Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Mitigating Delays and Unfairness in Appointment Systems

Jin Qi

informs
http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 2, February 2017, pp. 566–583
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Mitigating Delays and Unfairness in Appointment Systems

Jin Qi[a]

[a] Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Contact: jinqi@ust.hk (JQ)

Received: July 23, 2013
Accepted: August 31, 2015
Published Online in Articles in Advance:
March 25, 2016

https://doi.org/10.1287/mnsc.2015.2353

Abstract. We consider an appointment system where heterogeneous participants are sequenced and scheduled for service. Because service times are uncertain, the aims are to mitigate the unpleasantness experienced by the participants in the system when their waiting times or delays exceed acceptable thresholds and to address fairness in the balancing of service levels among participants. To evaluate uncertain delays, we propose the Delay Unpleasantness Measure, which takes into account the frequency and intensity of delays above a threshold, and introduce the concept of lexicographic min-max fairness to design appointment systems from the perspective of the worst-off participants. We focus our study in the healthcare industry in balancing physicians' overtime and patients' waiting times in which patients are distinguished by their service time characterizations. The model can be adapted in the robust setting when the underlying probability distribution is not fully available. To capture the correlation between uncertain service times, we suggest using the mean absolute deviations as the descriptive statistics in the distributional uncertainty set to preserve the linearity of the model. The optimal sequencing and scheduling decisions can be derived by solving a sequence of mixed-integer programming problems, and we report the insights from our computational studies.

History: Accepted by Yinyu Ye, optimization.

Funding: This research work was supported in part by NUS Global Asia Institute, Data Analytics Program and the Hong Kong RGC Project [No. T32-102/14N].

Keywords: appointment scheduling • robust optimization • lexicographic min-max fairness

1. Introduction

In any service system, because of the uncertainty in service times, waiting times or delays experienced by the participants are inevitable. However, the long waiting time that can occur for a scheduled appointment is an annoyance and often an indicator of poor quality of service. We focus our study in the healthcare industry, where the participants are patients and the physician. Decisions associated with appointment systems include the sequencing of patients and the scheduling of their appointment times, where these patients are distinguished by their service time characteristics. The goal of this paper is to design an appointment system that mitigates the unpleasantness experienced by the patients while waiting for treatment and by the physician in having to work overtime. The model is applicable in outpatient clinics to design consultation slots and in operating theatres to deliver an efficient and smooth schedule.

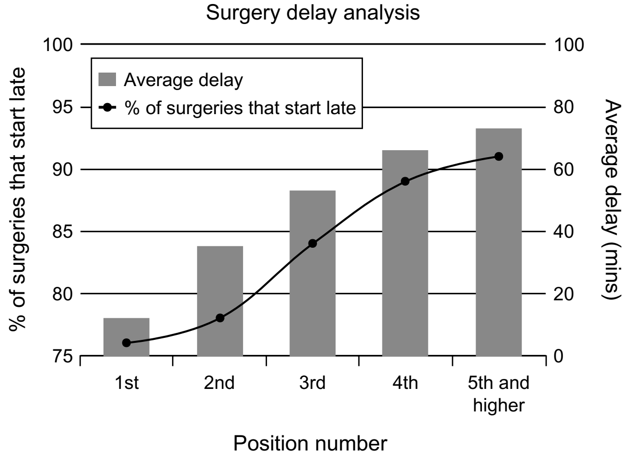
The study of appointment systems stems from the pioneering work of Bailey (1952). Before that, service providers typically allocated each patient a slot with the same fixed time length. Bailey (1952) designs an appointment scheduling rule that assigns two patients to the first slot, followed by other patients' arrivals evenly spaced. This minor change effectively reduces the physician's idle time by overcoming the problem of patient no-shows or lateness without compromising the patients' waiting time. Since then, many researchers have started to explore optimal appointment system settings under various conditions. For comprehensive literature reviews, we refer readers to Çayırlı and Veral (2003) and Gupta and Denton (2008), who highlight the current status and challenges in resolving appointment problems.

In general, patient access decisions involve two-stage planning (Patrick and Aubin 2013, Demeulemeester et al. 2013). The first stage is advance scheduling, which decides how many patients to assign within a fixed session, whereas the second stage, appointment scheduling, allocates time slot for each patient. Mak et al. (2015) has clearly explained the two-stage practice in operating theatres. In this paper, our appointment scheduling refers to the second stage, where the information about patients who need an appointment is known, and all the decisions must be made prior to the commencement of the session. Although the appointment for outpatient services is generally made in a dynamic fashion, this model serves as a reference table with designed time slots for different types of patients, which helps to create ideal appointment templates (Froehle and Magazine 2013). Now, we begin by discussing several concerns related to appointment system design problems.

The first concern regards characterizing the patients' experience of waiting, which is an integral aspect of service quality in a hospital environment. One commonly used service quality measure for describing this preference on uncertain waiting time is the expectation. However, the expected waiting time criterion may not adequately distinguish patients' attitudes toward uncertain delays because it corresponds to the average delay experienced by the patient over a potentially infinite number of visits under the same identical conditions. From the patients' perspectives, the unpleasantness of the waiting process may not increase proportionally with the length of the waiting time (Camacho et al. 2006), and a certain waiting time is considered acceptable among patients (Cartwright and Windsor 1992, McCarthy et al. 2000). In the survey conducted by Hill and Joonas (2005), 86% of respondents consider 30 minutes or fewer as an acceptable threshold. Huang (1994) empirically shows that for patients arriving at the appointment time, they appear reasonably satisfied if they wait no more than an average of 37 minutes, and their patience may steeply decline when the service delay exceeds this threshold. From the service providers' perspectives, their key performance indicator lies in the percentage of patients seen within a certain time threshold, instead of the total expected waiting time. For example, patients in the United States and United Kingdom can expect to be seen within 30 minutes of their given appointment time (National Health Service, Herzlinger 2006). The Ministry of Health Malaysia has also proposed one of the key performance indicators. In the dental clinics, for those patients that do not arrive late, the percentage of patients that can be seen within 30 minutes of the appointment time should be no less than 50% (Toh and Sern 2011). In an operating theatre of a local hospital in Singapore, the staff tracks the percentage of patients whose surgery is delayed more than 30 minutes to achieve the target at less than 30%. Following these empirical results, we can use a reasonable unpleasantness tolerance threshold to describe the patient's satisfaction with waiting processes and take the frequency of delays above this threshold as an alternative service quality measure. Nonetheless, several nonnegligible drawbacks have hampered the wide application of this measure. One disadvantage lies in the intensity of delay, for its inability to distinguish waiting processes with the same frequency of surpassing the patient's tolerance threshold but with different lengths of delay. Moreover, the computational intractability of this probability measure also arises as a result of a lack of convexity. Thus, we need to establish a new service quality measure that can to some extent reflect people's real attitudes toward the delay process and, in particular, can account for both the frequency and intensity of the delay over the threshold.

The optimization criterion for an appointment system involves multiple participants including patients and physicians. Currently, the majority of studies take a weighted average of the combinations among patients' waiting time and the physician's idle time and overtime as an optimization criterion, and they exploit different methods to solve. Three main streams are based on queueing theory (Wang 1993, 1999; Green and Savin 2008; Hassin and Mendel 2008), stochastic programming (Robinson and Chen 2003, Denton and Gupta 2003), and robust optimization (Mittal and Stiller 2011; Kong et al. 2013; Mak et al. 2014, 2015) frameworks. However, the second concern is that because the decisions are very sensitive to the prescribed weight for each participant, how to provide an accurate interpretation and estimation of these weights is a crucial issue (Mondschein and Weintraub 2003). Additionally, minimizing a weighted combination of expectations of patients' waiting time the physician's idle time and overtime fails to accommodate the fairness issue highlighted by Çayırlı and Veral (2003). In layman's terms, fairness regards distinguishing a strategy of keeping, say, 20 patients each waiting for 2 minutes and its counterpart of keeping only one of them waiting for 40 minutes (Klassen and Rohleder 1996). Çayırlı and Veral (2003) highlighted the phenomenon that the current appointment system is unfair to the patient at the last position, as waiting time tends to progressively build up. Figure 1 shows the current practice in an operating room in a local hospital in Singapore. Both the average delay and percentage of surgeries that start late increase with the position at which the surgery is scheduled. The notion of "fairness" has been widely studied in economics literatures (Young 1995, Sen and Foster 1997) and industrial applications, especially resource allocation problems (see Bertsimas et al. 2011 and references therein), but few papers focus on

Figure 1. Analysis of the Delay in Surgeries in a Local Hospital in Singapore



the appointment scheduling problems. For this reason, an effective appointment system should be able to guarantee the uniformity of qualities across multiple participants.

The third concern is the difficulties of eliciting the exact probability distributions for patients' service times. To cope with it, robust optimization techniques have also been applied to appointment problems (Mittal and Stiller 2011; Kong et al. 2013; Mak et al. 2015, 2014). In these papers, the optimization criteria are based on a weighted sum of the patients' expected waiting times and the physician's idle time and overtime. Mittal and Stiller (2011) consider the scheduling problem where only the bound support of service time is provided. They present a global balancing heuristic and prove that it delivers an optimal schedule under certain mild conditions. Kong et al. (2013) assume that the lower bound, mean, and covariance of the service time are known and formulate a robust min-max problem, which can be approximately solved by semidefinite programming. Mak et al. (2015) study the scheduling problem by assuming the knowledge of marginal moments of uncertain service time and derive a computationally tractable conic programming formulation.

In general, service times among different types of patients, such as new and returning patients, or different types of surgeries are not necessarily homogeneous. By exploiting the information of patients' classification, appointment systems would inevitably rely on the sequence of these various types of patients (Denton et al. 2007). Following the literature, we refer to the determination of appointment time and the sequence as the scheduling and sequencing decisions, respectively. However, because of the difficulty of the problems, few papers investigate the sequencing and scheduling decisions simultaneously. Weiss (1990) was the first to examine this problem and provides analytical results for a two-patient case with general service time distributions; however, the conclusions cannot be extended to the case of multiple patients. Wang (1999) addresses the problem with a specific assumption that patients' service times follow exponential distributions with different rates and infers that the optimal service sequence is in the descending order of service rates. Vanden Bosch and Dietz (2000, 2001) classify the patients into different categories according to their service times, which follow different phase-type distributions and approximately solve the problem with heuristics. Denton et al. (2007) jointly formulate the sequencing and scheduling problem into a two-stage stochastic programming model and suggest an interchange heuristic with the sampling average approximation technique. Gupta (2007) uses stochastic programming to model this problem and mainly highlights the complication of the problem by investigating the case with two patients only. Mak et al. (2014) suggest using inventory approximation to solve the sequencing problem. They also explore the sequence when the mean and variance of the service time are known and show that the smallest variance first rule under certain conditions is optimal (Mak et al. 2015).

Our paper is most relevant to Kong et al. (2013) and Mak et al. (2015) in terms of problem setting and robustness considerations. However, the main difference is that we take the fairness issue as the principle aim and develop solution techniques for both the sequencing and scheduling problems by using different descriptive statistics. We summarize our contributions as follows.

- We propose a new service quality measure named the Delay Unpleasantness Measure (DUM) to demonstrate the dependency of an individual participant's attitude toward the delay process based on his or her corresponding acceptable levels. Unlike the probability measure, the DUM collectively accounts for the frequency and intensity of delay over a threshold.
- We introduce the concept of lexicographic min-max fairness to tackle the fairness concern arising in the appointment system design. We lexicographically minimize the largest DUM, the second-largest DUM, and so on, to guarantee the best performance that every participant could achieve simultaneously. As far as we are aware, this is the first analytical paper to consider the subject of fairness as the principle aim in appointment system design.
- When the sequence is predetermined, we formulate a scheduling model that can be adapted in the robust setting. Unlike the common assumption in the literature, which imposes the independence among patients' uncertain service times, we capture the correlation by proposing the mean absolute deviation of summation over service times as the information that can mimic the physician's endogeneity behavior and also help retain the linearity of the model. The optimal scheduling decisions are derived by solving a small sequence of linear optimization problems.
- We extend the scheduling model to incorporate sequencing decisions when patients are heterogeneous and propose the solution methodology.

The rest of this paper is organized as follows. In Section 2, we show how a participant's behavior in the delay process can be characterized by the DUM. In Section 3, we introduce the concept of lexicographic min-max fairness and propose the solution procedure under the DUM. In Section 4, we propose a scheduling model for appointment systems by assuming that the patients' sequence is fixed and demonstrate how the resulting model can be solved. In Section 5, we extend our model to solve both sequencing and scheduling problems. In Section 6, we perform several computational studies with encouraging results on the DUM regarding the fairness concern. Finally, in Section 7, we provide conclusions and managerial insights.

1.1. Notations

We denote scalars by plain characters and use boldface lowercase characters to represent vectors, for example, $\mathbf{x} = (x_1, x_2, \ldots, x_N)'$. Given a vector $\mathbf{x}$, we write $(y_n, \mathbf{x}_{-n})$ for the vector with only the nth component change; i.e., vector $(y_n, \mathbf{x}_{-n}) = (x_1, \ldots, x_{n-1}, y_n, x_{n+1}, \ldots, x_N)$. We use boldface uppercase characters to represent matrix, for example, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)'$. Set $[i; j]$ represents positive running indices from i to j. We represent uncertain quantities by the tilde ("$\tilde{\ }$") sign and model random variable $\tilde{x}$ by a state-space Ω and a σ-algebra of events $\mathcal{F}$ in Ω. Furthermore, we use $\mathcal{L}$ as the space of real-valued random variables. Comparison $\tilde{x} \geq \tilde{y}$ represents statewise dominance; i.e., $x(\omega) \geq y(\omega)$ for all $\omega \in \Omega$. In addition, $\tilde{\mathbf{x}} \geq \tilde{\mathbf{y}}$ represents $\tilde{x}_n \geq \tilde{y}_n$ for all $n \in [1; N]$. To incorporate ambiguity, instead of specifying the true distribution $\mathbb{P}$ on $(\Omega, \mathcal{F})$, we assume that the true distribution belongs to a certain distributional uncertainty set $\mathbb{F}$, i.e., $\mathbb{P} \in \mathbb{F}$. Note that with this assumption, full knowledge of the underlying distribution is a special case, where $\mathbb{F} = \{\mathbb{P}\}$. We also denote by $\mathrm{E}_{\mathbb{P}}(\tilde{x})$ the expectation of $\tilde{x}$ under probability distribution $\mathbb{P}$.

2. Delay Unpleasantness Measure

In this section, we introduce a new service quality measure to evaluate the uncertain waiting time (service delay) of patients and the overtime (off-work delay) of physicians. We start by defining the DUM for the individual participant (patient or physician) in the appointment system.

We assume that each participant has his or her own tolerance threshold for waiting time. The tolerance threshold is an exogenous factor and varies according to participants' demographic profiles. For example, the tolerable threshold of elderly patients is much longer (Moschis et al. 2003). In addition, because the service times for returning patients are relatively short, in certain cases, they may deserve a shortened waiting process, which corresponds to a small threshold. We can use survey or interview methods to study patients' thresholds based on different medical departments, ages, frequency of visits, etc. (see, for instance, McCarthy et al. 2000, Hill and Joonas 2005). The physician's tolerance threshold can also be specified according to the physician's preference or can be transformed equivalently from the budget for the overtime. We let τ represent the participant's tolerance threshold and $\tilde{w}$ represent an uncertain delay and define the DUM as follows.

Definition 1 (Delay Unpleasantness Measure). Given a tolerance threshold $\tau \in \Re_+$ and an uncertain delay $\tilde{w} \in \mathcal{L}$, if we only know that its true distribution $\mathbb{P}$ belongs to a distributional uncertainty set $\mathbb{F}$, the DUM is a function $\rho_\tau\colon \mathcal{L} \to [0, 1]$ defined as

$$\rho_\tau(\tilde{w}) \triangleq \inf\{\alpha \geq 0 \mid \varphi_\alpha(\tilde{w}) \leq \tau\}$$

(or 1 if no such α exists), where

$$\varphi_\alpha(\tilde{w}) = \min_{\nu \in \Re}\left(\nu + \frac{1}{\alpha}\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w} - \nu)^+)\right), \quad \alpha \in (0, 1].$$

Function $\varphi_\alpha(\tilde{w})$ is the worst-case conditional value-at-risk (CVaR) (Zhu and Fukushima 2009, Natarajan et al. 2010). CVaR (Rockafellar and Uryasev 2000) is a coherent risk measure with the specific focus on the tail distribution and has become a major reference in the area of financial mathematics. In hospital settings, Dehlendorff et al. (2010) use simulation models and suggest that CVaR is a reliable measure for the waiting time. In Definition 1, $\varphi_\alpha(\tilde{w})$ denotes the worst-case expected waiting time in the conditional distribution of its upper α tail (Rockafellar 2007). Therefore, roughly speaking, DUM represents the smallest upper 100α percentile, such that the worst-case average of α longest delay is no more than patient's tolerable threshold. Hence, it is desirable to have an uncertain delay with smallest DUM value, since it implies that even for those long realizations, their expectation can still be no more than the tolerance threshold. This definition is similar to Shortfall aspiration level criterion in Chen and Sim (2009) and Definition 5 in Brown and Sim (2009) in the monetary context. Several properties of DUM are listed in Proposition 1.

Proposition 1. *The DUM, ρ_τ, has the following properties.*

(a) *Monotonicity: if $\tilde{w}_1 \geq \tilde{w}_2$, then $\rho_\tau(\tilde{w}_1) \geq \rho_\tau(\tilde{w}_2)$.*

(b) *Threshold satisficing: if $\tilde{w} \leq \tau$, then $\rho_\tau(\tilde{w}) = 0$.*

(c) *Tardiness intolerance: if $\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}(\tilde{w}) > \tau$, then $\rho_\tau(\tilde{w}) = 1$.*

(d) *Upper bound of tardiness probability: $\rho_\tau(\tilde{w}) \geq \mathbb{P}(\tilde{w} > \tau)$ for all $\mathbb{P} \in \mathbb{F}$.*

(e) *If $\mathbb{P}(\tilde{w} < \tau) > 0$ for all $\mathbb{P} \in \mathbb{F}$, then*

$$\rho_\tau(\tilde{w}) = \inf_{a>0} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((a(\tilde{w} - \tau) + 1)^+).$$

Property (a) captures the participant's essential preference for a shorter delay; i.e., if the waiting time $\tilde{w}_1$ is statewise greater than its counterpart $\tilde{w}_2$, then the former is not more preferred under the DUM. Property (b) indicates the participant's desire to be served within the threshold, and any uncertain delay that always meets the threshold is most preferred. By contrast, Property (c) indicates that the intolerance to any delay always exceeds the threshold in expectation. Property (d) suggests a close relationship between the DUM and frequency of delay over the threshold. We can guarantee that the frequency of delay over the threshold is less than the corresponding DUM. It is also shown to be the best convex conservative approximation of the frequency of delay over the threshold (Nemirovski and Shapiro 2006). Property (e) demonstrates that the DUM can be written as a form of an optimized expected utility, where the utility function is convex.

Next, we provide a simple illustration of the DUM. There are two options A and B on delay that are characterized as follows:

$$\tilde{w}_A = \begin{cases} 10 \text{ minutes}, & \text{with probability } 0.89, \\ 35 \text{ minutes}, & \text{with probability } 0.11, \end{cases}$$

$$\tilde{w}_B = \begin{cases} 10 \text{ minutes}, & \text{with probability } 0.9, \\ 60 \text{ minutes}, & \text{with probability } 0.1. \end{cases}$$

When the tolerance threshold $\tau = 30$ minutes, the outcome of minimizing the frequency of delay over a threshold suggests option B is better than A because $\mathbb{P}(\tilde{w}_B > 30) = 0.1 < \mathbb{P}(\tilde{w}_A > 30) = 0.11$, which indicates that this quality measure only focuses on the violation probability without taking the delay level into consideration. Instead, the use of the DUM can avoid these disadvantages, with its outcome suggesting that option A is more preferable than B because $\rho_{30}(\tilde{w}_A) = 11/80 \le 1/4 = \rho_{30}(\tilde{w}_B)$.

3. Lexicographic Min-Max Fairness

The service quality of an appointment system depends on all of the participants' experiences with delays. Hence, for a service system with N participants, we hope to find a feasible solution to deliver short delays for everyone. Hence, with the monotonicity property of the DUM, we formulate a multiple criteria optimization problem, in which all of the participants' DUMs are minimized; i.e.,

$$\min_{\tilde{\mathbf{w}} \in \mathcal{W}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}),$$

where $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) = (\rho_{\tau_1}(\tilde{w}_1), \ldots, \rho_{\tau_N}(\tilde{w}_N))$, and $\mathcal{W}$ represents the space of feasible delays experienced by the participants. However, because reducing one participant's delay may increase another participant's delay, we cannot minimize each participant's DUM simultaneously. To mitigate the unfairness and avoid discriminating a subset of participants in terms of their waiting experiences in the appointment system, we adopt the lexicographic min-max fairness solution approach (Young 1995). We first provide the definition of the lexicographic order.

Definition 2 (Lexicographic Order). Given $\boldsymbol{\tau} \in \Re_+^N$, we let $\rho_i(\tilde{\mathbf{w}})$ and $\rho_i(\tilde{\mathbf{v}})$, $\tilde{\mathbf{w}}, \tilde{\mathbf{v}} \in \mathcal{W}$, be the ith largest elements of $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}})$ and $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$, respectively. We say $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}})$ is lexicographically equivalent to $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$, denoted by

$$\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) =_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$$

if and only if $\rho_h(\tilde{\mathbf{w}}) = \rho_h(\tilde{\mathbf{v}})$ for all $h \in [1;N]$. Moreover, $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}})$ is lexicographically less than $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$, denoted by

$$\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) \prec_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$$

if and only if there exists $i^* \in [1;N]$ such that $\rho_h(\tilde{\mathbf{w}}) = \rho_h(\tilde{\mathbf{v}})$ for $h \in [1; i^* - 1]$ and $\rho_{i^*}(\tilde{\mathbf{w}}) < \rho_{i^*}(\tilde{\mathbf{v}})$. Similarly, we denote

$$\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) \preceq_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$$

if either $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) =_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$ or $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) \prec_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$.

The lexicographic order shows that the participant with the largest value of DUM has the highest priority in preference ranking among solutions in $\mathcal{W}$. Subsequently, if these values among different solutions are the same, then the next largest value is used in deciding preferences. We explore several characteristics of lexicographic ordering of participants' DUM and link them to issues of fairness in an appointment system.

Proposition 2. *The following properties hold for* $\tilde{\mathbf{w}}$, $\tilde{\mathbf{v}} \in \mathcal{W}$.

(a) *Monotonicity: if* $\tilde{\mathbf{w}} \le \tilde{\mathbf{v}}$, *then*

$$\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) \preceq_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}}).$$

(b) *Threshold satisficing: let* $\mathcal{S} \subset [1;N]$ *and* $\bar{\mathcal{S}}$ *be the complement set. Suppose* $\tilde{v}_j = \tilde{w}_j$ *for all* $j \in \mathcal{S}$ *and* $\tilde{v}_j \le \tilde{w}_j \le \tau_j$ *for all* $j \in \bar{\mathcal{S}}$; *then*

$$\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) =_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}}).$$

(c) *Discrimination resistance: let*

$$\mathcal{S}_1 = \{i \in [1;N] \mid \rho_{\tau_i}(\tilde{w}_i) = 1\} \quad \text{and}$$
$$\mathcal{S}_2 = \{i \in [1;N] \mid \rho_{\tau_i}(\tilde{v}_i) = 1\}.$$

Suppose $|\mathcal{S}_1| < |\mathcal{S}_2|$; *then*

$$\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) \prec_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}}).$$

Property (a) ensures consistency so that the reduction in delays for all of the participants is favorably valued. Property (b) ensures that for the participants whose delays are always within their thresholds, any improvement of their delays does not contribute to the lexicographic order. A participant is discriminated if the appointment system cannot guarantee his or her average waiting time below the threshold, which corresponds to the DUM taking a value of 1. Hence, Property (c) induces preferences for solutions that have fewer participants being discriminated. This property is in accord with the hospital's key performance indicator, to keep the number of patients who experience the worst waiting process as small as possible. In the context of the earlier example provided by Klassen and Rohleder (1996), if each patient's tolerable threshold is 30 minutes, the number of patients whose DUMs are equal to 1 is 1 for the strategy that keeps only one patient waiting for 40 minutes, whereas that for the other strategy is 0.

Because the lexicographic order is complete, we can rank solutions and replace the multiple criteria optimization by the following lexicographic minimization problem:

$$\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}^*) = \operatorname{lex\,min}_{\tilde{\mathbf{w}}\in\mathscr{W}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}),$$

where the optimal solution $\tilde{\mathbf{w}}^* \in \mathscr{W}$ satisfies

$$\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}^*) \preceq_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}}) \quad \forall\, \tilde{\mathbf{v}} \in \mathscr{W}.$$

Although this may not be a standard mathematical programming problem, we can obtain the optimal solution by solving a sequence of optimization problems (see Isermann 1982, Ogryczak et al. 2005) as follows:

Algorithm (Lexicographic minimization procedure)

1. Set $h := 1, \mathscr{G}_0 := [1;N]$,

$$\alpha_1 := \min_{\tilde{\mathbf{w}}\in\mathscr{W}} \max_{n\in\mathscr{G}_0} \rho_{\tau_n}(\tilde{w}_n) = \max_{n\in\mathscr{G}_0} \rho_{\tau_n}(\tilde{w}^*_{1n}),$$
$$\mathscr{I}_1 := \min\{j \in \mathscr{G}_0 \colon \rho_{\tau_j}(\tilde{w}^*_{1j}) = \alpha_1\},$$

where $\tilde{w}^*_{1n}$, $n \in \mathscr{G}_0$ is the corresponding optimal solution to derive α_1.

2. Set $\mathscr{G}_h := \mathscr{G}_{h-1} \backslash \mathscr{I}_h$. If $\mathscr{G}_h = \varnothing$, the algorithm terminates and outputs the solution. Otherwise, set $h := h+1$,

$$\begin{aligned}\alpha_h &:= \min_{\tilde{\mathbf{w}}\in\mathscr{W}} \left\{ \max_{n\in\mathscr{G}_{h-1}} \rho_{\tau_n}(\tilde{w}_n) \,\middle|\, \max_{n\in\mathscr{I}_i} \rho_{\tau_n}(\tilde{w}_n) \le \alpha_i, i \in [1;h-1] \right\} \\ &= \max_{n\in\mathscr{G}_{h-1}} \rho_{\tau_n}(\tilde{w}^*_{hn}), \\ \mathscr{I}_h &:= \min\{j \in \mathscr{G}_{h-1} \colon \rho_{\tau_j}(\tilde{w}^*_{hj}) = \alpha_h\},\end{aligned}$$

where $\tilde{w}^*_{hn}$, $n \in \mathscr{G}_{h-1}$ is the corresponding optimal solution to derive α_h.

3. Go to step 2.

In this algorithm, we minimize the maximum DUM among a set of participants and elicit the participants that attain the largest value. Hence, the optimal solution, $\tilde{\mathbf{w}}^* \in \mathscr{W}$, satisfies

$$\rho_{\tau_n}(\tilde{w}^*_n) = \alpha_i, \quad n \in \mathscr{I}_i,$$

for all $i \in [1;h]$.

Observe that the problem to derive α_h is the same as

$$\begin{aligned}\alpha_h = \min\ & \alpha \\ \text{s.t.}\ \ & \rho_{\tau_n}(\tilde{w}_n) \le \alpha, \quad n \in \mathscr{G}_{h-1}, \\ & \rho_{\tau_n}(\tilde{w}_n) \le \alpha_i, \quad n \in \mathscr{I}_i,\ i \in [1;h-1], \\ & \tilde{\mathbf{w}} \in \mathscr{W}.\end{aligned}$$

According to the definition of $\rho_{\tau_n}(\tilde{w}_n)$, we can equivalently solve

$$\begin{aligned}\inf\ & \alpha \\ \text{s.t.}\ & \nu_n + \frac{1}{\alpha} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+) \le \tau_n, \quad n \in \mathscr{G}_{h-1}, \\ & \nu_n + \frac{1}{\alpha_i} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+) \le \tau_n, \\ & \qquad\qquad n \in \mathscr{I}_i,\ i \in [1;h-1], \\ & \alpha \in (0,1], \\ & \tilde{\mathbf{w}} \in \mathscr{W}.\end{aligned} \tag{1}$$

Although the problem is nonlinear in α, we observe that both function $(1/\alpha)\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+)$ and the objective function are monotonic in α, and hence, we can use a binary search procedure to find the optimal solution in which α is minimized. Because the lexicographic minimization procedure requires solving a sequence of similar problems, from now on, we focus on solving problem (1) as a representative instance.

Remark 1. In the lexicographic minimization algorithm, when $h = 1, \mathscr{G}_0 = [1;N]$, the problem is equivalent to minimizing the maximum DUM over all of the participants. Intuitively, the minimization of the maximum DUM can already address the fairness issue; however, it may yield multiple optimal solutions because we only pay attention to the worst-off participants. Among these optimal solutions, some may not be Pareto optimal. In other words, for an optimal solution $\tilde{\mathbf{w}}$, there may exist other feasible solution $\tilde{\mathbf{v}}$ such that $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}}) \preceq_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}})$ and $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}}) \neq \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}})$. Hence, we propose using the lexicographic minimization algorithm to resolve this issue.

Remark 2. When the probability distribution is known, i.e., $\mathbb{F} = \{\mathbb{P}\}$, we define the model that minimizes the weighted summation of expected delays as follows:

$$\sum_{n\in[1;N]} \gamma_n \mathrm{E}_{\mathbb{P}}(\tilde{w}^{**}_n) = \min_{\tilde{\mathbf{w}}\in\mathscr{W}} \sum_{n\in[1;N]} \gamma_n \mathrm{E}_{\mathbb{P}}(\tilde{w}_n),$$

where γ_n, $n \in [1;N]$ is the weight assigned to the nth participant, and $\tilde{w}^{**}_n$, $n \in [1;N]$ is the corresponding optimal solution. Let $\tau^*_n = \mathrm{E}_{\mathbb{P}}(\tilde{w}^{**}_n)$, and it is easy to observe that $\tilde{w}^{**}_n$, $n \in [1;N]$ is also an optimal solution of the following problem, which minimizes the largest DUM over all of the participants:

$$\begin{aligned}\inf\ & \alpha \\ \text{s.t.}\ & \nu_n + \frac{1}{\alpha} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+) \le \tau^*_n, \quad n \in [1;N], \\ & \alpha \in (0,1], \\ & \tilde{\mathbf{w}} \in \mathscr{W}.\end{aligned}$$

4. The Scheduling Model

In this section, we consider a scheduling problem with one physician serving N patients and where the service sequence is predetermined. Hence, we only need to determine the appointment time for each patient. In particular, we provide the formulation of feasible set $\mathscr{W}$ and discuss the solution procedures when different information on the service time is given.

We consider a static case in which all of the scheduling decisions have to be made before the commencement of the session. We assume that patients arrive on time. According to the data collection of Harper and Gamlin (2003) and Zhu et al. (2011), the majority of patients arrive earlier than expected. This assumption avoids the complexity of modeling as a result of potential change in sequence. The physician is also assumed to start his or her session promptly. Hence, the first patient experiences no delay. In this model, patients may be heterogeneous and are characterized by their service time distributions and tolerance thresholds. We next specify the parameters for the model:

- N, the total number of patients to be scheduled;
- L, session length predetermined for the consultation of N patients;
- τ_n, the tolerance threshold of delay for the patient at nth position, $n \in [1;N]$;
- τ_{N+1}, physician's tolerance of his or her overtime;
- $\tilde{s}_n$, the nth patient's service time, $n \in [1;N]$;
- $\tilde{w}_n$, the nth patient's waiting time, $n \in [1;N]$;
- $\tilde{w}_{N+1}$, physician's overtime; and
- x_n, decision variable, appointment time for the nth patient. For notational simplicity, we let $x_1 = 0$, $x_{N+1} = L$, and its vector notation $\mathbf{x} = (x_1, \ldots, x_N, x_{N+1})'$.

Then, the feasible set of delay $\mathscr{W}$ can be formulated as

$$\mathscr{W} = \left\{ \tilde{\mathbf{w}} \;\middle|\; \begin{array}{l} \tilde{w}_1 = 0, \\ \tilde{w}_n = \max\{x_{n-1} + \tilde{w}_{n-1} + \tilde{s}_{n-1} - x_n, 0\}, \\ \qquad\qquad\qquad\qquad n \in [2;N+1], \\ \mathbf{x} \in \mathscr{X}, \end{array} \right\}$$

where set $\mathscr{X}$ is defined as

$$\mathscr{X} = \left\{ \mathbf{x} \;\middle|\; \begin{array}{l} x_1 = 0, \\ x_{n-1} \le x_n, \quad n \in [2;N+1], \\ x_{N+1} = L. \end{array} \right\}$$

The first two constraints in set $\mathscr{W}$ recursively calculate the delays experienced by the patients and the physician, whereas set $\mathscr{X}$ ensures sequencing compliance. Denton and Gupta (2003) further simplify the delay formulation as

$$\tilde{w}_n = \max\left\{0, \tilde{t}_{n-1}, \ldots, \sum_{k=1}^{n-1} \tilde{t}_k\right\}, \quad n \in [2;N+1], \tag{2}$$

where $\tilde{t}_n = \tilde{s}_n - (x_{n+1} - x_n)$, $n \in [1;N]$ represents the difference between the real service time and scheduled interval for the nth patient.

To derive the optimal scheduling decisions, we are now ready to formulate problem (1) as

$$\begin{aligned} &\inf \ \alpha \\ &\text{s.t.} \ \ \nu_n + \frac{1}{\alpha} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+) \le \tau_n, \quad n \in \mathscr{G}_{h-1}, \\ &\qquad \nu_n + \frac{1}{\alpha_i} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+) \le \tau_n, \\ &\qquad\qquad\qquad\qquad\qquad n \in \mathscr{I}_i,\ i \in [1;h-1], \\ &\qquad \tilde{w}_n = \max\left\{0, \tilde{t}_{n-1}, \ldots, \sum_{k=1}^{n-1} \tilde{t}_k\right\}, \quad n \in [2;N+1], \\ &\qquad \tilde{t}_n = \tilde{s}_n - (x_{n+1} - x_n), \quad n \in [1;N], \\ &\qquad \alpha \in (0,1], \\ &\qquad \mathbf{x} \in \mathscr{X}. \end{aligned} \tag{3}$$

Because the first patient experiences no delay, we have $\rho_{\tau_1}(\tilde{w}_1) = 0$ for any nonnegative threshold τ_1. Therefore, we can define $\mathscr{G}_0 = [2;N+1]$. Given α, α_i, $i \in [1;h-1]$, the most difficult part is in the simplification of function $\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+)$, which is complicated by the recursive property of uncertain waiting times. Based on Equations (2), we observe that

$$\begin{aligned} &\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+) \\ &\quad = \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left(\left(\max\left\{0, \tilde{t}_{n-1}, \ldots, \sum_{k=1}^{n-1} \tilde{t}_k\right\} - \nu_n\right)^+\right) \\ &\quad = \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\Bigg(\max\Bigg\{0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - \nu_n, \ldots, \\ &\qquad\qquad\qquad \sum_{k=1}^{n-1} (\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n\Bigg\}\Bigg). \end{aligned}$$

The calculation of this function inevitably depends on the information we possess about the uncertain service time $\tilde{s}_n$, $n \in [1;N]$. Next, we classify the information set we can have on $\tilde{s}_n$ and provide different reformulation and solution techniques.

Remark 3. In this model, we do not account for the physician's total idle time, which can be formulated as $L + \tilde{w}_{N+1} - \sum_{n=1}^{N} \tilde{s}_n$. By proposing a physician's overtime threshold, we indeed indirectly control the total idle time.

4.1. Stochastic Optimization Approach

When the uncertain service time follows a known discrete distribution, i.e., $\mathbb{F} = \{\mathbb{P}\}$, in which there are M sets of service times, $\{s_1^m, \ldots, s_N^m\}$, each occurring with probability p_m, $m \in [1;M]$, we have

$$\begin{aligned} &\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\Bigg(\max\Bigg\{0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - \nu_n, \ldots, \\ &\qquad\qquad \sum_{k=1}^{n-1} (\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n\Bigg\}\Bigg) \end{aligned}$$

$$= \sum_{m=1}^{M} p_m \max\Big\{0, -v_n, s_{n-1}^m - (x_n - x_{n-1}) - v_n, \ldots, \sum_{k=1}^{n-1}(s_k^m - (x_{k+1} - x_k)) - v_n\Big\}.$$

Hence, by adding decision variables q_{mn}, $m \in [1;M]$, $n \in [2;N+1]$, problem (3) is equivalent to

$$\begin{aligned}
&\inf\ \alpha\\
&\text{s.t.}\ \ v_n + \frac{1}{\alpha}\sum_{m=1}^{M} p_m q_{mn} \le \tau_n, \quad n \in \mathcal{G}_{h-1},\\
&\qquad v_n + \frac{1}{\alpha_i}\sum_{m=1}^{M} p_m q_{mn} \le \tau_n, \quad n \in \mathcal{I}_i,\ i \in [1;h-1],\\
&\qquad q_{mn} + v_n \ge 0, \quad n \in [2;N+1],\ m \in [1;M],\\
&\qquad q_{mn} + v_n + x_n - x_l \ge \sum_{k=l}^{n-1} s_k^m,\\
&\qquad\qquad l \in [1;n-1],\ n \in [2;N+1],\ m \in [1;M],\\
&\qquad q_{mn} \ge 0, \quad n \in [2;N+1],\ m \in [1;M],\\
&\qquad \alpha \in (0,1],\\
&\qquad \mathbf{x} \in \mathcal{X}.
\end{aligned}$$

Whenever α is fixed, the feasible set is a polyhedron comprising $O(MN)$ decision variables and $O(MN^2)$ constraints. In practice, this approach is amiable to empirical distributions where M is relatively small.

4.2. Distributionally Robust Optimization Approach

We also propose a distributionally robust optimization approach with the goal of preserving linearity of the model. We assume that the uncertainty set of service time distributions is characterized by their bounded supports $\mathbb{P}(\tilde{s}_k \in [\underline{s}_k, \bar{s}_k]) = 1$, means $\mathrm{E}_{\mathbb{P}}(\tilde{s}_k) = \mu_k$, $\mu_k \in (\underline{s}_k, \bar{s}_k)$, and bounds of mean absolute deviation $\mathrm{E}_{\mathbb{P}}(|\tilde{s}_k - \mu_k|) \le \sigma_k$, $\sigma_k > 0$ for all $k \in [1;N]$. Similar to the standard deviation, the mean absolute deviation is used to capture the dispersion of an uncertain parameter about its mean. In particular, for the normal distribution, the ratio between mean absolute deviation and standard deviation is equal to $\sqrt{2/\pi}$.

Intuitively, the worst-case probability distributions may result in highly correlated service times, which may not be realistic and may lead to conservative solutions (Mak et al. 2015). To impose correlation, the conventional approach is to specify covariance within the distributional uncertainty set, i.e., the descriptive statistics of $\mathrm{E}_{\mathbb{P}}((\tilde{s}_r - \mu_r)(\tilde{s}_k - \mu_k))$ for all $r, k \in [1;N]$, $r \le k$. However, this will necessarily lead to nonlinear optimization models, which are harder to solve (Kong et al. 2013). To avoid such nonlinearity, we propose a different approach to capturing correlation. We note that the delay of a participant may be influenced by the aggregation of uncertain service times of the earlier participants. When the physician sees a long delay, he or she may speed up to catch up with the schedule. Hence, in our distributional uncertainty set, we use the descriptive statistics of $\mathrm{E}_{\mathbb{P}}(|\sum_{m=r}^{k}((\tilde{s}_m - \mu_m)/\sigma_m)|)$ for all $r, k \in [1;N]$ and $r \le k$. Observe that

$$\mathrm{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^{k}\frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right) \le \sum_{m=r}^{k}\mathrm{E}_{\mathbb{P}}\left(\left|\frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right) \le k - r + 1,$$

in which the first equality is achieved under perfect correlation. As a proxy for modeling correlation, we impose the constraints

$$\mathrm{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^{k}\frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right) \le \epsilon_{rk}, \quad r, k \in [1;N],\ r \le k,$$

where $\epsilon_{rk} \in (0, k-r+1]$. Intuitively, we define $\epsilon_{kk} = 1$, which is equivalent to the information $\mathrm{E}_{\mathbb{P}}(|\tilde{s}_k - \mu_k|) \le \sigma_k$. These constraints set the bound for the dispersion of the total uncertain service times for $k - r + 1$ consecutive patients, which enable us to mimic the physician's endogeneity behavior and specify a less conservative uncertainty set while keeping the model linear. Now, the distributional uncertainty set can be written as

$$\mathbb{F} = \left\{\mathbb{P} \,\middle|\, \mathrm{E}_{\mathbb{P}}(\tilde{s}_k) = \mu_k,\ \mathbb{P}(\tilde{s}_k \in [\underline{s}_k, \bar{s}_k]) = 1,\ \mathrm{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^{k}\frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right) \le \epsilon_{rk},\ r, k \in [1;N],\ r \le k\right\}.$$

For convenience, we standardize the service time as $\tilde{z}_k = (\tilde{s}_k - \mu_k)/\sigma_k$ and define $\mathbb{F}_z$ as

$$\mathbb{F}_z = \left\{\mathbb{P} \,\middle|\, \mathrm{E}_{\mathbb{P}}(\tilde{z}_k) = 0,\ \mathbb{P}(\tilde{z}_k \in [\underline{z}_k, \bar{z}_k]) = 1,\ \mathrm{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^{k}\tilde{z}_m\right|\right) \le \epsilon_{rk},\ r, k \in [1;N],\ r \le k\right\},$$

where $\underline{z}_k = (\underline{s}_k - \mu_k)/\sigma_k$ and $\bar{z}_k = (\bar{s}_k - \mu_k)/\sigma_k$ for all $k \in [1;N]$. We calculate function $\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - v_n)^+)$ in the following proposition.

Proposition 3. *Given a scheduling decision* $\mathbf{x} \in \mathcal{X}$ *and* $n \in [2;N+1]$,

$$Z_P = \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left(\max\Big\{0, -v_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - v_n, \ldots, \sum_{k=1}^{n-1}(\tilde{s}_k - (x_{k+1} - x_k)) - v_n\Big\}\right)$$

can be formulated as the following linear optimization problem:

$$Z_D = \inf\left\{f_0 + \sum_{k=1}^{n-1}\sum_{r=1}^{k}\epsilon_{rk} g_{rk}\right\}$$

$$\text{s.t.}\quad f_0 + \sum_{k=1}^{n-1}(\underline{z}_k u_k^0 - \bar{z}_k v_k^0) \ge 0,$$

$$
\begin{aligned}
&f_0+v_n+\sum_{k=1}^{n-1}(\underline{z}_k u_k^n-\underline{z}_k v_k^n)\ge 0,\\
&f_0+v_n+x_n-x_l+\sum_{k=1}^{n-1}(\underline{z}_k u_k^l-\bar{z}_k v_k^l)\ge\sum_{k=l}^{n-1}\mu_k,\\
&\qquad\qquad l\in[1;n-1],\\
&u_k^l-v_k^l+\sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^l-c_{rm}^l)-f_k=0,\\
&\qquad\qquad k\in[1;n-1],\ l=0,n,\\
&u_k^l-v_k^l+\sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^l-c_{rm}^l)-f_k=0,\\
&\qquad\qquad k,l\in[1;n-1],\ k\le l-1,\\
&u_k^l-v_k^l+\sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^l-c_{rm}^l)-f_k=-\sigma_k,\\
&\qquad\qquad k,l\in[1;n-1],\ l\le k,\\
&b_{rk}^l+c_{rk}^l-g_{rk}=0,\\
&\qquad\qquad r,k\in[1;n-1],\ r\le k,\ l\in[0;n],\\
&u_k^l,v_k^l,b_{rk}^l,c_{rk}^l,g_{rk}\ge 0,\\
&\qquad\qquad r,k\in[1;n-1],\ r\le k,\ l\in[0;n].
\end{aligned}\tag{4}
$$

Correspondingly, problem (3) is equivalent to the following optimization problem:

$$
\begin{aligned}
\inf\ &\alpha\\
\text{s.t.}\ &v_n+\frac{1}{\alpha}\left(f_0^n+\sum_{k=1}^{n-1}\sum_{r=1}^{k}\epsilon_{rk}g_{rk}^n\right)\le\tau_n,\quad n\in\mathcal{G}_{h-1},\\
&v_n+\frac{1}{\alpha_i}\left(f_0^n+\sum_{k=1}^{n-1}\sum_{r=1}^{k}\epsilon_{rk}g_{rk}^n\right)\le\tau_n,\\
&\qquad\qquad n\in\mathcal{I}_i,\ i\in[1;h-1],\\
&f_0^n+\sum_{k=1}^{n-1}(\underline{z}_k u_k^{0n}-\bar{z}_k v_k^{0n})\ge 0,\quad n\in[2;N+1],\\
&f_0^n+v_n+\sum_{k=1}^{n-1}(\underline{z}_k u_k^{nn}-\bar{z}_k v_k^{nn})\ge 0,\\
&\qquad\qquad n\in[2;N+1],\\
&f_0^n+v_n+x_n-x_l+\sum_{k=1}^{n-1}(\underline{z}_k u_k^{ln}-\bar{z}_k v_k^{ln})\ge\sum_{k=l}^{n-1}\mu_k,\\
&\qquad\qquad l\in[1;n-1],\ n\in[2;N+1],\\
&u_k^{ln}-v_k^{ln}+\sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln}-c_{rm}^{ln})-f_k^n=0,\\
&\qquad\qquad k\in[1;n-1],\ l=0,n,\ n\in[2;N+1],\\
&u_k^{ln}-v_k^{ln}+\sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln}-c_{rm}^{ln})-f_k^n=0,\\
&\qquad\qquad k,l\in[1;n-1],\ k\le l-1,\ n\in[2;N+1],\\
&u_k^{ln}-v_k^{ln}+\sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln}-c_{rm}^{ln})-f_k^n=-\sigma_k,\\
&\qquad\qquad k,l\in[1;n-1],\ k\ge l,\ n\in[2;N+1],\\
&b_{rk}^{ln}+c_{rk}^{ln}-g_{rk}^n=0,\quad r,k\in[1;n-1],\\
&\qquad\qquad r\le k,\ l\in[0;n],\ n\in[2;N+1],\\
&u_k^{ln},v_k^{ln},b_{rk}^{ln},c_{rk}^{ln},g_{rk}^n\ge 0,\quad r,k\in[1;n-1],\\
&\qquad\qquad r\le k,\ l\in[0;n],\ n\in[2;N+1],\\
&\alpha\in(0,1],\\
&\mathbf{x}\in\mathscr{X}.
\end{aligned}\tag{5}
$$

The decision variables $f_0^n, f_k^n, u_k^{ln}, v_k^{ln}, b_{rk}^{ln}, c_{rk}^{ln}, g_{rk}^n$ for all $r,k\in[1;n-1], r\le k, l\in[0;n], n\in[2;N+1]$ are auxiliary variables for the optimization model. Problem (5) is quite complicated at first glance; however, for any $\alpha\in(0,1]$, we observe that the problem reduces to a linear feasibility problem including $O(N^4)$ continuous decision variables and $O(N^4)$ constraints. We assume that the onus is on the decision maker to select the threshold values so that problem (5) is feasible at $\alpha=1$. Otherwise, the delay thresholds are not attainable in expectation and should be adjusted accordingly to reflect what is realistically achievable in practice. When α decreases to 0, $\varphi_\alpha(\tilde{w})$ approaches the upper limit of $\tilde{w}$.

It is worthwhile to point out that the above scheduling formulation preserves linearity and greatly reduces the computational complexity. Each approach only requires solving a sequence of linear optimization problems.

5. The Sequencing and Scheduling Model

We now generalize the scheduling model to incorporate the realistic situation with sequencing decisions for heterogeneous patients. We classify all of the patients into J types. The same type of patient has identically distributed service time and the same tolerance threshold. Next, we clarify some extra parameters and decision variables:

- N_j, the number of jth type of patients, where $\sum_{j=1}^J N_j=N$;
- β_j, the tolerance threshold for the jth type of patients, $j\in[1;J]$;
- $\tilde{s}_{nj}$, the jth type of patient's service time if he or she is scheduled at the nth position, and $\tilde{s}_{nj}$, $n\in[1;N]$ are identically and independently distributed;
- y_{nj}, the binary decision variable—if the jth type of patient is scheduled at the nth position, then $y_{nj}=1$; otherwise, $y_{nj}=0$. Its matrix form is $\mathbf{Y}=(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N)'\in\{0,1\}^{N\times J}$.

Correspondingly, with the sequencing decisions, the patient at position $n\in[1;N]$ has an uncertain service time $\sum_{j=1}^J\tilde{s}_{nj}y_{nj}$ and tolerance threshold $\tau_n=\sum_{j=1}^J\beta_j y_{nj}$. We can formulate problem (1) with both sequencing and scheduling decisions as follows:

$$
\begin{aligned}
\inf\ &\alpha\\
\text{s.t.}\ &v_n+\frac{1}{\alpha}\sup_{\mathbb{P}\in\mathbb{F}}\mathrm{E}_{\mathbb{P}}((\tilde{w}_n-v_n)^+)\le\tau_n,\quad n\in\mathcal{G}_{h-1},\\
&v_n+\frac{1}{\alpha_i}\sup_{\mathbb{P}\in\mathbb{F}}\mathrm{E}_{\mathbb{P}}((\tilde{w}_n-v_n)^+)\le\tau_n,\\
&\qquad\qquad n\in\mathcal{I}_i, i\in[1;h-1],
\end{aligned}
$$

$$\begin{aligned}
&\tilde{w}_n = \max\left\{0, \tilde{t}_{n-1}, \dots, \sum_{k=1}^{n-1} \tilde{t}_k\right\}, \quad n \in [2; N+1],\\
&\tilde{t}_n = \sum_{j=1}^{J} \tilde{s}_{nj} y_{nj} - (x_{n+1} - x_n), \quad n \in [1; N],\\
&\alpha \in (0, 1],\\
&(\boldsymbol{\tau}, \mathbf{x}, \mathbf{Y}) \in \mathscr{Y}, \qquad (6)
\end{aligned}$$

in which

$$\mathscr{Y} = \left\{ (\boldsymbol{\tau}, \mathbf{x}, \mathbf{Y}) \;\middle|\; \begin{aligned}
&\sum_{j=1}^{J} \beta_j y_{nj} = \tau_n, \quad n \in [1; N],\\
&\sum_{n=1}^{N} y_{nj} = N_j, \quad j \in [1; J],\\
&\sum_{j=1}^{J} y_{nj} = 1, \quad n \in [1; N],\\
&y_{nj} \in \{0, 1\}, \quad n \in [1; N],\ j \in [1; J],\\
&\mathbf{x} \in \mathscr{X}.
\end{aligned} \right\}$$

Set $\mathscr{Y}$ guarantees that each patient is assigned to a position, and each position is allotted to only one patient.

To solve this problem, we can implement similar procedures described in Section 4. The difference lies in the calculation of function $\sup_{\mathbb{P} \in \mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}_n - \nu_n)^+)$, which is equivalent to

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathbb{F}} \mathrm{E}_{\mathbb{P}}\Bigg(\max\Bigg\{&0, -\nu_n, \sum_{j=1}^{J} \tilde{s}_{n-1, j} y_{n-1, j} - (x_n - x_{n-1}) - \nu_n,\\
&\dots, \sum_{k=1}^{n-1}\left(\sum_{j=1}^{J} \tilde{s}_{kj} y_{kj} - (x_{k+1} - x_k)\right) - \nu_n\Bigg\}\Bigg). \qquad (7)
\end{aligned}$$

For known discrete distribution cases in which there are M sets of service time, $(s_{nj}^m)_{n \in [1; N], j \in [1; J]}$ with probability $p_m, m \in [1; M]$, we add decision variables $q_{mn}, n \in [2; N+1], m \in [1; M]$. Problem (6) is equivalent to

$$\begin{aligned}
\inf\ & \alpha\\
\text{s.t.}\ & \nu_n + \frac{1}{\alpha} \sum_{m=1}^{M} p_m q_{mn} \leq \tau_n, \quad n \in \mathscr{G}_{h-1},\\
& \nu_n + \frac{1}{\alpha_i} \sum_{m=1}^{M} p_m q_{mn} \leq \tau_n, \quad n \in \mathscr{I}_i, i \in [1; h-1],\\
& q_{mn} + \nu_n \geq 0, \quad n \in [2; N+1], m \in [1; M],\\
& q_{mn} + \nu_n + x_n - x_l - \sum_{k=l}^{n-1} \sum_{j=1}^{J} s_{kj}^m y_{kj} \geq 0,\\
& \qquad l \in [1; n-1],\ n \in [2; N+1],\ m \in [1; M],\\
& q_{mn} \geq 0, \quad n \in [2; N+1], m \in [1; M],\\
& \alpha \in (0, 1],\\
& (\boldsymbol{\tau}, \mathbf{x}, \mathbf{Y}) \in \mathscr{Y}.
\end{aligned}$$

Similarly, the binary search algorithm is used to find the optimal solution. For any fixed $\alpha \in (0, 1]$, the problem becomes a mixed-integer programming problem, including $N \times J$ binary decision variables, $O(MN)$ continuous decision variables, and $O(MN^2)$ constraints.

Similar to the scheduling case, we can specify the descriptive statistics for the service time $\tilde{s}_{nj}$ as $\mathrm{E}_{\mathbb{P}}(\tilde{s}_{nj}) = \mu_j$ and $\mathrm{E}_{\mathbb{P}}(|\tilde{s}_{nj} - \mu_j|) \leq \sigma_j$. However, we cannot designate the mean absolute deviation for the total standardized service time of k consecutive patients because the sequence information is missing. Therefore, to obtain an amicably tractable robust optimization model, we assume that the uncertain service times $\tilde{s}_{1j}, \dots, \tilde{s}_{Nj}$ are respectively affinely dependent on a set of factors, $\tilde{z}_1, \dots, \tilde{z}_N$ for all patient types $j \in [1; J]$, and the centrality and dispersion of $\tilde{s}_{nj}$ are characterized by the patient type; i.e.,

$$\tilde{s}_{nj} = \tilde{z}_n \sigma_j + \mu_j,$$

for all $n \in [1; N]$ and $j \in [1; J]$. Furthermore, the factors have the same support, and its distributional uncertainty set is given as follows:

$$\begin{aligned}
\mathbb{F}_z = \Bigg\{\mathbb{P} \;\Bigg|\; &\mathrm{E}_{\mathbb{P}}(\tilde{z}_k) = 0, \mathbb{P}(\tilde{z}_k \in [\underline{z}, \bar{z}]) = 1,\\
&\mathrm{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^{k} \tilde{z}_m\right|\right) \leq \epsilon_{rk},\ r, k \in [1; N],\ r \leq k\Bigg\}.
\end{aligned}$$

With this linear formulation, problem (7) is written as

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathbb{F}_z} \mathrm{E}_{\mathbb{P}}\Bigg(\max\Bigg\{&0, -\nu_n, \sum_{j=1}^{J} (\tilde{z}_{n-1} \sigma_j + \mu_j) y_{n-1, j}\\
&- (x_n - x_{n-1}) - \nu_n, \dots,\\
&\sum_{k=1}^{n-1}\left(\sum_{j=1}^{J} (\tilde{z}_k \sigma_j + \mu_j) y_{kj} - (x_{k+1} - x_k)\right) - \nu_n\Bigg\}\Bigg). \qquad (8)
\end{aligned}$$

Proposition 4. *For any fixed decisions* $(\boldsymbol{\tau}, \mathbf{x}, \mathbf{Y}) \in \mathscr{Y}$ *and* $n \in [2; N+1]$, *problem* (8) *can be formulated as the following linear optimization problem*:

$$\begin{aligned}
\min\ & f_0 + \sum_{k=1}^{n-1} \sum_{r=1}^{k} \epsilon_{rk} g_{rk}\\
\text{s.t.}\ & f_0 + \sum_{k=1}^{n-1} (\underline{z} u_k^0 - \bar{z} v_k^0) \geq 0,\\
& f_0 + \nu_n + \sum_{k=1}^{n-1} (\underline{z} u_k^n - \bar{z} v_k^n) \geq 0,\\
& f_0 + \nu_n + x_n - x_l - \sum_{k=l}^{n-1} \sum_{j=1}^{J} \mu_j y_{kj}\\
& \quad + \sum_{k=1}^{n-1} (\underline{z} u_k^l - \bar{z} v_k^l) \geq 0, \quad l \in [1; n-1],\\
& u_k^l - v_k^l + \sum_{m=k}^{n-1} \sum_{r=1}^{k} (b_{rm}^l - c_{rm}^l) - f_k = 0,\\
& \qquad k \in [1; n-1], l = 0, n,
\end{aligned}$$

$$
\begin{aligned}
& u_k^l - v_k^l + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^l - c_{rm}^l) - f_k = 0, \\
& \qquad\qquad k,l \in [1;n-1],\ k \le l-1, \\
& u_k^l - v_k^l + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^l - c_{rm}^l) - f_k + \sum_{j=1}^{J}\sigma_j y_{kj} = 0, \\
& \qquad\qquad k,l \in [1;n-1],\ k \ge l, \\
& b_{rk}^l + c_{rk}^l - g_{rk} = 0, \\
& \qquad\qquad r,k \in [1;n-1],\ r \le k,\ l \in [0;n], \\
& u_k^l, v_k^l, b_{rk}^l, c_{rk}^l, g_{rk} \ge 0, \\
& \qquad\qquad r,k \in [1;n-1],\ r \le k,\ l \in [0;n].
\end{aligned}
$$

Henceforth, problem (6) is equivalent to

$$
\begin{aligned}
\inf\ & \alpha \\
\text{s.t.}\ & \nu_n + \frac{1}{\alpha}\left(f_0^n + \sum_{k=1}^{n-1}\sum_{r=1}^{k}\epsilon_{rk} g_{rk}^n\right) \le \tau_n, \quad n \in \mathcal{G}_{h-1}, \\
& \nu_n + \frac{1}{\alpha_i}\left(f_0^n + \sum_{k=1}^{n-1}\sum_{r=1}^{k}\epsilon_{rk} g_{rk}^n\right) \le \tau_n, \\
& \qquad\qquad n \in \mathcal{I}_i,\ i \in [1;h-1], \\
& f_0^n + \sum_{k=1}^{n-1}(\underline{z} u_k^{0n} - \bar{z} v_k^{0n}) \ge 0, \quad n \in [2;N+1], \\
& f_0^n + \nu_n + \sum_{k=1}^{n-1}(\underline{z} u_k^{nn} - \bar{z} v_k^{nn}) \ge 0, \quad n \in [2;N+1], \\
& f_0^n + \nu_n + x_n - x_l - \sum_{k=l}^{n-1}\sum_{j=1}^{J}\mu_j y_{kj} + \sum_{k=1}^{n-1}(\underline{z} u_k^{ln} - \bar{z} v_k^{ln}) \ge 0, \\
& \qquad\qquad l \in [1;n-1],\ n \in [2;N+1], \\
& u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = 0, \\
& \qquad\qquad k \in [1;n-1],\ l = 0, n,\ n \in [2;N+1], \\
& u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = 0, \\
& \qquad\qquad k,l \in [1;n-1],\ k \le l-1,\ n \in [2;N+1], \\
& u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n + \sum_{j=1}^{J}\sigma_j y_{kj} = 0, \\
& \qquad\qquad k,l \in [1;n-1],\ k \ge l,\ n \in [2;N+1], \\
& b_{rk}^{ln} + c_{rk}^{ln} - g_{rk}^n = 0, \\
& \qquad r,k \in [1;n-1],\ r \le k,\ l \in [0;n],\ n \in [2;N+1], \\
& u_k^{ln}, v_k^{ln}, b_{rk}^{ln}, c_{rk}^{ln}, g_{rk}^n \ge 0, \\
& \qquad r,k \in [1;n-1],\ r \le k,\ l \in [0;n],\ n \in [2;N+1], \\
& \alpha \in (0,1], \\
& (\boldsymbol{\tau}, \mathbf{x}, \mathbf{Y}) \in \mathcal{Y}.
\end{aligned}
$$

Given $\alpha \in (0,1]$, the sequencing and scheduling problem reduces to checking the feasibility of a mixed-integer optimization problem with $N \times J$ binary decision variables, $O(N^4)$ continuous decision variables, and $O(N^4)$ constraints.

6. Computational Results

In this section, we carry out three computational studies. In the first study, we investigate the problem of scheduling homogeneous patients and compare performances under two strategies: (1) lexicographic minimization of DUM (L-DUM) and (2) minimization of total expected delays (TED). The second study explores the performance of the appointment scheduling model under distributional ambiguity. In the third study, we solve a sequencing and scheduling problem for two patient types and provide some practical insights. We also provide a simple test with the real data from an operating theatre in a local hospital. The program is coded in Python and run on an Intel Core i7 PC with a 3.40 GHz central processing unit using CPLEX 12 as the integer linear program solver.

6.1. Comparison of Performance Measures

We compare the performance of two appointment system models, the L-DUM model and the TED model, which is formulated as follows:

$$
\begin{aligned}
\min\ & \left\{ \sum_{n=2}^{N} \gamma_p \mathrm{E}_{\mathbb{P}}(\tilde{w}_n) + \gamma_d \mathrm{E}_{\mathbb{P}}(\tilde{w}_{N+1}) \right\} \\
\text{s.t.}\ & \tilde{w}_n = \max\left\{0, \tilde{t}_{n-1}, \ldots, \sum_{k=1}^{n-1}\tilde{t}_k\right\}, \quad n \in [2;N+1], \\
& \tilde{t}_n = \tilde{s}_n - (x_{n+1} - x_n), \quad n \in [1;N], \\
& \mathbf{x} \in \mathcal{X}.
\end{aligned}
$$

6.1.1. Uniform Distribution. We consider the case of scheduling seven homogeneous patients who have the same delay thresholds, and service times are independently uniformly distributed. Stochastic optimization is used for both the L-DUM model and the TED model with the sample size set at 2,000. We first study in detail an instance and analyze the performance by varying the patients' and physician's delay thresholds. Then, we randomly generate 100 instances and investigate their average performances. For each instance, we (a) generate the corresponding parameters for uniform distributions, (b) sample the possible realizations of service time combination, (c) solve the scheduling problem by the L-DUM and the TED strategies, and (d) compute each participant's corresponding delay to summarize the performances.

In the first instance, the uniform distribution is specified with bound support $[0,2]$. Total session length is 7. We obtain the scheduling decisions with different weights and thresholds in Table 1.

In general, the optimal scheduling decisions under the L-DUM model exhibit a much shorter scheduled service time for the first patient compared with the TED model, especially when the physician's threshold is relatively short with respect to the patient's threshold. This decision delivers managerial insights consistent with Bailey (1952) because a short physician's threshold indicates a reduction in the physician's idle time during the session. By scheduling the patients

Table 1. Patients' Optimal Appointment Times Under Two Scheduling Methods

	$\text{TED}(\gamma_p, \gamma_d)$[a]		$\text{L-DUM}(\tau_p, \tau_d)$[b]			
	$(1,1)$	$(1,2)$	$(1,1)$	$(1.5,1.5)$	$(2,2)$	$(1,2)$
Patient 1	0	0	0	0	0	0
Patient 2	1.03	0.96	0.22	0.03	0.03	0.64
Patient 3	2.32	2.16	1.31	1.22	0.90	1.98
Patient 4	3.61	3.36	2.41	2.33	2.24	3.16
Patient 5	4.89	4.56	3.55	3.57	3.40	4.42
Patient 6	6.14	5.75	4.72	4.68	4.64	5.67
Patient 7	7.00	6.81	5.83	5.80	5.49	6.78

[a] γ_p, patients' weights; γ_d, physician's weight.
[b] τ_p, patients' delay thresholds; τ_d, physician's delay threshold.

earlier, this policy allows us to mitigate delays for the patients in the last several positions by sacrificing a certain level of delays for the first several patients. When the physician's tolerance threshold is large, for example, in the case of L-DUM$(1,2)$, the optimal solution schedules the last patient a little later to avoid the risk of a long wait. However, it is hard to describe the general scheduling rules in the L-DUM model because the optimal decision greatly depends on the specified tolerance thresholds.

Table 2 summarizes the delay performance of the worst-off participants (including all of the patients and the physician). As the findings are similar, for convenience and clarity, we report the numerical performance for the case in which the patients' and physician's thresholds take the value of one. In terms of total expected delays, we observe that the TED method performs better than the L-DUM model. However, this performance comes at the price of sacrificing the service levels of some participants. From the fairness perspective, when we pay particular attention to the most discriminated participants, our model makes a significant improvement over the TED model. When every participant has equal weight, the maximal average delay reduces from 1.56 to 0.83, and the frequency of delay over the threshold improves from 73% to 41%. When the weight assigned to the physician increases, the delay performance of the worst-off participant improves. However, this performance does not always improve with the increase of the physician's weight and how to specify the suitable weight for different parameter settings is a difficult problem.

6.1.2. Empirical Data.

We also test our model using empirical consultation data collected from a specialist outpatient clinic in a local hospital in Singapore from March to May 2012. In practice, each patient is allocated 15 minutes, and approximately 20 patients are scheduled each day. We consider the historical data during March and April (802 samples) as the information to make scheduling decisions, whereas the data in May (435 samples) are used for performance testing. The statistics of the service time are summarized in Table 3. To make a fair comparison, we directly use the stochastic optimization approach described before with these samples without fitting any distributions or using any descriptive statistics.

Our appointment design problem is to schedule 10 patients within a session length of 150 minutes. The performance derived with similar procedures is listed

Table 2. Delay Performance Under Two Scheduling Methods (Uniform Distribution)

	Delay performance of the worst-off participants						
	Expected delay	Frequency of delay over the threshold[a] (%)	Standard deviation of delay	Expected delay over the threshold[b]	VaR @95%[c]	VaR @99%	Total expected delays
L-DUM$(1,1)$[d]	0.83	41	0.87	0.26	2.54	3.49	5.50
TED$(1,1)$[e]	1.56	73	0.83	0.68	2.99	3.70	3.46
TED$(1,2)$	1.34	63	0.84	0.52	2.83	3.63	3.59
L-DUM$(1.5,1.5)$	1.00	25	0.87	0.14	2.54	3.49	5.76
TED$(1,1)$	1.56	51	0.83	0.36	2.99	3.69	3.46
TED$(1,2)$	1.34	41	0.84	0.26	2.83	3.63	3.59
L-DUM$(2,2)$	1.02	13	0.86	0.07	2.53	3.49	6.01
TED$(1,1)$	1.56	28	0.83	0.17	2.99	3.69	3.46
TED$(1,2)$	1.34	20	0.84	0.12	2.83	3.63	3.59
L-DUM$(1,2)$	1.15	24	0.82	0.10	2.65	3.52	3.98
TED$(1,1)$	1.56	28	0.83	0.17	2.99	3.69	3.46
TED$(1,2)$	1.34	20	0.84	0.12	2.83	3.63	3.59

[a] Frequency of delay over the threshold: $\mathbb{P}(\tilde{w} > \tau)$.
[b] Expected delay over the threshold: $\mathrm{E}_{\mathbb{P}}((\tilde{w} - \tau)^+)$.
[c] $\text{VaR@95\%} = \inf\{\nu \in \Re \mid \mathbb{P}(\tilde{w} > \nu) \leq 1 - 95\%\}$.
[d] $\text{L-DUM}(\tau_p, \tau_d)$.
[e] $\text{TED}(\gamma_p, \gamma_d)$.

Table 3. Statistics of Service Time from Empirical Data

Statistics	Average	Maximum	Minimum	Mean absolute deviation	Standard deviation
Minutes	13.84	107	1	6.52	9.41

in Table 4, which also manifests our conclusions for uniform distributions.

In general, compared with the TED method, the L-DUM model tries to schedule the first several patients as early as possible and provides a less discriminating solution that mitigates the unpleasantness of delays in the appointment system.

6.2. Distributional Ambiguity

In this experiment, we study the performance of the L-DUM model under distributional ambiguity. We schedule seven homogeneous patients and compare the delay performance of the worst-off participants (including all of the patients and the physician) under three scheduling decisions. The first two are derived by the stochastic optimization approach and the distributionally robust optimization approach in the L-DUM model. Sampling average approximation is used for the stochastic optimization approach, and the information on bound support, mean, and mean absolute deviation for the robust optimization approach is calculated accordingly. The third scheduling decision is derived from the TED method by using a sampling average approximation scheme. Total session length is 7. We consider two types of distributions: uniform distribution U$(0,2)$ and beta distribution $3 \times \text{Beta}(2,4)$. The sample size remains the same. The delay performance is listed in Table 5.

We observe that the performances of the stochastic optimization approach and the robust optimization approach in the L-DUM model are very close and much better than that of the TED method. With the distributional uncertainty set we proposed, the L-DUM model provides a comparatively good performance that is immunized against distributional ambiguity. It is particularly noteworthy that the computation time for the distributional robust optimization approach is relatively short. To solve each minimization problem, the stochastic optimization approach requires 44 seconds, whereas the distributional robust optimization approach only requires 8 seconds.

6.3. Two Sequencing and Scheduling Examples

We also investigate the sequencing and scheduling problem with heterogeneous patients. By calculating the optimal solutions, we hope to deliver some useful insights for managers to make decisions in a unified manner. For simplicity, we only consider two patient types: new and returning patients. Their demographics are collected from real clinic data and shown in Table 6, and the information on the mean absolute deviation is given as, for $i<k, i,k \in [1;N]$,

$$\epsilon_{ik} = \begin{cases} 1.71 & i=k-1, \\ 2.20 & i=k-2, \\ 2.52 & i=k-3. \end{cases}$$

The sequencing and scheduling decisions are illustrated in Figure 2. For decades, researchers have debated whether to first schedule returning patients (smallest variance) or new ones (largest variance). Our computational study actually suggests that such a universal rule may not be optimal, and the decisions may differ when the participants' tolerable thresholds vary. For instance, as shown in the first graph of Figure 2, we generally observe that if the physician's tolerance threshold is low, his or her delay can better be mitigated under the L-DUM model if a new patient, who may have a longer and more uncertain service time, is scheduled early. However, if patients' waiting tolerance is low, for example, in a pediatrics clinic, the L-DUM method arranges for the new patient to arrive at the

Table 4. Delay Performance Under Two Scheduling Decisions (Empirical Data)

	Delay performance of the worst-off participants						
	Expected delay	Frequency of delay over the threshold (%)	Standard deviation of delay	Expected delay over the threshold	VaR @95%	VaR @99%	Total expected delays
L-DUM(15, 15)	13.37	37	17.33	4.61	47.57	91	94.88
TED(1, 1)	24.12	63	18.57	11.21	51	104	66.65
TED(1, 2)	18.63	40	18.42	8.16	51	99	66.89
L-DUM(25, 25)	14.45	16	17.29	2.95	47.59	91	98.60
TED(1, 1)	24.12	35	18.57	6.51	51	104	66.65
TED(1, 2)	18.63	23	18.41	4.60	51	99	66.89
L-DUM(35, 35)	15.07	9	17.25	1.81	46.93	91	107.09
TED(1, 1)	24.12	19	18.57	3.60	51	104	66.65
TED(1, 2)	18.63	16	18.42	2.58	51	99	66.89

Note. See Table 2 notes for definitions.

Table 5. Delay Performance Under Three Scheduling Decisions

		Delay performance of the worst-off participants				
Distribution	Approach	Expected delay	Frequency of delay over the threshold (%)	Standard deviation of delay	Expected delay over the threshold	Total expected delays
Uniform	L-DUMs(1.2, 1.2)[a]	0.90	35	0.86	0.21	5.62
	L-DUMr(1.2, 1.2)	1.00	40	0.89	0.24	6.15
	TED(1, 1)	1.56	64	0.82	0.52	3.46
	L-DUMs(1.4, 1.4)	0.99	29	0.87	0.16	5.73
	L-DUMr(1.4, 1.4)	1.02	31	0.89	0.19	6.26
	TED(1, 1)	1.56	55	0.83	0.41	3.46
	L-DUMs(1.6, 1.6)	1.00	21	0.86	0.12	5.78
	L-DUMr(1.6, 1.6)	1.12	28	0.91	0.17	6.53
	TED(1, 1)	1.56	46	0.83	0.30	3.46
Beta	L-DUMs(1.2, 1.2)	0.89	28	0.84	0.18	5.18
	L-DUMr(1.2, 1.2)	1.00	34	0.86	0.21	5.79
	TED(1, 1)	1.47	58	0.80	0.45	3.18
	L-DUMs(1.4, 1.4)	0.93	20	0.84	0.14	5.29
	L-DUMr(1.4, 1.4)	1.02	29	0.86	0.16	5.89
	TED(1, 1)	1.46	48	0.79	0.34	3.18
	L-DUMs(1.6, 1.6)	0.83	16	0.84	0.10	5.24
	L-DUMr(1.6, 1.6)	1.14	26	0.88	0.14	6.20
	TED(1, 1)	1.46	39	0.79	0.26	3.18

[a] L-DUMs represents the stochastic optimization approach, and L-DUMr represents the robust optimization approach.

Table 6. Characterization of Heterogeneous Patients

Type	N_j	μ_j	σ_j	$[\underline{z}, \bar{z}]$
New patient ($j = 1$)	1	18	7	$[-2, 12]$
Returning patient ($j = 2$)	3	13	6	$[-2, 12]$

Table 7. Number of Days with a Different Number of Surgeries in an Operating Room

Number of surgeries per day	1	2	3	4	5	6	7
Number of days	76	123	159	116	71	11	4

Figure 2. Sequencing and Scheduling Decisions with Various Tolerances

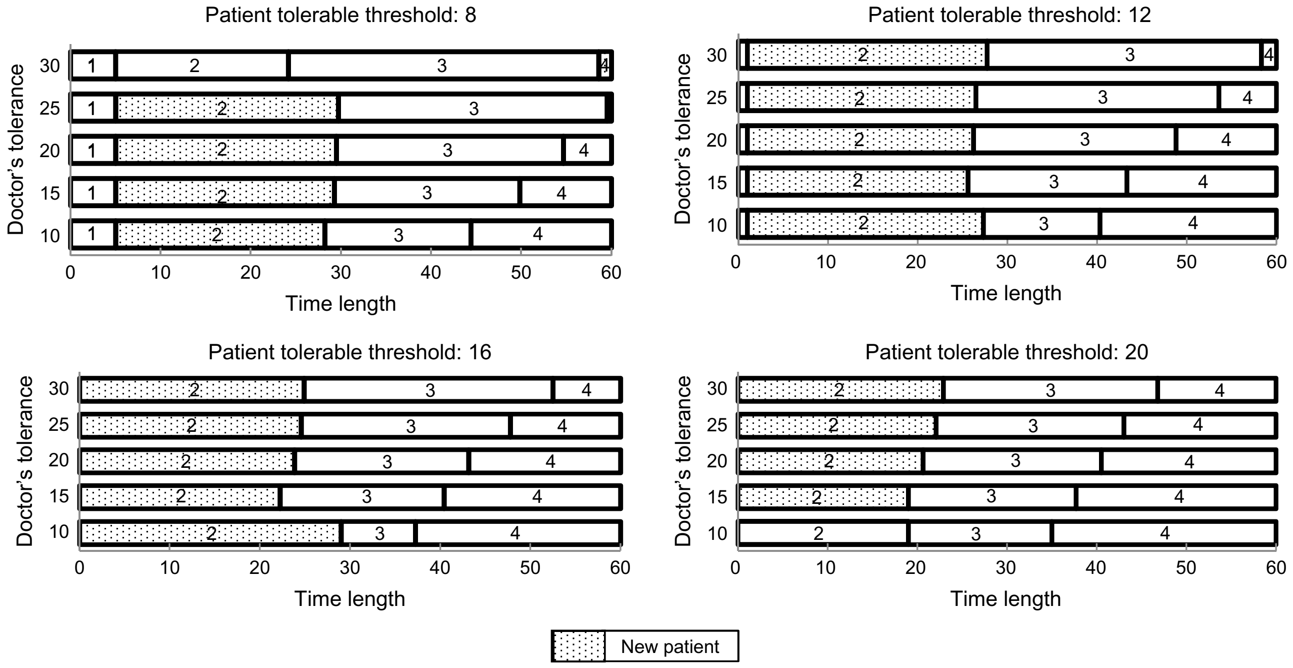


Table 8. Delay Performance of an Operating Theatre

		Position					
Performance	Method	1st	2nd	3rd	4th	5th and higher	Physician
Average delay	L-DUMs(20, 60)	0	4.38	3.29	2.82	4.35	10.38
	TED(1, 1)	0	1.67	2.10	2.73	6.94	10.02
	TED(1, 2)	0	2.52	5.06	5.87	10.57	7.69
Percentage of delay over the threshold	L-DUMs(20, 60)	0	6.7%	7.1%	5.1%	11.1%	5.0%
	TED(1, 1)	0	3.0%	4.6%	5.1%	13.6%	6.0%
	TED(1, 2)	0	4.4%	8.9%	10.1%	18.5%	4.6%

last position, such that his or her uncertain service time will not influence other patients' waiting times. Our program can easily solve a sequencing and scheduling problem for 10 patients within seconds.

We also use the real data collected from an operating room in a local hospital in Singapore to test our model. The operating room we consider is used only for the elective surgeries that are scheduled several days in advance. We have access to the data from 2011 to 2013 (560 days' record). Our data set consists of the types of surgeries, real operation durations, and the scheduled starting time on each day. On average, 3.1 surgeries are performed each day, and the numbers of days with different numbers of surgeries are summarized in Table 7.

We classify the surgeries based on the operation codes shown by the Ministry of Health, Singapore, and summarize their operation durations as samples. Each day, given the surgeries booked on that day, we calculate the sequencing and scheduling decisions based on the L-DUM model and the TED model with the empirical distributions and then use the real operation durations to simulate and summarize the performance in Table 8. Compared with the TED model, our model greatly reduces the delays for the patients at the last positions and mitigates the unfair effect by reallocating the waiting time across patients according to the defined threshold.

7. Conclusion

This paper proposes a new quality measure called the Delay Unpleasantness Measure to describe an individual's dissatisfaction with a waiting process. Then, given a number of heterogeneous patients, we lexicographically minimize the worst DUM to make both sequencing and scheduling decisions to mitigate the delay and unfairness in the appointment system.

In real practice, the same types of patients are generally allocated slots with the same session length. Hence, their waiting times tend to build up progressively, and the waiting times for those patients at the last positions are much longer than those of patients scheduled earlier. To alleviate the unfair effect toward the last patients, our model allocates a much shorter time slot for the first patient, i.e., it schedules the second patient much earlier. With this small change, the last patient's waiting time is reduced tremendously. Furthermore, unlike the smallest variance first rule, our computational study suggests that the sequencing decisions greatly depend on both the patients' and the physician's thresholds. When the patients' tolerance thresholds are relatively small, we hope to clear the cases as soon as possible; thus, patients with smaller variance and short average service time should be served first, and vice versa.

Acknowledgments

The authors acknowledge the assistance of Chew San Lee from the National University of Singapore for conducting the survey research and providing suggestions for this paper and Joe Sim (CEO, National University of Hospital, Singapore) for his support of this project. The authors are also grateful to the associate editor and the anonymous referees for their careful readings and suggestions to improve the paper.

Appendix

Proof of Proposition 1. (a) *Monotonicity*: If $\tilde{w}_1 \le \tilde{w}_2$, we have, for any $\alpha \in (0,1]$, $\varphi_\alpha(\tilde{w}_1) \le \varphi_\alpha(\tilde{w}_2)$ as a result of the monotonicity property of the function $\varphi_\alpha(\tilde{w})$. Therefore, $\rho_\tau(\tilde{w}_1) \le \rho_\tau(\tilde{w}_2)$.

(b) *Threshold satisficing*: If $\tilde{w} \le \tau$, $\rho_\tau(\tilde{w}) \le \rho_\tau(\tau) = \inf\{\alpha \ge 0 \mid \varphi_\alpha(\tau) \le \tau\} = 0$. Since $\rho_\tau(\tilde{w}) \in [0,1]$, we can immediately conclude that $\rho_\tau(\tilde{w}) = 0$.

(c) *Tardiness intolerance*: We first prove that $\varphi_1(\tilde{w}) = \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}(\tilde{w})$. According to the definition of $\varphi_\alpha(\tilde{w})$, $\varphi_1(\tilde{w}) \le 0 + \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}-0)^+) = \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}(\tilde{w})$. Moreover, as

$$\begin{aligned}\varphi_1(\tilde{w}) &= \min_{\nu\in\Re}\left\{\sup_{\mathbb{P}\in\mathbb{F}}(\nu + (\tilde{w}-\nu)^+)\right\} \ge \min_{\nu\in\Re}\left\{\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}(\nu + \tilde{w} - \nu)\right\}\\ &= \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}(\tilde{w}),\end{aligned}$$

we have $\varphi_1(\tilde{w}) = \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}(\tilde{w})$. Therefore, $\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}(\tilde{w}) > \tau$ is equivalent to $\varphi_1(\tilde{w}) > \tau$. According to the monotonicity property of function $\varphi_\alpha(\tilde{w})$, there exists no $\alpha \ge 0$ satisfying $\varphi_\alpha(\tilde{w}) \le \tau$, which leads to $\rho_\tau(\tilde{w}) = 1$.

(d) The proof is similar to that of Theorem 3 in Brown and Sim (2009).

(e) Given $\mathbb{P}(\tilde{w} > \tau) > 0$ for all $\mathbb{P} \in \mathbb{F}$, we can obtain, for any $\nu \ge 0$, $\nu + (1/\alpha)\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}-\tau-\nu)^+) > 0$. Hence,

$$\begin{aligned}\rho_\tau(\tilde{w}) &= \inf\{\alpha \ge 0 \mid \varphi_\alpha(\tilde{w}) \le \tau\}\\ &= \inf\left\{\alpha \ge 0 \mid \exists\, \nu \in \Re, \nu + \frac{1}{\alpha}\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}-\tau-\nu)^+) \le 0\right\}\end{aligned}$$

$$
\begin{aligned}
&= \inf\left\{\alpha \ge 0 \mid \exists\, \nu < 0, -\nu \ge \frac{1}{\alpha} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((\tilde{w}-\tau-\nu)^{+})\right\}\\
&= \inf\left\{\alpha \ge 0 \mid \exists\, a > 0, \frac{1}{a} \ge \frac{1}{\alpha} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left(\left(\tilde{w}-\tau+\frac{1}{a}\right)^{+}\right)\right\}\\
&= \inf\left\{\alpha \ge 0 \mid \alpha \ge \inf_{a>0}\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((a(\tilde{w}-\tau)+1)^{+})\right\}\\
&= \inf_{a>0}\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}((a(\tilde{w}-\tau)+1)^{+}). \quad \square
\end{aligned}
$$

Proof of Proposition 2. (a) *Monotonicity*: If $\tilde{\mathbf{w}} \le \tilde{\mathbf{v}}$, i.e., $\tilde{w}_n \le \tilde{v}_n$ for all $n \in [1;N]$, with the monotonicity property of $\rho_{\tau_n}(\tilde{w}_n)$, we have, for all $n \in [1;N]$, $\rho_{\tau_n}(\tilde{w}_n) \le \rho_{\tau_n}(\tilde{v}_n)$. Therefore, $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) \preceq_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$.

(b) *Threshold satisficing*: Because $\tilde{w}_n = \tilde{v}_n$ for all $n \in \mathcal{S}$, we have $\rho_{\tau_n}(\tilde{w}_n) = \rho_{\tau_n}(\tilde{v}_n)$. For any $j \in \bar{\mathcal{S}}$, $\tilde{w}_j, \tilde{v}_j \le \tau_j$; then according to threshold satisficing of the DUM, $\rho_{\tau_j}(\tilde{w}_j) = \rho_{\tau_j}(\tilde{v}_j) = 0$. Therefore, $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) =_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$.

(c) *Discrimination resistance*: If $|\mathcal{S}_1| < |\mathcal{S}_2|$, we have $\rho_i(\tilde{\mathbf{w}}) = \rho_i(\tilde{\mathbf{v}})$ for all $i \in [1;|\mathcal{S}_1|]$. For $i = |\mathcal{S}_1|+1 \le |\mathcal{S}_2|$, we have $\rho_i(\tilde{\mathbf{w}}) < 1 = \rho_i(\tilde{\mathbf{v}})$. Therefore, $\boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{w}}) \prec_{\text{lex}} \boldsymbol{\rho}_{\boldsymbol{\tau}}(\tilde{\mathbf{v}})$. $\square$

Proof of Proposition 3. To justify our claim, we first notice that the calculation of function

$$
\begin{aligned}
Z_P = \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\Bigg(\max\Bigg\{&0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1})\\
&- \nu_n, \ldots, \sum_{k=1}^{n-1}(\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n\Bigg\}\Bigg)
\end{aligned}
$$

can be equivalently written as an optimization problem as follows:

$$
\begin{aligned}
Z_P = \sup \mathrm{E}_{\mathbb{P}}\Bigg(\max\Bigg\{&0, -\nu_n, \sigma_{n-1}\tilde{z}_{n-1} + \mu_{n-1} - (x_n - x_{n-1})\\
&- \nu_n, \ldots, \sum_{k=1}^{n-1}(\sigma_k\tilde{z}_k + \mu_k - (x_{k+1} - x_k)) - \nu_n\Bigg\}\Bigg)\\
\text{s.t.}\quad & \mathrm{E}_{\mathbb{P}}(\tilde{z}_k) = 0, \quad k \in [1;n-1],\\
& \mathrm{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^{k}\tilde{z}_m\right|\right) \le \epsilon_{rk}, \quad r,k \in [1;n-1], r \le k,\\
& \mathbb{P}\{\tilde{z}_k \in [\underline{z}_k, \bar{z}_k], k \in [1;n-1]\} = 1.
\end{aligned} \tag{9}
$$

Its dual form can be written as

$$
\begin{aligned}
Z_1 = \min\ & f_0 + \sum_{k=1}^{n-1}\sum_{r=1}^{k}\epsilon_{rk}g_{rk}\\
\text{s.t.}\quad & f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1}\sum_{r=1}^{k} g_{rk}\left|\sum_{m=r}^{k}\tilde{z}_m\right| \ge 0,\\
&\qquad \forall\, z_k \in [\underline{z}_k, \bar{z}_k],\ k \in [1;n-1],\\
& f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1}\sum_{r=1}^{k} g_{rk}\left|\sum_{m=r}^{k} z_m\right| \ge -\nu_n,\\
&\qquad \forall\, z_k \in [\underline{z}_k, \bar{z}_k],\ k \in [1;n-1],\\
& f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1}\sum_{r=1}^{k} g_{rk}\left|\sum_{m=r}^{k} z_m\right|\\
&\quad \ge \sum_{k=l}^{n-1}(\sigma_k z_k + \mu_k - (x_{k+1} - x_k)) - \nu_n,\\
&\qquad \forall\, z_k \in [\underline{z}_k, \bar{z}_k],\ k,l \in [1;n-1],\\
& g_{rk} \ge 0, \quad r,k \in [1;n-1],\ r \le k,
\end{aligned} \tag{10}
$$

in which weak duality holds (see Isii 1963), and hence, $Z_P \le Z_1$. Observe that each constraint in problem (10) is the robust counterpart of a linear optimization problem with a bounded box uncertainty set. Hence, problem (10) is feasible and the objective is finite; i.e., $Z_1 < \infty$. Moreover, the dual form of the linear optimization problem,

$$
\begin{aligned}
\min\ & \sum_{k=1}^{l-1} f_k z_k + \sum_{k=l}^{n-1}(f_k - \sigma_k)z_k + \sum_{k=1}^{n-1}\sum_{r=1}^{k} g_{rk}\left|\sum_{m=r}^{k} z_m\right|\\
\text{s.t.}\ & z_k \ge \underline{z}_k, \quad k \in [1;n-1],\\
& z_k \le \bar{z}_k, \quad k \in [1;n-1],
\end{aligned}
$$

is equivalently written as

$$
\begin{aligned}
\max\ & \sum_{k=1}^{n-1}(\underline{z}_k u_k - \bar{z}_k v_k)\\
\text{s.t.}\ & u_k - v_k + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm} - c_{rm}) = f_k, \quad k \in [1;l-1],\\
& u_k - v_k + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm} - c_{rm}) = f_k - \sigma_k, \quad k \in [l;n-1],\\
& b_{rk} + c_{rk} = g_{rk}, \quad r,k \in [1;n-1],\ r \le k,\\
& u_k, v_k, b_{rk}, c_{rk} \ge 0, \quad r,k \in [1;n-1],\ r \le k.
\end{aligned}
$$

Combining all of these parts together, we can derive the optimization problem (4) in the proposition, and $Z_P \le Z_1 = Z_D$. To show that strong duality holds for the primal problem (9) and the dual problem (4), we cannot directly use the result of Isii (1963). To prove this, we derive the dual of problem (4) as

$$
\begin{aligned}
Z_2 = \max\Bigg\{&-\lambda_n\nu_n + \sum_{l=1}^{n-1}\lambda_l\left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1}\mu_k\right)\\
&+ \sum_{l=1}^{n-1}\sum_{k=l}^{n-1}\kappa_{lk}\sigma_k\Bigg\}\\
\text{s.t.}\quad & \sum_{l=0}^{n}\lambda_l = 1,\\
& \sum_{l=0}^{n}\kappa_{lk} = 0, \quad k \in [1;n-1],\\
& -\kappa_{lk} + \lambda_l\underline{z}_k \le 0, \quad k \in [1;n-1],\ l \in [0;n],\\
& \kappa_{lk} - \lambda_l\bar{z}_k \le 0, \quad k \in [1;n-1], l \in [0;n],\\
& -\eta_{rk}^{l} + \sum_{m=r}^{k}\kappa_{lm} \le 0,\\
&\qquad r,k \in [1;n-1],\ r \le k,\ l \in [0;n],\\
& -\eta_{rk}^{l} - \sum_{m=r}^{k}\kappa_{lm} \le 0,\\
&\qquad r,k \in [1;n-1],\ r \le k,\ l \in [0;n],\\
& \sum_{l=0}^{n}\eta_{rk}^{l} \le \epsilon_{rk}, \quad r,k \in [1;n-1], r \le k,\\
& \lambda_l \ge 0, \quad l \in [0;n].
\end{aligned} \tag{11}
$$

As strong duality holds in this linear optimization problem, we have $Z_D = Z_2 \in \mathfrak{R}$. Because $\mu_k \in (\underline{s}_k, \bar{s}_k)$, we have $0 \in (\underline{z}_k, \bar{z}_k)$ for all $k \in [1;n-1]$. Therefore, solution $\lambda_l = 1/(n+1), \kappa_{lk} = 0, \eta_{rk}^{l} = \epsilon_{rk}/(n+2), r,k \in [1;n-1], r \le k, l \in [0;n]$ is strictly feasible. Since problem (11) is a linear optimization problem with a finite objective and nonempty relative interior, there exists a sequence of interior feasible solutions whose objectives asymptotically converge to optimum. Hence, we have

$$
\begin{aligned}
Z_2 = \sup & -\lambda_n\nu_n + \sum_{l=1}^{n-1}\lambda_l\left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1}\mu_k\right) + \sum_{l=1}^{n-1}\sum_{k=l}^{n-1}\kappa_{lk}\sigma_k \\
\text{s.t.} \quad & \sum_{l=0}^{n}\lambda_l = 1, \\
& \sum_{l=0}^{n}\kappa_{lk} = 0, \quad k\in[1;n-1], \\
& -\kappa_{lk} + \lambda_l\underline{z}_k \le 0, \quad k\in[1;n-1],\, l\in[0;n], \\
& \kappa_{lk} - \lambda_l\bar{z}_k \le 0, \quad k\in[1;n-1],\, l\in[0;n], \\
& -\eta_{rk}^{l} - \sum_{m=r}^{k}\kappa_{lm} \le 0, \quad r,k\in[1;n-1],\, r\le k,\, l\in[0;n], \\
& -\eta_{rk}^{l} + \sum_{m=r}^{k}\kappa_{lm} \le 0, \quad r,k\in[1;n-1],\, r\le k,\, l\in[0;n], \\
& \sum_{l=0}^{n}\eta_{rk}^{l} \le \epsilon_{rk}, \quad r,k\in[1;n-1],\, r\le k, \\
& \lambda_l > 0, \quad l\in[0;n].
\end{aligned}
$$

Because $\lambda_l > 0$, by defining $\zeta_{lk} = \kappa_{lk}/\lambda_l, l\in[0;n], k\in[1;n-1]$, the above problem is equivalent to

$$
\begin{aligned}
Z_2 = \sup & -\lambda_n\nu_n + \sum_{l=1}^{n-1}\lambda_l\left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1}(\mu_k + \zeta_{lk}\sigma_k)\right) \\
\text{s.t.} \quad & \sum_{l=0}^{n}\lambda_l = 1, \\
& \sum_{l=0}^{n}\lambda_l\zeta_{lk} = 0, \quad k\in[1;n-1], \\
& -\zeta_{lk} \le -\underline{z}_k, \quad k\in[1;n-1],\, l\in[0;n], \\
& \zeta_{lk} \le \bar{z}_k, \quad k\in[1;n-1],\, l\in[0;n], \\
& -\eta_{rk}^{l} - \sum_{m=r}^{k}\zeta_{lm}\lambda_l \le 0, \\
& \qquad r,k\in[1;n-1],\, r\le k,\, l\in[0;n], \\
& -\eta_{rk}^{l} + \sum_{m=r}^{k}\zeta_{lm}\lambda_l \le 0, \\
& \qquad r,k\in[1;n-1],\, r\le k,\, l\in[0;n], \\
& \sum_{l=0}^{n}\eta_{rk}^{l} \le \epsilon_{rk}, \quad r,k\in[1;n-1],\, r\le k, \\
& \lambda_l > 0, \quad l\in[0;n], \\
= \sup & -\lambda_n\nu_n + \sum_{l=1}^{n-1}\lambda_l\left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1}(\mu_k + \zeta_{lk}\sigma_k)\right) \\
\text{s.t.} \quad & \sum_{l=0}^{n}\lambda_l\zeta_{lk} = 0, \quad k\in[1;n-1], \\
& \sum_{l=0}^{n}\lambda_l\left|\sum_{m=r}^{k}\zeta_{lm}\right| \le \epsilon_{rk}, \quad r,k\in[1;n-1],\, r\le k, \\
& \sum_{l=0}^{n}\lambda_l = 1, \\
& \zeta_{lk}\in[\underline{z}_k, \bar{z}_k], \quad k\in[1;n-1],\, l\in[0;n], \\
& \lambda_l > 0, \quad l\in[0;n]. \qquad (12)
\end{aligned}
$$

We observe that the feasible solution in problem (12) can be translated to $\tilde{z}_k$ being discretely distributed such that it takes values of ζ_{lk} with probability λ_l, $l\in[0;n]$ for all $k\in[1;n-1]$. Moreover, the objective of problem (12) satisfies

$$
\begin{aligned}
& -\lambda_n\nu_n + \sum_{l=1}^{n-1}\lambda_l\left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1}(\mu_k + \zeta_{lk}\sigma_k)\right) \\
& \quad \le \sum_{l=0}^{n}\lambda_l\Bigg(\max\Bigg\{0, -\nu_n, \sigma_{n-1}\zeta_{l,n-1} + \mu_{n-1} - (x_n - x_{n-1}) \\
& \qquad\qquad - \nu_n, \ldots, \sum_{k=1}^{n-1}(\sigma_k\zeta_{lk} + \mu_k - (x_{k+1} - x_k)) - \nu_n\Bigg\}\Bigg).
\end{aligned}
$$

Therefore, $Z_P \le Z_1 = Z_D = Z_2 \le Z_P$, and strong duality follows. □

References

Bailey NTJ (1952) A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *J. Roy. Statist. Soc. Ser. A* 14:185–199.

Bertsimas D, Farias VF, Trichakis N (2011) The price of fairness. *Oper. Res.* 59:17–31.

Brown DB, Sim M (2009) Satisficing measures for analysis of risky positions. *Management Science* 55:71–84.

Camacho F, Anderson R, Safrit A, Jones AS, Hoffmann P (2006) The relationship between patient's perceived waiting time and office-based practice satisfaction. *North Carolina Medical J.* 67:409–413.

Cartwright A, Windsor J (1992) *Outpatients and Their Doctors: A Study of Patients, Potential Patients, General Practitioners and Hospital Doctors* (Department of Health, Institute for Social Studies in Medical Care, London).

Çayırlı T, Veral E (2003) Outpatient scheduling in health care: A review of literature. *Production Oper. Management* 12:519–549.

Chen W, Sim M (2009) Goal driven optimization. *Oper. Res.* 57: 342–357.

Dehlendorff C, Kulahci M, Merser S, Andersen KK (2010) Conditional value at risk as a measure for waiting time in simulations of hospital units. *Quality Tech. Quant. Management* 7:321–336.

Demeulemeester E, Beliën J, Cardoen B, Samudra M (2013) Operating room planning and scheduling. Denton BT, ed. *Handbook of Healthcare Operations Management: Methods and Applications* (Springer, New York), 121–152.

Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35:1003–1016.

Denton B, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Sci.* 10:13–24.

Froehle CM, Magazine MJ (2013) Improving scheduling and flow in complex outpatient clinics. Denton BT, ed. *Handbook of Healthcare Operations Management: Methods and Applications* (Springer, New York), 229–250.

Green L, Savin S (2008) Reducing delays for medical appointments: A queuing approach. *Oper. Res.* 56:1526–1538.

Gupta D (2007) Surgical suites' operations management. *Production Oper. Management* 16:689–700.

Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* 40:800–819.

Harper PR, Gamlin HM (2003) Reduced outpatient waiting times with improved appointment scheduling: A simulation modeling approach. *OR Spectrum* 25:207–222.

Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management Sci.* 54:565–572.

Herzlinger RE (2006) Why innovation in health care is so hard. *Harvard Bus. Rev.* 84:58–66.

Hill CJ, Joonas K (2005) The impact of unacceptable wait time on health care patients' attitudes and actions. *Health Marketing Quart.* 23:69–87.

Huang XM (1994) Patient attitude towards waiting in an outpatient clinic and its applications. *Health Services Management Res.* 7:2–8.

Isermann H (1982) Linear lexicographic optimization. *OR Spektrum* 4:223–228.

Isii K (1963) On the sharpness of Chebyshev-type inequalities. *Ann. Institute Statist. Math.* 12:185–197.

Klassen KJ, Rohleder TR (1996) Scheduling outpatient appointments in a dynamic environment. *J. Oper. Management* 14:83–101.

Kong QX, Lee CY, Teo CP, Zheng ZC (2013) Scheduling arrivals to stochastic service delivery system using copositive cones. *Oper. Res.* 61:711–726.

Mak HY, Rong Y, Zhang J (2014) Sequencing appointments for service systems using inventory approximations. *Manufacturing Service Oper. Management* 16:251–262.

Mak HY, Rong Y, Zhang J (2015) Appointment scheduling with limited distributional information. *Management Sci.* 61:316–334.

McCarthy K, McGee HM, O'Boyle CA (2000) Outpatient clinic waiting times and non-attendance as indicators of quality. *Psych., Health Medicine* 5:287–293.

Mittal S, Stiller S (2011) Robust appointment scheduling. Working paper, Massachusetts Institute of Technology, Cambridge.

Mondschein S, Weintraub G (2003) Appointment policies in service operations: A critical analysis of the economic framework. *Production Oper. Management* 12:266–286.

Moschis GP, Bellinger DN, Curasi CF (2003) What influences the mature customer? *Marketing Health Care Services* 23:16–21.

Natarajan K, Sim M, Uichanco J (2010) Tractable robust expected utility and risk models for portfolio optimization. *Mathematical Finance* 20:695–731.

National Health Service. A guide to the National Health Service. http://www.publications.doh.gov.uk/pub/docs/doh/nhstxt.pdf.

Nemirovski A, Shapiro A (2006) Convex approximations of chance contrained programs. *SIAM J. Optim.* 17:969–996.

Ogryczak W, Pióro M, Tomaszewski, A (2005) Telecommunications network design and max-min optimization problem. *J. Telecomm. Inform. Tech.* 3:43–56.

Patrick J, Aubin A (2013) Models and methods for improving patient access. Denton BT, ed. *Handbook of Healthcare Operations Management* (Springer, New York), 403–420.

Robinson LW, Chen RR (2003) Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Trans.* 35:295–307.

Rockafellar RT (2007) Coherent approaches to risk in optimization under uncertainty. Klastorin T, ed. *OR Tools and Applications—Glimpses of Future Technologies*, INFORMS Tutorials in Operations Research (INFORMS, Hanover, MD), 38–61.

Rockafellar RT, Uryasev SP (2000) Optimization of conditional value-at-risk. *J. Risk* 2:21–41.

Sen A, Foster JE (1997) *On Economic Inequality* (Oxford University Press, Oxford, UK).

Toh LS, Sern CW (2011) Patient waiting time as a key performance indicator at orthodontic specialist clinics in Selangor. *Malaysian J. Public Health Medicine* 11:60–69.

Vanden Bosch PM, Dietz DC (2000) Minimizing expected waiting in a medical appointment system. *IIE Trans.* 32:841–848.

Vanden Bosch PM, Dietz DC (2001) Scheduling and sequencing arrivals to an appointment system. *J. Service Res.* 4:15–25.

Wang PP (1993) Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Res. Logist.* 40:345–360.

Wang PP (1999) Sequencing and scheduling N customers for a stochastic server. *Eur. J. Oper. Res.* 119:729–738.

Weiss EN (1990) Models for determining estimated start times and case orderings in hospital operation rooms. *IIE Trans.* 22:143–150.

Young HP (1995) *Equity: In Theory and Practice* (Princeton University Press, Princeton, NJ).

Zhu SS, Fukushima M (2009) Worst-case conditional value-at-risk with application to robust portfolio management. *Oper. Res.* 57:1155–1168.

Zhu ZC, Heng BH, Teow KL (2011) Reducing consultation waiting time and overtime in outpatient clinic: Challenges and solutions. Kolker A, Story P, eds. *Management Engineering for Effective Healthcare Delivery: Principles and Applications* (IGI Global, Hershey, PA), 229–245.