policy adoption threatens causal inference for many legal and policy settings (see, e.g., Grambsch, 2008).

Equity Implications and Staff Division. A 2014 Executive Order mandated that departments adopt initiatives with equity and social justice in mind. Public Health was, as a result, particularly conscientious of potential inequities of the grading system along income, ethnicity, and cuisine. This posed another challenge to randomization, as any single randomization (particularly at the cluster level) might be seen to unfairly burden some set of establishments. Grading also proved controversial

10King County Executive Order, ACO 9-2 (AEO), Advancing Equity and Social Justice Through Development of a Strategic Innovation Priority Plan and Executive Department Actions (2014).
Figure 2: Score distribution in San Diego, where 90 points is the cutoff for an ‘A’ grade, in 2011.

among inspection staff, and Public Health wanted to equitably share the implementation burden across two field offices.

**Budgetary Constraints and Inspector Behavior.** The county began with essentially no budget for evaluation. This is common for adoptions of grading systems, with none to date building in a rigorous evaluation. As a result, observational studies focus on whether the citation of violations decreases after the enactment of grading. In addition to regression-to-the-mean effects, such designs are limited as they conflate two distinct effects: (a) the effect of grading on whether inspectors cite violations for otherwise identical practices, and (b) the effect of grading on restaurant sanitation practices. To see why reliance on inspection scores alone can be problematic, Figure 2 displays the score distribution of establishments in San Diego. Higher scores represent better inspection performance, and the vertical line at 90 represents the cutoff to receive an ‘A’ grade. Given considerable discretion and ambiguity in the health code, it is implausible that establishments are cleaning up precisely to the threshold. The much more likely explanation is that inspectors use the discretion to avoid a confrontation, a dynamic widely noted in the public health literature (Bohnke and Graham, 2000) and of concern to senior management in King County. Kovacs et al. (2018), for instance, show that “fudging” inspection scores to meet the grading threshold increases when there are social ties between the inspector and establishment. As one San Diego inspector noted, “Some inspectors will give out a B for an 89 . . . usually warn somebody at that point. It’s a judgment
4 Solutions via Research Design

4.1 Stepped-Wedge (Randomized Rollout) Design

Because of close communication around the grading initiative, our team was aware of considerable operational challenges in launching the grading system. These operational challenges helped persuade the county to roll out the grading system in stages. The chief benefit from the county’s perspective would be that a smaller group of inspectors could be on-boarded and then learn from the pilot jurisdiction to train the rest of the staff. One point of negotiation was the duration of the rollout and the number of stages. There was no ability to extend the rollout beyond 2017 and operationally, the county felt that sequencing over four quarters would be ideal.

After this commitment, we adopted a stepped-wedge randomized trial \((Gambia Hepatitis Study Group, 1987; Hussey and Hughes, 2007)\). Conventionally, stepped-wedge designs randomize the order in which an intervention is implemented across pre-existing groups (e.g., municipalities). All groups end up being treated, but the treatment effect is identified based on (randomized) group-time variation in the intervention:

\[
Y_{ij} = \mu + \alpha_i + \beta_j + \theta T_{ij} + \epsilon_{ij}
\]  

where \(Y_{ij}\) is the foodborne illness rate for group \(i\) at time \(j\), \(\alpha_i\) is a fixed effect for group \(i\), \(\beta_j\) is a fixed effect for time \(j\), \(\theta\) is the treatment effect, and \(T_{ij}\) equals 1 for groups in time periods when they are assigned the treatment, and 0 otherwise. Once a group has received the treatment, it will continue to receive the treatment for all subsequent time periods. If grading is implemented over the course of four quarters in four regions, we should be able to observe reductions in foodborne illnesses corresponding to when grading is rolled out in each region, given the nearly immediate 20% effect reported for L.A. The design helps to account for a quarter-specific shocks (e.g., an outbreak).


\(^{12}\) Adapting the database and electronic tablets used for field inspections, in particular, proved challenging, as did the training of inspectors around how to communicate the design of the grading system.
across King County, as well as region-specific (but time-invariant) differences, and randomization allows for a “reasoned basis for inference” as we spell out below (Fisher, 1925, 1935; Ho and Imai, 2006; Imbens and Rubin, 2015).

4.2 Improving Power with Synthetic Regions

While the application of a stepped wedge design is more novel for law and public policy, it can also be underpowered. Power limitations are exacerbated by the fact that researchers commonly roll out the intervention across existing administrative units (e.g., municipalities). There is, hence, no guarantee that these units are comparable in baseline outcomes. To illustrate this, the left panel of Figure 3 plots five existing administrative regions in King County (top) and the associated time series for foodborne illnesses (bottom)\textsuperscript{13}. While the time series of the different regions exhibit

\textsuperscript{13}These regions are King County district court regions. It is worth noting that these regions do not map cleanly to ZIP codes, another downside of using existing boundaries when constrained to particular primitive units. We assigned border ZIP codes in a way that preserved contiguity between regions and was most faithful to the county’s district map (see https://www.kingcounty.gov/~/media/depts/elections/elections/maps/)
Table 1: Statistical power at a 20% effect size and size (false positive rate or Type I error) of the stepped-wedge quarterly rollout test using administrative regions (Admin.) or synthetic regions (Synth.), using $\alpha=0.05$ in a one-tailed test for reduction in foodborne illness. Three inferential methods are presented: a standard parametric method using cluster robust standard errors (Wooldridge, 2001), the wild bootstrap approach designed to account for a small number of clusters (Cameron et al., 2008; Esarey and Menger, 2017), and the randomization inference approach described in Section 4.3 (Imbens and Rubin, 2015). There are five administrative regions, which can be rolled out in 120 different ways, and 2 years, so power and size are calculated over 240 tests. Similarly, because there are five possible synthetic region allocations, each of which can be rolled out in 24 different ways, and 2 years, power and size are calculated over 240 tests.

We sought to improve power of the design by leveraging insights from synthetic control methods (Abadie et al., 2015, 2010). The basic insight is that by creating synthetic rollout regions, we may be able to construct more plausible comparison groups that exhibit parallel foodborne illness trends. Because inspectors are principally assigned based on ZIP code, we begin with ZIP codes as the primitive units. Our goal is to aggregate 85 ZIP codes in King County to four regions that are comparable on foodborne illness trends so as to maximize power. An additional degree of freedom comes from the fact that while all ZIP codes would be included in the implementation of grading, individual ZIP codes could be excluded from the evaluation of grading. This trades off higher internal validity for lower external validity, and brings our total target regions to five (one for each administrative unit and four for synthetic regions).
quarter, plus a leave-out group).

To improve performance, we imposed several other constraints on region identification. First, we require regions to be geographically contiguous. This minimizes the chance that diners cross over treated and non-treated areas, as it is well-established that most restaurant competition is local (Parsa et al., 2011). Second, for similar reasons, for 31 municipalities in King County, we required ZIP codes in the same municipality (other than Seattle) to remain in the same cluster. Third, because Seattle is a large municipality that spans 37 ZIP codes, with lower likelihood of contamination across the city, we subdivided Seattle into municipal districts based on boundaries of Seattle’s Office of Planning and Community Development (Assefa, 2010). Fourth, because foodborne illness has a relatively low baseline incidence of roughly 42 cases per 100,000 people (annually from 2011 to 2016), we imposed a population threshold of 50,000 for each region.

The sample space of all possible ways to group 85 ZIP codes in King County into five mutually exclusive groups is large:

\[
\binom{85}{5} = \frac{1}{5!} \left[ 5^{85} - \left( \binom{5}{1} 4^{85} - \binom{5}{2} 3^{85} + \binom{5}{3} 2^{85} - \binom{5}{4} 1^{85} \right) \right] \sim 2.15 \times 10^{57}
\]

Searching this sample space is computationally intensive and simple random sampling is inefficient. We hence developed an algorithm to efficiently identify regions subject to our constraints.

The algorithm begins by randomly seeding four ZIP codes and randomly adding contiguous ZIP codes until population threshold (50,000 people) is met in each region. Then we randomly propose adding neighboring ZIP codes based on whether the addition improves the mean absolute value of pairwise differences in yearly foodborne illness rates in the pretreatment period (2011-2016):

\[
MAV_a = \frac{1}{IJ(I-1)} \sum_{i=0}^{I} \sum_{l \neq i}^{J} \sum_{j=1}^{J} |Y_{ij}^a - Y_{lj}^a| 
\]  

(2)

14 We do so by matching ZIP codes to municipalities using Coven (2012). These municipalities represent postal office classifications, and hence differ slightly from conventional understandings of King County municipalities.
15 Because the neighborhood district boundaries are defined by census tract, we assigned each ZIP code to the district based on area.
16 King County has a population of roughly 2 million, and we tested population floors between 50,000 and 200,000. We also tested varying the minimum number of restaurants. We found that a population threshold of 50,000 provided the best pretreatment balance. Population was matched at the ZCTA level from the 2014 American Community Survey 5-year estimates.
where $Y_{ij}^a$ is the aggregated rate of foodborne illness in year $j$ ($j = 1...J$) in region $i$ ($i = 1...I$) of allocation $a$. Figure 3 shows an example of one run of the algorithm, which iterated through 49 possible region allocations. We ran this algorithm 2.3 million times, producing an average of 38 possible region allocations per run. We discard duplicate allocations. Because allocations within a run can be quite similar, we also thin potential region allocations by storing every fifth step. This results in 17.5 million possible allocations, which we subset to the most balanced ones as we explain below in Subsection 4.4.

The second column of Table 1 shows how the construction of such synthetic regions can improve statistical power. Instead of 70% power with administrative units, statistical power with synthetic regions increases to 95%. That said, size remains poor at 25% with conventional parametric estimates.

4.3 Randomization Inference

The reason for poor size with conventional methods lies in the small number of clusters (Cameron et al., 2008; Cameron and Miller, 2014). The middle panel of Table 1 applies the wild cluster bootstrap of Cameron et al. (2008), which seeks to correct for this deficiency. Even this technique, however, results in false positive rates three times larger than $\alpha$ of 0.05 and reduces power substantially. We hence study the use of randomization inference, which can account precisely for treatment assignment scheme of randomizing from (a) potential regional allocations, and (b) potential rollout orders with four regions within each allocation ($24 = 4!$ in our scheme). Under the sharp null hypothesis, outcomes are fixed: $(Y_{ij}|T_{ij} = 1) = (Y_{ij}|T_{ij} = 0)$ for the observed treatment vector $T$ and foodborne illness rate $Y_{ij}$ in region $i$ and quarter $j$. The only stochastic component comes from random treatment assignment.

Let $\Omega$ be the set of size all possible treatment vectors. The size of $\Omega$ is $K = M \times 4!$, where $M$ is the number of possible region allocations and $4!$ is the possible number rollout orders given a region allocation. We also tried a stochastic (as opposed to greedy) version of the algorithm, where a decrease in MAV due to adding a ZIP code would be stochastically accepted, but found little improvement.
Figure 4: Region growth algorithm progression. After starting regions are formed with random seeds in step 1, ZIP codes in contiguous municipalities are added to the regions if they minimize the criterion in equation 2. The algorithm stops when no available ZIP codes can further minimize the criterion.
allocation. Let $W(t)$ represent the test statistic, namely $\theta$ from the stepped-wedge regression in Equation (1) given a treatment vector $t$. Under the sharp null, this test statistic can be calculated across all possible $K$ randomizations, and the set $W(t)$ forms the randomization distribution. The one-tailed $p$-value can hence be calculated as:

$$Pr(W(T) \leq W(t)) = \frac{\sum_{t \in \Omega} \mathbb{1}(W(T) \leq W(t))}{K}$$

(3)

The stepped-wedge test statistic allows us to adjust region- and quarter-specific effects, while randomization inference leverages the rollout design over regions and time to test the null hypothesis that foodborne illness rates were invariant to the adoption of grading. The intuition is that we examine how extreme the observed test statistic is relative to all the other ways that grading could have been implemented under our randomization scheme. In that sense, randomization provides the “reasoned basis for inference” (Fisher, 1925, 1935; Imbens and Rubin, 2015; Ho and Imai, 2006; Rosenbaum, 2002b,a).

The right panel of Table 1 shows that this method of inference has desirable properties in our setting. Randomization inference controls false positive rate at roughly the $\alpha = 0.05$, which stands in sharp contrast to parametric and bootstrap methods.

4.4 Selecting Potential Region Allocation

**Power.** Having established that randomization inference with synthetic regions has desirable properties, we now spell out the selection of $M$ potential rollout regions from the 17.5 million possible region allocations. As noted before, $M$ region allocations $\times$ 24 orders (per allocation) form the “randomization distribution” (i.e., the reference distribution). The left panel of Figure 5 plots a random allocation of the 17.5 million, showing that there is still substantial imbalance in parts of the full distribution.

To select our region allocations, we hence ranked allocations by yearly MAV (originally used to
Equity Constraints. With a well-powered design, we explored addressing equity constraints articulated by the county: that regions should not contain lopsided numbers of inspectors to train; that the operational burden of implementation should be evenly across the two offices; that no single region should have an overwhelming majority of gradeable restaurants; and that grading resources

18 We also considered re-running the region growth algorithm with this improved metric as the optimization criterion, but found that it was unnecessary. First, our existing allocations ranked by the MSE criterion provided over 80% power at $m \leq 40$, as shown in the left panel of Figure 5. Second, a random sample of allocations revealed that the yearly MAV criterion is highly correlated with quarterly MAV ($r = 0.89$) and the quarterly MSE ($r = 0.77$). See Figure 12 in Appendix F. Third, due to institutional approvals, we were under considerable practical time constraints, as we received the foodborne illness data in November 2016 and were expected to deliver the final regions by early December 2016. We note that future work may well improve on the efficiency of the sampling algorithm, but for our purposes the basic goal of improving power via synthetic regions was met.

19 When the randomization distribution was large, we used Monte Carlo simulation with 120 simulations for a given set of region allocations, 24 rollout orders for a single region allocation, two effect sizes (0% and 20%), and 2 years.
should be “fairly” allocated to areas of particular income levels or urbanization. We hence trimmed region allocations by rejecting allocations with severe demographic imbalances across regions, as displayed in Table 3 of Appendix D. For external validity, we also excluded any allocation that included fewer than 50% of King County ZIP codes in the evaluation region.

After trimming, randomization inference remained adequately powered at 85% with Type I error controlled at 4%. Figure 6 shows the final randomization distribution with ZIP codes sized relative to their populations for visibility. Not all rollout orders are shown, but there are a total of 120 ways in which the grading system could have been rolled out. The top right box outlines the allocation and rollout order drawn on December 13, 2016.

4.5 Baseline Risk Factor Study

As the evaluation plan unfolded, enthusiasm for the study grew. In particular, to study the effects on *citation* behavior, we explored partnering with the FDA around their risk factor analysis. Under the Government Performance and Results Act, FDA trains an independent corps of federal inspectors to observe risk factors across the country, independent of the local health code and without citations to operators. Because of the concern that inspectors would be loathe to cite violations with the additional pressure of grading, such a risk factor analysis could be invaluable to measuring outcomes independent of how grading affects citation behavior. Unfortunately, FDA’s study plan permitted only a very limited number of such inspections in King County. In late 2016, the county identified additional funding for contractor inspectors, who would be hired from outside the county, trained in part by FDA, and carry out risk factor inspections independent of the King County inspection corps. Based on a power analysis, we created a custom risk factor checklist that would enable these inspectors to carry out observations quickly (see Appendix E), in lieu of the typical 45 minute to 1-hour visit, and randomized establishments for this team to visit. One downside to this risk factor data collection is that hiring constraints prevented commencing it before 2017. At the same time, political pressure meant grading had to begin in January 2017.
Figure 6: Final randomization distribution (K=120). ZIP codes are displayed as rectangles proportional to their population. The top of the figure shows how this rectangle transformation looks for all King County ZIP codes. The top 5 ranked region allocations are shown along the columns, and the 24 possible order permutations are shown along the rows (darker colored ZIP codes get grading earlier than lighter colored ZIP codes). ZIP codes not included in the evaluation are shown in grey. The allocation randomly selected for the actual rollout is outlined in red in the top right.
This meant that only post-treatment data was available for region 1, but we stratified the sampling scheme to conduct balanced inspections before and after grading implementation for the remaining three regions.

* * * *

It is important to note that all of these extensive design efforts transpired before grading began. The analysis plan follows readily from the design, as there is no way to change the regions (and underlying randomization distribution) after implementation. We view this pre-commitment as a virtue of prioritizing the experimental design (Rubin, 2008).

5 Results

Although our design was based around the analysis of foodborne illnesses (notifiable conditions), for completeness, we present analyses of the effect of the grading system on all available outcomes. Subsection 5.1 examines on measures of foodborne disease incidence. Subsection 5.2 assesses the effect on more direct measures of food handling practices from the risk factor study. Subsection 5.3 discusses effects on foodborne illness complaints and their resolution.

5.1 Foodborne Disease Incidence

Hospitalizations. We begin by fitting Equation 1 to quarterly hospitalization rates per region, using the patient ZIP code from Washington’s Comprehensive Hospital Abstract Reporting System (CHARS) available up through 2017. The left panel of Figure 7 provides the time series of each of the regions, with darker lines indicating regions randomized to implement grading earlier. The dashed line represents the start of the quarterly restaurant grading rollout. If grading had as immediate and large an effect as suggested by earlier work, the darker lines should exhibit earlier decreases in disease incidence. The time series, however, confirm that hospitalizations are quite stochastic. As shown in Table 2 we find a non-statistically significant increase in foodborne hospitalizations of 0.01 per 10,000 (one-tailed \(p =0.57\)), with a 95% confidence interval of [-0.09,
Table 2: Results leveraging the synthetic region design described in Section 4.2. For reported illnesses, the ZIP samples are the evaluation ZIPs. For other estimates, the sample comprises all ZIP codes. 95% CI is the confidence interval derived from inverting the randomization inference test.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Effect</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Foodborne Illness (Rate)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitalizations</td>
<td>0.01</td>
<td>0.57</td>
<td>[-0.09, 0.11]</td>
</tr>
<tr>
<td>Reported Illnesses</td>
<td>0.15</td>
<td>0.90</td>
<td>[-0.07, 0.45]</td>
</tr>
<tr>
<td><strong>Risk Factors (IRR)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grading in region</td>
<td>1.03</td>
<td>0.65</td>
<td>[0.88, 1.16]</td>
</tr>
<tr>
<td>Grading in restaurant</td>
<td>0.95</td>
<td>0.45</td>
<td>[0.81, 1.07]</td>
</tr>
<tr>
<td><strong>Illness Complaints (Rate)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Complaints</td>
<td>0.13</td>
<td>0.18</td>
<td>[-0.03, 0.30]</td>
</tr>
<tr>
<td>Probable/Confirmed Outbreaks</td>
<td>0.09</td>
<td>0.03</td>
<td>[0.01, 0.14]</td>
</tr>
</tbody>
</table>

0.11 hospitalizations per 10,000

**Illnesses.** The middle panel of Figure 7 displays the time series in the four regions of reported foodborne illness rates (per 10,000) from 2011 to the first quarter of 2018. Trends are much more stable than for hospitalizations. Yet contrary to claims about the benefits of grading, the earlier implementing regions, if anything, experienced higher surges in foodborne illness in 2017. King County experienced a rise in foodborne illness in the third quarter of 2017 and the first implementing region, if anything, exhibited a higher uptick in foodborne illness than the other regions. This finding is inconsistent with claims that restaurant grading can mitigate the impact of an outbreak by improving food sanitation practices. Using the model in Equation 1, we find a non-statistically significant increase of 0.15 cases per 10,000 associated with the rollout of restaurant grading, or an increase of 14% over the pretreatment quarterly average baseline (one-tailed p=0.90). Inverting the randomization test provides us with a 95% confidence interval of [-0.07, 0.45] cases per 10,000.

---

20 To maximize hospitalization volume, we use all ZIP codes within the four regions.
21 This uptick appears to have been driven by campylobacter. Two small campylobacter outbreaks in this time period have been investigated by the county, one related to a restaurant and one from home cooking (see [https://www.kingcounty.gov/depts/health/communicable-diseases/disease-control/outbreak/private-event-august-2017.aspx](https://www.kingcounty.gov/depts/health/communicable-diseases/disease-control/outbreak/private-event-august-2017.aspx) and [https://www.kingcounty.gov/depts/health/communicable-diseases/disease-control/outbreak/cafe-juanita.aspx](https://www.kingcounty.gov/depts/health/communicable-diseases/disease-control/outbreak/cafe-juanita.aspx)).
22 The units are limited to the evaluation ZIP codes in each of the four regions.
## Synthetic Controls from Outside King County

As an additional check, we also take a synthetic controls approach to compare King County’s foodborne hospitalization rate (per 10,000) to that of other counties in Washington State ([Abadie et al., 2010][23]). We create a “synthetic control” county that is comparable King County in pretreatment foodborne hospitalization trends from 2000-16. If restaurant grading caused a reduction in hospitalizations of 20%, we should observe a drop in King County relative to its synthetic control.[24]

Figure 8 presents results. The left panel shows that the synthetic control tracks King County’s pretreatment foodborne hospitalization trends very closely (MSPE\(_{\text{pre}}\) = 0.003). The group is comprised primarily of King County’s immediate neighbors, Snohomish and Pierce counties, which makes substantive sense. Contrary to a large grading effect, we observe no appreciable difference.

---

[23]Reported illness data at the county level for 2017 as of yet unavailable from the Washington State Department of Health.

[24]More formally, let \(Y^1 = \{y^1_1, \ldots, y^1_J\}\) represent the hospitalization rates in King County from year \(j \in \{j, \ldots, J\}\), and \(Y^0\) represent an \(I \times J\) matrix of hospitalization rates for \(I\) control counties over \(J\) years. The synthetic control calculates weights \(w_i\) for control county \(i\) that minimize the pretreatment mean squared prediction error:

\[
\text{MSPE}_{\text{pre}} = \frac{1}{n_{\text{pre}}} \sum_{j<2017} (y^1_j - y^0_j)^2.
\]

where \(y^0_j = \sum_{i=1}^I Y^0_{i,j} w_i\) is weighted foodborne hospitalization rate for the synthetic control in year \(j\), \(y^1_j\) is the foodborne hospitalization rate for King County in year \(j\), and \(n_{\text{pre}}\) is the number of pretreatment years. We use permutation inference to test for a treatment effect with the following test statistic \(\tau\):

\[
\tau = \frac{1}{n_{\text{post}}} \sum_{j \geq 2017} (y^1_j - y^0_j) - \frac{1}{n_{\text{pre}}} \sum_{j < 2017} (y^1_j - y^0_j)
\]

where \(n_{\text{post}}\) is the number of post-treatment years.
Figure 8: Synthetic control results using CHARS hospitalization data for all counties in Washington. A synthetic control county was developed for King County using the weights displayed in the lower left corner of the left panel. The difference between King County’s mean-squared prediction error pre- and post-treatment is not statistically different from the placebo distribution of 24 comparison counties (one-tailed $p = 0.38$).

between King County and comparison counties in 2017. The right panel shows the hospitalization gap between King County and its synthetic control in red, along with the hospitalization gaps for all placebo treated counties. The hospitalization gap is comparable to that of placebo-treated Washington counties, and we fail to reject the null hypothesis using permutation inference with a difference-in-difference test statistic (one-tailed $p = 0.38$).

Taking these findings together, we hence detect no statistically significant effects of restaurant grading on foodborne disease incidence.

5.2 Risk Factors

Because foodborne illnesses are often underreported, we now turn to direct evidence of “risk factors” in restaurants. If the threat of a poor restaurant grade improves sanitation practices, we should be able to observe these directly in food handling practices (e.g., handwashing, cross-contamination). Several studies examine whether restaurant grading reduced the number of violation citations, but the grading system may directly affect whether an inspector cites a violation for the same behavior (Boehnke and Graham, 2000; Kovacs et al., 2018). Evidence of sharp disparities around grading thresholds, as displayed in Figure 2, suggest that grading may have a perverse pub-

\[\text{To conduct permutation inference, we limit the reference distribution to counties where MSPE}_{\text{pre}} \text{ is within } 30 \text{ times that of King County. } 30 \text{ is the minimum threshold at which we can obtain a reference distribution large enough to reject the null hypothesis at } \alpha = 0.05.\]
lic health effect: creating the appearance of all ‘A’-graded establishments, while actually decreasing operator incentives to correct behavior. Because this was a major concern for Public Health, we examine the effect of grading on risk factors, as measured by the independent risk factor study. Recall that independent contractors randomly sampled restaurants in each region with a short-form inspection checklist that we designed for this process (see Appendix E).

We fit the following set of quasi-likelihood Poisson models for the count of violations cited

\[ y_{ijk} = \sum_{l=1}^{L} y_{ijkl} \] over all violations \( l = 1, \ldots, L \):

\[
\log E(y_{ijk}) = \log n_{ijk} + \alpha_i + \beta_j + \gamma_k + \theta T_{ij} + \epsilon_{ijk}
\] (8)

where \( n_{ijk} \) is an offset for the total number of possible violations observed for establishment \( k \) in region \( i \) and year-quarter \( j \), \( \alpha_i \) are region fixed effects, \( \beta_j \) are year-quarter fixed effects, \( \gamma_k \) are inspector fixed effects, \( T_{ij} \) is a dummy variable taking on the value of 1 if restaurant \( k \) in region \( i \) is subject to grading in year-quarter \( j \) and 0 otherwise. We fit an analogous model that examines the effect of an establishment actually being graded:

\[
\log E(y_{ijk}) = \log n_{ijk} + \alpha_i + \beta_j + \gamma_k + \theta G_{ijk} + \epsilon_{ijk}
\] (9)

where \( G_{ijk} \) is an indicator variable for whether restaurant \( k \) in region \( i \) has received a graded inspection (and presumably, a grade placard) as of year-quarter \( j \).
The middle rows of Table 2 present results. We find that the average citation count of restaurants in graded regions increased by 3% over that of restaurants in non-graded regions, but this difference is not statistically significant (one-tailed $p = 0.65$). Figure 9 corroborates this by showing relative stability in citation rates over the observation window. Restaurants that had a graded inspection exhibited a 5% decrease in their average citation count, but this difference is again not statistically significant (one-tailed $p = 0.45$). In sum, we find no evidence that restaurant grading caused restaurants to improve sanitation practices. It is important to note that these results are independent of any direct citation effect of grading, as these inspectors were not administering the King County health code and issued no reports to operators after observation.

We illustrate the possibility of such citation effects by plotting the time series of low-risk (blue) and high-risk (red) violations in the right panels of Figure 9. Recall that in contrast to other grading systems, King County relies only on high-risk (red) violations to grade establishments. While blue violations remain cited at comparable rates before and after grading, red violations decrease substantially. Given considerable ambiguity around what constitutes a high-risk classification (see Ho, 2017a; Ho et al., 2018) and the stability of risk factors and blue citations, these findings suggest that grading reduces the citation of existing high-risk violations.

This analysis also shows limitations to identifying the causal effect of grading by studying only health code violations (e.g., Wong et al., 2015). Such designs are akin to studying whether a grade

\begin{equation}
Y_{ijkt} = \theta_1 \text{Grading}_{ij} + \theta_2 \text{Grading}_{ij} \times \text{Points}_t + \alpha_{it} + \beta_{jyear}t + \gamma_{kt} + \epsilon_{ijkt}
\end{equation}

where $Y_{ijkt}$ is the number of violation points in for establishment $i$ in year-quarter $j$ from inspector $k$ for violation type $t$ (red or blue), $\alpha_{it}$ represents fixed effects for establishment $i$ and violation type $t$, $\beta_{jyear}t$ represents fixed effects for the year of year-quarter $j$ and violation type $t$, $\gamma_{kt}$ represents fixed effects for inspector $k$ and violation type $t$, $\theta_1$ represents the grading rollout treatment for establishment $i$ in year-quarter $j$ for blue points, and $\theta_2$ represents the grading rollout treatment for establishment $i$ in year-quarter $j$ for red points, as Points$_t$ is dummy variable that takes on the value of 1 if violation type $t$ is red and 0 otherwise.

This difference, however, is not statistically significant when using fixed effects for year-quarter $j$ and type $t$ ($\theta_2 = -0.51$, SE = 0.38). This is because the largest drop in red points relative to blue points occurred simultaneously for all regions at the beginning of 2017, rather than in a quarterly fashion that tracked the rollout of grading over the course of 2017. The beginning of 2017 also coincided with an area rotation for inspectors, so these results are most plausible with inspectors citing fewer violations in anticipation of grading. The stability of the blue violations, given uncertainty about the red and blue distinction, suggests that the effect is driven by changes in inspector behavior, not restaurant behavior.
cutoff for military enlistment has “effects” on student performance, when cutoffs may have direct effects on instructor grading (Rojstaczer and Healy, 2012).

5.3 Restaurant complaint data

Finally, we investigate whether grading affected consumer foodborne illness complaints or the rate at which the county substantiates complaints to classify an outbreak. We refer to an outbreak as substantiated when the county deemed the outbreak “probable” or “confirmed” after investigation of complaints. We aggregate complaints and outbreaks investigated by the county by region and quarter. We fit the following model for (1) all consumer complaints and (2) probable/confirmed outbreaks tied to consumer complaints:

\[
Y_{ij} = \alpha_i + \beta_j + \theta T_{ij} + \epsilon_{ij}
\]

where \(Y_{ij}\) is the complaint rate or outbreak rate of region \(i\) in year-quarter \(j\), \(\alpha_i\) are fixed effects for regions, \(\beta_j\) are fixed effects for year-quarters, and, \(T_{ij}\) is a dummy variable taking on the value of 1 if region \(i\) is subject to grading in year-quarter \(j\) and 0 otherwise.

The bottom rows of Table 2 present results. We find that complaints increased by 0.13 per 10,000 people alongside the rollout of restaurant grading, but the association was not statistically significant (95% CI [-0.03, 0.30]). Restaurant grading was, however, statistically significantly associated with the detection of 0.09 more outbreaks per 10,000 people (two-sided \(p=0.03\), 95% CI...
The left panel of Figure 10 shows the time series of complaints from 2011 to 2018, which depicts a secular increase in complaints. The middle panel shows that substantiated outbreaks increased starting in 2017 and the right panel shows that this was particularly driven by complaints about graded restaurants in region 1, which drives the statistically significant outbreak effect.

We consider several interpretations of these results. First, the interpretation most favorable to the grading system is that it increased the volume of complaints to Public Health. Grading placards contained QR codes that led to a Public Health website linking to instructions on how to report foodborne illness. Because underreporting makes foodborne surveillance challenging, such reports may have facilitated the investigation of otherwise isolated complaints. There are, however, weak points to this interpretation. The increase in the overall volume of complaints was not statistically significant, and the same increases in substantiated complaints are not as evident in regions that implemented grading in subsequent quarters.

A second possibility is that media coverage of the launch of the grading system independently affected how consumers engaged with the department. This effect may have been short-term and primarily manifested itself in the first quarter in the region where placards became available. This interaction between media and popular attention on the launch and availability of placards would explain the patterns we see, but suggests that the impact on public engagement may not be sustained.

A third possibility is that these effects are simply artifacts of shellfish norovirus outbreaks that disproportionately affected region 1 in the first quarter. The number of outbreaks driving these effects, for instance, is low. Yet we observe few prior outbreaks – particularly for low-grade illnesses such as norovirus – that exhibit such a sharp uptake in independent complaints. Moreover, the probability that an outbreak is substantiated is directly affected by public reporting of food poisoning cases. The fact that region 1 had a higher substantiation rate may be a function of increased public engagement.
Lastly, it is also possible that King County’s decision to post ongoing foodborne illness investigations in late 2015, as well as the hiring of a full-time foodborne illness / enteric disease investigator in late 2016, affected the probability that a complaint was substantiated. That said, these factors would seem to affect all grading regions equally, and we observe some evidence that the increase in substantiated complaints tracks the staggered rollout. But perhaps the investigative resources interacted with the greater public consciousness of food safety from the grading placards, providing a boost to investigations early on in the implementation. We have marginal evidence that consumers, on average, reported cases one day earlier than before grading.\textsuperscript{27} Earlier reports make for easier epidemiological investigation, which might explain the higher rate of substantiating complaints.\textsuperscript{28}

Our results provide suggestive evidence that the grading system increased engagement with public health surveillance, but we are not able to conclusively resolve which of these mechanisms may have driven the apparent increase in substantiated complaints and outbreaks.

6 Discussion

We hope to have demonstrated through this field experiment how research design can mitigate perceived limitations of RCTs.

We believe this model of academic-agency collaboration – whereby academics can help offset the challenges to evaluation – is fertile ground for promoting evidence-based law and policy. While county resources for planning the grading system were substantial (see Section \textsuperscript{2}), the county budget for the randomized \textit{evaluation} was insubstantial.\textsuperscript{29} The only substantial line item was the expense of contractor inspectors for the independent risk factor analysis, but the evaluation of health outcomes could have proceeded without that element.\textsuperscript{30} The chief reasons for why this

\textsuperscript{27}Using randomization inference for time from meal to reporting, trimming outliers over 2 standard deviations from the mean reporting time, yields a \( p \)-value of 0.07.

\textsuperscript{28}We thank Sargis Pogosjans for suggesting this analysis.

\textsuperscript{29}The main costs for the core RCT component were: (a) \$1,450 for hospital discharge data from Washington State; (b) additional staff time for training due to the staggered roll-out.

\textsuperscript{30}The budget allowed for 900 hours for three contractors, comprising roughly \$86k.
bargain could be struck were that: (a) the county committed in principle to rigorously evaluating the policy initiative and agency leadership was committed to managing the process; (b) our academic team, in turn, was willing to invest considerable research efforts to develop a tailored research design that addressed practical, political, and ethical concerns; and (c) administrative data, which had never been analyzed for evaluating these kinds of policy initiatives, was available.

Our trial provides critical evidence about the value, implementation, and administration of grading systems. Most importantly, it grounds expectations citizens, policymakers, and academics should have of these kinds of disclosure initiatives. Contrary to often cited data about the health benefits, we find no evidence that foodborne illnesses decreased. Our trial also shows an unintended consequence of grading, namely the citation effect documented in Subsection 5.2, which is consistent with grade inflation in other jurisdictions (Hoo 2012, Kovacs et al. 2018). While these findings may inform the decision of whether to adopt a grading system, our results also inform prospective decisions with the confines of a given grading system. Management and training of staff should focus on avoiding the unintended citation effect. The results also strongly support King County’s methodology to calibrate the grading thresholds each year to measure relative performance in a way that is more robust to citation effects (Ashwood et al. 2016). The evidence about increased public engagement suggests that grading systems may have the most salutary effects by public education about food safety principles, but that sustaining such public engagement remains a challenge.

To be sure, our case study also illustrates limitations for how research design can facilitate RCTs of policy implementation. First, one challenge to customized research designs is scalability. It took considerable learning, mutual respect, understanding, and artful negotiation among key stakeholders to craft our solution. From that perspective, navigating political barriers is a nonpecuniary cost to conducting a randomized evaluation in this context. Nonetheless, we believe this is a fruitful avenue, as there remains significant under-provision of randomized trials in vast areas of law and policy (Abramowicz et al. 2011, Gerber et al. 2014, Green and Thorley 2014, Margetts 2011). Good observational design can also involve considerable research investment (Rubin 2008,
and natural experiments are often unavailable for critical policy questions of interest. One of the virtues of a large-scale field experiment — given the amount of planning invested in this research collaboration — is that these kinds of trials are less likely to be left to filedrawers \cite{Rosenthal1979}.

Second, another potential limitation of our study is external validity. Our evaluation only provides valid estimates of the effect in King County. Unlike other grading systems, King County’s system is based on multiple routine inspections, uses only high risk violation, adjusts for inspector differences, and does not use a letter grade. The treatment effect we recover may hence not generalize to other jurisdictions. While this may seem like a substantial limitation, it turns out that grading systems in fact vary dramatically across nearly all jurisdictions \cite{Ho2012, p. 603, table 2}. Some jurisdictions do not post grades until appeals are resolved; others disclose only whether a restaurant passed an inspection; some conduct only one routine inspection per restaurant annually; and yet others have no meaningful grade variation at all \cite{id}. The concern about external validity is then more about the challenge of policy learning given the highly decentralized nature of retail food safety enforcement, which generates such dizzying variation, in the United States. Taken to the extreme, this concern undercuts the notion of states as laboratories of democracy. Nor is this problem one observational studies can address (LA’s effect may be limited to LA’s system and population), and there are in fact few alternative credible observational designs \cite{Ho2017}. The most promising path forward, in light of the rapid adoption of grading systems, would be to adapt our design to evaluate implementation across many jurisdictions.

Third, one limitation of the stepped-wedge design is that the treatment might be different in earlier periods than in later periods. For instance, if the agency is learning about how to implement the policy (e.g., informing operators), grading in the first period may not have the same effect as grading in a later period. We hence view a stepped-wedge design as complementary to implementation analyses better known in the public administration field. Such analyses should reveal operational challenges in rolling out a policy. Fortunately, the extent of planning in King County
meant that operational issues were limited. If such challenges exist, greater effects should be detectable in subsequent regions. Ideally, when a policy is anticipated to be difficult to implement, the design would include more stages of rollout, thereby facilitating both implementation analyses and increasing statistical power of the evaluation.

Fourth, although our design attempted to best account for potential nonindependence across restaurants, it is theoretically possible that treatment effect spillovers weakened power. Yet our use of subregions in a county that covers over 2,300 square miles, combined with the well-established fact that restaurant competition is predominantly local, should give some assurance that spillovers negate the ability to detect a grading effect. Moreover, the synthetic control analysis that compares King County to other Washington counties also provide no evidence that King County’s foodborne illness rate dropped by 20% upon adoption.

Last, there are surely ways in which our algorithm could have been refined. For instance, rather than making a binary decision on whether to include a ZIP code in an area, continuous weights could be used, although the search space would become quite large. (In the current setup, the possible number of partitions already exceeds the estimated number of seconds since the Big Bang.) Practically, time was a considerable constraint. After institutional approvals and data transfer, our team had little time to provide the regions to the county to plan operations.

In short, while we acknowledge these potential limitations, we believe there are few better alternatives than a randomized evaluation to credibly assessing the benefits of such substantial policy initiatives. Research design – in this case a synthetic cluster randomized stepped-wedge design – holds great promise for overcoming practical barriers to evidence-based law and policy.

---

31 The main operational issue was about the validity of grade calculations, having to do with the database syncing issues, and compliance by operators to posting the grade. These were documented only in 2018, and hence do not invalidate the stepped-wedge design, and limited in scope. If, however, the county changed the system, for instance by introducing substantial penalties for noncompliance with placard posting, our evaluation may not speak to this effect.
References


Assefa, Samuel. 2010. “Census Geography to Neighborhood Equivalency Files.”


Buck, Stuart, and Josh McGee. 2015. “Why government needs more randomized controlled trials: Refuting the myths.” *Laura and John Arnold Foundation*.


Coalition for Evidence-Based Policy. 2015. “Demonstrating How Low-Cost Randomized Controlled Trials Can Drive Effective Social Spending: Project Overview and Request for Proposals.” Project Overview and Request for Proposals.


Esarey, Justin, and Andrew Menger. 2017. “Practical and Effective Approaches to Dealing with Clustered Data.” Political Science Research and Methods, First View pp. 1–35.


House of Commons Science and Technology Committee. 2006. “Scientific Advice, Risk and Evidence Based Policy Making.”


Mdege, Noreen D., Mei-See Man, Celia A. Taylor (nee Brown), and David J. Torgerson. 2011. “Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation.” *Journal of Clinical Epidemiology* 64(9):936 – 948.


Public Health – Seattle & King County. 2014. “King County Food Protection Program Review: Final Report.” *Public Health – Seattle & King County*.


A Sample Placard

Figure 11: Example of food safety placard.
B Selection of Discharge Codes

MEMORANDUM

From: [Redacted]
To: Food Safety Group, Centers for Disease Control and Prevention (CDC)
Date: September 20, 2016
Subject: Foodborne Illness Selection for Prospective Evaluation

Based on earlier conversations, we understand that CDC does not currently retain a list of all foodborne illness diagnostic codes. As you know, Seattle & King County will be implementing a restaurant grading system in 2017 and in order to evaluate the effects of this policy change, we aimed to identify foodborne illnesses that restaurant sanitation practices are likely to affect.

We write to inform you of what our research has revealed and also seek your input on whether you agree with our selection of diseases. We hope that this research may be independently helpful to CDC’s monitoring efforts of food safety.

Based on our review of the public health literature, we propose to focus on illnesses that are:

(a) foodborne in at least 80% of cases (based on Mead1 or Scallen2), and

(b) travel-related in less than or equal to 20% of cases (based on Scallen3).

We include the travel restriction, as we do not expect restaurant grading to impact the incidence of conditions frequently acquired abroad. The travel restriction, however, may not be appropriate for monitoring non-restaurant related food safety conditions.

For reference, the Appendix summarizes data at the pathogen level in Table A and at the ICD code level in Table B. Our review has concluded that the ICD-9 and ICD-10 codes corresponding to these illnesses are:

<table>
<thead>
<tr>
<th>Pathogen</th>
<th>ICD-9</th>
<th>ICD-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salmonella (non-typhoidal)</td>
<td>0030 - 0039</td>
<td>A020 - A029</td>
</tr>
<tr>
<td>Campylobacter</td>
<td>00843</td>
<td>A045</td>
</tr>
<tr>
<td>E. Coli (shiga-toxin producing)</td>
<td>00804</td>
<td>A043</td>
</tr>
<tr>
<td>Vibriosis</td>
<td>0054, 00581</td>
<td>A053, A055</td>
</tr>
<tr>
<td>Listeriosis</td>
<td>0270</td>
<td>A320 - A329</td>
</tr>
<tr>
<td>Yersiniosis</td>
<td>00844</td>
<td>A046, A282</td>
</tr>
<tr>
<td>Trichinosis</td>
<td>124</td>
<td>B75</td>
</tr>
<tr>
<td>Other foodborne bacterial illnesses</td>
<td>0050-0053, 00589, 0059</td>
<td>A050-A052, A058, A059</td>
</tr>
</tbody>
</table>

In addition, in the ICD-10 system, B9622, B9623, B962 are secondary diagnostic codes that identify shiga-toxin producing strains of E. coli as the cause of illness and B967 is a secondary code that identifies Clostridium perfringens as the cause of illness.

We hope this list is helpful and would welcome any input you might be able to provide.
C Implementation Regions

In order to minimize the operational burden of implementing the grading system, we optimized the how ZIP codes that are not in the set of evaluation regions were added to regions. We balance the number restaurants subject to grading:

\[ R_a = \frac{1}{I(I-1)} \sum_{i=1}^{I} \sum_{l \neq i} |n_i^a - n_l^a| \]  \hspace{1cm} (12)

where \( n_i^a \) and \( n_l^a \) are the number restaurants in region \( i \) \((i = 1...I)\) or region \( l \) given allocation \( a \). The right panel of Figure 3 shows the resulting evaluation and implementation regions that were delivered to the county for the actual grading rollout.
### D Trimming

<table>
<thead>
<tr>
<th>Unit</th>
<th>Covariate</th>
<th>Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>(a) Population</td>
<td>Within 2.5&lt;sup&gt;th&lt;/sup&gt; and 97.5&lt;sup&gt;th&lt;/sup&gt; pctl</td>
</tr>
<tr>
<td></td>
<td>(b) Number of gradeable restaurants</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(c) Number of full-time employees assigned to gradeable restaurants</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(d) Mean household income</td>
<td></td>
</tr>
<tr>
<td>Allocation</td>
<td>(e) Percentage of included ZIPs that are in the Seattle municipal area</td>
<td>Within 2.5&lt;sup&gt;th&lt;/sup&gt; and 97.5&lt;sup&gt;th&lt;/sup&gt; pctl</td>
</tr>
<tr>
<td></td>
<td>(f) Percentage of restaurants in included ZIPs that are inspected by the Central Office</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(g) Percentage of total ZIPs in the county that are included in the allocation</td>
<td>At least 50%</td>
</tr>
</tbody>
</table>

Table 3: Balance covariates used to ensure that the comparison regions were operationally and politically feasible for the county. Population and household income are the 5-year estimates from the 2014 American Community Survey. Restaurant covariates are calculated from the baseline period 2014-2015. Gradeable restaurants are restaurants eligible to receive a grade according to King County’s internal classification. The Seattle municipal area includes 37 ZIP codes.
1) Handwashing: This item is marked IN Compliance only when employees are observed using proper handwashing techniques at appropriate times. Note: in the rare occasion where there is no need to wash hands and therefore handwashing is not observed, this item may be marked IN.
   a. This item is marked IN Compliance only when employees are observed using proper handwashing techniques.
   b. This item is marked IN Compliance only when employees are observed washing their hands at appropriate times.

   If either (a) or (b) are marked OUT, (1) must be marked OUT.

2) BHC: This item is marked IN Compliance only when employees are observed using suitable utensils or gloves to prevent bare hand (or arm) contact with RTE foods or are observed following an alternative procedure from an approved plan that allows bare hand contact with RTE foods. Note: there are very few such approved plans in KC.

3) Cross Contamination: This item is marked IN Compliance only when raw animal foods are separated from RTE foods during storage, preparation, and display. Cross Contamination due to soiled equipment or lack of hand washing is not included here.

4) Clean/Sanitize: This item is marked OUT of Compliance when food contact surfaces are not cleaned and sanitized before uses with different types of meat (e.g. placing beef on a cutting board contaminated with raw chicken) or when changing from raw animal foods to RTE food. This item is marked IN Compliance when food contact surfaces and utensils are clean to sight and to touch and sanitized before use.

5) TAAC: If PHF is above 41°F (45°F for raw shell eggs) or below 135°F (130°F or roasts) and the food is not in cold or hot holding equipment, and the establishment is not using time as a public health control (TAAC), mark this item OUT of Compliance. This item should also be marked OUT if the establishment is using TAAC and:
   - The food is unlabeled or improperly labeled, or
   - The food is not served or discarded within 4 hours after removal from temperature control, or
   - Approved written procedures are not available.

   This item may be marked NA when the establishment does not use TAAC.

   This item may be marked NO when the establishment uses TAAC, but not at the time of inspection.

6) Cold Holding: Except for PHF that are in the cooling process or have been recently out at room temperature during active preparation (less than 2 hours), this item should be marked OUT of Compliance if any PHF that is being stored under refrigeration (mechanical or using ice as a coolant) are found to exceed critical limits by 1 degree Fahrenheit or more.
   a. Enter the number of mechanical refrigeration units that are being used to cold hold PHF.
   b. Enter the number of mechanical refrigeration units that are holding PHF 1–5°F above critical limits.
   c. Enter the number of mechanical refrigeration units that are holding PHF 6°F or above critical limits.
7) **Hot Holding:** This item should be marked OUT of Compliance when PHF are stored in hot holding equipment and is found to be 134°F or below [130°F for whole beef roasts, corned beef roasts, pork roasts, and cured port roasts such as ham that are cooked according to the cooking chart found at WAC 246-215-03400 (2)].

This item may be marked NO when the establishment does not hold PHF, but none are being held hot during the time of inspection.

This item may be marked NA when the establishment does not hot hold PHF.

a. Enter the number of hot holding units that are being used to hot hold PHF.

b. Enter the number of hot holding units that are holding PHF 1-5°F below critical limits.

c. Enter the number of hot holding units that are holding PHF 6°F or below critical limits.

8) **Cooling:** This item should be marked OUT of Compliance if any of the cooling methods (time/temperature, shallow pan, or pieces of meat) are found to be out of compliance. See WAC 246-215-03515 (1). For this study, we are not assessing items (2), (3), or (4). For establishments using time/temperature, the item should be marked OUT of Compliance if the food does not cool from 135°F to 70°F in two hours or less; or if the food does not cool from 135°F to 4°F in six hours or less. Discussions with the PIC and Food Workers along with observations should be used to determine compliance.

This item may be marked NO when the establishment does cool PHF, but proper cooling temperature and time parameters cannot be determined during the length of the inspection.

This item may be marked NA when the establishment does not cool cooked PHF.

a. Enter (+) if the establishment uses the Time and Temperature method as specified in the Food Code, WAC 246-215-03515 (1) (a) and (b). Enter (-) if this method is not used.

b. Enter (+) if the establishment uses the shallow pan method as specified in the Food Code, WAC 246-215-03515 (c) (i). Enter (-) if the shallow pan method is not used.

c. Enter (+) if the establishment cools pieces of meat as specified in the Food Code, WAC 246-215-03515 (c) (ii).
F Quarterly vs Yearly Criterion

Figure 12: Criterion values in a random sample of results from the region algorithm optimized on yearly mean-absolute value (see Equation 2). Balance calculated at the yearly level is highly correlated with balance calculated at the quarterly level.