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}

Problem definition: Overbooking is commonly employed by outpatient clinics to counteract no-shows. State-of-the-art appointment scheduling systems are composed of a machine learning component, which predicts the individual patients' no-show probability, and an optimization component, which uses these predictions to schedule appointments. The goal is to minimize the schedule cost, computed as a weighted sum of patients' waiting time and the provider's overtime and idle time. ***Academic/Practical Relevance:*** Despite its widespread use, we show that the objective of minimizing schedule cost may cause the patients at higher risk of no-show to experience longer waits at the clinic than the other patients. This may translate into undesirable racial disparities, as the patients' no-show risk is typically correlated with their race. ***Methodology:*** We analytically study racial disparity in this context. Then, we propose new objective functions that minimize both schedule cost and racial disparity, and that can be readily adopted by researchers and practitioners. We develop a "race-aware" objective, which instead of minimizing the waiting times of all patients, minimizes the waiting times of the racial group expected to wait the longest.

We also develop “race-unaware” methodologies that do not consider race explicitly. We validate our findings both on simulated and real-world data. ***Results:*** 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 state-of-the-art scheduling systems cause black patients to wait about 30% longer than non-black patients. Our race-aware methodology achieves both goals of eliminating racial disparity and obtaining a similar schedule cost as that obtained by the state-of-the-art scheduling method, whereas the race-unaware methodologies fail to obtain both efficiency and fairness. ***Managerial Implications:*** Our work uncovers that the traditional objective of minimizing schedule cost may lead to unintended racial disparities. Both efficiency and fairness can be achieved by adopting a race-aware objective.

Keywords: Appointment Scheduling, Machine Learning, Racial Bias

Version date: January 22, 2021

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).

In a typical scheduling environment, clinics schedule patients into appointment slots with the goal of minimizing a weighted sum of the patients’ waiting time and the provider’s idle time and overtime. 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 appointment slots. Although overbooking increases the expected number of showing patients and decreases idle time, it also introduces the undesirable effects of patient waiting time and provider overtime. Waiting time is incurred if a patient’s visit starts later than the scheduled time, whereas provider overtime is incurred if the provider needs to work beyond the nominal end of the clinic session in order to finish seeing all patients.

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.

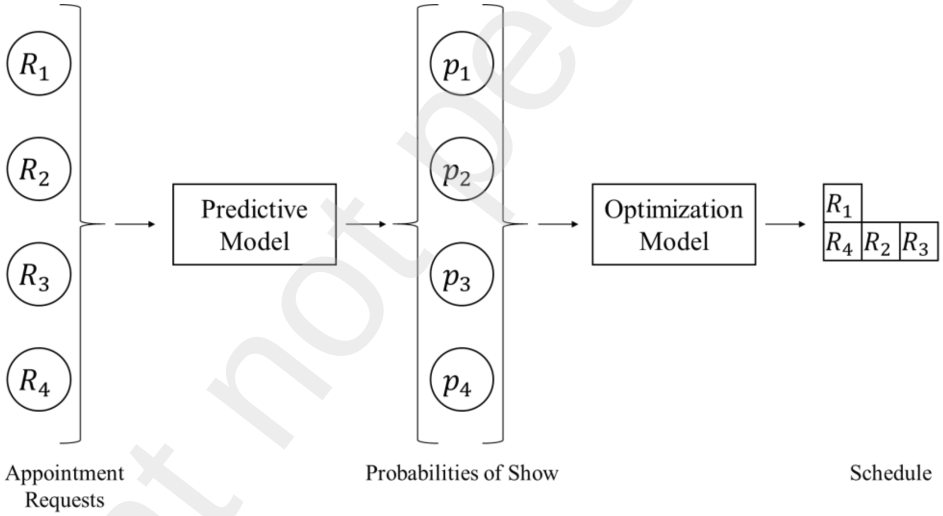


Figure 1: The predictive overbooking framework in a case with four appointment requests.

Zacharias and Pinedo (2014) noted that any schedule can be viewed as "a concatenation of alternating vertical and horizontal segments", where a vertical segment (OB) is an overbooked slot (graphically depicted with a vertical stack of slots) and a horizontal segment (Z, H_1, H_2, …, H_n) is a sequence of slots that are not overbooked (see Figure 2). We use a different notation for the first slot of the horizontal segment, Z, because that slot has different properties related to the waiting time compared to the rest of the slots in the horizontal segment, H. Throughout the paper, we use the term "segment" to denote a pair of vertical and horizontal segments. Some patients may be also scheduled before the first segment of the clinic session, in what we call Priority (P) slots, which by definition guarantee a zero-minute wait.

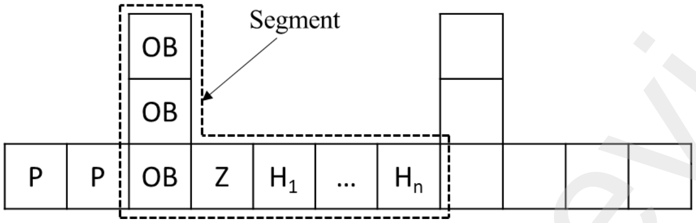


Figure 2: Appointment slot categories

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 all patients scheduled in OB have a lower (i.e., not higher) show probability than the patient scheduled in Z, the patient scheduled in Z has a lower show probability than the patient in H_1, and so on until the patient scheduled in H_n, who has the highest show probability among the patients in the segment. We will prove that if the optimal schedule includes P-slots, it is optimal to assign them to the patients with the highest show probability.

While that scheduling strategy is optimal from a cost-minimization point of view, it may have unintended ethical consequences. In section 3, we develop four Propositions that show that the longest wait in a segment is either experienced in OB or in Z appointment slots, which for this reason we deem to be "undesirable"; conversely, we deem H-slots and P-slots to be "desirable". If no-show behavior is correlated with a patient's race, then the patients in the group characterized with a higher risk of no-shows will be disproportionately scheduled is undesirable appointment slots.

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 that correlation, the predictive overbooking framework may result in significantly longer waiting times for a racial group. The disparity in waiting time between racial groups is especially unjust due to the evidence that people of color generally have inferior access to healthcare, receive poorer quality care, and experience worse healthcare outcomes (Centers for Disease Control, 2013). In this paper, we recognize the unintended consequences caused by the predictive overbooking framework. The interplay between predictive analytics and decision making is well-known to cause racial disparity in health care. Some of the latest examples, including this paper, have been recently cited in a United Nations report (Achiume, 2020), and are discussed in Section 2. As 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 seek to address those unintended ethical consequences, and provide solutions to move towards an equitable scheduling solution.

Motivated by the data set of an outpatient clinic whose black patients have a higher no-show rate than non-black patients, in this paper we assess the disparate impact that the predictive overbooking framework has on the waiting times of the black and non-black patients in the data set, and explore solution methods to reduce that racial disparity. While it is beyond the scope of this paper to offer an extensive discussion of the underlying ethical argument for correcting disparate racial impacts, our study is premised on the view that it is fundamentally unethical to punish black patients if they are predicted to have a lower show rate. Academic studies have shown that black patients may be less likely to be able to make it to appointments than white patients due to socioeconomic obstacles deeply rooted in historical racial discrimination (Williams et al. 2010). Therefore, any appointment schedule that results in black patients being given inferior scheduling slots because they "deserve" that slot, would in essence be penalizing those patients for the discrimination and socioeconomic conditions that they have historically suffered.

We first develop analytical insights to prove that the predictive overbooking framework tends to result in longer waiting times for the racial group at higher risk of no-show. After identifying the racial biases at work with the current predominant scheduling methodologies, we develop two types of solutions to eliminate scheduling bias. In our first solution, which we label "race aware", we consider race explicitly: instead of minimizing the waiting time of all patients, as the traditional objective does, our first solution minimizes the waiting time of the racial group that waits the longest. In addition to the race-aware solution, we also consider "race-unaware" alternative solutions that do not take race explicitly into account, but that still attempt to reduce racial bias. Among our solution methods, the race-aware solution is the only one capable of achieving both goals of reducing racial disparity and obtaining the same schedule quality as that obtained by the state-of-the-art scheduling method; in contrast, the race-unaware solutions fail to obtain both efficiency and fairness. 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. Literature Review and Current Regulations

2.1 Related Research

There are three topic areas which we will review for this paper: racial disparity in health care, empirical studies in patient no-show probabilities, and appointment scheduling with individual no-show probabilities. First, we discuss literature on the emerging subject of racial disparity in health care. In the United States, the ongoing Covid-19 pandemic has been far more deadly among Black and Latino communities than among White communities (Hooper et al., 2020). This is just the latest example of systemic racial disparities that affect the health of the most vulnerable communities. For decades, government and academic studies have highlighted healthcare inequality for black Americans that persists to this day. In January 1984, the "Heckler Report" (after Secretary of the US Department of Health and Human Services, Margaret Heckler) documented "significant progress" in terms of the health status of the American people as a whole, yet identified "continuing disparity in the burden of death and illness experienced by Blacks and other minority Americans as compared with our nation's population as a whole" (Heckler, 1985). Nelson (2002) concluded that ethnic and racial minorities in the United States were less likely to receive preventive health care than Whites and more likely to receive lower quality of care overall, even when accounting for socioeconomic factors such as income, neighborhood location, comorbidities, and health insurance (Hostetter, 2018). In 2020, the Office of Disease Prevention and Health Promotion cited the causes behind unequal health care access and outcomes for minorities as emanating from social determinants, i.e. "conditions in the environments in which people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks."

More recently, academic studies have focused on the ways that technology have exacerbated inequality and, in particular, unequal health care outcomes for black patients. Benjamin (2016) has characterized modern technologies as "one of the most effective conduits for reproducing racial inequality". She rejects the idea that we are living in a "post-racial" world and argues that unconscious and/or deliberate bias is built into technology such as artificial intelligence. 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".

Despite the unintended consequences resulting from the widespread adoption of algorithms, technology can also be a mean to reduce inequities. For example, Cui et al. (2020) found that guests with African American-sounding names were significantly less likely to be accepted by Airbnb hosts than guests with White–sounding names; however, they also found that this discrepancy disappears if the guest's past Airbnb reviews are positive. Rajkomar et al. (2018) provide the recommendation to seek to obtain "equal outcome", and not just equal prediction performance across the groups. In particular, they 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". Ganju et al. (2020) find information systems can be effectively used to reduce the disparity in amputation rate between black and nonblack patients suffering from diabetes mellitus.

Obermeyer et al. (2019) uncovered that predicting future health costs to decide which patients to provide better care to leads to racial bias, because sick black patients have a lower health care cost than healthier white patients. That occurs because black patients are more likely to access health care through the emergency department, whereas white patients are more likely to access it through elective procedures and surgeries, which are more expensive. Obermeyer et al. (2019) reduced this bias by using a race-unaware approach consisting of predicting future health rather than future costs. In other words, in that paper they found that selecting the right objective (maximizing health instead of minimizing cost) is critical to avoid disparities. In this paper, we develop both race-aware and race-unaware approaches to reduce racial bias in an outpatient appointment schedule.

The second area of research includes empirical studies in patient no-show probabilities. There have been several studies that analyze the correlation between no-show probabilities and patient characteristics, and many have found 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)". Huang and Hanauer (2014) found that African-Americans, who made up 5.3% of their dataset, were the least likely to show for general pediatric appointments. Miller et al. (2014) focused on patients who repeatedly no-showed, and found that younger, black, lower-income patients encountered the most barriers to care. The conclusion that patients of lower socioeconomic status were less likely to attend appointments was also found by Hamilton et al. (2002). Kaplan-Lewis and Percac-Lima (2013) found that black and Hispanic patients in an underserved population were more likely to no-show.

Several papers have also used patient interviews to identify the causes of no-shows. Kaplan-Lewis and Percac-Lima (2013) found that the two most common reasons for no-show were forgetting the appointment and miscommunication of the appointment time; no statistically significant differences were found between races in regard to the reason for no-show. Campbell et al. (2000), Martin et al. (2005), Neal et al. (2005), and Corfield et al. (2008) confirmed these results by finding that patients often attempted to cancel but were unable to reach the clinic. Lacy et al. (2004) identified that some patients fail to show up because they perceived the appointment to be uncomfortable, because the clinic did not respect them, or because they did not perceive their no-show as disruptive for the clinic.

The third area of research that we review 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 those 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, those studies fail to recognize that predictive overbooking may result in racially biased decisions. In this study, we extend the body of knowledge in the scheduling literature, and design a scheduling model that results in racially-equitable scheduling practices.

2.2 Legal and Ethical Considerations

Given that our race-aware solution method employs the patients' race, in this section we discuss the current regulations concerning the use of personal information in appointment scheduling. While the explicit use of race and other sensitive information has been regulated in contexts like lending, housing, and hiring (Hersch and Shinall, 2015; Massey, 2015), it is still largely unregulated in health care. In September 2019, the U.S Food and Drug Administration (FDA) released draft guidelines to regulate decision support systems in health care (https://www.fda.gov/media/109618/download, accessed on January 15, 2021). The guidelines show that the FDA intends to regulate software used for "critical" tasks like cancer diagnosis, but not software used for other "non-critical" tasks. Murray et al. (2020) argue that appointment scheduling could be classified as a non-critical medical task. Thus, the use of sensitive information to support scheduling decisions is currently allowed in practice. For instance, Murray et al. (2020) also point out that a popular appointment scheduling software allows the provider to explicitly use a patient's ethnicity, religion, and body mass index to predict their probability of show and inform scheduling decisions. That practice will inevitably penalize the most vulnerable populations. We note that the various solutions we propose (race-aware and race-unaware) are currently untested under the law as a solution to correcting racial bias in appointment scheduling.

In proposing alternative solutions to the problem of racial bias in medical appointment scheduling, we are conscious that we are treading into controversial waters. Some might argue that our race-aware solution is inappropriate because it takes race explicitly into account. For those who might fall into this camp, we have developed a number of alternative approaches that are race-unaware. While the authors believe that using race-aware policies in medical appointment scheduling is morally warranted, we are appreciative that others may hold a contrary view. Our results show that the race-aware solution is superior to the other race-unaware solutions that we developed not only in reducing racial disparity but also in obtaining high-quality schedules.

3. Clinic Model, Assumptions, and Properties

3.1. Clinic Model

We consider an outpatient clinic where one provider sees patients sequentially during the day. The input consists of a set of N appointment requests with individual show probabilities, p_i $(i = 1, \dots, N)$. We typically refer to the show probabilities, but occasionally refer to the no-show probabilities, $(1 - p_i)$. The scheduling problem consists of assigning each appointment request to one of F appointment slots. If $N > F$, there is overbooking, that is, at least one slot will be assigned more than one appointment request. 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-show (lower average show probability). All of our proofs are valid for any number of racial groups, but the current work focuses on the two-group case.

Throughout the paper, we consider a static scheduling problem where all appointment requests and patient information is known in advance. Existing work provides techniques to readily adapt static scheduling methodologies to the more realistic case of sequential scheduling, where a clinic schedules patients one at a time as the requests come in without knowledge of future patient information. An example of such a technique is proposed by Samorani and Harris (2019), who employ static schedules to guide sequential scheduling decisions in the case where patients have individual characteristics. Additionally, although some clinics operate with scheduling templates where the shape of the schedule is fixed (e.g., so that which slots are overbooked are fixed), in this paper the resulting schedule shape may change every day, depending on the specific set of individual show probabilities.

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 Section 6). 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 case of ties (e.g., if two patients who are 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 seen in the order they check in. 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).

3.2. 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 c_W for every time unit of delay; idle time cost is incurred at a rate of c_I for every time unit of idle time; and overtime cost is incurred whenever the provider finishes seeing the patients after the nominal end of the clinic session (i.e., after F time units from the start), at a rate of c_O 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. The TOF is as follows:

$$TOF = min\, E[cost] = min\left(\sum_{i=1}^{N}(c_W \cdot E[WT_i]) + c_I \cdot E[IT] + c_O \cdot E[OT]\right), \tag{1}$$

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

3.3. Analytical Connection between TOF and Racial Disparity

We now articulate why TOF is likely to lead to one racial group (G_1) experiencing waiting times longer than another (G_2). 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.

The goal of this section is to identify properties that suggest that, because of their lower show probabilities, patients in G_1 may be more likely to wait longer than patients in G_2. Our analysis shows that G_2 patients are more likely to be booked in P-slots (Proposition 1), which are commonly present in optimal schedules and guarantee a zero-minute wait (Proposition 2). Conversely, G_1 patients are more likely to be scheduled in OB-slots. We show that the longest wait within a segment is either in OB or Z (Propositions 3 and 4). Thus, throughout the paper we label slots OB and Z as "undesirable". In contrast, we label H- and P-slots "desirable", because they do not result in the longest wait (in fact, P-slots guarantee a zero-minute wait). Although our findings suggest that racial disparity is likely, they do not necessarily imply that it occurs in all cases. Determining whether disparity ultimately occurs requires fixing the number of patients in G_1 and G_2, as well as the number of appointment slots, F. It also requires strong assumptions on the distribution of the patients' show probabilities. That type of analysis does not generalize well, as it can only detect properties specific to a combination of parameters. We leave to future research the task of finding properties valid for realistic combinations of the abovementioned parameters.

To develop our analytical findings, we assume that the patients' individual show probabilities are sampled from a random distribution, such that a patient with a show probability p belongs to G_2 with probability $P(G_2|p)$, and to G_1 with probability $1 - P(G_2|p)$. Without choosing a specific distribution for the individual show probabilities, we formalize the fact that G_2 patients have a higher show probability than patients in G_1 by assuming that $P(G_2|p)$ is monotonically increasing with p. Due to this assumption, proving that patients with a lower show probability also have a longer CWT than patients with a higher show probability implies that patients in G_1, on average, have a longer CWT than patients in G_2 (see Online Supplement for the proof). Let us start by studying properties of the P-slots. We prove that for a fixed schedule shape, the patients with the highest show probability are scheduled in P-slots, if these slots exist. All proofs are in the online Supplement.

PROPOSITION 1: If the optimal schedule includes P-slots, then those slots are assigned to the patients with the highest show probability.

Proposition 1 is important in the context of disparity: since it is more likely for G_2 patients to be among those with the highest show probability, it is also more likely for G_2 patients to be scheduled in P-slots. We now show that P-slots are more commonly present in optimal schedules than previously thought. While Zacharias and Pinedo (2014)'s methodology reserved the P-slots only for those patients with a show probability equal to 1, here we find that the minimum show probability that makes it optimal for a patient to be scheduled in a P-slot may be significantly lower than that (e.g., 0.833 as shown after Corollary 1). To analytically find conditions that make it optimal for a patient to be scheduled in P, we develop Proposition 2 in the simplified case where all overbooking occurs in one slot, similar to the proofs in Zacharias and Pinedo (2014). All other Propositions do not make this assumption.

PROPOSITION 2: Let S be a schedule composed of N patients in F slots where all overbooking occurs in slot V (with V < F); that is, S is composed of a single segment (slots V to F), possibly preceded by a sequence of P-slots (slots 1 to V-1). Let P_0 be the probability of a zero backlog at the beginning of slot F, P_1 the probability of a backlog of one time unit at the beginning of slot F, and $E[B|B \geq 2]$ the expected backlog at the beginning of slot F, conditional to it being at least two time units. If the patient scheduled in slot F has a show probability greater than $\frac{(c_I+c_O)(1-P_0)}{(c_I+c_O)(1-P_0)+c_W\left(P_1+E[B|B\geq 2]\,(1-P_0-P_1)\right)}$, then there exists a better-quality schedule with an additional P-slot allocated to that patient.

Proposition 2 implies that if the show probability of the patient with the largest show probability is above a certain threshold, then that patient should not be scheduled in the last slot of the day; s/he should be scheduled in P instead. While it is easy to see that the threshold is less than 1, it is hard to interpret it or to estimate its value without knowing the show probability of all patients. To facilitate interpretation, consider the case where the schedule has only one double-booked slot:

COROLLARY 1 (one double-booked slot): Let S be a schedule composed of N *patients in* $F = N - 1$ *slots. That is, one slot is assigned to two patients and all other slots to only one. If S is optimal, then no patient with a show probability greater than* $\frac{c_I + c_O}{c_I + c_O + c_W}$ *is scheduled after the overbooked slot.*

As expected, the threshold to be scheduled in a P-slot is higher if the overtime and idle time cost coefficients (c_O and c_I) are large in relation to the waiting time cost coefficient (c_W). This confirms LaGanga and Lawrence (2012)'s finding that the earliest slots are more heavily overbooked if the overtime cost is the greater cost. Proposition 2 and Corollary 1 are important in the context of disparity because they suggest that even patients with show probabilities significantly smaller than 1 are scheduled in a P-slot. To see this intuition, consider the main parameters used by Zacharias and Pinedo (2014), c_I = 1, c_W = 0.5, and c_O = 1.5, which are also the main parameters used in Section 6.2 of this paper. If a schedule modeled with these parameters has only one double-booked slot, it is optimal to move all patients with a show probability greater than 0.833 to P-slots. In the data set analyzed in Section 6.1, 24.4% of the appointment requests fall into this category, and would therefore be scheduled in P-slots. Of the patients with a show probability greater than 0.833, only 17.7% are black, even though black patients make up 39.7% of the population. This result confirms that G_2 are more frequently scheduled in P-slots than G_1 patients.

After the P-slots are filled, Zacharias and Pinedo (2014) proved that the patients with the next highest show probability are scheduled in H-slots (from right to left), then in the Z-slot; finally, the patients with the lowest show probability are scheduled in the OB-slots. Thus, the patients with the highest show probability have priority access to P-slots and the right-most H-slots. Next, we show that the right-most H-slots result in a shorter CWT than the other H-slots and Z-slot of the same horizontal segment.

PROPOSITION 3: If i and j are two patients scheduled, respectively, in slots t and u $(u > t)$ of a horizontal segment, then j's CWT is shorter than or equal to i's CWT.

Because the patients in a horizontal segment are scheduled by increasing show probability, it is more likely to observe G_1 patients in the left-most slots and G_2 patients in the right-most slots. Consequently, this Proposition implies that, on average, patients belonging to G_1 tend to have a longer CWT than the patients belonging to G_2 scheduled in the same horizontal segment. Lastly, let us compare the CWT experienced in OB and Z-slots.

PROPOSITION 4: Let i be a patient with show probability p_i scheduled in an OB-slot and j the patient scheduled in the following Z-slot. Let s be the expected number of shows among all patients in an OB-slot except for patient i, conditional to observing at least one such show. If $p_i \leq 1 - \frac{s}{2}$, then patient i has a longer CWT than patient j. Specifically, for double and triple booking:

- *(double booking): If a patient i, scheduled in an OB-slot with only one other patient, has a show probability p_i < 0.5, then s/he has a longer CWT than the patient scheduled in Z.*
- *(triple booking): If a patient i, scheduled in an OB-slot with two other patients whose show probabilities are p_1 and p_2, has a show probability $p_i \leq \frac{p_1+p_2-2p_1p_2}{2p_1+2p_2-2p_1p_2}$, then s/he has a longer CWT than the patient scheduled in Z.*

Although it is impossible to conclusively determine whether the longest wait is experienced in the OB-slot or in the Z-slot, Proposition 4 is important in the context of disparity because it further suggests that patients with a lower show probability are penalized. In particular, it shows that the lower the show probability of a patient in an OB-slot, the higher the likelihood that they will experience a longer waiting time than the patient scheduled in the Z-slot. Note that the threshold of the triple-booking case is smaller than the 0.5 threshold of the double-booking case. Intuitively, that threshold becomes smaller as the number of patients scheduled in OB increases, because the patient in the Z-slot becomes more likely to be the one waiting longer if more patients are overbooked right before him/her. Thus, Proposition 4 suggests that the lower the average show probability of the patients in G_1, the larger the disparity in waiting times between the two groups.

In summary, Propositions 3 and 4 imply that the longest CWT in a segment is experienced by the lowest-show-probability patient scheduled in the OB-slot (who is more likely to belong to G_1 than G_2), or by the patient scheduled in the Z-slot. Thus, we deem OB-slots and Z-slots "undesirable"; conversely, we deem P-slots and H-slots "desirable". Proposition 1, together with Zacharias and Pinedo (2014)'s findings, imply that patients in G_1 are more likely to be scheduled in undesirable slots, whereas patients in G_2 are more likely to be scheduled in desirable slots. Our computational experiments of Section 6 will confirm that this translates into a disparity in waiting times between the patients in G_1 and those in G_2.

4. 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 the disparity between the waiting times experienced by the different patient groups. In this section, we use the expression "patient group" rather than "racial group" because we will explore different definitions of groups, each resulting in a different scheduling strategy.

First, in order to compare the waiting times experienced by different patient groups, we 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:

$$E\left[W_y\right] = \frac{\sum_{i \in G_y} E[WT_i]}{E\left[\#shows\ in\ G_y\right]}, y \in \{1,2,\dots,Y\} \tag{2}$$

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 $E\left[\#shows\ in\ G_y\right]$ is the expected number of showing patients in G_y. This formulation for computing a group's waiting time is well aligned with computing the average waiting time among the showing patients of that group. As shown in Section 5, 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 the disparity among patient groups. In the case of only two patient groups, the disparity can be defined as the absolute value of the difference between the expected waiting times of the groups. Explicitly adding a "disparity" component to TOF has two problems. First, we would need to decide the "weight" of the disparity component 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 would be unclear how to define disparity in the case of more than two groups.

Guided by those considerations, we propose an objective function, which we call "Unbiased Objective Function" (UOF), which does not require new weighting parameters and works for any number of groups. Instead of minimizing the disparity among groups, UOF minimizes the waiting time of the group waiting the longest:

$$UOF = \min(c_W \cdot W^{max} + c_I \cdot E[IT] + c_O \cdot E[OT]) \qquad (3),$$

where

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

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, and equal to 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.

As we will discuss in Section 6, depending on the definition of the Y groups in (4), UOF can be adapted to implement different objective functions, all aiming at striking a balance between schedule efficiency and fairness in terms of waiting times. Next, we analytically show that the schedule that minimizes UOF can be found efficiently.

5. Analytical Properties of UOF

We now present optimality conditions for the appointment scheduling problem with UOF as our objective function:

PROPOSITION 5:

(i) *If each slot is permitted to contain at most four expected shows, 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 patient group are sorted by increasing show probability.*

The limitation of (*i*) to four expected shows per slot, although necessary for the proof, has no impact on practice, as it would never be optimal to schedule in a single slot so many patients that four of them are expected to show up. Thanks to these properties, the solution space is dramatically reduced, making it possible to optimize UOF efficiently through a complete enumeration procedure.

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. Then, we measure quality and fairness obtained by employing TOF and UOF.

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 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 1 reports some summary statistics by race; all other variables are listed below.

Table 1: 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 average show probabilities of the two groups are significantly different (pval < 0.0001). 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 age, 2.8 the

patient's insurance type, 2.9 the patient's number of past no-shows, 2.10 the patient's past no-show rate, 2.11 the patient's number of past appointments, 2.12 the patient's past average lateness, 2.13 the patient's "time in the system" (computed as the time elapsed from the moment when the patient was registered to the appointment date), 2.14 the diagnosis code.

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, and Multilayer Perceptron (Han et al., 2011). For the non-probabilistic classifiers, we derived the show probabilities using Platt's method (Platt, 1999), which consists of building a logistic regression model that predicts the binary show outcome given the show score. Note that some of our features are linearly dependent on each other (e.g., number of prior appointments and number of prior no-shows). Given that multicollinearity may negatively affect Logistic Regression, we employed a Lasso regression (Tibshirani, 1996) prior to building logistic regression models. Lasso is particularly well-suited in this case given the large number of features (over 4,000, most of them dummies). The Lasso parameter α was chosen each time through cross-validation on the current training data with α ranging from 0.001 to 0.2000 in increments of 0.001.

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 used 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 2.

Table 2: Cross-validated AUC and Brier's score on the training set. The best-performing models in terms of Brier's score are in bold.

Classification Technique	All Features (Level-2)		All features except socio-economic indicators (Level-1)	
	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
Lasso + Logistic Regression	0.677	**0.180**	0.638	**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, we also build models that do not use any socio-economic features (i.e., features 2.1 to 2.8 are excluded). We do this to obtain show probabilities uncorrelated with race, thus, limiting racial disparity in the schedule. We refer to the model built with all features as Level-2, and the model built with all features except the socio-economic features as Level-1. For both sets of features, the best-performing technique in terms of Brier's score is Lasso + Logistic Regression. Thus, 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 were 0.635 and 0.186 using the Level-1 model, and 0.675 and 0.179 using the Level-2 model. Those values are similar to the values obtained on the training set, which suggests that there is no overfitting.

6.2. Scheduling Results

We now employ the estimated show probabilities 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. We consider the following combinations: $(N, F) \in \{(6,4), (7,5), (8,6), (10,7)\}$. We chose those values for two reasons. First, those values are appropriate choices for the show rate of our data set, which is 73.4%. Second, our data set contains the appointments of 16 providers, each seeing a daily average number of patients between 1.60 and 8.05, with an overall average across all providers of 4.99

daily patients. We note that clinics may see more patients by having multiple sessions in the same day (e.g., a morning and an afternoon session). 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 and in the features used to predict the patients' show probabilities. We compare the performance of the following strategies:

- **TOF-2**: This is the state-of-the-art strategy which employs the Level-2 model (the model that uses all features) to derive probabilities of show and TOF to schedule appointments. We expect this strategy to obtain the best-quality schedules (as measured by TOF), but also the largest disparity in waiting times between black and non-black patients.
- **TOF-1**: In an effort to limit the disparity, this strategy employs the Level-1 model (the model that excludes the socio-economic features) to derive probabilities of show and TOF to schedule appointments.
- **TOF-0**: Under this strategy, no prediction is made. All appointment requests are assumed to show with probability 0.733, the show rate in the training set. Appointments are scheduled by minimizing TOF. Because this strategy does not employ patient-level show predictions, we expect it to obtain no racial disparity, but also to obtain the worst-quality schedules.
- **Race-aware UOF (UOF-R)**: This strategy employs Level-2 show predictions and UOF to schedule patients. We define two groups for UOF: G_1 is the group of black patients and G_2 the group of non-black patients. That is, this strategy employs show predictions that are correlated with race, but then it limits racial disparity by employing UOF as objective. The suitability of this objective within the context of appointment scheduling requires a deeper ethical discussion, which is beyond the scope of this paper. For practitioners who prefer avoiding using a "race-aware" method, we develop two more methods that are "race-unaware" because they do not explicitly use race to make decisions.
- **No-show-based UOF (UOF-NS)**: Similar to UOF-R, this strategy employs Level-2 show predictions and UOF to schedule patients. However, instead of defining the two groups based on the patients' race,

here we define them based on their no-show risk: G_1 is the group of patients at highest risk of no-show and G_2 is the group of patients at lowest risk of no-show. For any scheduling problem, these two groups are found by clustering the N no-show probabilities given as input through the K-means algorithm (Xu and Wunsch, 2008) with K = 2. This strategy is similar to UOF-R in that it also addresses the racial disparity in the scheduling stage rather than in the prediction stage, but unlike UOF-R, it is race-unaware, because it does not employ race to make decisions.

- **Min-max UOF (UOF-MM)**: Similar to UOF-R, this strategy employs Level-2 show predictions and UOF to schedule patients. The goal of this race-unaware strategy is to minimize the waiting time experienced by the individual patient waiting the longest (i.e., a min-max approach). According to Rea et al. (2021a and 2021b), this objective maximizes "equality" because it aims to equalize all patients' waiting times. In relation to equation (4), this strategy is implemented by defining $Y = N$ one-patient groups (i.e., one group for each patient with each group containing exactly one patient). Like UOF-NS, this strategy also represents an attempt to address disparity at the scheduling stage without employing racial information.

After building the schedules under each strategy, we consider the appointments' actual show outcome from the data, and record the schedules' cost, overtime, and idle time. The costs are evaluated by computing the clinic cost through (1), i.e. TOF, because that metric reflects the actual scheduling cost incurred by the clinic. We also record the waiting times of black and non-black patients, as well as the proportion of black and non-black patients who show for their appointment and experience a wait of at least 30 minutes. We compute the racial disparity as the extra waiting time, calculated as a percentage, experienced by the group that waits longer. For each strategy, we also compute the percent deviation of the average cost obtained under that strategy and the cost obtained using the state-of-the-art strategy, TOF-2. Following the parameter choice of Zacharias and Pinedo (2014) and Robinson and Chen (2010, 2011), we fix the idle time cost rate to c_I = 1, the overtime cost rate to c_O = 1.5, and we consider different values of the waiting time cost rate

c_W between 0 and 1. The results in Table 3 are obtained by solving the instances with $(N, F) \in \{(6,4), (7,5)\}$ to optimality and the instances with $(N, F) \in \{(8,6), (10,7)\}$ with a 10-second run-time limit. Solving the instances with $(N, F) = (10,7)$ was further simplified by utilizing only 200 scenarios of no-shows randomly generated from the 2^10 total possible scenarios. Finally, to better interpret the results, we assume that each appointment slot is 30 minutes. Table 3 reports our results for c_W = 0.5. Table 4 reports our results after relaxing the assumption of constant service times, and assuming a coefficient of variation of 0.2, as suggested by Samorani and Ganguly (2016) for the case of 30-minute appointment slots.

Table 3 confirms that TOF-2 obtains the lowest cost and the largest racial disparity in waiting time, with black patients spending 33.12% longer than non-black patients in the waiting room. At the other end of the spectrum is TOF-0, which avoids racial disparity by not using individual patients' no-show predictions. However, the clinic cost of TOF-0 is 13.70% higher than that obtained by TOF-2, on average.

By employing only a subset of features, TOF-1 obtains a schedule cost and a racial disparity that are between those of TOF-0 and TOF-2. Notably, TOF-1 results both in a cost significantly greater than that obtained by TOF-2, and in a disparity significantly greater than zero. Our results on TOF-1 confirm the observation by Murray et al. (2020) that eliminating all socio-economic features still results in significant racial disparity. The reason lies in the presence of features other than socio-economic factors, e.g. the patient's prior no-show history, which are also correlated with race. The two race-unaware strategies that we propose, UOF-NS and UOF-MM, obtain schedules that are either significantly more costly than those obtained by TOF-2 or that have significant disparity.

Our race-aware methodology, UOF-R, obtains schedules that do not have any significant racial disparity and whose cost is generally not statistically different from that obtained by the state-of-the-art method, TOF-2. This result suggests that while the schedule that optimizes TOF has a significant racial disparity, there are other near-optimal schedules that are very similar to the optimal one, but do not have any racial disparity. For example, by exchanging the position of a black patient and a non-black patient that have similar show probabilities, one could obtain a schedule with a similar cost but with a different racial disparity.

Table 3: Results on four sets of problems with different number of appointment requests, $\boldsymbol{N}$, and appointment slots, $\boldsymbol{F}$. Waiting time, idle time and overtime cost coefficients are: c_W = 0.5, c_I = 1, $c_O = 1.5$, respectively. All times are in minutes. In **bold**, racial disparity significantly greater than zero and schedule cost statistically greater than the cost obtained by TOF-2 (pvalue < 0.01).

		TOF with a predictive model based on			No-Show UOF (UOF-NS)	Min-max UOF (UOF-MM)	Race-aware UOF (UOF-R)
		No predictive model (TOF-0)	All features except socio-econ. indicators (TOF-1)	All features (state-of-the-art method) (TOF-2)			
$N = 6, F = 4$	Overtime	24.79	23.84	24.14	23.33	22.84	23.98
	Idle time	11.83	10.87	11.17	10.36	9.87	11.02
	Black Patients' wait	15.54	15.02	15.24	15.27	15.85	13.14
	Non-black patients' wait	15.08	13.77	11.26	13.75	14.33	13.13
	Wait≥30 min. (%B vs %NB)	48 v 47	45 v 42	45 v 34	46 v 42	48 v 44	40 v 39
	Racial Disparity	3.05%	**9.08%**	**35.35%**	**11.05%**	**10.61%**	0.08%
	Cost Δ% (vs TOF-2)	**9.52%**	**3.17%**	0.00%	**1.98%**	**1.98%**	0.79%
$N = 7, F = 5$	Overtime	19.65	20.59	20.73	20.18	19.33	20.69
	Idle time	14.58	15.52	15.66	15.11	14.26	15.62
	Black Patients' wait	16.85	13.39	13.47	13.82	14.09	11.8
	Non-black patients' wait	16.49	12.14	9.94	12.44	12.99	11.67
	Wait≥30 min. (%B vs %NB)	51 v 50	41 v 38	41 v 30	41 v 38	44 v 41	36 v 36
	Racial Disparity	2.18%	**10.30%**	**35.51%**	**11.09%**	**8.47%**	1.11%
	Cost Δ% (vs TOF-2)	**15.08%**	**4.37%**	0.00%	**4.37%**	**2.78%**	1.98%
$N = 8, F = 6$	Overtime	17.84	18.02	18.19	17.24	16.58	18.02
	Idle time	20.39	20.57	20.74	19.79	19.13	20.57
	Black Patients' wait	14.88	12.28	12.31	12.82	12.81	10.99
	Non-black patients' wait	14.82	11.2	9.11	12.05	11.95	11.26
	Wait≥30 min. (%B vs %NB)	46 v 46	39 v 35	38 v 29	39 v 37	41 v 39	35 v 35
	Racial Disparity	0.40%	**9.64%**	**35.13%**	**6.39%**	**7.20%**	2.46%
	Cost Δ% (vs TOF-2)	**16.09%**	**4.60%**	0.00%	**4.98%**	**2.30%**	**3.07%**
$N = 10, F = 7$	Overtime	32.03	31.93	31.8	30.81	28.53	31.57
	Idle time	20.35	20.24	20.12	19.13	16.85	19.89
	Black Patients' wait	19.02	16.74	16.71	17.07	19.58	14.96
	Non-black patients' wait	18.81	15.23	13.21	15.51	18.73	15.14
	Wait≥30 min. (%B vs %NB)	53 v 52	47 v 44	46 v 39	47 v 44	53 v 52	43 v 42
	Racial Disparity	1.12%	**9.91%**	**26.50%**	**10.06%**	**4.54%**	1.20%
	Cost Δ% (vs TOF-2)	**14.11%**	**4.21%**	0.00%	**2.97%**	**7.18%**	1.49%
	Average Racial Disparity	1.69%	9.73%	33.12%	9.65%	7.70%	1.21%
	Avg Cost Δ%	13.70%	4.09%	0.00%	3.58%	3.56%	1.83%

Table 4: Results on four sets of problems with different number of appointment requests, $\boldsymbol{N}$, and appointment slots, $\boldsymbol{F}$, with stochastic service times. Waiting time, idle time and overtime cost coefficients are: $\boldsymbol{c_W} = 0.5$, $\boldsymbol{c_I} = 1$, $\boldsymbol{c_O} = 1.5$, respectively. All times are in minutes. In **bold**, racial disparity significantly greater than zero and schedule cost statistically greater than the cost obtained by TOF-2 (pvalue < 0.01).

		TOF with a predictive model based on			No-Show UOF (UOF-NS)	Min-max UOF (UOF-MM)	Race-aware UOF (UOF-R)
		No predictive model (TOF-0)	All features except socio-econ. indicators (TOF-1)	All features (state-of-the-art method) (TOF-2)			
$N = 6, F = 4$	Overtime	27.53	26.95	27.55	26.42	25.82	27.29
	Idle time	14.57	13.99	14.58	13.46	12.85	14.32
	Black Patients' wait	16.78	16.91	17.53	16.95	17.3	14.93
	Non-black patients' wait	16.26	15.27	13.12	15.35	15.68	15.05
	Wait≥30 min. (%B vs %NB)	44 v 43	43 v 39	43 v 33	43 v 39	44 v 40	38 v 37
	Racial Disparity	3.20%	**10.74%**	**33.61%**	**10.42%**	**10.33%**	0.80%
	Cost Δ% (vs TOF-2)	**4.41%**	1.02%	0.00%	-0.09%	-1.02%	0.07%
$N = 7, F = 5$	Overtime	22.75	24.38	24.85	23.9	22.95	24.77
	Idle time	17.68	19.31	19.78	18.83	17.88	19.7
	Black Patients' wait	18.34	15.56	16.16	15.69	15.82	13.73
	Non-black patients' wait	18.03	14.14	12.21	14.38	14.7	14.09
	Wait≥30 min. (%B vs %NB)	47 v 46	39 v 36	40 v 30	39 v 35	40 v 37	34 v 33
	Racial Disparity	1.72%	**10.04%**	**32.35%**	**9.11%**	**7.62%**	2.62%
	Cost Δ% (vs TOF-2)	**7.17%**	**1.63%**	0.00%	0.98%	-0.97%	0.76%
$N = 8, F = 6$	Overtime	21.26	22.09	22.46	21.31	20.46	22.17
	Idle time	23.81	24.64	25.01	23.86	23.01	24.72
	Black Patients' wait	16.48	14.61	15.29	14.91	14.75	13.11
	Non-black patients' wait	16.47	13.34	11.7	14.23	13.92	13.72
	Wait≥30 min. (%B vs %NB)	43 v 43	36 v 32	37 v 27	37 v 34	37 v 35	33 v 33
	Racial Disparity	0.06%	**9.52%**	**30.68**%	**4.78%**	**5.96%**	**4.65%**
	Cost Δ% (vs TOF-2)	**7.41%**	**1.54%**	0.00%	**1.61%**	-1.35%	0.62%
$N = 10, F = 7$	Overtime	35.8	36.32	36.57	35.26	32.37	36.24
	Idle time	24.12	24.64	24.89	23.58	20.69	24.56
	Black Patients' wait	21.07	19.54	20.07	19.74	21.45	17.63
	Non-black patients' wait	20.89	17.86	16.14	18.26	20.72	18.0
	Wait≥30 min. (%B vs %NB)	49 v 49	44 v 41	45 v 37	45 v 41	49 v 48	40 v 41
	Racial Disparity	0.86%	**9.41%**	**24.35%**	**8.11%**	**3.52%**	2.10%
	Cost Δ% (vs TOF-2)	**7.26%**	**1.87%**	0.00%	0.83%	**1.45%**	0.21%
	Average Racial Disparity	1.46%	9.93%	30.25%	8.10%	6.86%	2.54%
	Avg Cost Δ%	6.56%	1.52%	0.00%	0.83%	-0.47%	0.42%

Table 4 confirms that the disparity is also present when service times are stochastic. Note that because all methodologies assume constant service times when building the schedule, the solution found by TOF-2 may not be the best-quality solution anymore in the case of stochastic service times.

In order to identify the reasons underlying the positive performance of UOF-R, we recorded the slot types where each appointment (including both shows and no-shows) was scheduled. Table 5 reports the percentage of black and non-black patients that each strategy schedules in each type of appointment slots, for each problem instance size (N, F).

Table 5: Percentage of black versus percentage of non-black patients (including shows and no-shows) scheduled in undesirable appointment slots (OB, Z, and OB&Z) for each strategy.

	%Black vs %non-black	TOF-0	TOF-1	TOF-2	UOF-MM	UOF-NS	UOF-R
$N = 6$, $F = 4$	OB&Z	0 v 0	3 v 2	10 v 6	3 v 1	7 v 6	5 v 10
	OB	67 v 66	72 v 59	71 v 49	77 v 56	73 v 51	74 v 47
	Z	16 v 17	10 v 14	7 v 12	9 v 16	10 v 17	7 v 13
	H	16 v 17	9 v 17	8 v 24	7 v 18	6 v 17	10 v 21
	P	0 v 0	5 v 8	4 v 9	4 v 9	4 v 8	4 v 8
$N = 7$, $F = 5$	OB&Z	0 v 0	1 v 0	4 v 2	2 v 1	3 v 2	1 v 2
	OB	57 v 57	66 v 50	69 v 43	71 v 45	70 v 44	72 v 44
	Z	28 v 29	11 v 13	9 v 12	9 v 14	11 v 17	8 v 16
	H	14 v 14	15 v 26	12 v 31	12 v 29	10 v 26	14 v 27
	P	0 v 0	7 v 10	6 v 11	6 v 11	6 v 11	6 v 11
$N = 8$, $F = 6$	OB&Z	0 v 0	0 v 0	2 v 1	1 v 0	2 v 1	1 v 1
	OB	50 v 50	60 v 43	65 v 37	66 v 38	64 v 38	66 v 37
	Z	25 v 25	12 v 13	10 v 12	10 v 13	12 v 17	7 v 17
	H	25 v 25	20 v 33	16 v 38	16 v 36	15 v 32	20 v 33
	P	0 v 0	8 v 11	7 v 12	7 v 12	7 v 12	6 v 12
$N = 10$, $F = 7$	OB&Z	3 v 3	10 v 8	13 v 8	6 v 3	10 v 9	7 v 11
	OB	57 v 57	58 v 46	61 v 39	64 v 48	63 v 42	63 v 38
	Z	20 v 20	12 v 14	9 v 13	12 v 17	11 v 17	8 v 14
	H	19 v 19	16 v 26	13 v 33	14 v 25	13 v 25	17 v 29
	P	0 v 0	5 v 7	4 v 8	4 v 7	4 v 7	5 v 7

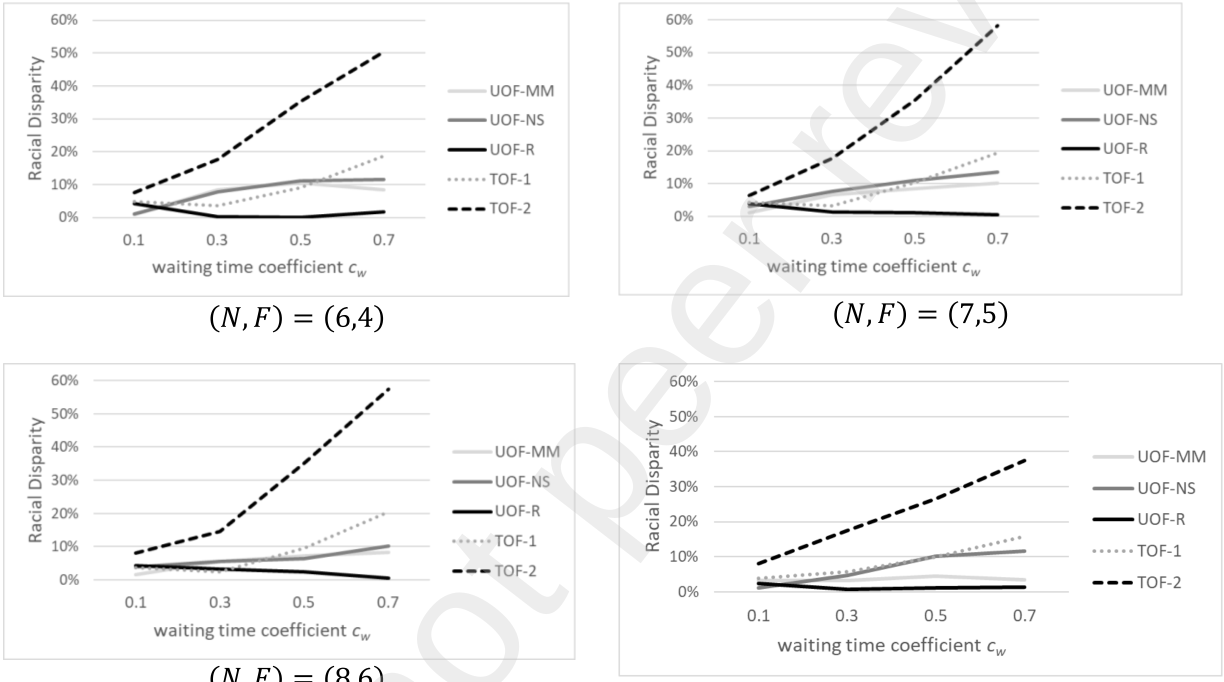
An appointment slot is labeled OB&Z if it is both overbooked and following another overbooked slot. The first column of Table 5 shows that TOF-0 schedules the same proportion of black and non-black patients into desirable appointment slots (P, H); this is unsurprising given that TOF-0 assumes that all patients have the same show probability. TOF-2, as expected, disproportionately schedules non-black patients into desirable appointment slots and black patients into undesirable appointment slots (OB, Z, OB&Z). For example, with N = 6 and F = 4, 8%+4%=12% of black patients are scheduled in desirable slots, compared to 24%+9%=33% of non-black patients. The numbers of TOF-1 are, once again, between those of TOF-0 and TOF-2.

Interestingly, similar to TOF-2, UOF-R also overbooks black patients much more often than non-black patients. For example, with N = 6 and F = 4, 74%+5%=79% of black patients are overbooked compared to 47%+10%=57% of non-black patients. However, while the percentage of black patients assigned an OB&Z-slot is 10% under TOF-2, it is only 5% under UOF-R; in contrast, the same percentage for non-black patients goes from 6% under TOF-2 to 10% under UOF-R. Although UOF-R is more likely to schedule black patients than non-black patients in OB-slots, it does so without resulting in black patients waiting longer than non-black patients, as shown by the average waiting times and the probability of long waits reported in Tables 3 and 4. This result shows that overbooking does not necessarily translate into longer waits; it will do so depending on the other patients scheduled with or before the focal patient.

Finally, note that TOF-0 results in some racial disparity under all parameter configurations. Although this disparity is always small and insignificant, its consistent presence under TOF-0 is counter intuitive, because under TOF-0 patients have an equal chance of being scheduled in any slot. When patients are scheduled in an OB slot, their conditional waiting time is positively correlated to the show probability of the other patients in the same slot (see Proposition 4). Consequently, when both black and nonblack patients are scheduled in the same OB slot, black patients will tend to have a longer conditional waiting time, on average, than nonblack patients, because of the two racial groups' different show rates. While throughout our experiments this structural disparity is always insignificant, we leave to future research the task to study under which circumstances it may play a bigger role.

6.3. Varying the waiting time coefficient

While the results in the previous subsection are all obtained by fixing the waiting time coefficient, c_W, to 0.5, in this subsection we study the robustness of our results for other values of c_W. To this end, we re-execute the same computational experiments as in the previous section (i.e., the experiments whose results are reported in Table 3) for $c_W \in \{0.1, 0.3, 0.5, 0.7\}$. For each strategy and each schedule size, Figures 3 and 4, respectively, illustrate the impact of the waiting time coefficient on racial disparity and on the schedule cost gap with respect to TOF-2. To make it easier to interpret the charts, we do not report the racial disparity obtained by TOF-0, because it is always very close to 0%, or the $\Delta_{cost}\%$ of TOF-2, because it is equal to 0% by definition.



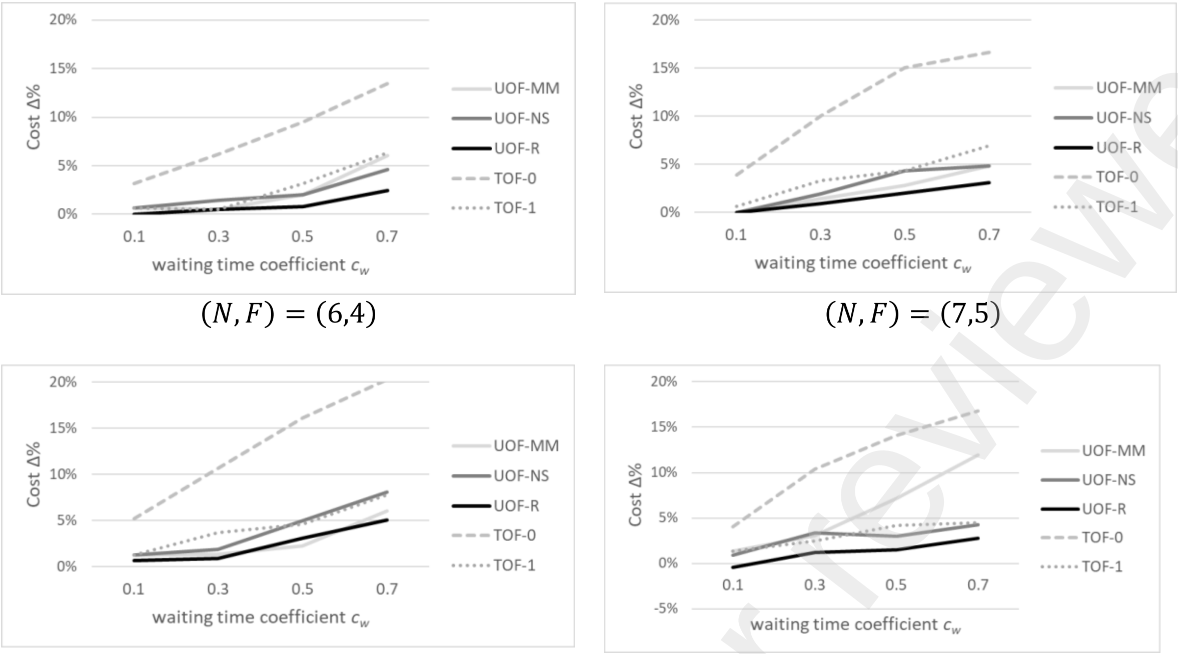
$(N, F) = (6,4)$

$(N, F) = (7,5)$

$(N, F) = (8,6)$

$(N, F) = (10,7)$

Figure 3: Effect of the waiting time coefficient, $\boldsymbol{c_W}$, on racial disparity, computed as percent absolute difference between waiting times experienced by black and non-black patients.



$(N, F) = (6,4)$

$(N, F) = (7,5)$

$(N, F) = (8,6)$

$(N, F) = (10,7)$

Figure 4: Effect of the waiting time coefficient, $\boldsymbol{c_W}$, on cost **Δ**%, computed as percent difference between the schedule cost obtained by each strategy and the schedule cost obtained by TOF-2.

Our results suggest that all methods achieve similar racial disparity and schedule cost for small values of $\boldsymbol{c_W}$. This is intuitive. Clinics that put a small weight on patient wait time tend to heavily overbook at the beginning of the session under all scheduling strategies. Scheduling multiple patients in the same few appointment slots eventually results in patients suffering similar wait times, and in clinics experiencing similar costs. As an extreme example, consider a schedule where all patients are scheduled in the first slot of the day: the difference in wait times would be small and the difference in costs would be zero. As $\boldsymbol{c_W}$ grows, the racial disparity of TOF-2 grows dramatically, highlighting the unsuitability of this strategy in contexts where limiting disparity is a concern. Among the other strategies, UOF-R is the only strategy capable of limiting racial disparity for all schedule sizes; it is also the strategy that consistently obtains the performance closest to that obtained by TOF-2. Among the race-unaware strategies, UOF-MM is the strategy that in most cases strikes the best balance between fairness and cost gap.

6.4. Stochastic Number of Patients

In this section, we check that our results hold when we relax the assumptions of a clinic scheduling a constant number of appointment requests, N. In the previous computational experiments, we assumed that a clinic scheduled a constant number of N appointment requests into a constant number of F appointment slots; in this section, we consider the case where F is fixed but N varies. We re-execute the same computational experiments as in Section 6.2, limitedly to the case with F = 5 appointment slots and let the value of N be taken from the set {6,7,8} with uniform probability. The results are reported in Table 6.

Table 6 suggests that the findings of Section 6.2 hold even if N varies: (1) TOF-2 results in a very large racial disparity, (2) UOF-R results in very little disparity and a small cost increase when compared to TOF-2, and (3) UOF-MM results in both a larger disparity and cost than UOF-R.

Table 6: Results with $\boldsymbol{F = 5}$ appointment slots and $\boldsymbol{N} \in \{\boldsymbol{6, 7, 8}\}$ patients, with each value of $\boldsymbol{N}$ equally likely. Waiting time, idle time and overtime cost coefficients are: $\boldsymbol{c_W}$ = 0.5, $\boldsymbol{c_I}$ = 1, $\boldsymbol{c_O = 1.5}$, respectively. All times are in minutes. In **bold**, racial disparity significantly greater than zero and schedule cost statistically greater than the cost obtained by TOF-2 (pvalue < 0.01).

		TOF with a predictive model based on			No-Show UOF (UOF-NS)	Min-max UOF (UOF-MM)	Race-aware UOF (UOF-R)
		No predictive model (TOF-0)	All features except socio-econ. indicators (TOF-1)	All features (state-of-the-art method) (TOF-2)			
$N \in \{6,7,8\}, F = 5$	Overtime	20.92	21.71	22.16	21.54	20.4	22.1
	Idle time	15.49	16.29	16.74	16.11	14.97	16.68
	Black Patients' wait	17.59	14.56	14.73	14.84	16.03	12.84
	Non-black patients' wait	17.16	13.02	10.81	13.24	15.26	12.78
	Wait≥30 min. (%B vs %NB)	50 v 50	43 v 39	42 v 33	43 v 38	45 v 43	38 v 36
	Racial Disparity	2.51%	**11.83%**	**36.26%**	**12.08%**	**5.05%**	0.47%
	Cost Δ% (vs TOF-2)	**12.50%**	**2.94%**	0.00%	**3.31%**	**5.15%**	**1.84%**

7. Generalization of our Results

In the previous section, we found that TOF is an unsuitable objective for those clinics concerned about limiting racial disparity. The goal of this section is to further compare the performance of two of our proposed methods, UOF-MM and UOF-R, and study the effect of those parameters that cannot be easily modified in the real-world data used in the previous section: the population show rate, the proportion of patients belonging to G_1 and G_2, and the distribution of the two groups' show probabilities. To this end, we generate a large number of artificial scheduling problems using different parameter combinations, employ UOF-MM and UOF-R to find the optimal schedules, and compare the cost and disparity obtained.

The scheduling problems are generated as follows. We only consider scheduling problems of N = 7 patients to be scheduled in F slots, where $F \in \{4,5,6\}$. The population show rate is set to $q = \frac{F}{N}$ to achieve a balance between supply and demand. A set of N appointment requests is generated by assuming that each request has a probability r of belonging to G_1 and $1 - r$ of belonging to G_2, where $r \in \{0.25,0.50,0.75\}$. The average show probability of the patients in G_2 is set to $p_2 = q + \delta$, where $\delta \in \{0.025,0.100\}$, while the average show probability of G_1 is set to $p_1 = \frac{q-p_2+p_2 r}{r}$, so that the overall show probability is equal to the population show rate, q. We generate each patient's individual show probability by sampling it from a beta distribution, $\beta(p_1, v)$, if the patient belongs to G_1, or $\beta(p_2, v)$ if the patient belongs to G_2. The variance, v, is set to either v_L or v_H, where $(v_L, v_H) = (0.05,0.10)$ if G_2's average show probability, p_2, is less than 90%, and $(v_L, v_H) = (0.01,0.03)$ if p_2 is greater than 90%. We vary the values of v_L and v_H for larger values of p_2 because the β-distribution may not be defined in some cases. The waiting time, idle time, and overtime cost per time unit are set to $c_W = 0.5$, c_I = 1 and c_O = 1.5, respectively.

We generated 200 random problems for each parameter combination, and we solved them using UOF-MM and UOF-R. For each parameter combination, Table 7 reports the average waiting time (WT_1 and WT_2) experienced by the showing patients in G_1 and G_2, respectively, as well as the provider's average overtime (OT) across the 200 problem instances generated under that parameter combination. We assume service times of 30 minutes. Table 7 also reports the percentage cost decrease obtained by using UOF-R when compared to UOF-MM.

In most cases, UOF-R performs better than UOF-MM both in terms of waiting time disparity and clinic cost. The waiting times obtained by UOF-R are generally shorter and more similar across G_1 and G_2 than those obtained by UOF-MM, and the overall cost is, in most cases, significantly lower using UOF-R (bold entries in Table 7). In contrast, in those cases where UOF-MM obtains a significantly smaller cost than UOF-MM (underlined entries in Table 7), it also obtains a much larger disparity in waiting times.

Table 7: Computational results on simulated scheduling problems. All times are in minutes. Cost is in bold (underlined) if significantly smaller (greater) under UOF-R (pval < 0.01).

					UOF-R			UOF-MM			Cost improvement obtained by UOF-R
F	r	q	p_1, p_2	v	WT_1	WT_2	OT	WT_1	WT_2	OT	
4	.25	.57	0.27,0.67	0.05	22	22	19	28	25	18	**1.9%**
4	.25	.57	0.27,0.67	0.1	16	16	17	24	17	16	0.0%
4	.25	.57	0.49,0.59	0.05	23	23	17	29	28	16	**2.4%**
4	.25	.57	0.49,0.59	0.1	15	16	17	24	21	16	**3.1%**
4	.5	.57	0.47,0.67	0.05	23	22	17	29	26	17	**2.2%**
4	.5	.57	0.47,0.67	0.1	16	16	16	23	19	15	**3.1%**
4	.5	.57	0.54,0.59	0.05	24	24	19	30	29	18	**2.4%**
4	.5	.57	0.54,0.59	0.1	16	16	17	23	22	16	**4.0%**
4	.75	.57	0.53,0.67	0.05	24	23	18	29	29	18	**2.1%**
4	.75	.57	0.53,0.67	0.1	17	16	17	23	22	16	**3.4%**
4	.75	.57	0.56,0.59	0.05	24	23	18	30	29	17	**2.2%**
4	.75	.57	0.56,0.59	0.1	17	16	17	21	24	16	**3.1%**
5	.25	.71	0.41,0.81	0.05	16	14	15	24	16	14	-1.1%
5	.25	.71	0.41,0.81	0.1	12	9	16	19	8	15	-3.6%
5	.25	.71	0.63,0.73	0.05	15	15	17	24	19	16	**3.3%**
5	.25	.71	0.63,0.73	0.1	10	10	14	18	13	14	**3.6%**
5	.5	.71	0.61,0.81	0.05	15	14	15	23	18	14	**2.9%**
5	.5	.71	0.61,0.81	0.1	10	8	14	18	9	13	1.3%
5	.5	.71	0.68,0.73	0.05	15	14	16	23	20	15	**3.8%**
5	.5	.71	0.68,0.73	0.1	9	9	14	15	13	14	**5.6%**
5	.75	.71	0.68,0.81	0.05	15	13	16	22	19	15	**4.4%**
5	.75	.71	0.68,0.81	0.1	10	6	14	15	9	14	**4.6%**
5	.75	.71	0.70,0.73	0.05	16	14	17	22	22	16	**4.0%**
5	.75	.71	0.70,0.73	0.1	10	8	13	14	14	13	**4.8%**
6	.25	.86	0.55,0.95	0.01	13	9	12	18	7	10	-15.0%
6	.25	.86	0.55,0.95	0.03	11	4	13	16	2	11	-13.1%
6	.25	.86	0.78,0.88	0.05	6	5	10	13	7	10	**2.3%**
6	.25	.86	0.78,0.88	0.1	4	3	11	12	1	11	-3.9%
6	.5	.86	0.75,0.95	0.01	10	8	13	19	8	11	-8.7%
6	.5	.86	0.75,0.95	0.03	8	3	12	15	2	11	-6.2%
6	.5	.86	0.83,0.88	0.05	6	5	12	11	7	12	**3.3%**
6	.5	.86	0.83,0.88	0.1	3	2	8	5	1	8	0.0%
6	.75	.86	0.82,0.95	0.01	9	4	14	16	7	12	0.0%
6	.75	.86	0.82,0.95	0.03	7	1	13	12	2	12	**2.6%**
6	.75	.86	0.84,0.88	0.05	5	3	11	9	7	11	**5.9%**
6	.75	.86	0.84,0.88	0.1	4	1	11	4	2	11	0.6%

In agreement with the results of Section 6, the results of this section suggest that UOF-R is, in most cases, the best-performing method in terms of cost and racial disparity. Despite being statistically significant, the difference in costs between UOF-R and UOF-MM are generally modest, which makes UOF-MM a good race-unaware alternative to UOF-R.

8. 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. Due to the structural properties of TOF, TOF tends to penalize the group of patients with the lower show probability. Because probability of show tends to be correlated with the patients' racial group, we showed that black patients are likely to experience significantly longer waiting times than the rest of the patient population. In our simulations, black patients wait over 30% longer than non-black patients. Our results suggest that this disparity is not eliminated by removing socio-economic indicators from the data.

To reduce the disparity, we develop a "race-aware" objective function, UOF-R, which instead of minimizing everyone's waiting time (as TOF does), minimizes the waiting time of the racial group expected to wait longest. This strategy dramatically reduces racial disparity while obtaining a similar clinic cost to that obtained by TOF. We also developed "race-unaware" objective functions, such as UOF-MM, which minimizes the waiting time of the patient expected to wait longest. UOF-MM is a good alternative to TOF, as it obtains a smaller racial disparity at the price of a small increase in clinic costs, but it performs significantly worse than UOF-R both in terms of disparity and clinic costs. The methodologies introduced in this paper are flexible enough to extend beyond racial differences, and could be applied to bridge the gap between any type of inequality. Our results may be generally described as seeking to raise the minimum quality of healthcare for all patients in a clinic; a goal which can be universally agreed upon.

Given that this is the first study to address racial disparity in appointment scheduling systems, there are several topics that can be addressed with future research. The first is the need of reflection on the appropriateness of race-aware methodologies for healthcare. Is it ethical to take into consideration a patient’s race to minimize overall racial disparity? We purposely leave this question open, hoping to spark a discussion in the ethics community.

The second avenue for future research is considering aspects of social disparity other than racial disparity. For example, a patient’s income could be used to formulate the scheduling problem in terms of socio-economic disparity, whose goal would be to minimize differences between poorer and wealthier patients’ waiting times. A third avenue for research consists of developing race-unaware methodologies that achieve both efficiency and racial fairness. We attempted to solve this question by developing UOF-MM and UOF-NS, but these objectives neither achieve the same efficiency as the race-aware method nor fully resolve racial disparity. A fourth avenue is to implement our methodologies at an actual clinic and measure not only the impact on clinic efficiency (as done by Soltani et al. 2019), but also the impact on racial disparity.

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)?

Acknowledgements

The authors thank the editors and the referees for their valuable suggestions. We thank the Operations Management group at the University of Alberta School of Business for their feedback on an earlier version of the paper. Michele Samorani’s work was partly funded by the 2020 Leavey Grant.

References

Achiume, E.T., 2020. Racial discrimination and emerging digital technologies: a human rights analysis. United Nations 44th Human Rights Council, Geneva, 15th July 2020. Report number A/HRC/44/57.

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.

Benjamin, R., 2016. Innovating inequity: If race is a technology, postracialism is the Genius Bar. *Ethnic and racial studies*, *39*(13), pp.2227-2234.

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

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

Campbell, J.D., Chez, R.A., Queen, T., Barcelo, A. and Patron, E., 2000. The no-show rate in a high-risk obstetric clinic. *Journal of Women's Health & Gender-Based Medicine*, *9*(8), pp.891-895.

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

Corfield, L., Schizas, A., Williams, A. and Noorani, A., 2008. Non-attendance at the colorectal clinic: a prospective audit. *The Annals of The Royal College of Surgeons of England*, *90*(5), pp.377-380.

Cui, R., Li, J. and Zhang, D.J., 2020. Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Science*, *66*(3), pp.1071-1094.

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.

Ganju, K.K., Atasoy, H., McCullough, J. and Greenwood, B., 2020. The Role of Decision Support Systems in Attenuating Racial Biases in Healthcare Delivery. *Management Science*.

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.

Hamilton, W., Round, A. and Sharp, D., 2002. Patient, hospital, and general practitioner characteristics associated with non-attendance: a cohort study. *Br J Gen Pract*, *52*(477), pp.317-319.

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

Office of Disease Prevention and Health Promotion. 2020. Social Determinants of Health. Available from: https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

Heckler, M.M., 1985. *Report of the Secretary's Task Force Report on Black and Minority Health Volume I: Executive Summary*. Gvernment Printing Office.

Hersch, J. and Shinall, J.B., 2015. Fifty years later: The legacy of the Civil Rights Act of 1964. *Journal of Policy Analysis and Management*, *34*(2), pp.424-456.

Hooper, M.W., Nápoles, A.M. and Pérez-Stable, E.J., 2020. COVID-19 and racial/ethnic disparities. *Jama*.

Hostetter, M. and Klein, S., 2018. In focus: Reducing racial disparities in health care by confronting racism. *Commonwealth Fund, September*, *27*, pp.121-27.

Huang, Y. and Hanauer, D.A., 2014. Patient no-show predictive model development using multiple data sources for an effective overbooking approach. *Applied clinical informatics*, *5*(03), pp.836-860.

Kaplan-Lewis, E. and Percac-Lima, S., 2013. No-show to primary care appointments: why patients do not come. *Journal of primary care & community health*, *4*(4), pp.251-255.

Lacy, N.L., Paulman, A., Reuter, M.D. and Lovejoy, B., 2004. Why we don't come: patient perceptions on no-shows. *The Annals of Family Medicine*, *2*(6), pp.541-545.

LaGanga, L.R. and Lawrence, S.R., 2012. Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management*, *21*(5), pp.874-888.

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*, *28*(8), pp.2068-2086.

Martin, C., Perfect, T. and Mantle, G., 2005. Non-attendance in primary care: the views of patients and practices on its causes, impact and solutions. *Family Practice*, *22*(6), pp.638-643.

Massey, D.S., 2015, June. The Legacy of the 1968 Fair Housing Act. In *Sociological Forum* (Vol. 30, pp. 571-588).

Miller, A.J., Chae, E., Peterson, E. and Ko, A.B., 2015. Predictors of repeated "no-showing" to clinic appointments. *American journal of otolaryngology*, *36*(3), pp.411-414.

Murray, S.G., R.M. Watcher, R.J. Cucina. 2020. “Discrimination By Artificial Intelligence In A Commercial Electronic Health Record—A Case Study, " Health Affairs Blog, January 31, 2020. DOI: 10.1377/hblog20200128.626576

Nelson, A., 2002. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the National Medical Association*, *94*(8), p.666.

Neal, R.D., Hussain-Gambles, M., Allgar, V.L., Lawlor, D.A. and Dempsey, O., 2005. Reasons for and consequences of missed appointments in general practice in the UK: questionnaire survey and prospective review of medical records. *BMC Family Practice*, *6*(1), p.47.

Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), pp.447-453.

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, 169*(12), pp.866-872.

Rea, D., C. Froehle, S. Masterson, B. Stettler, G. Fermann, A. Pancioli. 2021a. Unequal but Fair: Incorporating Distributive Justice in Operational Allocation Models. *Production and Operations Management*. Forthcoming.

Rea, D., L. Lozano, C. Froehle. 2021b. *Algorithmic Justice: Fair Compensation Through Principled Compromises*. Working paper.

Robinson, L.W. and Chen, R.R., 2010. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management*, *12*(2), pp.330-346.

Robinson, L.W. and Chen, R.R., 2011. Estimating the implied value of the customer's waiting time. *Manufacturing & Service Operations Management*, *13*(1), pp.53-57.

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.

Samorani, M. and Harris, S., 2019. The Impact of Probabilistic Classifiers on Appointment Scheduling with No-Shows. In *Fortieth International Conference on Information Systems, Munich*.

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.

Soltani, M., Samorani, M. and Kolfal, B., 2019. Appointment scheduling with multiple providers and stochastic service times. *European Journal of Operational Research*, *277*(2), pp.667-683.

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.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), pp.267-288.

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.

Xu, R. and Wunsch, D., 2008. *Clustering* (Vol. 10). John Wiley & Sons.

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

Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling
Online Supplement – Proofs

Hypothesis from Section 3.3:
Assume the following:

- *(i)* *The patients' individual show probabilities are sampled from a random distribution, such that a patient with a show probability p belongs to G_2 with probability $r(p)$, and to G_1 with probability $1 - r(p)$,*
- *(ii)* *$r(p)$ is monotonically increasing with p, and*
- *(iii)* *Patients with a lower show probability have a longer conditional waiting time (CWT) than patients with a higher show probability.*

Under the above assumptions, patients in G_1, on average, have a longer CWT that patients in G_2.

PROOF: To facilitate the proof, we first introduce Lemma 1.

Lemma 1: If $\frac{A}{B} \geq \frac{C}{D}$ and $B \geq D$, then $\frac{A-C}{B-D} \geq \frac{C}{D}$.

PROOF of Lemma 1: Given $\frac{A}{B} \geq \frac{C}{D}$ and $B \geq D$, the following holds

$$AD \geq BC$$
$$AD - CD \geq BC - CD$$
$$(A - C)D \geq (B - D)C$$
$$\frac{A - C}{B - D} \geq \frac{C}{D} \blacksquare$$

Now we proceed with the proof of our hypothesis, by mathematical induction. In Case 1, we assume that the patients can only have two possible show probability values; in Case 2, we consider the case of n possible distinct values of show probability.

<u>Case 1</u>: We consider the simplest case with only two possible show probability values. Thus, let us assume that the patients are partitioned into two groups, each composed of f_1 and f_2 patients, respectively, such that all patients in group $i \in \{1,2\}$ have show probability p_i, with $p_1 \leq p_2$ without loss of generality. Among the f_1 patients with a show probability of p_1, the probability of a patient belonging to G_2 is $r_1 = r(p_1)$, and among the f_2 patients with a show probability of p_2, the probability of a patient belonging to G_2 is $r_2 = r(p_2)$. Let w_1 be the average CWT of patients with a show probability of p_1 and let w_2 be the average CWT for patients with a show probability of p_2. According to our assumptions we have $w_1 \geq w_2$ and $r_1 \leq r_2$ because $p_1 \leq p_2$.
The average waiting time for G_1 and G_2 can be computed as follows:

$$w_{G_2} = \frac{f_1 r_1 w_1 + f_2 r_2 w_2}{f_1 r_1 + f_2 r_2}$$

$$
\begin{aligned}
w_{G_1} &= \frac{f_1(1-r_1)w_1 + f_2(1-r_2)w_2}{f_1(1-r_1) + f_2(1-r_2)} \\
&= \frac{(f_1 w_1 + f_2 w_2) - (f_1 r_1 w_1 + f_2 r_2 w_2)}{(f_1 + f_2) - (f_1 r_1 + f_2 r_2)}
\end{aligned}
$$

Using Lemma 1, it is possible to prove $w_{G_1} \ge w_{G_2}$ by proving::

$\frac{f_1 w_1 + f_2 w_2}{f_1 + f_2} \ge \frac{f_1 r_1 w_1 + f_2 r_2 w_2}{f_1 r_1 + f_2 r_2}$ and $(f_1 + f_2) > (f_1 r_1 + f_2 r_2)$

First, note that $(f_1 + f_2) > (f_1 r_1 + f_2 r_2)$ because $0 < r_1, r_2 < 1$. Let us now show that:

$$
\frac{f_1 w_1 + f_2 w_2}{f_1 + f_2} \ge \frac{f_1 r_1 w_1 + f_2 r_2 w_2}{f_1 r_1 + f_2 r_2}
$$

$$
\frac{(f_1 w_1 + f_2 w_2)(f_1 r_1 + f_2 r_2) - (f_1 + f_2)(f_1 r_1 w_1 + f_2 r_2 w_2)}{(f_1 + f_2)(f_1 r_1 + f_2 r_2)} \ge 0
$$

$$
(f_1^2 r_1 w_1 + f_1 f_2 r_1 w_2 + f_1 f_2 r_2 w_1 + f_2^2 r_2 w_2) - (f_1^2 r_1 w_1 + f_1 f_2 r_2 w_2 + f_1 f_2 r_1 w_1 + f_2^2 r_2 w_2) \ge 0
$$

$$
f_1 f_2 (r_2 - r_1)(w_1 - w_2) \ge 0,
$$

which is true because $w_1 \ge w_2$, and $r_1 \le r_2$.

Case 2: We now consider the generalized case with $n-1$ possible show probability values. Thus, let us assume that the patients are partitioned into $n-1$ groups, each composed of f_i patients, such that all patients in group $i \in \{1,2,\dots,n-1\}$ have show probability p_i, with $p_1 \le \cdots \le p_{n-1}$ without loss of generality. Among the f_i patients with a show probability of p_i, the probability of a patient belonging to G_2 is $r_i = r(p_i)$. Our assumptions imply $r_1 \le \cdots \le r_{n-1}$. We proceed with the proof by induction. Suppose that $w_{G_1} \ge w_{G_2}$ if patients are partitioned into the $n-1$ groups described above; we will show that the same holds if the patients were partitioned into n groups. On the case with $n-1$ groups, Lemma 1 implies that:

$$
\frac{\sum_{i=1}^{n-1} f_i w_i}{\sum_{i=1}^{n-1} f_i} \ge \frac{\sum_{i=1}^{n-1} f_i r_i w_i}{\sum_{i=1}^{n-1} f_i r_i},
$$

Which is equivalent to:

$$
\left(\sum_{i=1}^{n-1} f_i w_i\right)\left(\sum_{i=1}^{n-1} f_i r_i\right) - \left(\sum_{i=1}^{n-1} f_i\right)\left(\sum_{i=1}^{n-1} f_i r_i w_i\right) \ge 0
$$

We want to prove that $w_{G_1} \ge w_{G_2}$ if a new group is added, where the new group (The n-th group) is composed of f_n patients, all having a show probability p_n, with $p_{n-1} \le p_n$. To prove that, we want to prove the following (by Lemma 1):

$$
\frac{\sum_{i=1}^{n} f_i w_i}{\sum_{i=1}^{n} f_i} \ge \frac{\sum_{i=1}^{n} f_i r_i w_i}{\sum_{i=1}^{n} f_i r_i},
$$

Which is equivalent to

$$
\begin{aligned}
&\left[\left(\sum_{i=1}^{n-1} f_i w_i\right)\left(\sum_{i=1}^{n-1} f_i r_i\right) - \left(\sum_{i=1}^{n-1} f_i\right)\left(\sum_{i=1}^{n-1} f_i r_i w_i\right)\right] + f_n r_n \left(\sum_{i=1}^{n-1} f_i w_i\right) + f_n w_n \left(\sum_{i=1}^{n-1} f_i r_i\right) \\
&\qquad \ge\ f_n r_n w_n \left(\sum_{i=1}^{n-1} f_i\right) + f_n \left(\sum_{i=1}^{n-1} f_i r_i w_i\right)
\end{aligned}
$$

Because we already proved that the first component of the left-hand side is greater than or equal to zero, we only need to prove the following:

$$f_n r_n \left(\sum_{i=1}^{n-1} f_i w_i\right) + f_n w_n \left(\sum_{i=1}^{n-1} f_i r_i\right) \geq \ f_n r_n w_n \left(\sum_{i=1}^{n-1} f_i\right) - f_n \left(\sum_{i=1}^{n-1} f_i r_i w_i\right),$$

Which is equivalent to:

$$fn \sum_{i=1}^{n-1} f_i (r_n - r_i)(w_i - w_n) \geq 0,$$

Which is true because $w_1 \geq \cdots \geq w_n$ and $r_1 \leq \cdots \leq r_n$

PROPOSITION 1: If the optimal schedule includes P-slots, then those slots are assigned to the patients with the highest show probability.

PROOF: Suppose that in the optimal schedule, a patient scheduled in a P-slot has a show probability p_1, and another patient scheduled in slot j, which is not a P-slot, has a show probability of p_2, where $p_1 < p_2$. By swapping the two patient slot assignments, we obtain a shorter idle time, overtime, and waiting time. The decrease in overtime and waiting times is obvious. To see why there is also a decrease in idle time, consider that the expected finish time can be expressed in two ways: first, as the number of slots F plus the expected overtime, $E[OT]$; second, as the expected number of shows, $E[shows]$, plus the expected idle time, $E[IT]$. So,

$$E[OT] + F = E[shows] + E[IT]$$

The exchange operation described above causes $E[OT]$ to decrease without affecting F or $E[shows]$. So, the exchange operation must also cause $E[IT]$ to decrease. ■

PROPOSITION 2: Let S be a schedule composed of N patients in F slots where all overbooking occurs in slot V (with V < F); that is, S is composed of a single segment (slots V to F), possibly preceded by a sequence of P-slots (slots 1 to V-1). Let P_0 be the probability of a zero backlog at the beginning of slot F, P_1 the probability of a backlog of one time unit at the beginning of slot F, and $E[B|B \geq 2]$ the expected backlog at the beginning of slot F, conditional to it being at least two time units. If the patient scheduled in slot F has a show probability greater than $\frac{(c_I+c_O)(1-P_0)}{(c_I+c_O)(1-P_0)+c_W\left(P_1+E[B|B\geq 2](1-P_0-P_1)\right)}$, then there exists a better-quality schedule with an additional P-slot allocated to that patient.

PROOF: Let i be the patient scheduled in slot F of schedule S with a show probability of p_i. Let X denote the portion of S from slots 1 to $F-1$, inclusive. Consider an alternative schedule S', obtained by moving patient i from the slot F to the first slot, and moving X back by one slot (see Figure A.1 below).

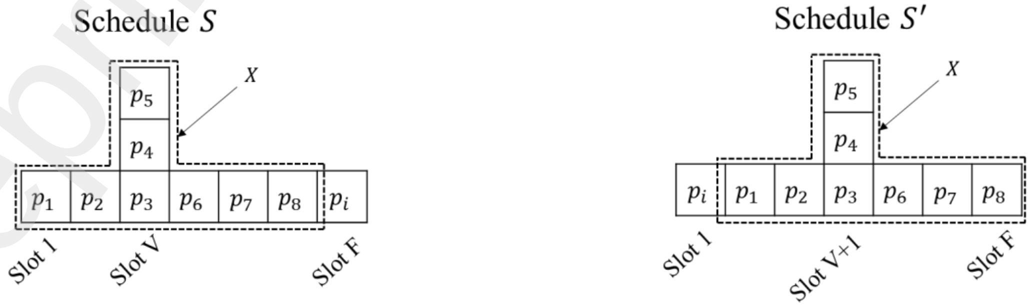


Figure A.1: Depiction of Schedule S and S'

Assume that $p_i > \frac{(c_I+c_O)(1-P_0)}{(c_I+c_O)(1-P_0)+c_W\left(P_1+E[B|B\geq 2](1-P_0-P_1)\right)}$; we will complete the proof by proving that $cost(S') < cost(S)$, thus S' is a better-quality schedule. Moving patient i does not affect the waiting time or the idle time in the slots spanned by X, thus, for both schedules, we only need to calculate the waiting time of i, the idle time experienced in the slot where i is scheduled, and the overtime.

For schedule S: If i shows up, the waiting time s/he suffers will be 1 with probability P_1 and $E[B|B \geq 2]$ with probability $1 - P_0 - P_1$. The idle time in slot F is one-time unit only if patient i does not show up (probability $1 - p_i$), and there is no backlog at the beginning of slot F. If i shows up, the overtime is $P_1 + E[B|B \geq 2](1 - P_0 - P_1)$, whereas the overtime if i does not show up is $E[B|B \geq 2] - 1$ (experienced only if the backlog at the beginning of slot F is of at least two patients, which has a probability of $(1 - P_0 - P_1)$).

For schedule S': Patient i's waiting time is zero. The idle time in slot 1 is one time unit if i does not show up (probability $1 - p_i$); the overtime is $P_1 + E[B|B \geq 2](1 - P_0 - P_1)$, regardless of whether or not patient i shows. A breakdown of the relevant costs is in the following table:

	$\boldsymbol{S}$	$\boldsymbol{S'}$
Patient i's waiting time	$p_i\left(P_1 + E[B\|B \geq 2](1 - P_0 - P_1)\right)$	0
Idle time in the slot where i is scheduled	$(1 - p_i)P_0$	$1 - p_i$
Overtime	$p_i\left(P_1 + E[B\|B \geq 2](1 - P_0 - P_1)\right) + (1 - p_i)\left((E[B\|B \geq 2] - 1)(1 - P_0 - P_1)\right)$	$P_1 + E[B\|B \geq 2](1 - P_0 - P_1)$

The cost of S' is less than the cost of S when $cost(S') - cost(S) < 0$, or ,

$$\begin{aligned} c_I(1 - p_i) + c_O\left(P_1 + E[B|B \geq 2](1 - P_0 - P_1)\right) - c_W p_i\left(P_1 + E[B|B \geq 2](1 - P_0 - P_1)\right) \\ - c_I P_0(1 - p_i) \\ - c_O\left(p_i\left(P_1 + E[B|B \geq 2](1 - P_0 - P_1)\right) \right. \\ \left. + (1 - p_i)\left((E[B|B \geq 2] - 1)(1 - P_0 - P_1)\right)\right) < 0 \end{aligned}$$

$$c_I(1 - p_i) + c_O(1 - P_0) - c_W p_i\left(P_1 + E[B|B \geq 2](1 - P_0 - P_1)\right) - c_I P_0(1 - p_i) - c_O p_i(1 - P_0) < 0$$

Solving for p_i we obtain

$$p_i > \frac{(c_I + c_O)(1 - P_0)}{(c_I + c_O)(1 - P_0) + c_W\left(P_1 + E[B|B \geq 2](1 - P_0 - P_1)\right)} \;\blacksquare$$

PROPOSITION 3: If i and j are two patients scheduled, respectively, in slots t and u ($u > t$) of a horizontal segment, then j's CWT is shorter than or equal to i'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 ■

PROPOSITION 4: Let i be a patient with show probability p_i scheduled in an OB-slot and j the patient scheduled in the following Z-slot. Let s be the expected number of shows among all patients in an OB-slot except for patient i, conditional to observing at least one such show. If $p_i \leq 1-\frac{s}{2}$, then patient i has a longer CWT than patient j. Specifically, for double and triple booking:

- *(double booking): If a patient i, scheduled in an OB-slot with only one other patient, has a show probability p_i < 0.5, then s/he has a longer CWT than the patient scheduled in Z.*
- *(triple booking): If a patient i, scheduled in an OB-slot with two other patients whose show probabilities are p_1 and p_2, has a show probability $p_i \leq \frac{p_1+p_2-2p_1p_2}{2p_1+2p_2-2p_1p_2}$, then s/he has a longer CWT than the patient scheduled in Z.*

PROOF: Let slot t be an OB-slot with m patients plus patient i. Let P_0 be the probability that none among the m patients scheduled in t shows up. Let $E[shows]$ be the expected number of shows among those m patients, and s be the expected number of shows among those m 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 $t+1$, a Z-slot. To complete the proof, we 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 m patients[1].

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

If none among the m 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 patient j will wait for s time units; if i does not show, then patient j will wait for $s-1$ time units:

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

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

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

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

So, $cwt_i \geq cwt_j$ if and only if $p_i \leq 1-\frac{s}{2}$

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 m patients:

[1] In general, if m patients with show probabilities $p_1, p_2, \dots, p_m$ are scheduled in the same slot, the expected number of patients seen before patient i is $\frac{\sum_{j=1,\dots,i-1,i+1,\dots,m} p_j}{2}$. That occurs because patients scheduled in the same slot are seen in random order; thus, each patient in that slot has a 50% chance of being seen before i.

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

If none among the m 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:

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

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

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

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

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

Double booking: Assume patient i is double booked in an OB-slot together with another patient, and j is the patient in the following slot, a Z-slot. We know that patient i waits longer than patient j if $p_i \leq 1 - \frac{s}{2}$. In that case, $s = 1$, because that is the expected number of shows among a group composed of one patient (the patient double booked with patient i) conditional to at least one patient showing. Thus, patient i waits longer than patient j if $p_i \leq 0.5$.

Triple booking: Assume patient i is triple booked in an OB-slot; let p_1, p_2, and p_i be the show probabilities of the patients in the slot; and j the patient in the following slot, a Z-slot. We know that patient i waits longer than patient j if $p_i \leq 1 - (1 - P_0)\left(\frac{s}{2}\right)$. Note that $(1 - P_0)s = E[shows]$, where P_0 is the probability that, without considering patient i, no patient shows up in the overbooked slot. So, substituting for s, $p_i \leq 1 - (1 - P_0)\left(\frac{s}{2}\right)$ is equivalent to

$$p_i - 1 + \frac{E[shows]}{2(1 - P_0)} \leq 0$$

Note that $E[shows] = p_1 + p_2$ and that $P_0 = (1 - p_1)(1 - p_2)$. Substituting for P_0 in the inequality above:

$$p_i \leq \frac{p_1 + p_2 - 2p_1p_2}{2p_1 + 2p_2 - 2p_1p_2} \blacksquare$$

PROPOSITION 5:

(iii) If slots are permitted to contain at most four expected shows, there exists a schedule which minimizes UOF with no empty slots.

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

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

Lemma 2: 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 2: 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 A.2 depicts the initial schedule, S1, and the schedule after patient i is moved, S2.

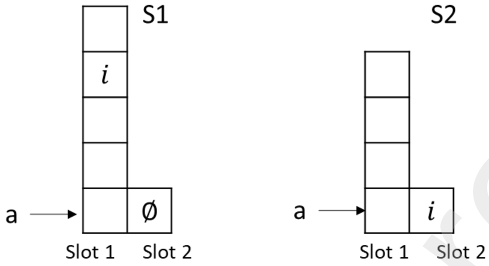


Figure A.2: Depiction of situations S1 and S2

By moving i to slot 2, the expected number of patients overflowing to the following slots will not change, so that move will not affect the expected waiting time of the following patients, or the overtime. Because neither the number of expected shows nor the overtime change, the overall idle time also does not change. Also, with that move, the waiting times of groups G_2, G_3, ... scheduled in the first slot may decrease, and will not increase. We will now show that this move also decreases the waiting time of i's group, G_1, under reasonable assumptions. From the main manuscript, the expected waiting time of group G_1 is calculated as:

$$E[W_1] = \frac{\sum_{i \in G_1} E[WT_i]}{E[\#shows\ in\ G_1]}$$

Because the move does not affect the denominator, we only need to check that the numerator decreases. Let a be the expected number of patients that overflow to slot 1 from the preceding slot (slot 0) in either schedule. Let n_1 be the expected number of G_1-patients who show in the first slot of S2 (i.e., i is excluded). Let W_2 be the sum of expected waiting times of all G_1-patients scheduled in slot 1 of S2, conditional to $a = 0$. Thus, $W_2 + an_1$ is the sum of the expected waiting times of all G_1-patients in the first slot of S2. Let n_2 be the expected number of patients belonging to all other groups showing in the first slot of S2. Let b be the expected number of patients that overflow from slot 1 to slot 2 in S2, and assume that patient i shows with probability p_i.

In S1, if i does not show up, then the expected sum of waiting times suffered by G_1-patients is equal to $W_2 + an_1$; if i shows up, the waiting time incurred by the G_1-patients other than i (whose expected number is n_1) increases by $\frac{1}{2}$ per showing patient, because i has a 50% chance of being seen before each of them, and i also waits $a + \frac{n_2+n_1}{2}$. Thus, before performing the move, the sum of G_1's waiting times is:

$$V_1^{S1} = (1 - p_i)(W_2 + an_1) + p_i \left((W_2 + an_1) + \frac{n_1}{2} + a + \frac{n_1 + n_2}{2} \right)$$

After moving i, the waiting time incurred by G_1-patients is $W_2 + an_1$ plus the waiting time suffered by i, if s/he shows up:

$$V_1^{S2} = W_2 + an_1 + p_i b$$

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

$$V_1^{S2} \leq V_1^{S1}$$

$$W_2 + an_1 + p_i b \leq (1 - p_i)(W_2 + an_1) + p_i\left((W_2 + an_1) + \frac{n_1}{2} + a + \frac{n_1 + n_2}{2}\right)$$

$$p_i b \leq p_i\left(\frac{n_1}{2} + a + \frac{n_1 + n_2}{2}\right)$$

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

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

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

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

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

$$n_2 \leq 2$$

Thus, if the expected number of non-G_1-patients showing in slot 1 is less than two, then moving i to slot 2 will not increase the objective function value. (end of proof of Lemma 2)

PROOF of (i): the following proof is similar to one of the proofs of Zacharias and Pinedo (2014) for TOF. Suppose that the schedule has empty slots. Let t be the first empty slot. Because there are more patients N than slots F, at least another slot has at least two patients.

Case 1: Suppose that the schedule prior to t has a slot with more than one patient. Let t_0 be the last slot before t with more than one patient. This implies that slots $t_0 + 1, \dots, t - 1$ have at most one patient assigned to them. If the patients scheduled in t_0 all belong to the same group, the proof is trivial. Let us assume that at least two groups are present in t_0. Because there are at most four shows and at least two groups in slot t_0, it is always possible to find a group G such that the expected shows among the non-G-patients is at most two. Consider the following move: find a group G in slot t_0 such that the expected number of shows belonging to the other groups in that slot are at most two. Then, take a patient i scheduled in slot t_0 belonging to group G, and move that patient to slot $t_0 + 1$. Next move all patients that were scheduled in slots $t_0 + 1, \dots, t - 1$, and reassign them to slots $t_0 + 2, \dots, t$ respectively, in the same order, one after the other. The expected number of patients at the end of slot t remains the same as before. This implies that the waiting time cost associated with slots $t + 1, \dots, F$ as well as the overtime cost, remain the same as before. Because neither the number of expected shows nor the overtime change, the overall idle time also does not change. The waiting time of the patients scheduled in slots t_0 and $t_0 + 1$ decreases due to Lemma 2. 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 total expected cost less than or equal to the schedule prior to the move.

Case 2: Suppose that the schedule prior to t does not have a slot with more than one customer. Thus, there is no patient overflowing from slot $t - 1$ to slot t. Consider the following altered schedule: move all patients 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 stays the same, but the overtime may decrease. After this move, there is at least one empty slot (slot F), and there 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 patient, perform the same move described in this case (case 2), obtaining a new schedule of at least the same quality as the previous one; else, perform the move described in case 1 to obtain a new schedule with no empty slot of at least the same quality as the original one.

PROOF of (*ii*): First, consider two patients i and j, both belonging to the same group, and scheduled in slots t and u respectively of the same horizontal segment (i.e., $(u > t)$ and 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 i and j. Suppose that i and j's show probabilities are p_i and p_j, respectively, and that $p_j \leq p_i$. We will show that swapping i and j will decrease the expected cost. Note that swapping i and j will not affect the waiting time of the patients scheduled after slot u, or the overtime of the schedule. Because neither the number of expected shows nor the overtime change, the idle time also does not change. Also, swapping i and j will decrease the waiting time experienced by all patients scheduled between i and j, because $p_j \leq p_i$, thereby decreasing the expected waiting times of other patient groups. To complete the proof, we must show that the sum of the waiting times experienced by i and j decreases. Let b_t be the number of expected patients overflowing to slot t, and b_u the number of expected patients overflowing to slot u assuming that i shows up, under the current schedule. Note that b_u depends on the show probabilities of the patients scheduled between slot t and u. The sum of the waiting times experienced by i and j before the move is:

$$W^{before} = p_i p_j (b_t + b_u) + p_i\big(1 - p_j\big) b_t + p_j (1 - p_i) \max(0, b_u - 1)$$

The waiting time after the move is:

$$W^{after} = p_i p_j (b_t + b_u) + p_j (1 - p_i) b_t + p_i\big(1 - p_j\big) \max(0, b_u - 1)$$

We now show that,

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

$$\begin{aligned} p_i p_j (b_t + b_u) + p_j (1 - p_i) b_t + p_i\big(1 - p_j\big) \max(0, b_u - 1) \\ \leq p_i p_j (b_t + b_u) + p_i\big(1 - p_j\big) b_t + p_j (1 - p_i) \max(0, b_u - 1) \end{aligned}$$

$$p_j (1 - p_i) b_t + p_i\big(1 - p_j\big) \max(0, b_u - 1) \leq p_i\big(1 - p_j\big) b_t + p_j (1 - p_i) \max(0, b_u - 1)$$

$$p_j b_t + p_i \max(0, b_u - 1) \leq p_i b_t + p_j \max(0, b_u - 1)$$

$$p_j (b_t - \max(0, b_u - 1)) \leq p_i (b_t - \max(0, b_u - 1))$$

which is true because $p_j \leq p_i$.

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

$$W^{before} = p_i p_j \left(b_t + \frac{n_t}{2} + b_u\right) + p_i(1 - p_j)\left(b_t + \frac{n_t}{2}\right) + p_j(1 - p_i)\max(0, b_u - 1)$$

The waiting time after the move is:

$$W^{after} = p_i p_j \left(b_t + \frac{n_t}{2} + b_u\right) + p_j(1 - p_i)\left(b_t + \frac{n_t}{2}\right) + p_i(1 - p_j)\max(0, b_u - 1)$$

We now show that,

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

$$p_i p_j \left(b_t + \frac{n_t}{2} + b_u\right) + p_j(1 - p_i)\left(b_t + \frac{n_t}{2}\right) + p_i(1 - p_j)\max(0, b_u - 1)$$

$$\leq p_i p_j \left(b_t + \frac{n_t}{2} + b_u\right) + p_i(1 - p_j)\left(b_t + \frac{n_t}{2}\right) + p_j(1 - p_i)\max(0, b_u - 1)$$

$$p_j\left(b_t + \frac{n_t}{2}\right) + p_i \max(0, b_u - 1) \leq p_i\left(b_t + \frac{n_t}{2}\right) + p_j \max(0, b_u - 1)$$

$$p_j\left(b_t + \frac{n_t}{2} - \max(0, b_u - 1)\right) \leq p_i\left(b_t + \frac{n_t}{2} - \max(0, b_u - 1)\right)$$

which is true because $p_j \leq p_i$. ■

References

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