Fair Decision-Making for Food Inspections

Shubham Singh
University of Illinois at Chicago
Chicago, Illinois, USA
ssing57@uic.edu

Bhuvni Shah
University of Illinois at Chicago
Chicago, Illinois, USA
bshah46@uic.edu

Chris Kanich
University of Illinois at Chicago
Chicago, Illinois, USA
ckanich@uic.edu

Ian A. Kash
University of Illinois at Chicago
Chicago, Illinois, USA
iankash@uic.edu

ABSTRACT

We revisit the application of predictive models by the Chicago Department of Public Health to schedule restaurant inspections and prioritize the detection of critical food code violations. We perform the first analysis of the model's fairness to the population served by the restaurants in terms of average time to find a critical violation. We find that the model treats inspections unequally based on the sanitarian who conducted the inspection and that, in turn, there are geographic disparities in the benefits of the model. We examine four alternate methods of model training and two alternative ways of scheduling using the model and find that the latter generate more desirable results. The challenges from this application point to important directions for future work around fairness with collective entities rather than individuals, the use of critical violations as a proxy, and the disconnect between fair classification and fairness in dynamic scheduling systems.

CCS CONCEPTS

• **Applied computing** → *Decision analysis*; **Computing in government**; • **Computing methodologies** → *Planning and scheduling*.

KEYWORDS

food inspections, fairness, scheduling

ACM Reference Format:
Shubham Singh, Bhuvni Shah, Chris Kanich, and Ian A. Kash. 2022. Fair Decision-Making for Food Inspections. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), October 6–9, 2022, Arlington, VA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3551624.3555289

1 INTRODUCTION

The Chicago Department of Public Health (CDPH) issues food safety guidelines and conducts inspections of more than 16,000 food establishments. Through these inspections, CDPH sanitarians educate owners and workers about food safety practices, inspect the premises and practices for safe food handling, and promote a healthy environment for food preparation. The City of Chicago records each of the food inspections on its public data portal.[1] As there are a limited number of sanitarians, a natural goal is to use data to prioritize performing the inspections that best protect the public health. Inspections which identify critical violations of the food code allow conditions posing the highest risk of causing a food-borne illness to be addressed. Thus, data scientists working for the city and their collaborators trained a machine learning (ML) model to predict the likelihood of an inspection resulting in a critical violation [31]. The trained ML model, which we refer to as the Schenk Jr. et al. model, was used to prioritize food inspections in a simulated study. An evaluation of the model showed that using it to schedule inspections achieves a 7-day improvement in the mean time to detect a critical violation compared to the actual inspection schedule followed by sanitarians [31].

In this paper, we first reexamine the model from a fairness perspective and assess how the improvement gained by employing the model is shared by different parts of the city. A key driver of geographic variation is that not all sanitarians report critical violations at the same rate, with a range of less than three percent to more than forty percent inspections cited with violations. Since the model prioritizes the restaurants inspected by sanitarians who report a high rate of critical violations, residents of the city regions where sanitarians cite critical violations at a higher-than-average rate tend to see inspections prioritized at the expense of the other regions.

We then explore approaches to prioritize food inspections in a fairer way. Our interventions span two broad classes of techniques: (a) those where we train a new model to predict critical violations in a fairer way and (b) *post-processing* approaches where we use the Schenk Jr. et al. model as-is but modify the way the model is used to achieve a fairer resource allocation. We examine four different approaches to training fair models, and find that they mitigate but do not eliminate the geographic unfairness that results when the models are used to schedule inspections. We consider two post-processing approaches which adjust the way sanitarian identities are used when scheduling inspections: one censors the sanitarian

EAAMO '22, October 6–9, 2022, Arlington, VA, USA

ACM ISBN 978-1-4503-9477-2/22/10…$15.00
https://doi.org/10.1145/3551624.3555289

[1] Chicago Data Portal: https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5

feature when predicting while the other uses the model to reschedule each sanitarian's inspections rather than globally rescheduling all inspections. Our results show that these post-processing approaches are more effective than retraining in achieving fair outcomes. After analyzing the fairness properties of the various approaches, we examine the trade off between efficiency and fairness. We find that the post-processing approaches enable an attractive trade-off between efficiency and fairness while the fair models are essentially Pareto dominated.

We conclude by discussing three key issues our results raise for future work. The inspection of restaurants is different from much of the fairness literature in that each entity has many stakeholders, so models based on simple binary protected attributes are not a good fit for some protected groups of interest. The use of critical violations as a proxy for public health raises important questions about what fairness means in this setting, particularly in light of the heterogeneity across sanitarians. Finally, we discuss the disconnect between training a model for classification and our goal of achieving fairness in a dynamic scheduling system.

1.1 Related Work

Food inspections remain an essential food safety practice to prevent food-borne illnesses. However, the variation in practices and guidelines across jurisdictions results in a lack of consistency in local and national food safety levels and the link between inspections and food safety is actively studied. One of the first studies done in Seattle-King County found that restaurants with lower food inspection scores[2] were likely to have more outbreaks than restaurants with higher scores [19]. Another study done in Miami–Dade County found no correlation between restaurant inspection scores and the outbreak of food-borne illnesses [8]. Jones et al. found inconsistencies between criteria for high risk establishments and establishments that resulted in outbreaks through a study in the state of Tennessee and called for a deeper examination into the restaurant inspection system [21]. Today, there are still calls for systemic change in local food safety inspection systems across the country due to the poor predictive power of the current inspection framework [2]. Technologists have also been exploring how to improve the food inspection process and searching for non-traditional data sources. Google search activity in an ML-based approach to identify higher risk establishments has shown some promise [30], but the addition of Yelp data did not improve the prediction of inspection scores [1].

More broadly our work sits at the intersection of two trends in the use of artificial intelligence techniques. First is the use of predictive analytics to bring algorithmic decision-making to government operations. We witness a rise in the use of ML algorithms for delivery of public services, digitization of court records, and management of government programs [7, 13]. Second is the increased openness of government data. A recent study points out the need for more transparency to counter the public distrust of AI and promote its use for the common good [24]. Such initiatives help authorities improve their operations and provide transparency to their decision-making process. Kaggle tournaments have been used to encourage public involvement in civic model creation [15] to create better public services [28]. We provide a case study of using open data to analyze the fairness of predictive analytics and provide interventions to improve it. Previous case studies of other domains include predictive policing [14] and child maltreatment [6].

[2]Sanitarians assigned each inspection a score out of 100, and a lower score indicated more critical violations.

2 BACKGROUND

Schenk Jr. et al. sought a predictive model to prioritize the routine food inspections conducted by sanitarians from the Chicago Department of Public Health (CDPH) [31]. Rather than relying solely on manual scheduling of food inspections by CDPH, in 2015 the Chicago Department of Innovation and Technology and data scientists from the Civic Consulting Alliance created a machine learning model to aid in this scheduling process.

The model specifically looks at the scheduling of routine food inspections which are conducted once or twice a year at each establishment *independent of any customer complaints*. The dataset used to train and test the model consists of information from 18,000 inspections over 4 years with the training set from September 2011 to April 2014 and the testing set from September 2014 to October 2014. The dataset is derived from several datasets from the Chicago Open source portal[3] including those about crime rates, sanitation, weather, and food inspections. The dataset also includes information about the sanitarians who conducted the inspections. In order to protect the individual identity and the privacy of the approximately three dozen sanitarians, they are grouped into six sanitarian clusters based on their *critical violation rate*, which is the percentage of inspections they conduct that result in critical violations. The clusters, which are used as features in the model, were named after the lines of Chicago rail transit system: Purple, Blue, Orange, Green, Yellow, and Brown. From this dataset, they train a logistic regression model using features including the sanitarian cluster conducting the inspection, past establishment violation records, and surrounding environment data (crime rates, cleanliness, temperature) to predict the likelihood of the inspection resulting in a critical violation.

The trained model is then used to output a risk score for each food inspection where a higher score signifies a higher risk of finding a critical violation. These risk scores are used to prioritize inspections: the highest risk score should be inspected first and the remainder inspected in decreasing order of the risk scores (model coefficients detailed in Table 2 in supplementary material).

Evaluating the model requires making an assumption on how it would be used. Schenk Jr. et al. [31] reassigned the dates of the inspections in the test set based on their predicted risk score, while preserving the number of inspections performed each day and the identity of the sanitarian performing each inspection (illustrated in supplementary material §A.5). The main goal of their work was to ensure that the inspections which resulted in critical violations were conducted as rapidly as possible. Therefore, each schedule of inspections (the original and the one created by the model) was evaluated by the average time required to detect critical violations. This was calculated by taking the mean of the number of days between date the inspection was scheduled and the first day of the test set window (i.e. September 1, 2014) for those inspections that

[3]https://data.cityofchicago.org/

resulted in a critical violation[4]. They found that, on average, the schedule based on their model detected critical violations 7-days faster than the original schedule when applied to the two-month test set. The dataset and code used in this analysis are available in the project's repository.[5]

Kannan, Shapiro, and Bilgic [22] provided an independent analysis of the model results and identified several issues with the model and its analysis. Their finding regarding one of the feature sets turns out to be particularly relevant for our analysis (we also revisit their other findings in §7). In particular, they argued that using the sanitarian clusters as a predictive feature unfairly changes the prediction risk for the establishment. Since the clusters were created by grouping sanitarians with similar violation rates, it was likely that establishments set to be inspected by sanitarian clusters with a high propensity to find violations were much more likely to have a high risk score for a potential violation, regardless of the other attributes. They view this outcome as being unfair to the restaurant. Because CDPH's intent in performing inspections is to protect the public health, our primary focus is on fairness *to customers* of the establishment. Nevertheless, we show that this differentiated behavior of sanitarians and its use by the model has important consequences for our fairness concerns on larger geographic areas across the city.

3 FAIRNESS OF THE SCHENK JR. ET AL. MODEL

We focus on the fairness of the way Schenk Jr. et al.[31] model is used to schedule inspections for the city and in turn, distributes public health benefits among Chicago's residents. To quantify fairness, we borrow from the existing definitions of the fairness literature and adapt them to the problem of food inspections. The first definition we focus on is Demographic Parity (DP) or Statistical Parity, defined as a classifier having equal positive predicted rates for advantaged ($A = 1$) and disadvantaged ($A = 0$) groups [5]. Given a prediction $\hat{Y}$, demographic parity is satisfied if

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1) \ . \tag{1}$$

We are interested in achieving similar amounts of time taken to complete food inspections across groups of interest, which makes our fairness objectives compatible with the evaluation metrics of the model. Since we consider multiple groups, A is categorical with values $\{a^0, \ldots, a^{n-1}\}$, where n is the total number of groups. Let T represent the random variable for the time to complete a random food inspection. Our interpretation of DP is

$$\mathbb{E}[T \mid A = a^i] = \mathbb{E}[T \mid A = a^{i+1}] \qquad \text{s.t. } 0 \leq i < n \ . \tag{2}$$

Eq. 2 states a fair schedule, on average, has equal times to conduct the inspections belonging to different groups. Another widely applicable definition of fairness is Equal Opportunity (EOpp), which is defined as the classifier having equal true positive rates for advantaged and disadvantaged groups [17]. Formally,

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1) \ . \tag{3}$$

[4]This approach makes several strong assumptions. We discuss concerns related to the efficiency and operational constraints of food inspection at the city level in §6.2. We discuss the use of critical violation detection as a proxy for public safety in §7.
[5]https://github.com/Chicago/food-inspections-evaluation

Our interpretation of equal opportunity requires having similar times to detect critical violations in food inspections across all groups. This can be written as

$$\mathbb{E}[T \mid A = a^i, Y = 1] = \mathbb{E}[T \mid A = a^{i+1}, Y = 1] \qquad \text{s.t. } 0 \leq i < n \ . \tag{4}$$

Eq. 4 states a schedule is fair if the inspections *where a critical violation was found* ($Y = 1$), on average, took equal amounts of time to be detected across the groups. These interpretations of DP and EOpp are consistent with work on extending these concepts beyond simple classification settings [4].

Throughout the rest of the paper, we consider an early detection of a critical violation to be a desirable outcome. Although this outcome helps in achieving the broader goal of protecting public health, there a multitude of groups (stakeholders) that stand to benefit from food inspections: CDPH, restaurant owners, and the customers. Since a restaurant serves a large number of people living or working in its neighborhood and they are affected by potential food-borne illness stemming from unsafe conditions, a major advantage of an early violation detection goes to the customers (we elaborate on the choice of stakeholder in §7). Thus, our primary interest is in applying these definitions to groups consisting of restaurants located in a particular region of the city.

3.1 Fairness along geographic lines

For our fairness analysis, we examine the effects of the model on different regions of the city, colloquially known as "sides". We explore how the residents of the different regions are affected by using the model to schedule the food inspections. Using the ZIP codes of the inspected restaurants in the dataset, we match their location to the nine sides[6] of the city. We largely follow the methodology used by the City of Chicago as described in §2 and fully detailed in [31]. However, to improve the robustness of the results we perform a cross-validated evaluation, rather than the evaluation on the last 60-days performed by Schenk Jr. et al.. The dataset contains 19 non-overlapping periods spanning 60 days from the first inspection date till the last. Out of these 19 evaluations periods, we exclude three evaluation periods that did not contain inspections for all 60 days. While we consider both the notions of fairness in Equations 2 and 4, since the overall goal of the system is to improve the detection of critical violations we put more emphasis on the second.

Fig. 1 shows the *difference* between the *average time taken to detect critical violations in a specific region* and the *overall average for that schedule* using solid colors. This corresponds to the extent to which EOpp is violated in that region, Eq. 4. It also shows the difference between the average time taken to conduct inspections (regardless of whether a critical violation is found) in a region and the overall average for that schedule. This corresponds to DP (Eq. 2) and is shown using light colors. The two schedules we consider here are: (a) "Default Schedule" (blue bars), which is the schedule of inspections that the sanitarians originally followed as they conducted inspections and (b) "Schenk Jr. Schedule" (orange bars), which is the schedule obtained using the Schenk Jr. et al. model risk scores and reordering the inspections based on the scores

[6]For a map, see https://en.wikipedia.org/wiki/Chicago#/media/File:Chicago_community_areas_map.svg. Accessed 06/12/2021

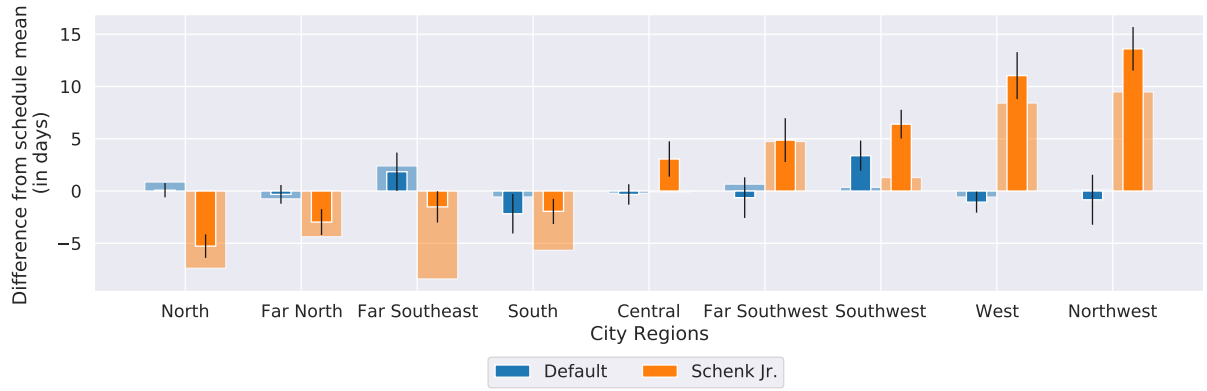

Figure 1: The figure illustrates the difference between the mean time to detect a critical violation in a particular region and the overall mean time for that schedule (EOpp) using the narrower, solid-colored bars. The wider, light colored bars represent the difference in time to conduct inspections regardless of whether a violation was found (DP). The labels represent the major regions of Chicago. Error bars indicate the standard error of EOpp from 16-fold cross validation. (The error bars for DP are similar and omitted for legibility.)

such that inspections with a high risk score (i.e. high predicted likelihood of being a critical violation) are conducted earlier. The bars indicating negative values signify that the detection times are quicker than the schedule mean (the group is better off than average) and the positive values show that the detection times are slower than the schedule mean (the group is worse off than average).

Considering the Default schedule, we observe all of the regions have detection times close to the schedule mean, consistent with a random schedule being perfectly fair. On the other hand, four out of nine sides have quicker detection times than the average under the Schenk Jr. schedule. For the remaining five that are worse off, two sides receive a far greater delay (at least 10 days) in detecting critical violations. The trends for the inspection times are similar and suggest that a large part of the gain in critical violation detection for the advantaged regions under the Schenk Jr. schedule comes from inspections in those regions being moved earlier as a whole rather than specifically the inspections most likely to find critical violations. The breakdown of the detection times by region underscores the disparate outcome the Schenk Jr. schedule would have on food inspections in different regions of the city. If used, an individual's place of residence can determine if they have an expedited or delayed routine inspection of food establishments in their neighborhood, which in turn impacts their likelihood of being subjected to a food-borne illness. Our fairness analysis along racial (§A.6) and economic lines (§A.7) only finds small effects for these groupings.

3.2 Exploring the cause of unfairness

An examination of the coefficients of the Schenk Jr. et al. model (§A.1) shows that the sanitarian conducting the food inspection is a key feature. As the dataset clusters multiple sanitarians together to protect their identity, we examine sanitarian behavior at the cluster level. We first inspect the critical violation rate for each of the sanitarian clusters. The critical violation rate is computed as the ratio

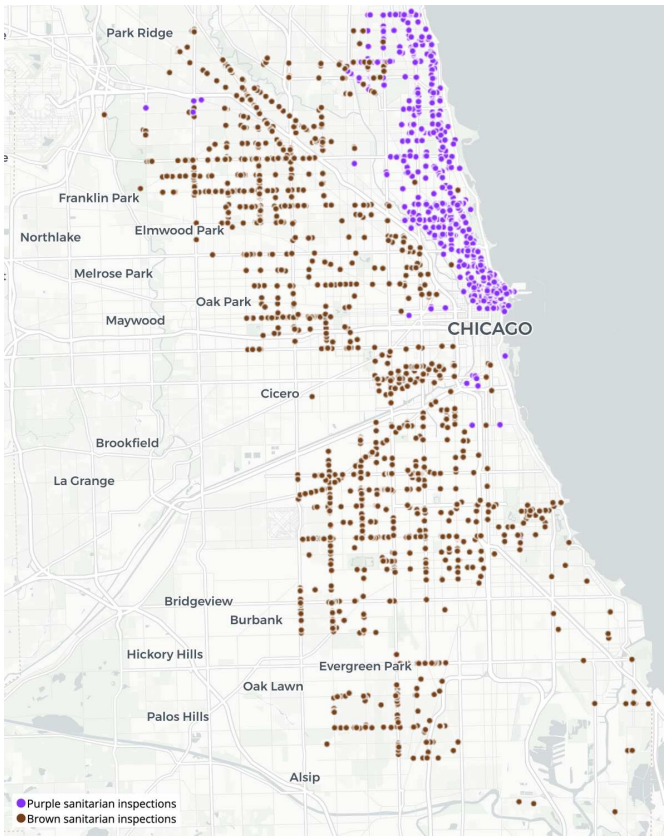

Figure 2: Map of Chicago with purple and brown dots representing the location of food inspections done by Purple (the highest critical violation rate) and Brown (the lowest critical violation rate) cluster sanitarians, respectively.

Table 1: Table showing the inspector clusters and their critical violation rate for the inspections conducting during the model evaluation period.

Sanitarian Cluster	Critical Violation Rate
Purple	40.83%
Blue	25.53%
Orange	13.76%
Green	9.68%
Yellow	5.94%
Brown	2.5%

of the number of inspections that resulted in a critical violation to the total number of inspections conducted by the sanitarians for a cluster. The critical violation rates for each sanitarian cluster vary widely, as shown in Table 1. The Purple sanitarian cluster has the highest rate of citing the restaurants with a critical violation at 41%. On the other hand, the Brown sanitarian cluster has the lowest critical violation rate of 2.5%. Through personal communication with the authors of [31], we learned that the approximately three dozen sanitarians are grouped into six clusters purely based on their critical violation rate. This variation in critical violation rate across sanitarians has at least three possible causes: different strictness among sanitarians, different characteristics of the restaurants inspected, and effects of one inspection on future inspections. In §A.3, we analyze a set of restaurants which had repeat routine inspections by two or more distinct sanitarian clusters. Essentially, we condition on the restaurants being inspected and observe a strong correlation between more critical violations cited and one of the inspections done by a high violation rate sanitarian (e.g. Purple cluster) regardless of the order. This confirms that model unfairness is driven by the sanitarians rather than properties of the restaurants a sanitarian cluster inspects or the timing and nature of inspections by different clusters.

To explore the effects of sanitarian critical violation rate on the unfair outcome for Chicago residents, we plot the location of the inspections on a map of Chicago using the latitudes and longitudes from the dataset. Fig. 2 shows the inspections done by the Purple cluster sanitarians and those done by the Brown cluster sanitarians. We are particularly interested in these two clusters because they represent the sanitarians with the highest and the lowest critical violation rates. We observe that the inspections conducted by Purple cluster sanitarians are concentrated in the North and Central parts of the city. In contrast, Brown cluster inspections are scattered around in the Northwest and Southwest parts of the city. Therefore, the residents living in the North and Central parts of the city are more advantaged by having a smaller time to detect a critical violation detection than the residents living in the other parts of the city. §A.2 shows the maps for all sanitarian clusters, highlighting the various ways they are scattered across the city.

Finally, we plot the difference in detection and inspections times from the schedule means broken down by sanitarian clusters (rather than by regions as was done in Fig. 1) under the Default and Schenk Jr. schedules in Fig. 3. Despite varying violation rates, under the

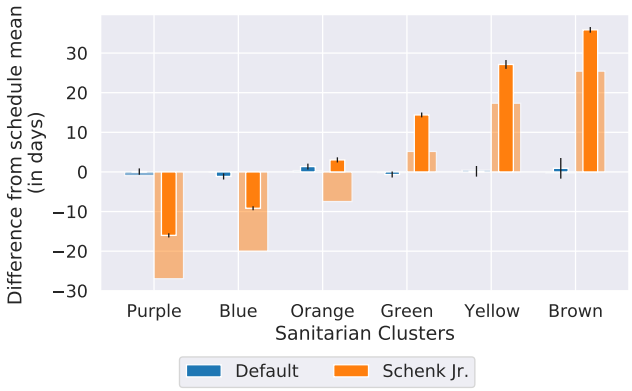

Figure 3: The light-colored bars illustrate the time to conduct an inspection (DP) and the solid-colored bars illustrate the time to detect a critical violation (EOpp) relative to the schedule mean. Lower values are better for the cluster. Error bars indicate the standard error for EOpp. Those for DP are omitted for legibility.

Default schedule all clusters of sanitarians both detect critical violations (solid blue bars) and conduct inspections (light blue bars) at around the same time on average. This shows inspections for different clusters were scheduled at roughly equal times, regardless of their results. On the other hand, under the Schenk Jr. schedule (light orange bars) the inspections are sorted in the order of the critical violation rates (Table 1). The inspections by the Purple sanitarian cluster are scheduled first, and those by the Brown sanitarian cluster are scheduled last. The average times to detect the critical violations follow a similar trend (dark orange bars). This provides further evidence that the model effectively schedules inspections done by the sanitarians in the order of their violation rate.

To summarize, our analysis suggests that the variation in the violation rates across sanitarian clusters and their significance as features in the Schenk Jr. et al. model is one of the major causes of geographic unfairness in the resulting schedule. In the remainder of the paper, we investigate mitigations for both the direct unfairness across sanitarian clusters and the resulting indirect unfairness across regions.

4 FAIRNESS THROUGH MODEL RETRAINING

In this section, we examine techniques aimed at achieving a fair allocation of food inspection times across sanitarian clusters and city regions by retraining the model in ways designed to result in fairer predictions of critical violations. We use the risk scores from the retrained models to reorder the inspections and measure how fair each approach is by computing the difference from the schedule mean in days. Our evaluation results preserve the originally assigned sanitarians clusters and the number of inspections done per day. For brevity, we show only the results for the time to detect a critical violation (EOpp, Eq. 2). Results for DP in §A.4 show similar trends.

Figure 4: A disaggregated view of the time to detect a critical violation under four schedules obtained using different model retraining techniques. The bars show the difference in detection times from the schedule mean across sanitarian clusters (Fig. 4a, top) and geographic groups (Fig. 4b, bottom) with error bars showing the standard error. (Best viewed in color.)

4.1 Remove Sanitarians from the Model

For our first approach, we intervene at the *pre-processing* stage. We train a logistic regression model, the same class of model used by Schenk Jr. et al., but do not give the model access to the sanitarian features. We use the scores from the model to reorder the inspections and call the resulting inspection schedule the "No-Sanitarian" schedule.

Fig. 4 shows the time to detect a critical violation for the No-sanitarian schedule in purple. Although the variation under the No-sanitarian schedule reduces in magnitude compared to the Schenk Jr. schedule, the detection times still differ across the sanitarian clusters (Fig. 4a). Inspections done by Purple cluster sanitarians get a higher priority and their mean times are faster than all other sanitarian clusters. Conversely, Brown cluster sanitarians take the most time to detect critical violations. We also see varied detection times across regions (Fig. 4b). In summary, we observe an improvement over the Schenk Jr. schedule for sanitarian clusters but not a definitive improvement for regions.

Our findings support those from the prior literature [3, 23, 27, 36] that removing a protected feature, in this case the sanitarian cluster, does not remove bias from the model. We believe that the correlation of the remaining dataset features with the sanitarian clusters allows the model to continue to discriminate.

4.2 Fair Regression with Polyvalent Protected Attributes

Now, we implement the approach proposed by Zafar et al. that adds fairness constraints to the logistic regression optimization [34]. Their fairness constraints support polyvalent (non-binary) protected features, like the sanitarian clusters in our case. The model enforces a constraint which limits the allowed covariance between the distance to the boundary of the classifier and the protected attributes on the logistic regression loss optimization. Intuitively, this should avoid the exploitation of correlations we saw with the No-sanitarian schedule. The allowed covariance is a parameter determining the trade-off between fairness and accuracy. We selected the covariance threshold ($c = 0.001$) as the one that produced the fairest outcomes after testing values in $\{0.0, 10^{-6}, 0.001, 0.01, 0.1\}$. The resulting scores from the trained model are used to rearrange the inspections and obtain the "Zafar Schedule".

Fig. 4 shows the results for the Zafar schedule in pink. The detection times vary less in comparison to the No-sanitarian schedule. In particular, the early detection times for Purple cluster sanitarians and later detection times for the Orange cluster are substantially reduced. We see the greatest improvement for the Brown cluster as their detection times are now essentially identical to the overall schedule mean. For geographic groups, the regions that were disadvantaged in the Schenk Jr. and No-sanitarian schedules, namely Far Southwest, Southwest, West, Northwest, see considerable improvements. Also, the detection times for the most advantaged regions (North and Far North) are now slightly worse than the schedule mean. This is consistent with our intuition that the ability of the Zafar et al. model to limit the covariance between the decision boundary and the protected attribute should allow it to eliminate the residual effects of the sanitarian features from the dataset and reach a better outcome than the No-sanitarian schedule.

4.3 Fair Regression with Binary Protected Attributes

Next, we explore the logistic regression model proposed by Rezaei et al.. It robustly optimizes log loss under an adversarial distribution constrained to lie near the distribution from the data and uses constraints to enforce fairness objectives [29]. Their work focuses on three common fairness objectives: Demographic Parity (DP) [5], Equal Opportunity (EOpp) [17], and Equality of Odds [17]. Since we examine EOpp in this section, we use their model for that objective. Rezaei et al. model requires the protected attributes to be binary, so we convert the sanitarian clusters from categorical to binary values by splitting them along their violation rates. We assign the majority protected attribute ($A = 1$) to the inspections conducted by Purple, Blue, and Orange cluster sanitarians which a have higher violation rate compared to the rest (Table 1). Similarly, we assign the remaining inspections done by Green, Yellow and Brown cluster sanitarians to the minority protected attribute ($A = 0$). The model allows a regularization parameter C, and we select its value ($C = 0.5$) that results in the fairest outcome from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

We report the results obtained from the Rezaei et al. model under the fairness constraint of EOpp and term them "Rezaei EOpp Schedule" in Fig. 4 in olive. We observe that the detection times become less fair compared to the Zafar schedule for the sanitarian clusters although the fairness for city regions is closer. We believe one of the reasons the Rezaei et al. model does not perform as well as the Zafar model is rooted in the loss of information when converting the sanitarian cluster values from categorical to binary sensitive values. For example, nothing prevents the model from delaying Orange cluster inspections to prioritize those of the Purple cluster as the two clusters have been combined. This emphasizes the importance of developing fair ML models which accept polyvalent protected attributes rather than limiting analysis to the binary case. Another reason is that the use of robust optimization means that not only is the model's ability to enforce fairness limited by the need to force it on other nearby models, but that for EOpp in particular there are additional technical complications due to the conditioning on true positives in the definition.

4.4 Group Proportional Fair Regression

Finally, we adopt Krishnaswamy et al.'s Proportional Fairness classifier [25]. Rather than protect specified attributes, they provide guarantees for arbitrary, unknown groups. This is achieved by training a randomized classifier which guarantees that, for each possible group, the expected utility is in proportion to that of the group's optimal classifier. The randomized classifier consists of multiple models that are weighted during the training. To get a single risk score to use when scheduling, we calculate the probability the inspection is predicted as critical (i.e. the sum of weights of classifiers that predict an inspection as critical). We call the schedule obtained from this method the "Krishnaswamy Schedule", shown in Fig. 4 in cyan. The results for the Krishnaswamy schedule are similar to the Zafar schedule and a substantial improvement over the Schenk Jr. Schedule. However, the variation in detection times (for both sanitarian clusters and city sides) is still not close to the near-perfect fairness achieved by the Default schedule.

To conclude, the approaches we discuss mitigate the sanitarian effect to an extent. However, we believe none of them offer a complete solution as even the fairest (Zafar and Krishnaswamy) still have substantial variation across regions.

5 FAIRNESS THROUGH MODEL USAGE

In this section, we examine two *post-processing* approaches to reduce model disparity. First, we explore suppressing the sanitarian features during model evaluation. Second, we study the effect of using the model output to schedule the inspections within the sanitarian clusters. As a reminder, we preserve the sanitarian cluster assigned to the inspections in the test set when rescheduling them. We present results for EOpp; similar results for DP are in §A.4.

5.1 Schenk Jr. Schedule with Sanitarians assigned later

A natural way to use the trained model in practice is by predicting the likelihood of an inspection being a critical violation in the absence of a specific sanitarian and doing those inspections first. We do this by keeping the Schenk Jr. et al. model as-is and setting the sanitarian features to be zero during the evaluation periods. We obtain a new schedule by sorting the inspections by the predicted scores and term it the "Sanitarian-blind schedule". This approach is distinct from the No-sanitarian schedule suggested in §4.1. That schedule results from eliminating all information about the sanitarian clusters during *training and rescheduling* phases. The Sanitarian-blind schedule does not modify the Schenk Jr. et al. model but receives no signal related to the sanitarian cluster assignment during *rescheduling*.

In Fig. 5, the detection times for Sanitarian-blind schedule are represented in green. Broadly, the Sanitarian-blind schedule distributes the detection times among sanitarian clusters similarly to the Krishnaswamy schedule. The Purple sanitarian cluster remains the most advantaged group and the Brown the most disadvantaged. The behavior can be attributed to the fact that while we have blinded the sanitarian features, some of the remaining features correlate with them, as discussed in §4.1. The detection times across regions in Fig. 5b reflect an analogous behavior.

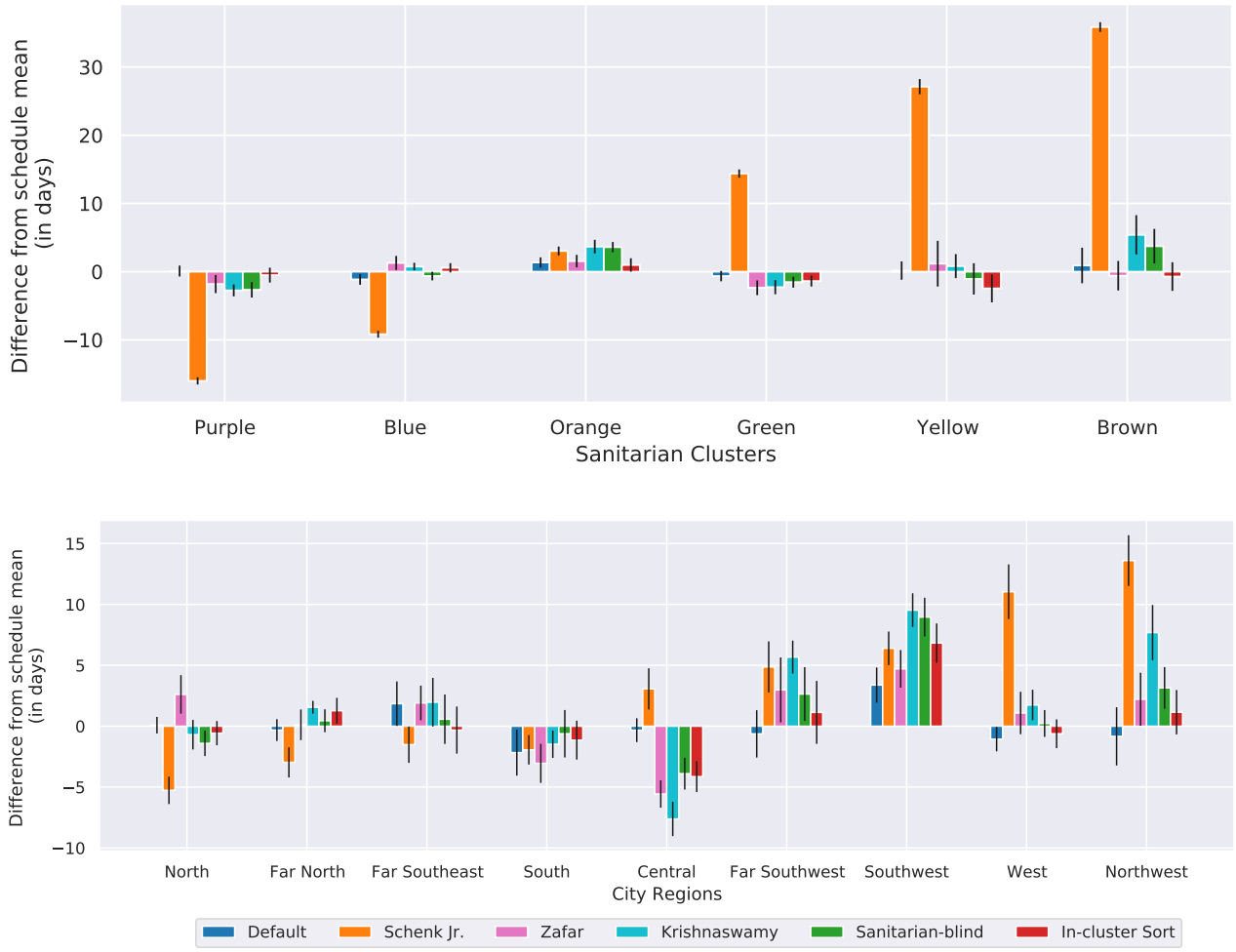

Figure 5: The times to detect violation under the schedules obtained by post-processing techniques. The bars show the difference from the schedule mean grouped by sanitarian clusters (Fig. 5a, top) and sides of Chicago (Fig. 5b, bottom) with error bars giving the standard error. (Best viewed in color.)

5.2 Schenk Jr. Schedule with In-cluster reordering

Another way we could use the model is to first assign each sanitarian a list of restaurant inspections to perform, then use the model to prioritize within each sanitarian's list. The scenario is essentially a "localized" version of the Schenk Jr. et al. objective [31]. We retain the trained model and its predicted scores using all the features for the evaluation periods. Under all the previous approaches, the inspections can be rearranged based on the predicted score without any constraints. For this approach, we consider all the inspections done by each sanitarian cluster separately, sort only those inspections, replace them in the Default schedule, and repeat for each sanitarian cluster. In other words, the resulting schedule keeps the number of inspections each sanitarian cluster conducts each day the same as in the Default schedule. See §A.5 for an illustrated example. We refer to this schedule as the "In-cluster Sort Schedule". Unlike the Sanitarian-blind schedule as described in §5.1, the In-cluster Sort schedule does not lose any information during the rescheduling stage and leverages the information gathered from the extra features available.

Fig. 5 illustrates the performance of the In-Cluster Sort schedule in red color. The results show the In-Cluster Sort produces a more equal outcome and the notable differences in the detection times for the Purple and Brown cluster sanitarians from the Krishnaswamy and Sanitarian-blind schedules have become negligible. Correspondingly, Fig. 4b depicts that the gap in detection times across North and Northwest sides has been bridged as well. These results are achieved despite the limitations of our data only allowing us to implement this intervention at the level of sanitarian clusters rather than at the intended level of individual sanitarians.

6 EFFICIENCY AND FEASIBILITY

In this section we move beyond fairness alone to consider other important aspects of selecting an approach. First we examine the trade-off between fairness and efficiency. Then we consider how the schedules obtained could be used given operational constraints. This raises important questions about the feasibility of the schedules, the choice of right performance metric, and the possibility that some of the efficiency advantages of some methods may be illusory.

6.1 Fairness and Efficiency Trade-off

We begin by defining our measures of efficiency and fairness. Starting from our definition of Equal Opportunity, we take as our notion of efficiency as the mean time to detect a critical violation: $\mu = \mathbb{E}[T \mid Y = 1]$. For fairness, we compute the same metric for each protected group (i.e. sanitarian cluster or region): $\mu_i = \mathbb{E}[T \mid A = a^i, Y = 1]$. We then sum the absolute distance of each of the n groups from the overall mean and use this as our fairness metric:

$$d = \sum_{i=0}^{n-1} |\mu_i - \mu| \ . \tag{5}$$

This approach is similar in spirit to quantifying the extent to which equal opportunity is violated in a classification setting by comparing the difference in the relevant probabilities between groups.[7]

In Fig. 6, we plot the efficiency on the y-axis and fairness on the x-axis. Lower values are better for both. The Default schedule is the most fair but the least efficient. In contrast, the Schenk Jr. schedule is the most efficient but the least fair to the sanitarian clusters. The Zafar and Rezaei algorithms have parameters which have the effect of trading off between efficiency and fairness, so for these we plot a range of parameter values ($c = \{0.001, 0.01, 0.1\}$ and $C = \{0.5, 0.2, 0.1, 0.05, 0.01, 0.005\}$ respectively) and illustrate the trade-off curve they enable with dashed lines. We use a dashed gray line to illustrate the Pareto frontier, the set of schedules that are not dominated in terms of both efficiency and fairness by (a convex combination of) other schedules. The two model usage approaches lie on or near the Pareto frontier for both sanitarian clusters and regions, indicating they represent trade-offs between efficiency and fairness that may be interesting in practice. Neither is clearly better than the other.

Some of the model retraining approaches are near the Pareto frontier for sanitarian clusters, but all are far from it for regions, making their desirability questionable. The Zafar schedule varies its efficiency for a relatively small change in fairness. As fairness decreases, the Zafar schedule overlaps with the No-sanitarian cluster. This is expected as with higher allowed covariance between decision function and protected attributes the model gets more ability to use the residual sanitarian features. The Krishnaswamy and Rezaei schedules appear largely dominated, with the exception of Rezaei toward the efficient but unfair part of the Pareto frontier for sanitarian clusters.

[7]Results for DP are not meaningful as all schedules preserve the number of inspections conducted each day so they have the same efficiency on average.

6.2 Operational Constraints

From §2, we know that the risk scores for Schenk Jr. et al. model are weighted by the sanitarian cluster. Since the inspections are sorted based on the risk score, the Schenk Jr. schedule assumes that all the inspections are fungible.

Consider a scenario when all the inspections done by Purple cluster sanitarians are scheduled first. Would it mean the other sanitarians conduct no inspections during that time? Do the Purple cluster sanitarians remain idle after conducting their inspections early on? If the inspections were reassigned to a different sanitarian cluster, would the result change? These questions point to some of the operational constraints encountered in practice and are not accounted for Schenk Jr. et al.'s methodology. A real scheduling approach needs to be able to account for factors such as limited capacity for a sanitarian to conduct inspections in a day both – time needed to conduct the inspections themselves and the time needed to travel from inspection to inspection. Efficiency gains which do not respect these constraints may be illusory.

Such considerations are another advantage of the post-processing techniques in §5. The Sanitarian-blind schedule works by placing the inspections in an order without needing an assigned sanitarian, allowing later assignment of sanitarians in that respect operational constraints. Likewise, the In-cluster Sort schedule ensures the number of inspections conducted by each cluster each day is reasonable, although it does not account for travel times.[8]

7 DISCUSSION

We have revisited the application of predictive models by CDPH to schedule restaurant inspections and performed the first analysis from the perspective of fairness to the population served by the restaurants. We found that the model treats inspections unequally based on the cluster of the sanitarian who conducted the inspection and that there are, as a result, geographic disparities in the benefits of the model. We examined approaches to using the original model in a fairer way and ways to train the model to achieve fairness and found more success with the former class of approaches.

While our analysis and conclusions are limited to a single data set from the city of Chicago and the particular algorithmic approaches tested, we believe this setting is representative of an important class of problems. Our communications with experts in food safety suggest that the resource allocation problems and wide differences in violation rates faced by Chicago are common in many jurisdictions. Beyond food safety, cities conduct other types of inspections including of structural inspections of buildings, fire safety, business licensing, and enforcement of environmental and accessibility regulations. Thus, we conclude by discussing three broad challenges our results point to for future work.

In contrast to much of the literature that focuses on the fair treatment of individuals, things being inspected typically have many stakeholders. In this work we have taken the simple approach of identifying restaurants with the people who live nearby, but this is certainly a rough approximation at best. There is a need for better methods to understand who is affected by inspection decisions and how. A related problem is understanding and quantifying the

[8]Ideally we would rearrange inspections at the level of individual sanitarians, but data limitations only allow us to treat inspections by cluster as fungible.

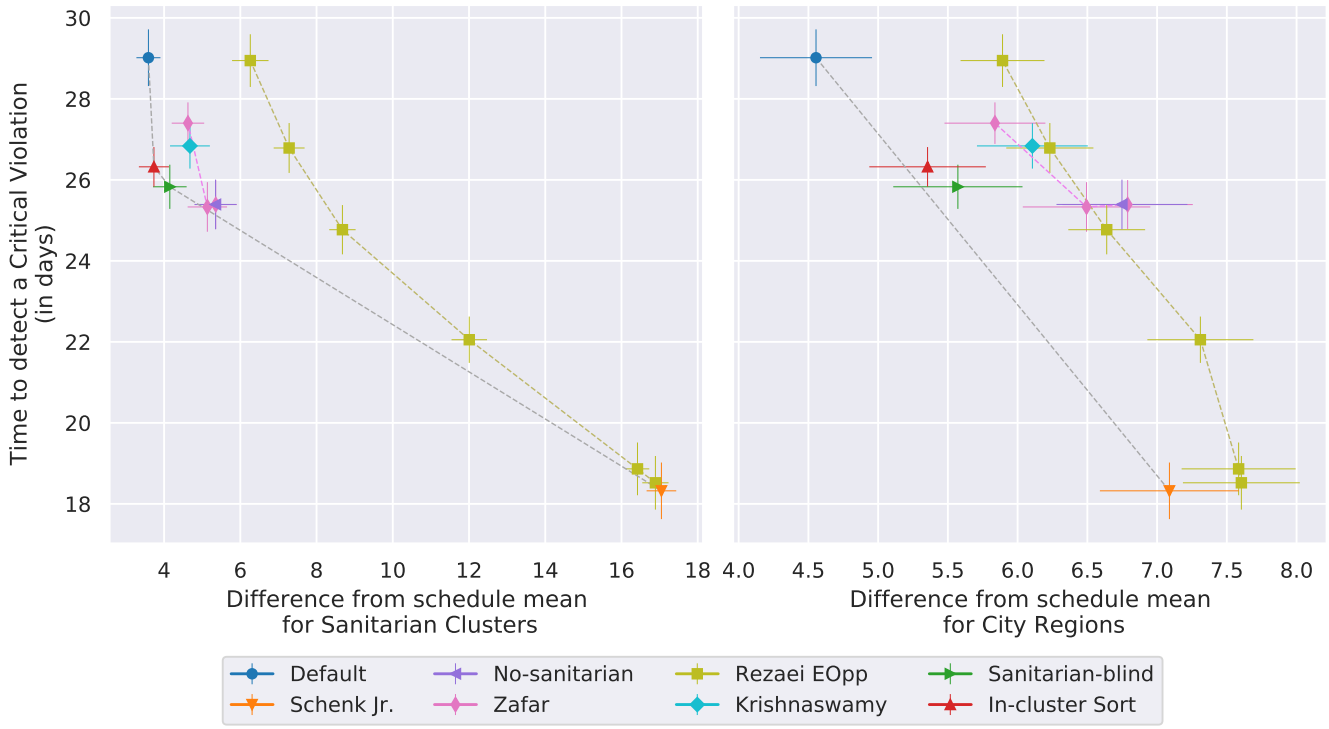

Figure 6: A scatter plot showing the trade-off between the time to detect a violation (EOpp, y-axis) and the fairness which is computed as the average of absolute distance from the mean across all group (x-axis), as defined in Eq. 5. A lower mean detection time and a lower distance from mean are desirable. Error bars give the standard error from 16 cross-validated runs.

effects of inspection scheduling across groups based on race or economic status. The approach we explored (see supplementary material §A.6,§A.7) found limited fairness effects for these groupings but it is unclear whether this is because the algorithms were in fact fair or the approach does a poor job of quantifying the fairness effects. Beyond simply measuring fairness, developing fair classification algorithms that can handle the sort of continuous-valued protected attributes that arise when the data captures the demographic breakdown of, e.g., a neighborhood is a largely unexplored challenge.

While the goal of inspections is to protect public health, their effectiveness is challenging to measure directly. We have followed Chicago's approach of using detecting critical violations of the food code as early as possible as a proxy. The use of proxies is common, and has caused notable issues in other domains (for example the use of arrests as a proxy for crime [14]). The risk of feedback loops has been pointed out in both this and other domains [6, 22]. However, we wish to stress that sanitarians have discretion in how they resolve issues they observe, ranging from punishment in the form of critical violations to education and helping restaurant owners correct issues in the course of the inspection. So a low violation rate is not necessarily indicative of a sanitarian simply missing issues. Prior work has found that factors such as the outcome of a previous inspection and the position of an inspection in an inspector's daily schedule may significantly impact the detection of violations in an inspection [18]. This raises difficult questions about what it means to be fair. Our approach of reordering within each sanitarian cluster ducks this issue to some extent, assuming what a critical violation "means" to a given sanitarian is consistent across time (although this may not eliminate all issues; see Finding 2 of [22]). However, questions remain including how this can provide fairness guarantees to individuals and whether all critical violations are equally bad. Given the range of violation rates, it seems likely that some restaurants with no critical violation inspected by Brown cluster sanitarians actually deserve more scrutiny than many restaurants with a critical violation inspected by Purple cluster sanitarians, meaning some of the increased performance of the original model may be illusory. [9] What is a better proxy for sanitarians who find critical violations only rarely? Should we be not just reordering inspections but actively shaping which sanitarian performs them to enhance fairness?

Finally, while the models we use are trained to perform classification, their use in this context is for ranking which in turn is used for scheduling. There is room to improve over our approach at all stages of this pipeline. Would it be good to instead learn a counterfactual "sanitarian-independent" violation probability, as is done when predicting clicks in search advertising [16] and has been explored in the literature on causal models in fairness [23]? Rather than trying to achieve fair classification or doing the ranking in

[9] This can also be viewed as an issue of unfairness to restaurants [22].

ways that address unfairness in the classification, are there better approaches that directly leverage ideas from the literature on fair rankings [32, 35] or the literature on fair classification in the context of larger systems [10]? We have treated scheduling as a single, static problem, but inspections occur on an ongoing basis. How should we understand and achieve fairness in the full, dynamic setting? This last question in particular points to potentially fruitful ways to study this domain in light of the literatures on fairness in reinforcement learning [20, 33] and overall fairness in comparison to local or immediate fairness [9, 11, 12, 26].

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Grants IIS-1939743 and CNS-1801644.

REFERENCES

[1] Kristen M Altenburger and Daniel E Ho. 2019. Is Yelp actually cleaning up the restaurant industry? A re-analysis on the relative usefulness of consumer reviews. In *The World Wide Web Conference*. 2543–2550.

[2] Zoe C Ashwood, Becky Elias, and Daniel E Ho. 2021. Improving the Reliability of Food Safety Disclosure: Restaurant Grading in Seattle and King County, Washington. *Journal of environmental health* 84, 2 (2021), 30–.

[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[4] Jack Blandin and Ian Kash. 2021. Fairness Through Counterfactual Utilities. *arXiv preprint arXiv:2108.05315* (2021).

[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*. 13–18. https://doi.org/10.1109/ICDMW.2009.83 ISSN: 2375-9259.

[6] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.

[7] Cary Coglianese and Lavi Ben Dor. 2020. AI in Adjudication and Administration. *Brooklyn Law Review, Forthcoming, U of Penn Law School, Public Law Research Paper* 19-41 (2020).

[8] M A Cruz, D J Katz, and J A Suarez. 2001. An assessment of the ability of routine restaurant inspections to predict food-borne outbreaks in Miami-Dade County, Florida. *American Journal of Public Health* 91, 5 (May 2001), 821–823. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1446663/

[9] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[10] Cynthia Dwork and Christina Ilvento. 2018. Fairness Under Composition. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[11] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. 2020. Individual fairness in pipelines. *arXiv preprint arXiv:2004.05167* (2020).

[12] Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna Gummadi, and Patrick Loiseau. 2019. The price of local fairness in multistage selection. *arXiv preprint arXiv:1906.06613* (2019).

[13] David Freeman Engstrom, Daniel E Ho, Catherine M Sharkey, and Mariano-Florentino Cuéllar. 2020. Government by algorithm: Artificial intelligence in federal administrative agencies. *Available at SSRN 3551505* (2020).

[14] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*. PMLR, 160–171.

[15] Edward L. Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. *American Economic Review* 106, 5 (May 2016), 114–18. https://doi.org/10.1257/aer.p20161027

[16] Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. 2010. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *ICML*.

[17] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[18] Maria R Ibanez and Michael W Toffel. 2020. How Scheduling Can Bias Quality Assessment: Evidence from Food-Safety Inspections. *Management science* 66, 6 (2020), 2396–2416.

[19] K Irwin, J Ballard, J Grendon, and J Kobayashi. 1989. Results of routine restaurant inspections can predict outbreaks of foodborne illness: the Seattle-King County experience. *American Journal of Public Health* 79, 5 (May 1989), 586–590. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1349498/

[20] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 1617–1626.

[21] Timothy F. Jones, Boris I. Pavlin, Bonnie J. LaFleur, L. Amanda Ingram, and William Schaffner. 2004. Restaurant inspection scores and foodborne disease. *Emerging Infectious Diseases* 10, 4 (April 2004), 688–692. https://doi.org/10.3201/eid1004.030343

[22] Vinesh Kannan, Matthew A Shapiro, and Mustafa Bilgic. 2019. Hindsight Analysis of the Chicago Food Inspection Forecasting Model. *arXiv preprint arXiv:1910.04906* (2019).

[23] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744* (2017).

[24] Bran Knowles and John T. Richards. 2021. The Sanction of Authority: Promoting Public Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 262–271. https://doi.org/10.1145/3442188.3445890

[25] Anilesh Krishnaswamy, Zhihao Jiang, Kangning Wang, Yu Cheng, and Kamesh Munagala. 2021. Fair for All: Best-effort Fairness Guarantees for Classification. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*, Arindam Banerjee and Kenji Fukumizu (Eds.). PMLR, 3259–3267. https://proceedings.mlr.press/v130/krishnaswamy21a.html

[26] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.

[27] Kristian Lum and James Johndrow. 2016. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077* (2016).

[28] Keegan McBride, Gerli Aavik, Maarja Toots, Tarmo Kalvet, and Robert Krimmer. 2019. How does open government data driven co-creation occur? Six factors and a 'perfect storm'; insights from Chicago's food inspection forecasting model. *Government Information Quarterly* 36, 1 (Jan. 2019), 88–97. https://doi.org/10.1016/j.giq.2018.11.006

[29] Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. 2020. Fairness for Robust Log Loss Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (Apr. 2020), 5511–5518. https://doi.org/10.1609/aaai.v34i04.6002

[30] Adam Sadilek, Stephanie Caty, Lauren DiPrete, Raed Mansour, Tom Schenk, Mark Bergtholdt, Ashish Jha, Prem Ramaswami, and Evgeniy Gabrilovich. 2018. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *npj Digital Medicine* 1, 1 (Nov. 2018), 1–7. https://doi.org/10.1038/s41746-018-0045-1 Number: 1 Publisher: Nature Publishing Group.

[31] T. Schenk Jr., G. Leynes, A. Solanki, S. Collins, G. Smart, B. Albright, and D. Crippin. 2015. *Forecasting restaurants with critical violations in Chicago*. Technical Report.

[32] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. *arXiv preprint arXiv:1902.04056* (2019).

[33] Min Wen, Osbert Bastani, and Ufuk Topcu. 2021. Algorithms for Fairness in Sequential Decision Making. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1144–1152.

[34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 962–970. https://proceedings.mlr.press/v54/zafar17a.html

[35] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in Ranking: A Survey. *arXiv preprint arXiv:2103.14000* (2021).

[36] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.