Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling

Michele Samorani
Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA
{msamorani@scu.edu}

Shannon L Harris
School of Business, Virginia Commonwealth University, Richmond, VA 23284, USA
{harriss10@vcu.edu}

Linda Goler Blount
Black Women's Health Imperative, Washington, DC 20003, USA
{lgblount@bwhi.org}

Haibing Lu
Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA
{hlu@scu.edu}

Michael A. Santoro
Leavey School of Business, Santa Clara University, Santa Clara, CA 95053, USA
{masantoro@scu.edu}

Machine learning is often employed in appointment scheduling to identify the patients with the greatest no-show risk, so as to schedule them into overbooked slots, and thereby maximize the clinic performance, as measured by a weighted sum of all patients' waiting time and the provider's overtime and idle time. However, if the patients with the greatest no-show risk belong to the same demographic group, then that demographic group will be scheduled in overbooked slots disproportionately to the general population. This is problematic because patients scheduled in those slots tend to have a worse service experience than the other patients, as measured by the time they spend in the waiting room. Such negative experience may decrease patient's engagement and, in turn, further increase no-shows. Motivated by the real-world case of a large specialty clinic whose black patients have a higher no-show probability than non-black patients, we demonstrate that combining machine learning with scheduling optimization causes racial disparity in terms of patient waiting time. Our solution to eliminate this disparity while maintaining the benefits derived from machine learning consists of explicitly including the objective of minimizing racial disparity. We validate our solution method both on simulated data and real-world data, and find that racial disparity can be completely eliminated with no significant increase in scheduling cost when compared to the traditional predictive overbooking framework.

Keywords: Appointment Scheduling, Machine Learning, Algorithmic Bias, Socio-economic Bias, Racial Bias

Revision date: October 9, 2019

1. Introduction

Providing affordable, inclusive, and timely access to quality healthcare has become one of the most pressing issues in our society (Dai and Tayur, 2019). Appointment scheduling in medical offices is one of the most common ways to access medical services, and has attracted considerable attention from the management science research community (Ahmadi-Javid et al. 2017).

Patient no-shows represent a major challenge for effective appointment scheduling at outpatient medical clinics, because they reduce provider utilization, ultimately resulting in delayed patient access to health care. A popular way to counteract no-shows is to overbook appointments. Although overbooking increases the expected number of showing patients and decreases idle time, it also introduces the undesirable effects of patient waiting time, incurred when a patient's visit starts late because of overcrowding, and provider overtime, incurred if the provider needs to work beyond the nominal end of the clinic session in order to finish seeing all patients. In a typical scheduling environment, clinics are interested in scheduling a given set of patients into their appointment slots, so as to minimize a weighted sum of the provider's idle time and overtime and the patients' waiting time.

Recent work in appointment scheduling indicates that clinic costs due to idle time, overtime, and patients' waiting time can be substantially reduced by combining machine learning and optimization into a framework called "predictive overbooking" (Figure 1). The predictive overbooking framework consists of a predictive model and an optimization model. Given a set of N appointment requests $(R_1, R_2, \dots, R_N)$, a predictive model predicts their individual probabilities of show $(p_1, p_2, \dots, p_N)$, and an optimization model is subsequently used to optimally schedule the appointment requests based upon the estimated probabilities. The objective of the optimization model is to minimize a weighted sum of patients' waiting time and provider overtime and idle time.

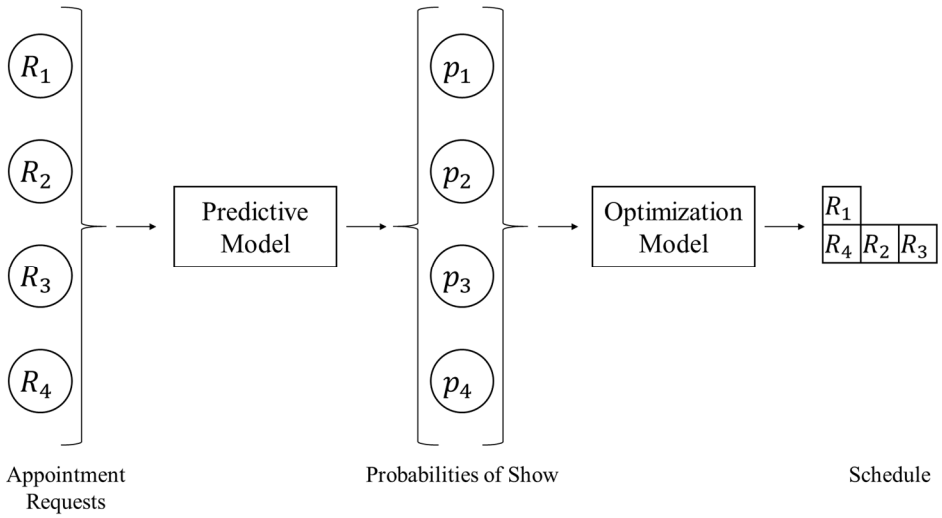

Figure 1: The predictive overbooking framework in a case with four appointment requests.

Zacharias and Pinedo (2014) noted that any schedule can be viewed as "a concatenation of alternating vertical and horizontal segments", where a vertical segment v is an overbooked slot (graphically depicted with a vertical stack of slots) and a horizontal segment h is a sequence of slots that are not overbooked (see Figure 2). Throughout the paper, we use the term "segment" to denote a pair of vertical and horizontal segments.

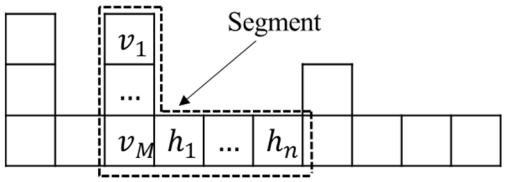

Figure 2: A segment in a schedule is made of a vertical segment v and a horizontal segment h

If it is assumed that all patients have the same medical priority, Zacharias and Pinedo (2014) proved that within a segment, it is optimal to schedule patients in increasing order of their show probability. In reference to Figure 2, this means that patients v_1 to v_M all have a lower (i.e., not higher) show probability than patient h_1, and that h_1 has a lower show probability than h_2, who has a lower probability than h_3, and so on until h_n, who has the highest show probability among the patients in this segment.

While this scheduling strategy is optimal from a pure cost-minimization point of view, it may have unintended ethical consequences. As the sociologist Ruha Benjamin puts it, "the road to inequity is paved with technical fixes" in the name of achieving "objectivity, efficiency, profitability, and progress" (Benjamin, 2019). We will prove that patients scheduled in an overbooked appointment slot (i.e., v_1 to v_M in Figure 2) or in the immediately following slot (i.e., h_1) tend to suffer a longer waiting time than patients scheduled in the rest of the horizontal segment (i.e., h_2 to h_n). Thus, if no-show behavior is correlated with the patients' race, then overbooking patients based upon their predicted show probabilities will lead to one racial group being disproportionately scheduled toward the left end of each segment, thereby facing longer waiting times than the other groups.

Although a large number of studies have found that race is correlated with no-show probability (see the survey by Dantas et al. 2018), no existing study has recognized that, because of this, the predictive overbooking framework may result in significantly longer waiting times for a racial group. This disparity in waiting time for black patients – who oftentimes are the patient class with the greatest probability of no-show – is especially unjust due to the evidence that those patients generally have inferior access to healthcare, receive poorer quality care, and experience worse healthcare outcomes (Centers for Disease Control, 2013).

While it is beyond the scope of this paper to offer an extensive discussion of the underlying ethical argument for correcting disparate racial impacts, it should be pointed out that it would be fundamentally

unethical to punish black patients for their lower show rate. This is because academic studies have shown that race is highly correlated with socioeconomic obstacles deeply rooted in historical racial discrimination which make black patients less likely to be able to make it to appointments than white patients (Williams et al. 2010). Therefore, any suggestion that patients with a high risk of no-show "deserve" to be given inferior scheduling slots would in essence be penalizing black people for the discrimination and socioeconomic conditions that they have historically suffered.

Motivated by the data set of an outpatient clinic whose black patients have a higher no-show rate than white patients, in this paper we assess the disparate impact that the predictive overbooking framework has on the waiting times of the different racial groups, and provide a solution method to remove racial disparity with modest effects on clinic costs.

We first develop analytical insights to prove that the predictive overbooking framework tend to result in longer waiting times for the racial group at higher risk of no-show; we also prove that this racial disparity is directly proportional to the prediction performance. Second, we develop an appointment scheduling methodology to eliminate racial disparity in waiting times by explicitly taking into account the racial groups in the objective function. Instead of minimizing the waiting time of all patients, our new objective minimizes the waiting time of the racial group that waits the longest. Our results suggest that doing so has the potential of completely removing racial disparity at a modest increase in scheduling cost.

2. Literature Review

There are three topic areas which we will review for this paper. The first area is the emerging subject of algorithmic bias in health care. Gianfrancesco et al. (2018) express the "concern that biases and deficiencies in the data used by machine learning algorithms may contribute to socioeconomic disparities in health care". Rajkomar et al. (2018) echo this concern and provide the recommendation to seek to obtain "equal outcome", and not just equal prediction performance across the groups. To do this, they also recommend to take into consideration the membership to a racial group explicitly instead of adopting "the commonly discussed fairness principle of unawareness which states that a model should not use the membership of the group as a feature".

The second area of research includes empirical studies in appointment scheduling. It is well known that race and ethnicity are correlated with the probability of patient no-shows. After surveying 105 empirical studies on no-shows, Dantas et al. (2018) concluded that "minority groups were consistently associated with increased no-show, but not surprisingly different groups were considered minorities in different countries (e.g., Hispanics and Afro-Americans in the United States)".

The third area of research includes studies that developed methodologies to schedule patients based on their individual no-show probabilities. Li et al. (2019), Samorani and LaGanga (2015), Srinivas and Ravindran (2018), Zacharias and Pinedo (2014), Samorani and Harris (2019) are some examples of a

growing body of literature that promotes the predictive overbooking procedure depicted in Figure 1. The goal of these papers is to minimize the clinic cost (typically, the patients' waiting time and the provider's overtime and idle time) using individual no-show probabilities. However, despite the presence of a large body of work in algorithmic bias in health care and the empirical evidence that race and no-show probabilities are correlated, these studies fail to recognize that predictive overbooking may result in racially biased decisions.

To the best of our knowledge, our work is the first one to measure and address the racial disparity that takes place in appointment scheduling.

2. Clinic Model, Assumptions, and Properties

2.1. Clinic Model

We consider an outpatient clinic where one provider sees patients sequentially during the clinic day. The input consists of a set of N appointment requests with individual show probabilities p_i $(i = 1, \dots, N)$. We typically utilize show probabilities in our model, but occasionally utilize the no-show probabilities, $(1 - p_i)$. The scheduling problem consists of assigning each appointment request to one of F appointment slots. Each patient belongs to one of two racial groups, G_1 or G_2. Throughout the paper, group G_1 denotes the group with the greater risk of no-shows. All of our proofs are valid for any number of racial groups, but the current work focuses on the two-group case.

If $N > F$, there is overbooking, that is, at least one slot will be assigned to more than one appointment request. We assume that all patients who show up are punctual, and that the provider sequentially sees the patients who show up; the time taken for each appointment is assumed constant and equal to the length of one appointment slot (this assumption is relaxed in Appendix C).

If no patient is present at the beginning of a slot, the provider stays idle for the duration of that slot. If more than one patient is present at the beginning of a slot, the provider sees the one with the earliest scheduled time, while all others wait for at least the duration of the slot. In the case of ties (e.g., if two patients overbooked in the same slot show up), the provider selects a patient at random among those sharing the earliest scheduled time. This way of breaking ties reflects the fact that patients typically check in sequentially even if they arrive at the same time, and are in the order they are scheduled. If there are patients present at the end of the regular clinic session, the provider will see them sequentially in overtime (i.e., slot $F + 1$, $F + 2$, etc).

2.2. The Traditional Objective Function (TOF)

We now introduce the traditional objective function (TOF) for the appointment scheduling problem. The TOF objective is to minimize a weighted sum of the patients' expected waiting time and the provider's expected overtime and idle time. Patient i incurs a waiting time cost if his/her appointment starts late, at a

rate of ω for every time unit of delay; idle time cost is incurred at a rate of ρ for every time unit of idle time, whereas overtime cost is incurred whenever the provider finishes seeing the patients after the nominal end time of the clinic session (i.e., after F time units from the start), at a rate of τ for every time unit of overtime. Without loss of generality, we fix the length of a time unit to the length of one appointment slot, and we fix the waiting time cost to $\omega = 1$. The TOF is as follows:

$$TOF = \min E[cost] = \min\left(\sum_{i=1}^{N}(\omega \cdot E[wt_i]) + \rho \cdot E[it] + \tau \cdot E[ot]\right), \quad (1)$$

where wt_i is the waiting time experienced by patient i and it and ot are the provider's idle tme and overtime, respectively. $E[\cdot]$ represents the expected value. When a patient does not show up, his/her waiting time is zero.

2.3. Analytical Connection between TOF and Racial Disparity

We now articulate why TOF is likely to lead to one racial group experiencing waiting times longer than another. First, Zacharias and Pinedo (2014)'s Corollary 2 proved that, in order to minimize TOF, within each schedule segment it is optimal to schedule the patients in increasing order of their show probability, with the patients at greatest risk of no-show scheduled in the vertical segment. Now we show that the patients at greatest risk of no-show are also those who tend to experience the longest waiting times. We first define a patient's "conditional waiting time" (CWT) as his/her waiting time conditional to showing up. The CWT is an important metric in this study, because any disparity between patients is assessed by measuring the waiting times of the patients that show up. All proofs are in Appendix A.

PROPOSITION 1: If i and j are two patients scheduled in an overbooked slot (vertical segment) and i has a lower show probability than j, then i has a longer CWT than j.

PROPOSITION 2: If i and j are two patients scheduled, respectively, in slots t and u ($u > t$) of a horizontal segment, then the i's CWT is greater than or equal to j's CWT.

The two propositions above show that the patients at highest risk of no-show are also those who tend to experience the longest waiting time within a vertical segment and within a horizontal segment. Now, we find sufficient conditions for which a patient scheduled in a vertical segment waits longer than a patient scheduled in the next slot.

LEMMA 1: A showing patient i scheduled in an overbooked slot expects to wait longer than any patient scheduled in the slot right after if $\frac{s^+}{2} + p_i \leq 1$, where p_i is patient i's show probability and s^+ is the number

of expected shows (conditional to observing at least one show) among all patients in the overbooked slot except patient i.

To generate easily interpretable insights, we reformulate Lemma 1 in the cases where the vertical segment contains two or three patients (i.e., a slot is at most triple-booked), which are the cases most commonly found in practice:

PROPOSITION 3 (double booking): A showing patient i *scheduled in slot* t *with one other patient, expects to wait longer than any patient scheduled in slot* $t+1$*, if s/he has a show probability* $p_i < 0.5$.

PROPOSITION 4 (triple booking): A showing patient i *scheduled in slot* t *with two other patients whose show probabilities are* p_1 *and* p_2*, expects to wait longer than any patient scheduled in slot* $t+1$*, if s/he has a show probability* $p_i \leq \frac{p_1+p_2-2p_1p_2}{2p_1+2p_2-2p_1p_2}$.

Zacharias and Pinedo (2014)'s proofs, together with our Propositions 2, 3, and 4 suggest that the patients with very low show probabilities, who usually belong to G_1, tend to be overbooked and, when they show up, end up waiting the longest. We now show that if a predictive model is used to predict no-shows, then the better the prediction performance, the more likely it is for G_1 patients to be overbooked. Limited to this proof, we consider the simplified case of a binary classifier, whose predictions are binary show outcomes.

PROPOSITION 5: If a binary classifier is used to predict the patients' show probabilities and TOF is used to schedule the appointments, then increasing its sensitivity or specificity will result in a larger proportion of G_1 *customers to be overbooked.*

We leave to future work the extension of this proof to the case where the classifier provides show probabilities instead of binary show outcomes.

In this section, we analytically showed that minimizing TOF, when coupled with machine learning, will disproportionately overbook patients from G_1. The lower G_1 patients' show probability, the more likely that they will wait longer than G_2 patients. Next, we will develop a new objective function that attempts to minimize this racial disparity.

3. The Unbiased Objective Function (UOF)

In this section, we develop an alternative objective function, the Unbiased Objective Function (UOF), whose goal is to strike a balance between minimizing the clinic cost (as TOF does) and minimizing racial disparity between the groups.

First, in order to compare the waiting times experienced by different racial groups, we need to define the expected waiting time of a group of patients. Defining it as the sum of the individual patients' expected waiting times (as in (1)) is unsuitable for the task of comparing the waiting time suffered by different patient groups of potentially different cardinalities, because computing the sum will penalize the least numerous group. Thus, given Y groups (G_1, G_2,…, G_Y),we compute the expected waiting time of group G_y as:

$$E[W_y] = \frac{\sum_{i \in G_y} E[wt_i]}{E[\#shows\ in\ G_y]}, y \in \{1,2,\dots,Y\} \quad (2)$$

where $E[wt_i]$ is the expected waiting time experienced by patient i, G_y contains the indices of the patients belonging to group G_y, and $E[\#shows\ in\ G_y]$ is the expected number of showing patients in G_y. Although there are other ways to compute a group's waiting time (see Appendix B), this formulation is better aligned with computing the average waiting time among the showing patients of that group. As shown in Section 4, this formulation also has properties that make it possible to efficiently solve the scheduling problem.

We now turn our attention to developing an objective function which, in addition to minimizing waiting time, idle time, and overtime, also minimizes racial disparity. In the case of only two racial groups, racial disparity can be defined as the absolute value of the difference between the expected waiting times of the groups. Explicitly adding a "racial disparity" component to TOF has two problems. First, we would need to decide the "weight" of this new objective in relation to the other three (waiting time, idle time, and overtime). An excessively large weight could lead to undesirable schedules, such as one where all patients wait a very long time, but the two groups wait approximately the same time. Second, it is unclear how to define disparity in the case of more than two groups.

Guided by these considerations, we propose an objective function, which we call "Unbiased Objective Function" (UOF), which does not need new parameters and works for any number of racial groups. Instead of minimizing the disparity among groups, UOF minimizes the waiting time of the group waiting the longest (i.e., a min-max objective function):

$$UOF = \min(\omega \cdot W^{max} + \rho \cdot E[it] + \tau \cdot E[ot]) \quad (3),$$

where

$$W^{max} = E[\#total\ shows] \max_{y \in \{1,\dots,Y\}} (E[W_y]) \quad (4),$$

Where W^{max} is the "scaled" waiting time of the group waiting the longest. It is scaled because the group's waiting time is multiplied by the expected number of shows among all patients, $E[\#total\ shows]$. Thanks to this scaling, W^{max} can be interpreted as the sum of all patients' waiting times, under the assumption that all patients' waiting time is the same as the average waiting time of the group that waits the longest. Note that if there is only one patient group, then the denominator of (2) is equal to $E[\#total\ shows]$, and UOF reduces to TOF.

4. Analytical Properties of UOF

We now derive optimality conditions for the appointment scheduling problem with UOF as objective:

PROPOSITION 6:

- *(i) Under reasonable conditions[1], there exists a schedule which minimizes UOF with no empty slots.*
- *(ii) There exists a schedule which minimizes UOF in which, within each segment, the patients of the same racial group are sorted by increasing show probability.*

Thanks to these properties, the solution space is dramatically reduced, making it possible to optimize UOF efficiently through a complete enumeration procedure.

5. Computational Results

In this section, we generate a large number of scheduling problems using different parameter combinations, find the optimal schedules that minimize TOF and UOF, and compare their quality in terms of cost and of racial disparity.

The scheduling problems are generated as follows. A set of N appointment requests is generated by assuming that each request has a probability of 50% to belong to either group G_1 or G_2. The average show probability of G_1 is set to $p_1 = q - \delta$, while the average show probability of G_2 is set to $p_2 = q + \delta$, where q is the population show rate and δ a parameter. Depending on their group, the patients' show probabilities are sampled from two beta distributions $\beta(p_1, v)$ and $\beta(p_2, v)$, where v is the variance. We consider the following parameter combinations: $(N, F) \in \{(4,2), (4,3), (6,4), (7,5), (8,6)\}$, which imply a population show rate q of 0.50, 0.75, 0.67, 0.71, and 0.75, respectively; $\delta \in \{0.05, 0.10\}$; and $v \in \{0.05, 0.10\}$. The idle time and overtime cost per time unit are set to twice and eight times the cost of waiting time, respectively (i.e., $\rho = 2$ and $\tau = 8$). To better interpret the results, we assume that each appointment slot is 30 minutes.

[1] If slots containing more than four expected shows are not allowed.

Table 1: Computational results on simulated scheduling problems.

Parameters			Schedule Cost		TOF Waiting Time in minutes		UOF Waiting Time in minutes		Overtime (minutes)		Disparity $(G_1 - G_2)$ (minutes)		Effect of UOF over TOF		
													Disparity Reduction	Cost Increase	Overtime increase (minutes)
N, F	p_1, p_2	v	TOF	UOF	G_1	G_2	G_1	G_2	TOF	UOF	TOF	UOF			
4,2	0.4,0.6	0.05	4.61	4.64	12.46	9.14	11.82	10.79	9.26	9.17	3.32	1.03	68.98%	0.65%	-0.09
4,2	0.4,0.6	0.1	4.65	4.71	9.89	7.08	9.98	8.6	9.88	9.8	2.81	1.38	50.89%	1.29%	-0.08
4,2	0.45,0.55	0.05	4.66	4.67	12.08	10.48	12.09	11.59	9.3	9.21	1.6	0.5	68.75%	0.21%	-0.09
4,2	0.45,0.55	0.1	4.81	4.86	9.61	8.38	10.01	9.46	10.3	10.21	1.23	0.55	55.28%	1.04%	-0.09
4,3	0.65,0.85	0.05	3.60	3.73	7.47	2.55	7.01	5.47	7.42	7.37	4.92	1.54	68.70%	3.61%	-0.05
4,3	0.65,0.85	0.1	3.48	3.55	5.54	1.28	4.46	2.61	7.65	7.7	4.26	1.85	56.57%	2.01%	0.05
4,3	0.7,0.8	0.05	4.01	4.09	6.91	4.9	7.02	6.49	8.4	8.34	2.01	0.53	73.63%	2.00%	-0.06
4,3	0.7,0.8	0.1	3.55	3.62	4.74	2.2	4.08	3.37	7.68	7.71	2.54	0.71	72.05%	1.97%	0.03
6,4	0.57,0.77	0.05	6.15	6.36	13.5	6.68	11	10.38	11.93	11.89	6.82	0.62	90.91%	3.41%	-0.04
6,4	0.57,0.77	0.1	5.52	5.76	8.99	3.51	7.13	6.4	11.62	11.73	5.48	0.73	86.68%	4.35%	0.11
6,4	0.62,0.72	0.05	6.34	6.45	12.64	8.66	11.07	10.84	12.12	12.06	3.98	0.23	94.22%	1.74%	-0.06
6,4	0.62,0.72	0.1	5.66	5.79	8.27	5.31	6.94	7.09	11.46	11.46	2.96	0.15	94.93%	2.30%	0.00
7,5	0.61,0.81	0.05	6.52	6.85	12.76	5.11	10.09	9.4	12.49	12.53	7.65	0.69	90.98%	5.06%	0.04
7,5	0.61,0.81	0.1	5.72	6.01	8.68	2.26	6.16	4.92	11.95	12.22	6.42	1.24	80.69%	5.07%	0.27
7,5	0.66,0.76	0.05	6.70	6.86	11.72	7.93	10.25	9.85	12.42	12.42	3.79	0.4	89.45%	2.39%	0.00
7,5	0.66,0.76	0.1	6.10	6.27	7.03	4.17	5.86	5.57	12.67	12.79	2.86	0.29	89.86%	2.79%	0.12
8,6	0.65,0.85	0.05	6.81	7.22	11.82	4.04	8.95	8.2	13.14	13.4	7.78	0.75	90.36%	6.02%	0.26
8,6	0.65,0.85	0.1	5.64	5.92	7.89	0.84	5.3	3.23	11.54	11.93	7.05	2.07	70.64%	4.96%	0.39
8,6	0.7,0.8	0.05	7.18	7.36	10.18	6.24	8.68	8.47	13.8	13.78	3.94	0.21	94.67%	2.51%	-0.02
8,6	0.7,0.8	0.1	5.78	5.97	5.72	2.47	4.74	3.9	11.79	11.91	3.25	0.84	74.15%	3.29%	0.12
												Average	78.12%	2.83%	0.04

We generated 200 random problems for each parameter combination, and we solved them using TOF and UOF. For each parameter combination, Table 1 reports the average performance obtained by TOF and UOF on the 200 problem instances generated under that parameter combination. When the schedule is generated by optimizing TOF, the waiting time incurred by G_1 is always longer than that incurred by G_2; the column "Disparity" reports their difference. In the schedule found by optimizing UOF, G_1's waiting time generally decreases, while G_2's increases so that the two groups roughly wait the same and the disparity decreases by 78.12% on average. The overtime is generally almost unaffected, while the cost increases by 2.83% on average.

Now that we have established that UOF can potentially reduce racial disparity across a large set of parameter combinations with a small impact on cost, we turn our attention to a real-world example.

6. Case Study on Real-World Data

In this section, we implement the predictive overbooking framework (Figure 1) using the data set from an existing outpatient clinic.

6.1. Predictive Model

The first step in the implementation of the predictive overbooking framework is to build a predictive model to estimate the show probabilities of a set of input appointment requests.

The data set considered in this study comes from a large specialty clinic in the East Coast. It contains approximately 40,000 appointments made over three years by 13,000 patients, most of whom identify themselves as "White" or "Black". The data set has one entry for each appointment, and includes information on the appointment as well as on the patient. The variables are listed below.

The dependent variable of the predictive model is a binary indicator of show. The population show rate is 73.4%, but there are large differences depending on the race. Table 2 reports some summary statistics by race.

Table 2: Summary statistics by race

	Rel. Frequency	Show rate
White	55.8%	78.1%
Black	39.4%	66.1%
Asian	1.3%	82.7%
Other	3.6%	75.5%

Analyzing further, there is also a difference in show probability when breaking down the show rates based upon other socio-economic factors such as employment status and marital status. The show rate is

66.2% among unemployed patients and 77.1% among patients with a full-time job; but the unemployment rate is higher among black patients (49.2%) than among white patients (30.4%). Similarly, the show rate is 78.6% among married patients and 68.7% among single patients; while only 28.4% of the black patients are married, 56.9% of the white patients are. Thus, if scheduling decisions are made based upon show probabilities that are calculated from socio-economic variables, a group of patients may still experience biased scheduling. Because the vast majority of patients are either black or white, we consider the two racial groups "black" and "non-black".

The features describing each appointment are the following:

1. **Appointment-level features**: 1.1 The appointment time, 1.2 the lead time to the appointment (the time elapsed from the moment when the appointment is requested to the moment when the appointment takes place), 1.3 the day of the week, 1.4 the ID of the specific building of the appointment;
2. **Patient-level features**: 2.1 The patient's marital status, 2.2 the patient's employment status, 2.3 the patient's employer, 2.4 the patient's city name, zip code, and county name, 2.5 the patient's preferred language, 2.6 the distance between the patient's home and the clinic, 2.7 the patient's number of past no-shows, 2.8 the patient's past no-show rate, 2.9 the patient's age, 2.10 the patient's number of past appointments, 2.11 the patient's past average lateness, 2.12 the patient's "time in the system" (computed as the time elapsed from the moment when the patient was registered to the appointment date), 2.13 the diagnosis code, 2.14 the patient's insurance type.

Note that race has been excluded from the set of features. To build a predictive model, we proceed as follows. First, we randomly partition the data into a training set (80% of the data) and a test set (20% of the data). The training set is used to derive a predictive model, while the test set is used to evaluate its predictive performance and assess any racial disparity.

We first execute a 10-fold cross validation on the training data set using the following classification techniques (all with the default parameters provided by the machine learning package scikit-learn): Random Forests, Gaussian Naïve Bayesian Networks, Logistic Regression, AdaBoost, Multilayer Perceptron. For the non-probabilistic classifiers, we derived the no-show probabilities using Platt's method (Platt, 1999), which consists of building a logistic regression model that predicts the binary no-show outcome given the no-show score.

At each iteration of the cross-validation procedure, we recorded the Area Under the receiver operating Curve (AUC) and Brier's score (Brier, 1950), two common metrics to evaluate the prediction quality of predicted probabilities. The former metric measures how well the classification technique ranks the appointments from the most likely to the least likely to no-show; the latter metric computes the mean squared difference between the predicted probabilities and the real binary outcome; the smaller the Brier's

score, the higher the quality of the probabilities. We select the classifier with the smallest Brier's score because Samorani and Harris (2019) show that the Brier score is a better indicator of scheduling performance than the AUC. The cross-validated prediction performance is reported in Table 3.

Table 3: Cross-validated AUC and Brier's score on the training set

Classification Technique	All Features		All features except socio-economic indicators	
	AUC	Brier's score	AUC	Brier's score
Random Forest	0.632	0.187	0.604	0.191
Gaussian Naïve Bayes	0.644	0.188	0.616	0.190
Logistic Regression	**0.668**	**0.182**	**0.636**	**0.187**
AdaBoost	0.679	0.194	0.655	0.195
Multi-layer perceptron	0.620	0.192	0.588	0.195

In addition to building predictive models using all features listed above (first two columns), we also build models that do not use any socio-economic indicator (features 2.1 to 2.6). We do this to obtain show probabilities uncorrelated to race, thus limiting racial disparity in the schedule.

For both sets of features, the best-performing technique is Logistic Regression. So, we build a Logistic Regression model using the entire training set, and then use it to predict the show probability of the appointments in the test set. The AUC and Brier's score obtained on the test set where 0.631 and 0.186 using the reduced set of features and 0.663 and 0.182 using the complete set of features. Those values are similar to those obtained on the training set, which suggests that there is no overfitting.

6.2. Scheduling Results

We now employ the show predictions obtained above to optimally schedule appointments. To this end, we simulate a large number of scheduling problems in which N appointment requests need to be scheduled in F slots. Thus, we consider the following combinations: $(N, F) \in \{(6,4), (7,5), (8,6)\}$, which are reasonable choices given the show rate of 73.4%. For each (N, F) combination, we generate 5,000 scheduling problems by randomly sampling with replacement N appointment requests from the test set. We solve each problem using strategies that differ in the objective function (TOF or UOF) and in how the show probabilities are derived (All features, All features except socio-economic indicators, No feature). Under the "No feature" strategy, no prediction is made and all appointment requests have the same show probability, 73.30%, which is training set show rate.

For each schedule obtained, we compute cost, overtime, idle time, and waiting times by considering the appointments' actual show outcome from the data. All schedules are evaluated computing the clinic cost through (1), i.e. TOF, because that metric reflects the actual scheduling cost incurred by the clinic. Throughout our experiments, we assume again an idle time cost rate of $\rho = 2$ and an overtime cost of rate of $\tau = 8$. Finally, to better interpret the results, we assume that each appointment slot is 30 minutes.

Table 4: Results on three sets of problems with different number of appointment requests $\boldsymbol{N}$ and appointment slots $\boldsymbol{F}$. All times are in minutes. In **bold**: statistically significant racial disparity (pvalue < 0.01). Underlined: statistically higher cost (pvalue < 0.01) than the cost obtained by the state-of-the-art method.

		TOF with a predictive model based on			UOF w/ all features
		No feature (i.e., without a predictive model)	All features except socio-economic indicators	All features (state-of-the-art method)	
	Overtime	22.07	21.61	21.79	21.77
	Idle time	9.11	8.64	8.82	8.81
$N = 6$	Black Patients' wait	19.09	18.56	18.03	16.63
$F = 4$	Non-black patients' wait	18.70	17.40	15.88	16.75
	Schedule Cost	9.28	8.97	8.86	8.86
	Racial Disparity	2.09%	**6.67%**	**13.54%**	0.72%
	Overtime	17.62	17.56	17.81	17.81
	Idle time	12.55	12.49	12.74	12.74
$N = 7$	Black Patients' wait	19.50	17.72	16.75	15.43
$F = 5$	Non-black patients' wait	19.20	16.82	15.14	15.81
	Schedule Cost	9.07	8.72	8.58	8.58
	Racial Disparity	1.56%	**5.35%**	**10.63%**	2.46%
	Overtime	14.12	14.36	14.50	14.50
	Idle time	16.67	16.91	17.05	17.05
$N = 8$	Black Patients' wait	19.20	16.71	15.57	14.52
$F = 6$	Non-black patients' wait	19.04	16.07	14.11	14.85
	Schedule Cost	8.93	8.54	8.35	8.36
	Racial Disparity	0.84%	**3.98%**	**10.35%**	2.27%
	Average Racial Disparity	1.50%	5.33%	11.51%	1.82%

For each combination of number of patients N and number of slots F, Table 4 reports the average value of the provider's idle time and overtime, the waiting time experienced by the black and non-black patients

that show up, the schedule cost, and the racial disparity, computed as the absolute percent difference between the waiting times experienced by the two racial groups.

Limitedly to TOF, our results show that using all features to predict show probabilities (which is the "state-of-the-art" method) results in the lowest cost but also in the largest racial disparity (11.51% average difference between black and non-black patients' waiting times). Unsurprisingly, not using a predictive model results in the largest cost and in an insignificant racial disparity, as each patient is assumed to have the same show probability.

Interestingly, eliminating all socio-economic indicators still results in significant racial disparity. The reason lies in the presence of features other than socio-economic indicators, like the patient's prior no-shows, which are also correlated with race.

The last column reports the results obtained by employing all features to predict show probabilities and UOF to schedule appointments. This method obtains schedules whose cost is not different from that obtained by the state-of-the-art method and without any significant racial disparity. Compared to the state-of-the-art method, UOF removes racial disparity by decreasing the black patients' waiting time and increasing the non-black patients' waiting time, so that the two quantities become similar. Appendix C reports the results after relaxing the assumption of constant service times.

7. Conclusion

This paper extends the body of work on predictive overbooking, which aims at scheduling appointments based on individual patients' show probabilities, in order to minimize TOF. Because the probability of show tends to depend on the patients' racial groups, and because of the structural properties of TOF that we uncovered, the traditional predictive overbooking framework disproportionately overbooks the racial group of patients with the lower show probability, who consequently experience significantly longer waiting times: in our simulations, black patients' waiting times are 11.51% longer than non-black patients'. In turn, negative experiences at a clinic might decrease the level of engagement of patients that are already at high risk of no-show, which may result in more no-shows in the future, and more racial disparity. Our results suggest that this disparity is not eliminated by removing socio-economic indicators from the data.

To reduce the disparity, we develop a different objective function, UOF, which instead of minimizing everyone's waiting time (as TOF does), minimizes the waiting time of the group expected to wait longest. This strategy eliminates racial disparity while obtaining a similar clinic cost to that obtained by the traditional method.

Opportunities for future research include developing methodologies for more than two race groups, investigating the performance of our method with different racial ratios, allowing the number of patients to be a decision variable instead of an exogenous parameter, implementing a sequential scheduling method

which schedules appointments as they come in. Also, while our method aims at minimizing the disparity of every single clinic session, perhaps it is sufficient to minimize the average disparity over a longer horizon, thereby letting some individual sessions to be affected by disparity.

Because this is the first work on racial disparity in appointment scheduling, there are also higher-level questions that create opportunities for future work. Does racial disparity manifest itself in ways other than longer waiting times (e.g., longer wait from appointment request to appointment day)? And, do these longer wait times adversely impact patient outcomes (e.g. patients leave rather than seek care)? Does racial disparity affect other aspects of health care access (e.g., the emergency room)? Should there be "racial fairness" or bias considerations that health care providers abide by when scheduling appointments or that machine learning developers adhere to when patient scheduling algorithms are created?

References

Ahmadi-Javid, A., Jalali, Z. and Klassen, K.J., 2017. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, *258*(1), pp.3-34.

Brier, G.W., 1950. "Verification of forecasts expressed in terms of probability". *Monthey Weather Review*, *78*(1), pp.1-3.

Benjamin, R., 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons. Page 12.

Centers for Disease Control and prevention (CDC). 2013, November 22. Health disparities &inequalities report—United States, 213. MMWR.62(suppl 3): 1-187.

Dai, T. and Tayur, S.R., 2019. Healthcare Operations Management: A Snapshot of Emerging Research. *Manufacturing & Service Operations Management, Forthcoming*.

Dantas, L.F., Fleck, J.L., Oliveira, F.L.C. and Hamacher, S., 2018. No-shows in appointment scheduling–a systematic literature review. *Health Policy*, *122*(4), pp.412-421.

Gianfrancesco, M.A., Tamang, S., Yazdany, J. and Schmajuk, G., 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, *178*(11), pp.1544-1547.

Li, Y., Tang, S.Y., Johnson, J. and Lubarsky, D.A., 2019. Individualized No-show Predictions: Effect on Clinic Overbooking and Appointment Reminders. *Production and Operations Management*.

Platt, J., 1999. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". *Advances in large margin classifiers*, *10*(3), pp.61-74.

Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. and Chin, M.H., 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*.

Samorani, M. and Harris, S., 2019. The Impact of Probabilistic Classifiers on Appointment Scheduling with No-shows. Working paper.

Samorani, M. and LaGanga, L.R., 2015. Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research*, *240*(1), pp.245-257.

Srinivas, S. and Ravindran, A.R., 2018. Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework. *Expert Systems with Applications*, *102*, pp.245-261.

Williams, D.R., Mohammed, S.A., Leavell, J. and Collins, C., 2010. Race, socioeconomic status and health: Complexities, ongoing challenges and research opportunities. *Annals of the New York Academy of Sciences*, *1186*, p.69.

Zacharias, C. and Pinedo, M., 2014. Appointment scheduling with no-shows and overbooking. *Production and Operations Management*, *23*(5), pp.788-801.

Appendix A – Proofs

PROPOSITION 1: If i and j are two patients scheduled in an overbooked slot (vertical segment) and i has a lower show probability than j, then i has a longer CWT than j

Proof: Let i and j be the indices of two patients with show probabilities p_i and p_j scheduled in slot t, respectively, and let $p_j > p_i$. Given that the show outcomes of the other patients are stochastic, let us consider one generic realization (b, n), where b is the actual backlog at the beginning of slot t and n is the actual number of shows among all other patients scheduled in t. We will show that i has a longer CWT than j for any realization (b, n). The CWT of patient i is equal to b plus the number of patients scheduled in the vertical segment that are expected to be seen before patient i. If j shows up, then there are $n + 1$ patients different from i who show up in the slot. Because patients scheduled in the same slot are seen in random order, each of these $n + 1$ patients has a probability of 50% to be seen before patient i. So, the expected number of patients seen before patient i is $\frac{n+1}{2}$. If j does not show up, then there are n patients different from i who show up in the slot. In this case, the expected number of patients seen before patient i is $\frac{n}{2}$.

$$cwt_i = b + p_j \frac{n+1}{2} + \frac{(1-p_j)n}{2} = b + \frac{p_j}{2} + \frac{n}{2}$$

Analogously,

$$cwt_j = b + p_i \frac{n+1}{2} + \frac{(1-p_i)n}{2} = b + \frac{p_i}{2} + \frac{n}{2}$$

Thus,

$$cwt_j < cwt_i \blacksquare$$

PROPOSITION 2: If i and j are two patients scheduled, respectively, in slots t and u ($u > t$) of a horizontal segment, then the i's CWT is greater than or equal to j's CWT

Proof: Consider a horizontal segment (i.e., no patients are overbooked). Let patient i be scheduled in slot t and patient j be scheduled in slot u, with $u > t$. Let b be the backlog at the beginning of slot t. Then, the CWT of patient i is equal to b. In contrast, the CWT of patient j is equal to b only if patient i, as well as all of the patients scheduled in slots $t + 1, \dots, u - 1$ show up; otherwise, it is less than b. So, the CWT of patient j is less than or equal to b ■

LEMMA 1: A showing patient i scheduled in an overbooked slot expects to wait longer than any patient scheduled in the slot right after if $\frac{s^+}{2} + p_i \leq 1$, where p_i is patient i's show probability and s^+ is the number

of expected shows (conditional to observing at least one show) among all patients in the overbooked slot except patient i.

Proof: Let slot t have n patients plus patient i. Let P_0 be the probability that none among the n patients scheduled in t shows up. Let s be the expected number of shows among those n patients and s^+ be the expected number of shows among those n patients, conditional to at least one of them showing up. Note that $s^+ \geq 1$ by definition. Let b be the actual backlog at the beginning of slot t and assume that patient j is scheduled in slot $t+1$. We want to find sufficient conditions for which patient i's CWT, cwt_i, is longer than patient j's CWT, cwt_j.

Case 1. Assume $b = 0$. The expected CWT of patient i, cwt_i, is equal to 0 if no other patient shows up in his/her slot; otherwise, it is equal to half of the expected shows among the n patients[2]:

$$cwt_i = (1 - P_0)\frac{s^+}{2}$$

If none among the n patients show up, then patient j's CWT is $cwt_j = 0$. Otherwise, s/he will wait the number of shows exceeding one, because there will be one patient serviced in slot t. If i shows, then s^+ patients will be in backlog, if i does not show, then there will be $s^+ - 1$:

$$cwt_j = (1 - P_0)\big(p_i s^+ + (1 - p_i)(s^+ - 1)\big)$$

Next we subtract cwt_j from cwt_i to determine when cwt_i is at greater than or equal to cwt_j.

$$\big(cwt_i - cwt_j\big) = (1 - P_0)\left(\frac{s^+}{2} - p_i s^+ - (1 - p_i)(s^+ - 1)\right) \geq 0$$

$$= 1 - p_i + \frac{s^+}{2} \geq 0$$

So, $W_i \geq W_j$ if and only if $\frac{s^+}{2} + p_i \leq 1$

Case 2. Assume $b \geq 1$. The expected conditional waiting time of patient i, cwt_i, is equal to b if no other patient shows up in his/her slot; otherwise, it is equal to b plus half of the expected shows among the n patients:

$$cwt_i = b + (1 - P_0)\frac{s^+}{2}$$

[2] In general, if m patients with show probabilities $p_1, p_2, \ldots, p_m$ are scheduled in the same slot, the expected number of patients seen before patient i is $\frac{\sum_{j \neq i} p_j}{2}$. The reason is that, because patients scheduled in the same slot are seen in a random order, each patient in that slot has a 50% chance of being seen before i.

If none among the n patients show up, then patient j's waiting time is $b-1$ if i doesn't show up and b if i shows up. Otherwise, s/he will wait b plus the number of shows exceeding one:

$$cwt_j = P_0(1-p_i)(b-1) + P_0 p_i b + (1-P_0)\big(b + p_i s^+ + (1-p_i)(s^+-1)\big)$$

$$cwt_j = P_0(1-p_i)(b-1) + P_0 p_i b + (1-P_0)(b + s^+ - 1 + p_i)$$

Next we subtract cwt_j from cwt_i to determine when cwt_i is at least as great as cwt_j.

$$\big(cwt_i - cwt_j\big) = b + (1-P_0)\frac{s^+}{2} - b - p_i - (1-P_0)s^+ + 1 \geq 0$$

$$= 1 - (1-P_0)\left(\frac{s^+}{2}\right) - p_i \geq 0$$

So, $cwt_i - cwt_j \geq 0$ in case 2 if and only if $(1-P_0)\left(\frac{s^+}{2}\right) + p_i \leq 1$.

Conclusion: $cwt_i \geq cwt_j$ if the conditions relative to both cases are true: $\frac{s^+}{2} + p_i \leq 1$ and $(1-P_0)\left(\frac{s^+}{2}\right) + p_i \leq 1$. It can be easily seen that if the former inequality is satisfied, so is the latter. Thus, if $\frac{s^+}{2} + p_i \leq 1$, then $cwt_i \geq cwt_j$ in all cases. ■

PROPOSITION 3 (double booking): A showing patient i scheduled in slot t with one other patient, expects to wait longer than any patient scheduled in slot $t+1$, if s/he has a show probability $p_i < 0.5$.

Proof: Let p_{other} and p_i be the patients double booked in a single slot, and j the patient in slot $t+1$. From Lemma 1, patient i waits longer than patient j if $\frac{s^+}{2} + p_i \leq 1$. In this case, s^+ is equal to one, because that is the expected number of shows among a group composed of one patient (patient "other") conditional to at least one patient showing. So, patient i waits longer than patient j if $p_i \leq 0.5$. ■

PROPOSITION 4 (triple booking): A showing patient i scheduled in slot t together with two other patients whose show probabilities are p_1 and p_2 expects to wait longer than any patient scheduled in slot $t+1$ if s/he has a show probability $p_i \leq \frac{p_1+p_2-2p_1p_2}{2p_1+2p_2-2p_1p_2}$.

Proof: Let p_1, p_2, and p_i be the patients triple booked in a single slot, and j the patient in the following slot. From Lemma 1, patient i waits longer than patient j if $\frac{s^+}{2} + p_i \leq 1$. Note that $(1-P_0)s^+ = s$ (shown at the end of this proof for any number of patients n), where P_0 is the probability that, without considering patient i, no patient shows up in the overbooked slot. So, substituting for s^+, $\frac{s^+}{2} + p_i \leq 1$ is equivalent to

$$\frac{s}{2(1-P_0)} + p_i - 1 \leq 0$$
$$s + 2(1-P_0)p_i - 2(1-P_0) \leq 0$$
$$s + 2p_i - 2p_iP_0 - 2 + 2P_0 \leq 0$$
$$s + 2p_i - 2 + P_0(2 - 2p_i) \leq 0$$

In addition to patient i, the overbooked slot has two patients with show probabilities p_1 and p_2. The probability of neither of them showing up is $P_0 = (1-p_1)(1-p_2)$. Substituting for P_0 in the inequality above:

$$s + 2p_i - 2 + (1-p_1)(1-p_2)(2-2p_i) \leq 0$$
$$s + 2p_i - 2 + (1-p_1-p_2+p_1p_2)(2-2p_i) \leq 0$$
$$s + 2p_i - 2 + 2 - 2p_1 - 2p_2 + 2p_1p_2 - 2p_i + 2p_1p_i + 2p_ip_2 - 2p_1p_2p_i \leq 0$$

Because s is the expected number of shows among patients 1 and 2, $s = p_1 + p_2$. Substituting again:

$$p_1 + p_2 - 2p_1 - 2p_2 + 2p_1p_2 + 2p_1p_i + 2p_ip_2 - 2p_1p_2p_i \leq 0$$
$$-p_1 - p_2 + 2p_1p_2 + 2p_1p_i + 2p_ip_2 - 2p_1p_2p_i \leq 0$$
$$2p_i(p_1 + p_2 - p_1p_2) \leq p_1 + p_2 - 2p_1p_2$$
$$p_i \leq \frac{p_1 + p_2 - 2p_1p_2}{2p_1 + 2p_2 - 2p_1p_2} \blacksquare$$

Here, we show that $(1-P_0)s^+ = s$. Let P_v the probability of v shows among the n patients scheduled in a slot. Then, we can write s^+ and P_0 as follows:

$$s^+ = \frac{\sum_{v=1}^{n} P_v \cdot v}{\sum_{v=1}^{n} P_v}$$
$$P_0 = 1 - \sum_{v=1}^{n} P_v$$

So, $(1-P_0)s^+ = \sum_{v=1}^{n} P_v \frac{\sum_{v=1}^{n} P_v \cdot v}{\sum_{v=1}^{n} P_v} = \sum_{v=1}^{n} P_v \cdot v = s$

PROPOSITION 5: If a binary classifier is used to predict the patients' show probabilities and TOF is used to schedule the appointments, then increasing its sensitivity or specificity will result in a larger proportion of G_1 customers to be overbooked.

Proof: We consider a fixed demand of n_1 and n_2 from two groups G_1 and G_2. Patients are divided into two classes N (no-shows) and S (shows). Patients in G_1 have a probability π_1 to belong to N, whereas patients in G_1 have a probability π_2 ($\pi_2 < \pi_1$) to belong to N. Among the n_1 patients in G_1, there are $n_1\pi_1$ expected

to be in N and $n_1(1-\pi_1)$ expected to be in S. Among the n_2 patients in G_2, there are $n_2\pi_2$ expected to be in N and $n_2(1-\pi_2)$ expected to be in S. Figure A1 depicts the patient population.

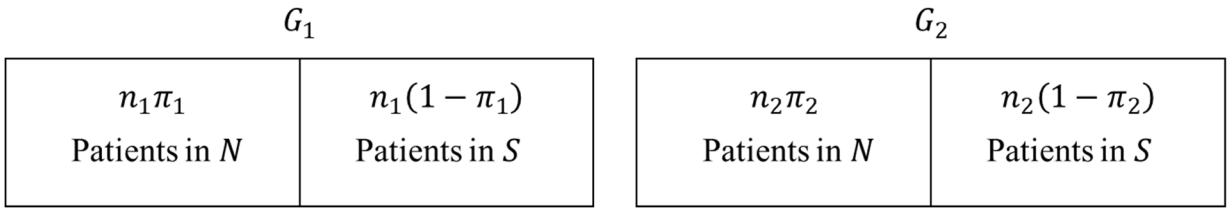

Figure A1: How the patient population is partitioned

Suppose that we have a binary classifier, whose task is to classify patients into classes N (the positive class, no-shows) and S (the negative class, shows). Because of Lemma 2 from Zacharias and Pinedo (2014), vertical segments will be filled with patients predicted in class N first; so patients predicted in class N are more likely to be overbooked. We now show that the among all patients predicted to be N, the proportion of G_1 patients increases if sensitivity or specificity increases.

The prediction performance of the classifier is represented by a sensitivity α and a specificity $1-\beta$. That is, the sensitivity, α, is the probability of correctly classifying a true no-show, whereas β is the probability of misclassifying a true show, thereby predicting him/her to be a no-show. For each group G_1 and G_2, the patients predicted to be in class N include 1) patients within that group that are correctly classified and 2) patients within that group in class S that are misclassified.

Therefore, the number of patients of G_1 predicted to belong to N is:

$$Pred(N|G_1) = n_1\pi_1\alpha + n_1(1-\pi_1)\beta$$

Similarly, the number of patients of G_2 predicted to belong to N is:

$$Pred(N|G_2) = n_2\pi_2\alpha + n_2(1-\pi_2)\beta$$

Out of all of the patients predicted to be in N, the proportion of patients belonging to G_1 is:

$$E = \frac{Pred(N|G_1)}{Pred(N|G_1) + Pred(N|G_2)}$$

$$E = \frac{n_1(\pi_1\alpha + (1-\pi_1)\beta)}{n_1(\pi_1\alpha + (1-\pi_1)\beta) + n_2(\pi_2\alpha + (1-\pi_2)\beta)}$$

We want to show that increasing the sensitivity, α, or the specificity, $1-\alpha$, results in an increase in the value of E. Let us take the derivatives $\frac{\delta E}{\delta \alpha}$ and $\frac{\delta E}{\delta \alpha}$.

$$\frac{\delta E}{\delta \alpha} = \frac{n_1 n_2 \beta(\pi_1 - \pi_2)}{\left(n_1(\pi_1\alpha + (1-\pi_1)\beta) + n_2(\pi_2\alpha + (1-\pi_2)\beta)\right)^2}$$

$$\frac{\delta E}{\delta \beta} = \frac{n_1 n_2 \alpha(\pi_2 - \pi_1)}{\left(n_1(\pi_1\alpha + (1-\pi_1)\beta) + n_2(\pi_2\alpha + (1-\pi_2)\beta)\right)^2}$$

The derivative $\frac{\delta E}{\delta \alpha}$ is positive if $n_1 n_2 \beta(\pi_1 - \pi_2) > 0$, which is true, because $\pi_2 < \pi_1$. So, E increases with the sensitivity. Similarly, the derivative $\frac{\delta E}{\delta \beta}$ is negative, which implies that increasing the specificity $1 - \beta$ will increase E. ■

PROPOSITION 6:

(i) Under reasonable conditions[3], there exists a schedule which minimizes UOF with no empty slots.

(ii) There exists a schedule which minimizes UOF in which, within each segment, the patients of the same racial group are sorted by increasing show probability.

To prove part *(i)*, we must first prove the following Lemma:

(beginning of Lemma)

Lemma: If an overbooked slot is followed by an empty slot, then moving a patient i to the next slot will not increase the objective function of UOF, as long as the expected shows among the patients in groups other than i's group are at most two.

Proof of Lemma: Let us assume that we have several groups of patients G_1, G_2, G_3,…, scheduled in one slot and that the next slot is empty. We now analyze the effect on the objective of moving patient i, who is assumed to belong to G_1 without loss of generality, to the adjacent empty slot. Figure A2 depicts the initial situation, S1, and the final situation, S2.

[3] If slots containing more than four expected shows are not allowed.

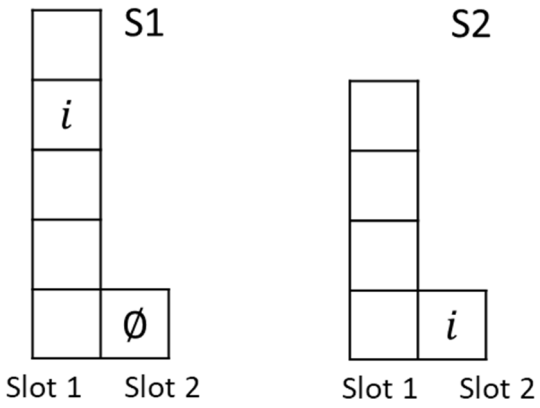

Figure A2: Depiction of situations S1 and S2

By moving i to slot 2, the expected number of patients overflowing to the following slots will not change, and so this move will not affect the expected waiting time of the following patients, the idle time incurred after slot 2, or the overtime. Note that this move is not going to affect the idle time incurred in slots 1 and 2 (moving p only changes the slot where idle time may be incurred). Also, with this move the waiting times of groups G_2, G_3, ... scheduled in the first slot may decrease but not increase. We will now show that this move also decreases the waiting time of i's group, G_1, under reasonable assumptions.

Let n_1 be the number of patients belonging to G_1 scheduled in the first slot of S2 (i.e., i is excluded); let n_2 be the number of patients belonging to all other groups scheduled in the first slot of S2. Let a be the expected number of patients that overflow to slot 1 from the preceding slot. Let b be the expected number of patients that overflow from the first slot to the second slot in S2; let $W2 + an_1$ be the sum of the expected waiting times of all patients in G_1 in the first slot of S2, and assume patient i shows with probability p_i. So, in S1, if i does not show up, then the expected sum of waiting times is equal to $W_2 + an_1$; if i shows up, the waiting time incurred by all showing patients in G_1 (whose expected number is n_1) increases by $\frac{1}{2}$ because i has 50% chance of being seen before them, and i also waits $a + \frac{n_2+n_1}{2}$ (as already seen in the proof of Proposition 1). So, before performing the move, the sum of G_1's waiting times is:

$$N_1^{before} = (1 - p_i)(W2 + an_1) + p_i\left((W2 + an_1) + \frac{n_1}{2} + a + \frac{n_1 + n_2}{2}\right)$$

After moving i, the waiting times incurred by G_1 are $W_2 + an_1$ plus the backlog suffered by p, if s/he shows up:

$$N_1^{after} = W2 + an_1 + pb$$

Moving i does not increase the waiting time of G_1 if and only if:

$$N_1^{after} \leq N_1^{before}$$

$$W2 + an_1 + pb \leq (1-p)(W2 + an_1) + p\left((W2 + an_1) + \frac{n_1}{2} + a + \frac{n_1 + n_2}{2}\right)$$

$$pb \leq -p(W2 + an_1) + p\left((W2 + an_1) + \frac{n_1}{2} + a + \frac{n_1 + n_2}{2}\right)$$

$$pb \leq p\left(\frac{n_1}{2} + a + \frac{n_1 + n_2}{2}\right)$$

$$pb \leq p\left(n_1 + \frac{n_2}{2} + a\right)$$

$$b \leq n_1 + \frac{n_2}{2} + a$$

Note that because the expected shows in the first slot of S2 are $n_1 + n_2$ and one of them is serviced in the first slot, the expected number of patients overflowing to the second slot, b, is less than or equal to $n_1 + n_2 - 1$:

$$b \leq n_1 + n_2 - 1$$

Thus, a sufficient condition for the move to not increase the objective is:

$$n_1 + n_2 - 1 \leq n_1 + \frac{n_2}{2} + a$$

$$n_2 \leq 2 + 2a$$

Thus, if $n_2 \leq 2$, then moving p to slot 2 will not increase the objective function value. (end of Lemma)

Let us prove (i). Suppose that the schedule has empty slots. Let t be the first empty slot. Since there are more customers N than slots F, at least another slot has at least two customers.

Case 1: Suppose that the schedule prior to t has a slot with more than one customer. Let t_0 be the last slot before t with more than one customer. This implies that slots $t_0, t_0 + 1, \dots, t - 1$ have at most one customer assigned to them. Consider the following move: find a group G in slot t_0 such that the expected number of shows belonging to the other group in that slot are at most two (if that slot contains only one group, then G is that group); then, take a customer i scheduled in slot t_0 belonging to group G and all customers assigned to slots $t_0, t_0 + 1, \dots, t - 1$ and reassign them to slots $t_0 + 1, \dots, t$ respectively, in the same order, one after the other (i.e., shift these customers one slot to the right in the schedule). The expected number of customers at the end of slot t remains the same as before. This implies that the waiting and idle time costs associated with slots $t + 1, \dots, F$ as well as the overtime cost remain the same as before. The waiting time of patients scheduled in slots t_0 and $t_0 + 1$ goes down because of the Lemma above. The waiting time of patients assigned to slots $t_0 + 2, \dots, t$ goes down, since every patient faces a lower expected backlog. Therefore, the altered schedule results in a not-higher total expected cost.

Case 2: Suppose that the schedule prior to t does not have a slot with more than one customer. So, there is no patient overflowing from slot $t-1$ to slot t. Consider the following altered schedule: move all customers assigned to slots $t+1, \dots, F$ to the left by one slot (i.e., to slots $t, \dots, F-1$). The expected cost will not increase because the waiting time of all patients stay the same but the overtime may decrease. After this move there is at least one empty slot (slot F) and may be more empty slots between slot t and slot F. Redefine t as the first empty slot of the new schedule. If the schedule prior to t does not have a slot with more than one customer, re-execute case 2, obtaining a new schedule of at least the same quality as the previous one; else, execute case 1 to obtain a new schedule with no empty slot of at least the same quality as the original one.

Let us prove (ii). *Proof*: First, consider two patients a and b, both belonging to the same group, and scheduled in slots t and u, respectively of the same horizontal segment (i.e., there is exactly one patient scheduled in each slot $t, t+1, \dots, u$). Without loss of generality, assume that there is no other patient between slots t and u that belong to the same group as a and b. Suppose that a and b's show probabilities are p_a and p_b, respectively, and that $p_a \geq p_b$. We will show that swapping a and b will decrease the expected cost. Note that the swapping move will not affect the waiting time of the patients scheduled after slot u or the idle time incurred after slot u; thus, it will also not affect the overtime. Also, the swapping move will decrease the waiting time experienced by all patients scheduled between a and b, because $p_b < p_a$, thereby decreasing the expected waiting times of other patient groups. So, all we need to show is that the sum of the waiting times experienced by a and b decreases. Let o_t be the number of expected patients overflowing to slot t, and o_u the number of expected patients overflowing to slot u assuming that a shows up, under the current schedule. Note that o_u depends on the show probabilities of the patients scheduled between slot t and u. The sum of the waiting times experienced by a and b before the move is:

$$W^{before} = p_a p_b (o_t + o_u) + p_a (1 - p_b) o_t + p_b (1 - p_a) \max(0, o_u - 1)$$

The waiting time after the move is:

$$W^{after} = p_a p_b (o_t + o_u) + p_b (1 - p_a) o_t + p_a (1 - p_b) \max(0, o_u - 1)$$

We now show that,

$$W^{after} \leq W^{before}$$

$$p_a p_b (o_t + o_u) + p_b (1 - p_a) o_t + p_a (1 - p_b) \max(0, o_u - 1)$$
$$\leq p_a p_b (o_t + o_u) + p_a (1 - p_b) o_t + p_b (1 - p_a) \max(0, o_u - 1)$$
$$p_b (1 - p_a) o_t + p_a (1 - p_b) \max(0, o_u - 1) \leq p_a (1 - p_b) o_t + p_b (1 - p_a) \max(0, o_u - 1)$$
$$p_b o_t - p_a p_b o_t + p_a \max(0, o_u - 1) - p_a p_b \max(0, o_u - 1)$$
$$\leq p_a o_t - p_a p_b o_t + p_b \max(0, o_u - 1) - p_a p_b \max(0, o_u - 1)$$

$$p_b o_t + p_a \max(0, o_u - 1) \leq p_a o_t + p_b \max(0, o_u - 1)$$
$$p_b(o_t - \max(0, o_u - 1)) \leq p_a(o_t - \max(0, o_u - 1))$$

True because $p_b \leq p_a$.

Second, consider two patients a and b, both belonging to the same group G and scheduled in the same segment as follows: a is scheduled in a vertical segment, that is, s/he is scheduled at slot t together with other patients (some belonging to group G, and others belonging to other groups); b is scheduled in the adjacent horizontal segment, that is, s/he is scheduled at slot u, and all slots $t+1, t+2, \dots, u$ have exactly one patient scheduled in them. Without loss of generality, no patient in slots $t+1, t+2, \dots, u-1$ belongs to group G. Suppose that a and b's show probabilities are p_a and p_b, respectively, and that $p_a \geq p_b$. We will show that swapping a and b will decrease the expected cost. Note that the swapping move will not affect the waiting time of the patients scheduled after slot u; thus, it will also not affect the overtime. Also, the swapping move will decrease the waiting time experienced by all patients scheduled between a and b, because $p_b < p_a$, thereby decreasing the expected waiting times of other patients (who may belong to any group). So, all we need to show is that the sum of the waiting times experienced by a and b decreases. Let o_t be the number of expected patients overflowing to slot t, n_t the number of expected shows in slot t excluding patient a, and o_u the number of expected patients overflowing to slot u assuming that a shows up, under the current schedule. Note that o_u depends on the show probabilities of the patients scheduled between slot t and u. The sum of the waiting times experienced by a and b before the move is:

$$W^{before} = p_a p_b \left(o_t + \frac{n_t}{2} + o_u\right) + p_a(1 - p_b)\left(o_t + \frac{n_t}{2}\right) + p_b(1 - p_a)\max(0, o_u - 1)$$

The waiting time after the move is:

$$W^{after} = p_a p_b \left(o_t + \frac{n_t}{2} + o_u\right) + p_b(1 - p_a)\left(o_t + \frac{n_t}{2}\right) + p_a(1 - p_b)\max(0, o_u - 1)$$

We now show that,

$$W^{after} \leq W^{before}$$

$$p_a p_b \left(o_t + \frac{n_t}{2} + o_u\right) + p_b(1 - p_a)\left(o_t + \frac{n_t}{2}\right) + p_a(1 - p_b)\max(0, o_u - 1)$$
$$\leq p_a p_b \left(o_t + \frac{n_t}{2} + o_u\right) + p_a(1 - p_b)\left(o_t + \frac{n_t}{2}\right) + p_b(1 - p_a)\max(0, o_u - 1)$$
$$p_b \left(o_t + \frac{n_t}{2}\right) + p_a \max(0, o_u - 1) \leq p_a \left(o_t + \frac{n_t}{2}\right) + p_b \max(0, o_u - 1)$$
$$p_b \left(o_t + \frac{n_t}{2} - \max(0, o_u - 1)\right) \leq p_a \left(o_t + \frac{n_t}{2} - \max(0, o_u - 1)\right)$$

True because $p_b \leq p_a$ ■

Appendix B – Alternative Measures of a Groups' Waiting Time

Measuring the expected waiting time of a group of patients is not trivial. Through an example, here we show that different measures lead to different conclusions as to which group is expected to wait longer. Consider the three-patient schedule illustrated in Figure B1, where two patients from G_1 are scheduled in slot 1 and have 20% probability of showing up, whereas a patient from G_2 is scheduled in slot 2 and has a 100% probability of showing up.

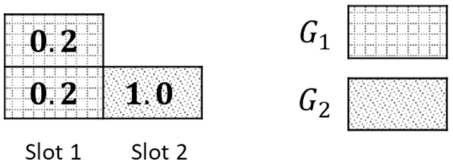

Figure B1: An example schedule

Table B1 lists all possible realizations. With a probability of 32%, exactly one patient from G_1 shows up (realization 1), in which case we observe one patient in G_1 and one patient in G_2 both experiencing a no waiting time. With a probability of 4%, both patients from G_1 show up (realization 2), in which case we observe one patient in G_1 waiting for one time units and two patient in G_2 waiting for 0 and 1 time units; in this case, the average waiting time within G_1 is 0.5 and the average waiting time within G_2 is one. Lastly, with a probability of 64%, no patients from G_1 shows up (realization 3), in which case we observe one patient in G_1 waiting for zero time units.

Table B1: Observed waits and group-level average waiting time for all realizations

Realiz.	Prob.	Observed waits		Expected Average Waiting time	
		Among G_1	Among G_2	Among G_1	Among G_2
1	.32	0	0	0	0
2	.04	1, 0	1	0.5	1
3	.64	NA	0	NA	0

Table B2 lists three possible ways to compute the waiting time of a group.

Table B2: Different methods for computing group-level waiting times

Method	G_1's waiting time	G_2's waiting time
Sum of waiting times (used in (1))	$.04 \times (1 + 0) = .04$	$.04 \times 1 = .04$
Expected Average Waiting time	$.04 \times .5 = .02$	$.04 \times 1 = .04$
Sum of waiting times divided by expected shows (used in (2))	$\frac{.04}{.2 + .2} = .1$	$\frac{.04}{1} = .04$

First, the "sum of waiting times" method consists of using the traditional formulation of waiting time used in (1) separately for each group, by computing the sum, weighted by the probability of realization, of all waiting times observed in a group. Using this formulation results in concluding that both groups expect to wait the same time: 0.04 time units. The limitation of this method is that it ignores the difference between the cardinalities of the groups, thereby underestimating the waiting time with the smallest expected cardinality (G_1 in our example).

The second method we analyze consists of computing the sum, weighted by the probability of realization, of the patients' expected waiting time (last two columns of Table B1). In 4% of the cases, we observe that the average waiting time of G_1 patients is 0.5, which results in a 0.1 expected average waiting time. This method leads to the conclusion that G_2 patients wait longer. The limitation of this method is that it assumes that not observing patients in a group is equivalent to observing that they wait 0.

The last method, which we developed in section 4.1, addresses this limitation by dividing the sum of the expected waiting times by the number of expected shows in that group. As explained in section 4.1, this method measures the expected wait suffered by a random patient in a group, conditional to showing up. A showing G_1 patient has a 20% chance of waiting 0.5, which results in a 0.1 expected wait; a showing G_2 patient has a 4% chance of waiting one time unit, which results in an expected wait of one time unit. This method leads to the conclusion that G_1 patients wait longer.

Appendix C – Stochastic Service Times

The computational study of Section 6 was performed under the assumption that service times take a constant amount of time equal to 30 minutes (i.e., the length of one appointment slot). Here, we relax that assumption, with the goals of (1) measuring the racial disparity that occurs when service times are stochastic, and (2) assessing the merit of our approach in this more realistic case.

In line with existing work, we assume that service times follow a lognormal distribution whose mean is equal to the length of the appointment slot. We consider the following parameter configurations:

- Configuration 1: Average service time = slot length = 20 minutes, coefficient of variation = 0.3
- Configuration 2: Average service time = slot length = 30 minutes, coefficient of variation = 0.2
- Configuration 3: Average service time = slot length = 60 minutes, coefficient of variation = 0.1

As in Samorani and Ganguly (2016), we assume that service times have a smaller coefficient of variation in clinics with longer service times, so that the probability of taking 10 minutes longer or 10 minutes shorter than expected is about 20% for 40-minute appointments.

Table C1: Results on Configuration 1 on three sets of problems with different number of appointment requests $\boldsymbol{N}$ and appointment slots $\boldsymbol{F}$. All times are in minutes. In **bold**: statistically significant racial disparity (pvalue < 0.01). Underlined: statistically higher cost than the cost obtained by UOF trained with all features (pvalue < 0.01).

		TOF with a predictive model based on			UOF w/ all features
		No feature (i.e., without a predictive model)	All features except socio-economic indicators	All features (state-of-the-art method)	
$N = 6$ $F = 4$	Overtime	17.28	17.06	17.35	17.42
	Idle time	8.63	8.41	8.70	8.78
	Black Patients' wait	13.84	13.58	13.39	12.27
	Non-black patients' wait	13.48	12.61	11.66	12.39
	Schedule Cost	10.79	10.53	10.53	10.58
	Racial Disparity	2.67%	**7.69%**	**14.84%**	0.98%
$N = 7$ $F = 5$	Overtime	14.80	15.05	15.44	15.44
	Idle time	11.42	11.67	12.06	12.06
	Black Patients' wait	14.40	13.27	12.87	11.59
	Non-black patients' wait	14.20	12.68	11.70	12.28
	Schedule Cost	10.77	10.53	10.52	10.54
	Racial Disparity	1.41%	**4.65%**	**10.00%**	5.95%

	Overtime	12.77	13.33	13.60	13.62
	Idle time	14.47	15.03	15.30	15.32
$N = 8$	Black Patients' wait	14.17	12.76	12.22	11.24
$F = 6$	Non-black patients' wait	14.09	12.43	11.28	11.87
	Schedule Cost	10.76	10.53	10.48	10.50
	Racial Disparity	0.57%	2.65%	**8.33%**	5.60%
	Average Racial Disparity	1.55%	5.00%	11.06%	4.18%

Table C2: Results on Configuration 2 on three sets of problems with different number of appointment requests $\boldsymbol{N}$ and appointment slots $\boldsymbol{F}$. All times are in minutes. In **bold**: statistically significant racial disparity (pvalue < 0.01). Underlined: statistically higher cost than the cost obtained by UOF trained with all features (pvalue < 0.01).

		TOF with a predictive model based on			UOF w/ all features
		No feature (i.e., without a predictive model)	All features except socio-economic indicators	All features (state-of-the-art method)	
	Overtime	24.59	24.19	24.56	24.63
	Idle time	11.62	11.23	11.59	11.66
$N = 6$	Black Patients' wait	20.20	19.77	19.41	17.82
$F = 4$	Non-black patients' wait	19.72	18.41	16.96	17.99
	Schedule Cost	10.27	9.99	9.95	9.99
	Racial Disparity	2.43%	**7.39%**	**14.45%**	0.95%
	Overtime	20.61	20.83	21.32	21.31
	Idle time	15.54	15.76	16.25	16.24
$N = 7$	Black Patients' wait	20.90	19.17	18.45	16.72
$F = 5$	Non-black patients' wait	20.62	18.30	16.75	17.56
	Schedule Cost	10.19	9.90	9.86	9.87
	Racial Disparity	1.36%	**4.75%**	**10.15%**	**5.02%**
	Overtime	17.36	18.03	18.32	18.34
	Idle time	19.91	20.58	20.87	20.89
$N = 8$	Black Patients' wait	20.56	18.32	17.40	16.06
$F = 6$	Non-black patients' wait	20.44	17.77	15.97	16.79
	Schedule Cost	10.12	9.84	9.75	9.76
	Racial Disparity	0.59%	3.10%	**8.95%**	**4.55%**
	Average Racial Disparity	1.46%	5.08%	11.18%	3.51%

Table C3: Results on Configuration 3 on three sets of problems with different number of appointment requests $\boldsymbol{N}$ and appointment slots $\boldsymbol{F}$. All times are in minutes. In **bold**: statistically significant racial disparity (pvalue < 0.01). Underlined: statistically higher cost than the cost obtained by UOF trained with all features (pvalue < 0.01).

		TOF with a predictive model based on			UOF w/ all features
		No feature (i.e., without a predictive model)	All features except socio-economic indicators	All features (state-of-the-art method)	
	Overtime	46.67	45.81	46.35	46.41
	Idle time	20.74	19.88	20.42	20.48
$N = 6$	Black Patients' wait	39.29	38.33	37.45	34.46
$F = 4$	Non-black patients' wait	38.43	35.82	32.84	34.75
	Schedule Cost	9.78	9.48	9.41	9.43
	Racial Disparity	2.24%	**7.01%**	**14.04%**	0.84%
	Overtime	38.26	38.42	39.16	39.14
	Idle time	28.12	28.28	29.02	29.00
$N = 7$	Black Patients' wait	40.42	36.91	35.22	32.16
$F = 5$	Non-black patients' wait	39.85	35.15	31.91	33.39
	Schedule Cost	9.63	9.31	9.22	9.23
	Racial Disparity	1.43%	**5.01%**	**10.37%**	3.82%
	Overtime	31.49	32.40	32.83	32.85
	Idle time	36.59	37.50	37.93	37.95
$N = 8$	Black Patients' wait	39.77	35.03	32.98	30.58
$F = 6$	Non-black patients' wait	39.50	33.84	30.09	31.65
	Schedule Cost	9.53	9.19	9.05	9.06
	Racial Disparity	0.68%	**3.52%**	**9.60%**	3.50%
	Average Racial Disparity	1.45%	5.18%	11.34%	2.72%

Appendix References

Samorani, M. and Ganguly, S., 2016. Optimal sequencing of unpunctual patients in high-service-level clinics. *Production and Operations Management*, *25*(2), pp.330-346.

Zacharias, C. and Pinedo, M., 2014. Appointment scheduling with no-shows and overbooking. *Production and Operations Management*, 23(5), pp.788-801.