Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

Mitigating Delays and Unfairness in Appointment Systems

Jin Qi, Melvyn Sim
NUS Business School, National University of Singapore

We consider an appointment system where heterogenous participants are sequenced and scheduled for service. As service times are uncertain, the aim is to mitigate the unpleasantness experienced by the participants in the system when their waiting times or delays exceed acceptable thresholds, and address fairness concerning balancing of service levels among participants. In evaluating uncertain delays, we propose the Delay Unpleasantness Measure (DUM) which takes into account of 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 context of outpatient clinics in balancing doctor's overtime and patients' waiting time 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 mean absolute deviation as descriptive statistics in the distributional uncertainty set to preserve linearity of the model. The optimal sequencing and scheduling decisions could be derived by solving a sequence of mixed-integer programming problems and we report the insights from our computational studies.

Key words: Appointment scheduling, robust Optimization, Lexicographic min-max fairness

1. Introduction

In an efficient service system, due to the uncertainty in service times, waiting times or delays experienced by the participants are inevitable. However, long waiting time that occurs in a scheduled appointment is an annoyance and leads to poor quality of service. We focus our study in the context of an outpatient clinic, where the participants are patients and the doctor. Decisions associated with the appointment systems include the sequencing of patients, who are distinguished by their service time characteristics, and the scheduling of their appointment times. The goal of this paper is to design an appointment system that mitigates the unpleasantness experienced by the patients in waiting for consultation and by the doctor in having to work overtime.

The study of appointment systems stems from the pioneering work of Bailey (1952). Before that, service providers typically allocate each patient a slot with the same fixed time length. Bailey (1952) designs an appointment scheduling rule which assigns two patients at the first slot, followed by other patients' arrivals evenly spaced. This minor change effectively reduces doctor's idle time by overcoming the problem of patients no-show or lateness without compromising on the patients'

waiting time. Since then, many researchers have started to explore the optimal appointment system settings under various conditions. For comprehensive literature review, we refer readers to Cayirli and Veral (2003) and Gupta and Denton (2008), which highlight the current status and challenges in resolving appointment problems.

As outpatient appointment system is complicated with various settings, this paper only contributes to the literature with static assumption, that is, the information about patients who need appointment is known, and all the decisions must be made prior to the commencement of a clinic session. Hence, all of the following analysis concentrates on the static case only. Now, we will begin with discussing several concerns related to appointment system design problems.

The first concern regards to characterizing patients' experience of waiting, which is an integral aspect of service quality in a hospital environment among others. One commonly used service quality measure for describing this preference on uncertain waiting time is the expectation, which corresponds to the average delay experienced by the patient over potentially infinite number of visits under the same identical conditions. However, the expected waiting time criterion may not adequately distinguish patients' attitudes towards uncertain delay. We conduct a simple online survey to elicit preference for uncertain waiting times, which is analogous to St. Petersberg paradox in the monetary context.[1] Among the 118 respondents, 107 respondents' choices are inconsistent with the expected waiting time criterion.[2] Accordingly, we investigate behavioral preference for waiting time. From patients' perspective, the unpleasantness on waiting process may not proportionally accord to the length of waiting time. Huang (1994) empirically shows that, for patients arriving on the appointment time, they appear reasonably satisfied if they wait no more than averagely 37 minutes. The evidence also manifests that their patience may steeply decline when the service delay exceeds this threshold. Following this empirical result, we could reasonably assume that individual patient may have an unpleasantness tolerance threshold, and therefore, we could take the frequency of delays above this threshold as an alternative service quality measure. From service providers' perspectives, this measure also offers patients a service commitment to meet their tolerance thresholds. For example, patients in UK can expect to be seen within 30 minutes of their given appointment time (National Health Service, UK). Nonetheless, several non-negligible drawbacks have hampered the wide application of this measure. One disadvantage lies in the intensity of delay, for its inability of distinguishing waiting processes with the same frequency of surpassing patients' tolerance threshold but with different length of delay. Moreover, the computational intractability of this probability measure also arises due to lack of convexity. Thence, we need to establish a new service quality measure which could in some extent reflect people's real attitudes towards delay process, in particular, could be account for both the frequency and intensity of delays over the threshold.

The optimization criterion for an appointment system involves multiple participants including patients and doctors. Currently, majority of studies take a weighted average of the combinations among patients' waiting time, doctor's idle time and overtime as an optimization criterion, and exploit different methods to solve. Three main streams are based on queueing theory (see Wang, 1993; Wang, 1999; Green and Savin, 2008; Hassin and Mendel, 2008), stochastic programming (see Robinson and Chen, 2003; Denton and Gupta, 2003), and robust optimization (see Mittal and Stiller, 2011; Kong et al. 2012; Mak et al. 2012) frameworks. However, as 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, doctor's idle time and overtime fails to accommodate the fairness issue highlighted by Cayirli and Veral (2003). In layman terms, fairness regards to 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). Cayirli and Veral (2003) have highlighted the phenomenon that current appointment system is unfair to the patient at the last position, as waiting time tends to progressively build up. The notion of "fairness" has been widely studied in economics literatures (see Young, 1995; Sen and Foster, 1997) and industrial applications, especially resource allocation problems (see Bertsimas et al. 2011 and references therein). For this reason, an effective appointment system should be able to guarantee the uniformity of qualities across multiple participants.

To cope with the difficulties of eliciting the exact probability distribution for patients' consultation time, robust optimization techniques have also been applied in appointment problems (see Mittal and Stiller, 2011; Kong et al. 2011; Mak et al. 2012). In these papers, the optimization criteria are based on a weighted sum of patients' expected waiting time, doctor's idle time and overtime. Mittal and Stiller (2011) consider the scheduling problem where only the bound support of service time is provided. To minimize the sum of waiting time cost and idle time cost, they present a global balancing heuristic, and prove that it will deliver an optimal schedule under certain mild condition. Kong et al. (2012) assume lower bound, mean, and covariance of the service time are known, and formulate a robust minmax problem, which could be solved by a semidefinite programming relaxation. Mak et al. (2012) investigate the scheduling problem by assuming the knowledge of marginal moments of uncertain service time, and derive a computationally tractable conic programming formulation.

In general, consultation times among differ types of patients such as new and repeated one are not necessarily homogenous. Since the doctor would be familiar with the medical history of repeated patients, their consultation times tend to be shorter than new ones. To exploit the information of patients' classification, appointment systems would inevitably rely on the sequencing decisions

on these various types of patients. Due to the difficulty of the problems, few papers investigate the sequencing and scheduling decisions simultaneously. Weiss (1990) is the first to examine this problem and provides analytical results for two patients case with general service time distribution, however, the conclusions could not be simply extended to multiple patients case. Wang (1999) addresses the problem with a specific assumption that patients' service time follows exponential distribution 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 that follow different phase-type distributions. They approximately solve the scheduling problem by shifting the appointment time to incrementally improve the objective value for a given sequence, and then swap the sequence pairwisely until it terminates. 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 problem by investigating the case with two patients only.

To fully characterize all the above perspectives in appointment system design, especially, to mitigate the delay and unfairness in the appointment system, this paper first proposes a new service quality measure named Delay Unpleasantness Measure (DUM) to demonstrate the dependency of individual participant's attitude towards his/her delay process based on their corresponding acceptable level. This acceptable level is an exogenous factor, and varies according to patients' demographic profile. For example, the tolerable threshold of children may be much shorter than that of old patients. Besides, as the consultation time for repeated patients is relatively short, in certain cases, they may deserve a shortened waiting process, which corresponds to a small threshold. Unlike probability measure, DUM collectively accounts for the frequency and intensity of delay over a threshold. Secondly, we present the concept of lexicographic min-max fairness to tackle the fairness concern arising in appointment system design. We lexicographically minimize the worst DUM, the second worst DUM, and so on. Thirdly, by assuming patients' sequence is predetermined, we develop a scheduling model that can be adapted in the robust setting. Different from the conventional distributional uncertainty set, in which covariance matrix is used to capture the correlation among uncertain service times, we propose mean absolute deviation of summation over service times as the information that could help retain linearity of the model. Therefore, the optimal decisions are derived by solving a small sequence of linear optimization problems. Fourthly, this model could be extended to incorporate sequencing decisions when patients are heterogeneous.

The rest of the paper is organized as follows. In Section 2, we show how a participant's behavior in 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 outpatient appointment systems by assuming 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.

Notations: We denote scalars by plain characters, and use boldface lowercase characters to represent vectors, for example, $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)'$. Given a vector $\boldsymbol{x}$, we write $(y_n, \boldsymbol{x}_{-n})$ for the vector with only the nth component change, i.e., vector $(y_n, \boldsymbol{x}_{-n}) = (x_1, \ldots, x_{n-1}, y_n, x_{n+1}, \ldots, x_N)$. We use boldface uppercase characters to represent matrix, for example, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)'$. Set $\{i, \ldots, j\}$ represents positive running indices from i to j.

We represent uncertain quantities by tilde ($\tilde{}$) sign and model random variable $\tilde{x}$ by a state-space Ω and a $\sigma-$algebra of events $\mathcal{F}$ in Ω. Besides, we use $\mathcal{L}$ as the space of real-valued random variables. Comparison $\tilde{x} \geq \tilde{y}$ represents state-wise dominance, i.e., $x(\omega) \geq y(\omega)$ for all $\omega \in \Omega$. In addition, $\tilde{\boldsymbol{x}} \geq \tilde{\boldsymbol{y}}$ represents $\tilde{x}_n \geq \tilde{y}_n$ for all $n \in \{1, \ldots, N\}$. 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., $\mathbb{P} \in \mathbb{F}$. Noting 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. A measure of delay unpleasantness

In this section, we will motivate and introduce a new service quality measure to evaluate uncertain waiting time (service delay) of patients and overtime (off-work delay) of doctors. We will start with defining Delay Unpleasantness Measure (DUM) for individual participant (patient or doctor) in the appointment system. We assume that each participant has his/her own tolerance threshold τ on waiting time, and the real uncertain delay is represented by $\tilde{w}$. DUM takes into account of both the frequency and intensity of delay over the threshold and is defined as follows.

Definition 1 *Given an uncertain delay $\tilde{w} \in \mathcal{L}$ and tolerance threshold $\tau \in \Re_+$, the Delay Unpleasantness Measure is a function $\rho_\tau : \mathcal{L} \to [0, 1]$ defined as*

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

(or 1 if no such α exists), where

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

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. Function $\varphi_\alpha(\tilde{w})$ is the worst-case Conditional Value-at-Risk (CVaR) (see Zhu and Fukushima, 2009 and Natarajan, 2010) when we only have information that the true distribution $\mathbb{P}$ lies in a distributional uncertainty set $\mathbb{F}$. CVaR (Rockafellar and Uraysev, 2000) is a measure with specific focus on the tail distribution, and has become a major reference in the area of financial mathematics with its endearing properties. It is also shown to be the best convex conservative approximation of frequency of delay over the threshold (Nemirovski and Shapiro, 2006). 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. Several properties of DUM are listed in Proposition 1.

Proposition 1 *A DUM, ρ_τ has the following properties:*

(a) Monotonicity: if $\tilde{w}_1 \leq \tilde{w}_2$, then $\rho_\tau(\tilde{w}_1) \leq \rho_\tau(\tilde{w}_2)$;

(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

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

Proof. (a) Monotonicity: if $\tilde{w}_1 \leq \tilde{w}_2$, we have for any $\alpha \in (0,1]$, $\varphi_\alpha(\tilde{w}_1) \leq \varphi_\alpha(\tilde{w}_2)$ because of monotonicity property of $\varphi_\alpha(\tilde{w})$ function. Therefore, $\rho_\tau(\tilde{w}_1) \leq \rho_\tau(\tilde{w}_2)$.

(b) Threshold Satisficing: if $\tilde{w} \leq \tau$, $\rho_\tau(\tilde{w}) \leq \rho_\tau(\tau) = \inf\{\alpha \geq 0 \,|\varphi_\alpha(\tau) \leq \tau\} = 0$. With the bound that $\rho_\tau(\tilde{w}) \in [0,1]$, we could immediately conclude $\rho_\tau(\tilde{w}) = 0$.

(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 $\varphi_\alpha(\tilde{w})$, $\varphi_1(\tilde{w}) \leq 0 + \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w}-0)^+\right) = \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}(\tilde{w})$. Moreover, since

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

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 monotonicity property of function $\varphi_\alpha(\tilde{w})$, there exists no $\alpha \geq 0$ satisfying $\varphi_\alpha(\tilde{w}) \leq \tau$, which leads to $\rho_\tau(\tilde{w}) = 1$.

(d) The proof can be referred to Theorem 3 in Brown and Sim (2009).

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

$$\begin{aligned}
\rho_\tau(\tilde{w}) &= \inf\{\alpha \geq 0 \,|\, \varphi_\alpha(\tilde{w}) \leq \tau\} \\
&= \inf\left\{\alpha \geq 0 \,\middle|\, \exists \nu \in \Re, \nu + \frac{1}{\alpha} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w} - \tau - \nu)^+\right) \leq 0\right\} \\
&= \inf\left\{\alpha \geq 0 \,\middle|\, \exists \nu < 0, -\nu \geq \frac{1}{\alpha} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w} - \tau - \nu)^+\right)\right\} \\
&= \inf\left\{\alpha \geq 0 \,\middle|\, \exists a > 0, \frac{1}{a} \geq \frac{1}{\alpha} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left(\left(\tilde{w} - \tau + \frac{1}{a}\right)^+\right)\right\} \\
&= \inf\left\{\alpha \geq 0 \,\middle|\, \alpha \geq \inf_{a>0} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((a(\tilde{w} - \tau) + 1)^+\right)\right\} \\
&= \inf_{a>0} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((a(\tilde{w} - \tau) + 1)^+\right).
\end{aligned}$$

□

Remark 1. Property (a) captures participant's essential preference to a shorter delay, i.e., if the waiting time $\tilde{w}_1$ is state-wise greater than its counterpart $\tilde{w}_2$, then the former is not more preferred under the DUM. Property (b) indicates participant's desire to be served within the threshold and any uncertain delay that always meets the deadline will be most preferred. In contrast, Property (c) indicates 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 a threshold. We could guarantee that the frequency of delay over the threshold is less than the corresponding DUM. 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. Given two options A and B on delay, where

$$\tilde{w}_A = \begin{cases} 10 \text{ minutes, with probability } 0.89; \\ 30 \text{ minutes, with probability } 0.11. \end{cases} \qquad \tilde{w}_B = \begin{cases} 10 \text{ minutes, with probability } 0.9; \\ 60 \text{ minutes, with probability } 0.1. \end{cases}$$

When the tolerance threshold $\tau = 29$ minutes, the outcome of minimizing frequency of delay over a threshold suggests option B is better than A with $\mathbb{P}(\tilde{w}_B > 29) = 0.1 < \mathbb{P}(\tilde{w}_A > 29) = 0.11$, 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 suggests that option A is more preferable than B as $\rho_{29}(\tilde{w}_A) = \frac{11}{95} \leq \frac{5}{19} = \rho_{29}(\tilde{w}_B)$.

3. Lexicographic min-max fairness

The service quality of an appointment system depends on the participants' experiences on delays and we can formulate this as a multiple criteria optimization problem in which participants' DUMs are minimized, i.e.

$$\min_{\tilde{\boldsymbol{w}} \in \mathcal{W}} \{\boldsymbol{\rho_\tau}(\tilde{\boldsymbol{w}})\}$$

where $\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}}) = (\rho_{\tau_1}(\tilde{w}_1), \ldots, \rho_{\tau_N}(\tilde{w}_N))$ and $\mathcal{W}$ represents the space of feasible waiting times experienced by the participants. Among the Pareto optimal solutions, we would like to mitigate unfairness and avoid discriminating a subset of participants in terms of their service experiences in the appointment system. We adopt the lexicographic min-max fairness solution approach (see Young, 1995).

DEFINITION 2 *Let $\rho_i(\tilde{\boldsymbol{w}})$ and $\rho_i(\tilde{\boldsymbol{v}})$, $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}} \in \mathcal{W}$ be the ith largest elements of $\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}})$ and $\boldsymbol{\rho_T}(\tilde{\boldsymbol{v}})$ respectively. We say $\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}})$ is lexicographically equivalent to $\boldsymbol{\rho_T}(\tilde{\boldsymbol{v}})$, denoted by*

$$\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}}) =_{\text{lex}} \boldsymbol{\rho_T}(\tilde{\boldsymbol{v}})$$

if and only if $\rho_h(\tilde{\boldsymbol{w}}) = \rho_h(\tilde{\boldsymbol{v}})$ for all $h \in \{1, \ldots, N\}$. Moreover, $\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}})$ is lexicographically less than $\boldsymbol{\rho_T}(\tilde{\boldsymbol{v}})$, denoted by

$$\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}}) \prec_{\text{lex}} \boldsymbol{\rho_T}(\tilde{\boldsymbol{v}})$$

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

$$\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}}) \preceq_{\text{lex}} \boldsymbol{\rho_T}(\tilde{\boldsymbol{v}})$$

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

The lexicographic ordering shows that the participant with the worst 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 worst value will be used in deciding preference. We explore some 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 $\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}} \in \mathcal{W}$:*

(a) Monotonicity: if $\tilde{\boldsymbol{w}} \leq \tilde{\boldsymbol{v}}$, then

$$\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}}) \preceq_{\text{lex}} \boldsymbol{\rho_T}(\tilde{\boldsymbol{v}}).$$

(b) Threshold Satisficing: let $\mathcal{S} \subset \{1, \ldots, N\}$ and $\bar{\mathcal{S}}$ be the complement set. Suppose $\tilde{v}_j = \tilde{w}_j$ for all $j \in \mathcal{S}$ and $\tilde{v}_j \leq \tilde{w}_j \leq \tau_j$ for all $j \in \bar{\mathcal{S}}$, then

$$\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}}) =_{\text{lex}} \boldsymbol{\rho_T}(\tilde{\boldsymbol{v}}).$$

(c) Discrimination Resistance: let

$$\mathcal{S}_1 = \{i \in \{1, \ldots, N\} \,|\rho_{\tau_i}(\tilde{w}_i) = 1\} \text{ and } \mathcal{S}_2 = \{i \in \{1, \ldots, N\} \,|\rho_{\tau_i}(\tilde{v}_i) = 1\}.$$

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

$$\boldsymbol{\rho_T}(\tilde{\boldsymbol{w}}) \prec_{\text{lex}} \boldsymbol{\rho_T}(\tilde{\boldsymbol{v}}).$$

Proof. (a) Monotonicity: if $\tilde{\boldsymbol{w}} \leq \tilde{\boldsymbol{v}}$, i.e., $\tilde{w}_n \leq \tilde{v}_n$ for all $n \in \{1,\ldots,N\}$, with the monotonicity property of $\rho_{\tau_n}(\tilde{w}_n)$, we have for all $n \in \{1,\ldots,N\}$, $\rho_{\tau_n}(\tilde{w}_n) \leq \rho_{\tau_n}(\tilde{v}_n)$. Therefore, $\boldsymbol{\rho_\tau}(\tilde{\boldsymbol{w}}) \preceq_{\text{lex}} \boldsymbol{\rho_\tau}(\tilde{\boldsymbol{v}})$.
(b) Threshold Satisficing: Since $\tilde{w}_n = \tilde{v}_n$ for all $n \in \mathcal{S}$, we have $\rho_{\tau_n}(\tilde{w}_n) = \rho_{\tau_n}(\tilde{v}_n)$. For any $j \in \bar{\mathcal{S}}$, $\tilde{w}_j, \tilde{v}_j \leq \tau_j$, then according to Threshold Satisficing of DUM, $\rho_{\tau_j}(\tilde{w}_j) = \rho_{\tau_j}(\tilde{v}_j) = 0$. Therefore, $\boldsymbol{\rho_\tau}(\tilde{\boldsymbol{w}}) =_{\text{lex}} \boldsymbol{\rho_\tau}(\tilde{\boldsymbol{v}})$.
(c) Discrimination Resistance: if $|\mathcal{S}_1| < |\mathcal{S}_2|$, we have $\rho_i(\tilde{\boldsymbol{w}}) = \rho_i(\tilde{\boldsymbol{v}})$ for all $i \in \{1,\ldots,|\mathcal{S}_1|\}$. For $i = |\mathcal{S}_1|+1 \leq |\mathcal{S}_2|$, we have $\rho_i(\tilde{\boldsymbol{w}}) < 1 = \rho_i(\tilde{\boldsymbol{v}})$. Therefore, $\boldsymbol{\rho_\tau}(\tilde{\boldsymbol{w}}) \prec_{\text{lex}} \boldsymbol{\rho_\tau}(\tilde{\boldsymbol{v}})$. □

REMARK 2. Monotonicity ensures consistency so that reduction in delays for all participants will be favorably valued. Threshold Satisficing property ensures that the participants whose delays are always within their thresholds, then any improvement of their delays do not contribute to the lexicographic ordering. A participant is discriminated if the appointment system cannot guarantee his/her average waiting time below the threshold, which corresponds to the DUM taking value of one. Hence, Discrimination Resistance 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 experiences the worst waiting process as small as possible. [3]

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

$$\text{lex}\min_{\tilde{\boldsymbol{w}}\in\mathcal{W}}\{\boldsymbol{\rho_\tau}(\tilde{\boldsymbol{w}})\},$$

where the optimal solution $\tilde{\boldsymbol{w}}^* \in \mathcal{W}$ satisfies

$$\boldsymbol{\rho_\tau}(\tilde{\boldsymbol{w}}^*) \preceq_{\text{lex}} \boldsymbol{\rho_\tau}(\tilde{\boldsymbol{v}}) \qquad \forall\, \tilde{\boldsymbol{v}} \in \mathcal{W}.$$

Though 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 and Ogryczak et al., 2005) as follows:

Algorithm: Lexicographic Minimization Procedure

1. Set $h := 1, \mathcal{G}_0 := \{1,\ldots,N\}$,

$$\alpha_1 := \min_{\tilde{\boldsymbol{w}}\in\mathcal{W}} \max_{n\in\mathcal{G}_0} \rho_{\tau_n}(\tilde{w}_n),$$
$$\mathcal{I}_1 := \left\{ j \in \mathcal{G}_0 \;:\; \min_{\tilde{\boldsymbol{w}}\in\mathcal{W}} \left\{ \rho_{\tau_j}(\tilde{w}_j) \middle| \max_{n\in\mathcal{G}_0} \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_1 \right\} = \alpha_1 \right\}.$$

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

$$\alpha_h := \min_{\tilde{\boldsymbol{w}}\in\mathcal{W}} \left\{ \max_{n\in\mathcal{G}_{h-1}} \rho_{\tau_n}(\tilde{w}_n) \;\middle|\; \max_{n\in\mathcal{I}_i} \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_i, i \in \{1,...,h-1\} \right\},$$
$$\mathcal{I}_h := \left\{ j \in \mathcal{G}_{h-1} \;:\; \min_{\tilde{\boldsymbol{w}}\in\mathcal{W}} \left\{ \rho_{\tau_j}(\tilde{v}_j) \middle| \begin{array}{l} \max\limits_{n\in\mathcal{G}_{h-1}} \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_h, \\ \max\limits_{n\in\mathcal{I}_i} \rho_{\tau_n}(\tilde{w}_n) \leq \alpha_i, i \in \{1,...,h-1\} \end{array} \right\} = \alpha_h \right\}.$$

3. Go to Step 2.

In this algorithm, we minimize the the maximum DUM among a set of participants and elicit the subset of participants that attain the worst value. Hence, the optimum solution, $\tilde{\boldsymbol{w}}^* \in \mathcal{W}$ satisfies

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

for all $i \in \{1, \ldots, h\}$. Observe that the problem to derive α_h is the same as

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

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

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

Though the problem is nonlinear in α, we observe that $\frac{1}{\alpha} \sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}} \left((\tilde{w}_n - \nu_n)^+ \right)$ is monotonic in α and hence we could use binary search procedure to find the optimal solution in which α is minimized. Similarly, we can determine $\mathcal{I}_i$ by performing a sequence of binary search procedures.

4. Appointment system design

We first consider an appointment scheduling problem with one doctor serving N patients under the following assumptions:

Assumptions

- Schedules have to be made before the commencement of the session.
- Patients may be heterogenous and are characterized by their service time distributions and tolerance thresholds.
- The consultation sequence of patients is pre-determined.
- Patients arrive on time.[4]
- Doctor will start his/her session promptly. Hence, the first patient experiences no delay.

Model parameters and decision variables

- N: total number of patients to be scheduled;
- L: session length pre-determined for the consultation of N patients;
- τ_n: the tolerance threshold of delay for the patient at nth position, $n \in \{1, \ldots, N\}$;
- τ_{N+1}: doctor's tolerance on his/her overtime;

- $\tilde{s}_n$: consultation time of the nth patient;
- $\tilde{w}_n$: waiting time of the nth patient, $n \in \{1, \ldots, N\}$;
- $\tilde{w}_{N+1}$: doctor's overtime;
- 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 $\boldsymbol{x} = (x_1, \ldots, x_N, x_{N+1})'$.

We first specify the feasible set of waiting times, $\mathcal{W}$ as follows:

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

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

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

The first two constraints in the set $\mathcal{W}$ recursively calculate the delays experienced by the patients and the doctor, while the set $\mathcal{X}$ ensures sequencing compliance. Accordingly as in Denton and Gupta (2003), we further simplify the formulation by defining the difference between the real service time and scheduled interval as $\tilde{t}_n$ for the nth patient

$$\tilde{t}_n = \tilde{s}_n - (x_{n+1} - x_n), \quad n \in \{1, ..., N\}. \tag{2}$$

It follows that the nth patient's waiting time and doctor's overtime can be represented by

$$\tilde{w}_n = \max\left\{0, \tilde{t}_{n-1}, \ldots, \sum_{k=1}^{n-1} \tilde{t}_k\right\}, \quad n \in \{2, ..., N+1\}. \tag{3}$$

Since the lexicographic minimization procedure requires solving a sequence of similar problems, we will focus on solving Problem (1) as a representative instance. To derive the optimal scheduling decisions, we formulate Problem (1) as

$$\begin{array}{ll} \inf\ \alpha & \\ \text{s.t. } \nu_n + \dfrac{1}{\alpha} \sup\limits_{\mathbb{P} \in \mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w}_n - \nu_n)^+\right) \le \tau_n, & n \in \mathcal{G}_{h-1}, \\ \quad \nu_n + \dfrac{1}{\alpha_i} \sup\limits_{\mathbb{P} \in \mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w}_n - \nu_n)^+\right) \le \tau_n, & n \in \mathcal{I}_i, i \in \{1, ..., h-1\}, \\ \quad \tilde{w}_n = \max\left\{0, \tilde{t}_{n-1}, \ldots, \sum\limits_{k=1}^{n-1} \tilde{t}_k\right\}, & n \in \{2, ..., N+1\}, \\ \quad \tilde{t}_n = \tilde{s}_n - (x_{n+1} - x_n), & n \in \{1, ..., N\}, \\ \quad \alpha \in (0, 1], & \\ \quad \boldsymbol{x} \in \mathcal{X}. & \end{array} \tag{4}$$

Since the first patient's waiting time is zero, we have $\rho_{\tau_1}(\tilde{w}_1) = 0$ for any nonnegative threshold τ_1. Therefore, we can define $\mathcal{G}_0 = \{2, \ldots, N+1\}$.

We first focus on the simplification of function $\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w}_n-\nu_n)^+\right)$, which is complicated by the recursive property of uncertain waiting times. In conjunction with Equations (2) and (3), we observe that

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

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

4.1. Stochastic optimization approach

For the case of 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, $m\in\{1,\ldots,M\}$, we have

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left(\max\left\{0,-\nu_n,\tilde{s}_{n-1}-(x_n-x_{n-1})-\nu_n,\ldots,\sum_{k=1}^{n-1}(\tilde{s}_k-(x_{k+1}-x_k))-\nu_n\right\}\right)\\
&=\sum_{m=1}^{M} p_m \max\left\{0,-\nu_n,s_{n-1}^m-(x_n-x_{n-1})-\nu_n,\ldots,\sum_{k=1}^{n-1}(s_k^m-(x_{k+1}-x_k))-\nu_n\right\}.
\end{aligned}$$

Therefore, by adding decision variables q_{mn}, $m\in\{1,\ldots,M\}, n\in\{2,\ldots,N+1\}$, Problem (4) is equivalent to

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

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 distributional robust optimization approach with the goal of preserving linearity of the model. We assume the family of service times distributions are characterized based on their bounded supports $\mathbb{P}(\tilde{s}_k \in [\underline{s}_k, \overline{s}_k]) = 1$, means $\mathrm{E}_{\mathbb{P}}(\tilde{s}_k) = \mu_k$, $\mu_k \in (\underline{s}_k, \overline{s}_k)$ and bounds of mean absolute deviation $\mathrm{E}_{\mathbb{P}}(|\tilde{s}_k - \mu_k|) \leq \sigma_k$, $\sigma_k > 0$ for all $k \in \{1, \ldots, N\}$. Intuitively, the worst case probability distributions may result in highly correlated service times, which may not be realistic and lead to conservative solutions. 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 $r, k \in \{1, \ldots, N\}$, $r \leq k$. However, this will necessarily lead to nonlinear optimization models, which are harder to solve (Kong et al. 2012, Mak et al. 2012). To avoid nonlinearlity, we propose a different approach of capturing correlation. We note that the waiting time of a participant may be influenced by the aggregation of uncertain service times of earlier participants. Hence, in our distributional uncertainty set, we use the descriptive statistics of $\mathrm{E}_{\mathbb{P}}\left(\left|\sum_{m=r}^{k} \frac{\tilde{s}_m - \mu_m}{\sigma_m}\right|\right)$ for all $r, k \in \{1, \ldots, N\}$ and $r \leq k$. Observe that

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

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

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

where $\epsilon_{rk} \in (0, k - r + 1]$. Without loss of generality, we define $\epsilon_{kk} = 1$ that is equivalent to the information $\mathrm{E}_{\mathbb{P}}(|\tilde{s}_k - \mu_k|) \leq \sigma_k$. These constraints set the bound for the dispersion of the total uncertain service times for $k - r + 1$ consecutive patients, and enable us to specify less conservative uncertainty set while keeping the model linear. Now, the distributional uncertainty set can be written as

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

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

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

and we have

$$\begin{aligned}
&\sup_{\mathbb{P} \in \mathbb{F}} \mathrm{E}_{\mathbb{P}}\left(\max\left\{0, -\nu_n, \tilde{s}_{n-1} - (x_n - x_{n-1}) - \nu_n, \ldots, \sum_{k=1}^{n-1}(\tilde{s}_k - (x_{k+1} - x_k)) - \nu_n\right\}\right) \\
&= \sup_{\mathbb{P} \in \mathbb{F}_z} \mathrm{E}_{\mathbb{P}}\left(\max\left\{0, -\nu_n, \sigma_{n-1}\tilde{z}_{n-1} + \mu_{n-1} - (x_n - x_{n-1}) - \nu_n, \ldots, \sum_{k=1}^{n-1}(\sigma_k \tilde{z}_k + \mu_k - (x_{k+1} - x_k)) - \nu_n\right\}\right)
\end{aligned}$$

PROPOSITION 3 *For a given $\boldsymbol{x} \in \mathcal{X}$ and $n \in \{2, \ldots, N+1\}$, the problem*

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

corresponds to the optimal value of the following linear optimization problem

$$\begin{array}{rll}
Z_D = \min & f_0 + \displaystyle\sum_{k=1}^{n-1} \sum_{r=1}^{k} \epsilon_{rk} g_{rk} & \\
s.t. & f_0 + \displaystyle\sum_{k=1}^{n-1} \left(\underline{z}_k u_k^0 - \overline{z}_k v_k^0 \right) \geq 0, & \\
& f_0 + \nu_n + \displaystyle\sum_{k=1}^{n-1} \left(\underline{z}_k u_k^n - \overline{z}_k v_k^n \right) \geq 0, & \\
& f_0 + \nu_n + x_n - x_l + \displaystyle\sum_{k=1}^{n-1} \left(\underline{z}_k u_k^l - \overline{z}_k v_k^l \right) \geq \sum_{k=l}^{n-1} \mu_k, \; l \in \{1, ..., n-1\}, & \\
& u_k^l - v_k^l + \displaystyle\sum_{m=k}^{n-1} \sum_{r=1}^{k} \left(b_{rm}^l - c_{rm}^l \right) - f_k = 0, & k \in \{1, ..., n-1\}, l = 0, n, \\
& u_k^l - v_k^l + \displaystyle\sum_{m=k}^{n-1} \sum_{r=1}^{k} \left(b_{rm}^l - c_{rm}^l \right) - f_k = 0, & k, l \in \{1, ..., n-1\}, k \leq l-1, \\
& u_k^l - v_k^l + \displaystyle\sum_{m=k}^{n-1} \sum_{r=1}^{k} \left(b_{rm}^l - c_{rm}^l \right) - f_k = -\sigma_k, & k, l \in \{1, ..., n-1\}, l \leq k, \\
& b_{rk}^l + c_{rk}^l - g_{rk} = 0, & r, k \in \{1, ..., n-1\}, r \leq k, l \in \{0, ..., n\}, \\
& u_k^l, v_k^l, b_{rk}^l, c_{rk}^l, g_{rk} \geq 0, & r, k \in \{1, ..., n-1\}, r \leq k, l \in \{0, ..., n\}.
\end{array} \tag{5}$$

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

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

can be equivalently written as an optimization problem as follows

$$\begin{array}{rll}
Z_P = \sup & \mathrm{E}_{\mathbb{P}} \left(\max \left\{ 0, -\nu_n, \sigma_{n-1} \tilde{z}_{n-1} + \mu_{n-1} - (x_n - x_{n-1}) - \nu_n, \ldots, \displaystyle\sum_{k=1}^{n-1} (\sigma_k \tilde{z}_k + \mu_k - (x_{k+1} - x_k)) - \nu_n \right\} \right) & \\
\text{s.t.} & \mathrm{E}_{\mathbb{P}} (\tilde{z}_k) = 0, & k \in \{1, ..., n-1\}, \\
& \mathrm{E}_{\mathbb{P}} \left(\left| \displaystyle\sum_{m=r}^{k} \tilde{z}_m \right| \right) \leq \epsilon_{rk}, & r, k \in \{1, ..., n-1\}, r \leq k, \\
& \mathbb{P} \{ \tilde{z}_k \in [\underline{z}_k, \overline{z}_k], k \in \{1, ..., n-1\} \} = 1. &
\end{array} \tag{6}$$

Its dual form can be written as

$$
\begin{aligned}
Z_1 = \min\ & f_0 + \sum_{k=1}^{n-1}\sum_{r=1}^{k} \epsilon_{rk} g_{rk} \\
\text{s.t. } & f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1}\sum_{r=1}^{k} g_{rk} \left| \sum_{m=r}^{k} \tilde{z}_m \right| \geq 0, && \forall\, z_k \in [\underline{z}_k, \overline{z}_k], k \in \{1,...,n-1\}, \\
& f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1}\sum_{r=1}^{k} g_{rk} \left| \sum_{m=r}^{k} z_m \right| \geq -\nu_n, && \forall\, z_k \in [\underline{z}_k, \overline{z}_k], k \in \{1,...,n-1\}, \\
& f_0 + \sum_{k=1}^{n-1} f_k z_k + \sum_{k=1}^{n-1}\sum_{r=1}^{k} g_{rk} \left| \sum_{m=r}^{k} z_m \right| \geq \sum_{k=l}^{n-1} (\sigma_k z_k + \mu_k - (x_{k+1} - x_k)) - \nu_n, \\
& && \forall\, z_k \in [\underline{z}_k, \overline{z}_k], k, l \in \{1,...,n-1\}, \\
& g_{rk} \geq 0, && r, k \in \{1,...,n-1\}, r \leq k,
\end{aligned}
\tag{7}
$$

in which weak duality holds (see Isii, 1963), and hence, $Z_P \leq Z_1$. Observe that each constraint in Problem (7) is the robust counterpart of a linear optimization problem with bounded box uncertainty set. Hence, Problem (7) is feasible and objective is finite, i.e., $Z_1 < \infty$. Moreover, the dual form of the linear optimization problem

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

is equivalently written as

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

Combining all these analysis parts together, we could derive the optimization problem (5) in the proposition, and $Z_P \leq Z_1 = Z_D$. To show that strong duality holds for the primal problem (6) and the dual problem (5), we cannot directly use the result of Isii (1963). To prove it, we derive the

dual of Problem (5) as

$$
\begin{aligned}
Z_2 = \max\ & -\lambda_n\nu_n + \sum_{l=1}^{n-1}\lambda_l\left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1}\mu_k\right) + \sum_{l=1}^{n-1}\sum_{k=l}^{n-1}\kappa_{lk}\sigma_k \\
\text{s.t. } & \sum_{l=0}^{n}\lambda_l = 1, \\
& \sum_{l=0}^{n}\kappa_{lk} = 0, && k\in\{1,...,n-1\}, \\
& -\kappa_{lk} + \lambda_l\underline{z}_k \le 0, && k\in\{1,...n-1\}, l\in\{0,...,n\}, \\
& \kappa_{lk} - \lambda_l\overline{z}_k \le 0, && k\in\{1,...n-1\}, l\in\{0,...,n\}, \\
& -\eta^l_{rk} + \sum_{m=r}^{k}\kappa_{lm} \le 0, && r,k\in\{1,...,n-1\}, r\le k, l\in\{0,...,n\}, \\
& -\eta^l_{rk} - \sum_{m=r}^{k}\kappa_{lm} \le 0, && r,k\in\{1,...,n-1\}, r\le k, l\in\{0,...,n\}, \\
& \sum_{l=0}^{n}\eta^l_{rk} \le \epsilon_{rk}, && r,k\in\{1,...,n-1\}, r\le k, \\
& \lambda_l \ge 0, && l\in\{0,...,n\}.
\end{aligned}
\tag{8}
$$

Since strong duality holds in this linear optimization problem, we have $Z_D = Z_2 \in \Re$. Since, $\mu_k \in (\underline{s}_k, \overline{s}_k)$, we have $0 \in (\underline{z}_k, \overline{z}_k)$ for all $k \in \{1,\ldots,n-1\}$. Therefore, solution $\lambda_l = \frac{1}{n+1}, \kappa_{lk} = 0, \eta^l_{rk} = \frac{\epsilon_{rk}}{n+2}, r,k\in\{1,\ldots,n-1\}, r\le k, l\in\{0,\ldots,n\}$ is strictly feasible. Since Problem (8) is a linear optimization problem with finite objective and non-empty relative interior, there exists a sequence of interior feasible solutions whose objectives asymptotically coverage to optimum. Hence, we have

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

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

$$
\begin{aligned}
Z_2 = \sup\ & -\lambda_n\nu_n + \sum_{l=1}^{n-1}\lambda_l\left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1}(\mu_k + \zeta_{lk}\sigma_k)\right) \\
\text{s.t.}\ & \sum_{l=0}^{n}\lambda_l = 1, \\
& \sum_{l=0}^{n}\lambda_l\zeta_{lk} = 0, && k\in\{1,...,n-1\}, \\
& -\zeta_{lk} \le -\underline{z}_k, && k\in\{1,...n-1\}, l\in\{0,...,n\}, \\
& \zeta_{lk} \le \overline{z}_k, && k\in\{1,...n-1\}, l\in\{0,...,n\}, \\
& -\eta_{rk}^l - \sum_{m=r}^{k}\zeta_{lm}\lambda_l \le 0, && r,k\in\{1,...,n-1\}, r\le k, l\in\{0,...,n\}, \\
& -\eta_{rk}^l + \sum_{m=r}^{k}\zeta_{lm}\lambda_l \le 0, && r,k\in\{1,...,n-1\}, r\le k, l\in\{0,...,n\}, \\
& \sum_{l=0}^{n}\eta_{rk}^l \le \epsilon_{rk}, && r,k\in\{1,...,n-1\}, r\le k, \\
& \lambda_l > 0, && l\in\{0,...,n\}, \\
= \sup\ & -\lambda_n\nu_n + \sum_{l=1}^{n-1}\lambda_l\left(-\nu_n - x_n + x_l + \sum_{k=l}^{n-1}(\mu_k + \zeta_{lk}\sigma_k)\right) \\
\text{s.t.}\ & \sum_{l=0}^{n}\lambda_l\zeta_{lk} = 0, && k\in\{1,...,n-1\}, \\
& \sum_{l=0}^{n}\lambda_l\left|\sum_{m=r}^{k}\zeta_{lm}\right| \le \epsilon_{rk}, && r,k\in\{1,...,n-1\}, r\le k, \\
& \sum_{l=0}^{n}\lambda_l = 1, \\
& \zeta_{lk}\in[\underline{z}_k, \overline{z}_k], && k\in\{1,...n-1\}, l\in\{0,...,n\}, \\
& \lambda_l > 0, && l\in\{0,...,n\}.
\end{aligned}
\tag{9}
$$

We observe that the feasible solution in Problem (9) can be translated to $\tilde{z}_k$ being discrete distributed that takes values of ζ_{lk} with probability λ_l, $l\in\{0,\ldots,n\}$ for all $k\in\{1,\ldots,n-1\}$. Moreover, the objective of Problem (9) satisfies

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

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

Correspondingly,

$$
\begin{aligned}
&\inf\ \alpha \\
&\text{s.t. } \nu_n + \frac{1}{\alpha}\left(f_0^n + \sum_{k=1}^{n-1}\sum_{r=1}^{k}\epsilon_{rk}g_{rk}^n\right) \le \tau_n, && n \in \mathcal{G}_{h-1}, \\
&\quad \nu_n + \frac{1}{\alpha_i}\left(f_0^n + \sum_{k=1}^{n-1}\sum_{r=1}^{k}\epsilon_{rk}g_{rk}^n\right) \le \tau_n, && n \in \mathcal{I}_i, i \in \{1,...h-1\}, \\
&\quad f_0^n + \sum_{k=1}^{n-1}\left(\underline{z}_k u_k^{0n} - \overline{z}_k v_k^{0n}\right) \ge 0, && n \in \{2,...,N+1\}, \\
&\quad f_0^n + \nu_n + \sum_{k=1}^{n-1}\left(\underline{z}_k u_k^{nn} - \overline{z}_k v_k^{nn}\right) \ge 0, && n \in \{2,...,N+1\}, \\
&\quad f_0^n + \nu_n + x_n - x_l + \sum_{k=1}^{n-1}\left(\underline{z}_k u_k^{ln} - \overline{z}_k v_k^{ln}\right) \ge \sum_{k=l}^{n-1}\mu_k, && l \in \{1,...,n-1\}, n \in \{2,...,N+1\}, \\
&\quad u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = 0, && k \in \{1,...,n-1\}, l = 0, n, n \in \{2,...,N+1\}, \\
&\quad u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = 0, && k, l \in \{1,...,n-1\}, k \le l-1, n \in \{2,...,N+1\}, \\
&\quad u_k^{ln} - v_k^{ln} + \sum_{m=k}^{n-1}\sum_{r=1}^{k}(b_{rm}^{ln} - c_{rm}^{ln}) - f_k^n = -\sigma_k, && k, l \in \{1,...,n-1\}, k \ge l, n \in \{2,...,N+1\}, \\
&\quad b_{rk}^{ln} + c_{rk}^{ln} - g_{rk}^n = 0, && r, k \in \{1,...,n-1\}, r \le k, l \in \{0,...,n\}, n \in \{2,...,N+1\}, \\
&\quad u_k^{ln}, v_k^{ln}, b_{rk}^{ln}, c_{rk}^{ln}, g_{rk}^n \ge 0, && r, k \in \{1,...,n-1\}, r \le k, l \in \{0,...,n\}, n \in \{2,...,N+1\}, \\
&\quad \alpha \in (0,1], \\
&\quad \boldsymbol{x} \in \mathcal{X}.
\end{aligned}
\tag{10}
$$

Problem (10) is quite complicated at a 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. When α decreases to zero, $\varphi_\alpha(\tilde{w})$ approaches the upper limit of $\tilde{w}$. We assume that it is onus of the decision maker to select the threshold values so that Problem (10) 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.

It is worthy pointing 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. Incorporating sequencing decisions

We now generalize the scheduling model to incorporate the realistic situation with sequencing decisions for heterogeneous patients. First, we clarify some extra parameters and decision variables.

- J: number of patient types. Patients with the same type have same mean μ_j, mean absolute deviation σ_j of the consultation time, and same tolerance threshold;

- N_j: number of jth type patients, where $\sum_{j=1}^{J} N_j = N$;
- β_j: the tolerance threshold of delay for jth type patients, $j \in \{1,\ldots,J\}$;
- $\tilde{s}_{nj}$: uncertain service time associated with the nth patient if he/she belongs to jth type;
- y_{nj}: binary decision variable, if the jth type patient is scheduled in the nth position, then $y_{nj}=1$, otherwise, $y_{nj}=0$. Its matrix form is $\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_N)' \in \{0,1\}^{N\times J}$.

Correspondingly, with the sequencing decisions, the patient at position $n \in \{1,\ldots,N\}$ has 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:

$$
\begin{aligned}
&\inf\ \alpha \\
&\text{s.t. } \nu_n + \frac{1}{\alpha}\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w}_n-\nu_n)^+\right) \le \tau_n,\ \ n\in\mathcal{G}_{h-1}, \\
&\quad \nu_n + \frac{1}{\alpha_i}\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w}_n-\nu_n)^+\right) \le \tau_n,\ n\in\mathcal{I}_i, i\in\{1,...,h-1\}, \\
&\quad \tilde{w}_n = \max\left\{0,\tilde{t}_{n-1},\ldots,\sum_{k=1}^{n-1}\tilde{t}_k\right\}, \quad n\in\{2,...,N+1\}, \\
&\quad \tilde{t}_n = \sum_{j=1}^{J}\tilde{s}_{nj}y_{nj} - (x_{n+1}-x_n), \qquad n\in\{1,...,N\}, \\
&\quad \alpha\in(0,1], \\
&\quad (\boldsymbol{\tau},\boldsymbol{x},\boldsymbol{Y})\in\mathcal{Y},
\end{aligned}
\tag{11}
$$

in which

$$
\mathcal{Y} = \left\{ (\boldsymbol{\tau},\boldsymbol{x},\boldsymbol{Y}) \left| \begin{array}{ll}
\displaystyle\sum_{j=1}^{J}\beta_j y_{nj} = \tau_n, & n\in\{1,...,N\}, \\
\displaystyle\sum_{n=1}^{N} y_{nj} = N_j, & j\in\{1,...,J\}, \\
\displaystyle\sum_{j=1}^{J} y_{nj} = 1, & n\in\{1,...,N\}, \\
y_{nj}\in\{0,1\}, & n\in\{1,...,N\}, j\in\{1,...,J\}, \\
\boldsymbol{x}\in\mathcal{X}. &
\end{array} \right. \right\}.
$$

Set $\mathcal{Y}$ guarantees that each patient is assigned to a position, and each position 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 $\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left((\tilde{w}_i-\nu_i)^+\right)$, which is equivalent to

$$
\sup_{\mathbb{P}\in\mathbb{F}} \mathrm{E}_{\mathbb{P}}\left(\max\left\{0,-\nu_n,\sum_{j=1}^{J}\tilde{s}_{n-1,j}y_{n-1,j}-(x_n-x_{n-1})-\nu_n,\ldots,\sum_{k=1}^{n-1}\left(\sum_{j=1}^{J}\tilde{s}_{kj}y_{kj}-(x_{k+1}-x_k)\right)-\nu_n\right\}\right) \tag{12}
$$

For known discrete distribution case in which there are M sets of service time, $\left(s^m_{nj}\right)_{n\in\{1,...,N\},j\in\{1,...,J\}}$ with probability $p_m, m\in\{1,\ldots,M\}$, Problem (12) can be formulated as

$$
\sum_{m=1}^{M} p_m \max\left\{0,-\nu_n,\sum_{j=1}^{J}s^m_{n-1,j}y_{n-1,j}-(x_n-x_{n-1})-\nu_n,\ldots,\sum_{k=1}^{n-1}\left(\sum_{j=1}^{J}s^m_{kj}y_{kj}-(x_{k+1}-x_k)\right)-\nu_n\right\}.
$$

By adding decision variables $q_{mn}, n \in \{2,\ldots,N+1\}, m \in \{1,\ldots,M\}$, Problem (11) is equivalent to

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

Similarly, binary search algorithm is used for finding 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.

To obtain an amicably tractable robust optimization model, we assume that the uncertain service times $\tilde{s}_{1j},\ldots,\tilde{s}_{Nj}$ are respectively affinely dependent on a set of factors, $\tilde{z}_1,\ldots,\tilde{z}_N$ for all patient types $j \in \{1,\ldots,J\}$. Moreover, 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 $n \in \{1,\ldots,N\}$ and $j \in \{1,\ldots,J\}$. Furthermore, the factors have the same support and its distributional uncertainty set is given as follows:

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

With this linear formulation, Problem (12) is written as

$$
\begin{aligned}
\sup_{\mathbb{P}\in\mathbb{F}_z} \mathrm{E}_{\mathbb{P}} \Bigg(\max \Bigg\{ & 0, -\nu_n, \sum_{j=1}^{J} (\tilde{z}_{n-1}\sigma_j + \mu_j) y_{n-1,j} - (x_n - x_{n-1}) - \nu_n, \ldots, \\
& \sum_{k=1}^{n-1} \left(\sum_{j=1}^{J} (\tilde{z}_k \sigma_j + \mu_j) y_{kj} - (x_{k+1} - x_k) \right) - \nu_n \Bigg\} \Bigg),
\end{aligned}
\tag{13}
$$

PROPOSITION 4 *For any fixed decisions* $(\boldsymbol{\tau}, \boldsymbol{x}, \boldsymbol{Y}) \in \mathcal{Y}$ *and* $n \in \{2, \ldots, N+1\}$*, Problem (13) corresponds to the optimal value of the following linear optimization problem*

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

Proof. The proof is similar to that of Proposition 3 and hence omitted. □

Henceforth, Problem (11) is equivalent to

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

Given $\alpha \in (0,1]$, the sequencing and scheduling problem reduces to check 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 Study

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 appointment scheduling model under distributional ambiguity. In the the third study, we solve a sequencing and scheduling problem for two patient types and provide some practical insights. The program is coded in python and run on a Intel Core i7 PC with a 3.40 GHz CPU by calling CPLEX 12 as ILP solver.

6.1. Comparison of quality measures

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

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

We consider the case of scheduling seven homogeneous patients who have the same delay thresholds. We assume patients' consultation times are independent and identically distributed with two-point distributions. Hence, we have a number $2^7 = 256$ of scenarios, which could allow us to enumerate all possible realizations, and calculate the exact optimal scheduling decisions. We first study in detail an instance and analyze the performance by varying patients' and doctor's delay thresholds. Afterwards, we randomly generate 100 instances and investigate their average performances. For each instance, we (a) generate the corresponding parameters for two-point distributions, (b) enumerate all 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, two-point distribution is specified with realizations 1 and 4, and mean as 2. Total session length is 16. We obtain the scheduling decisions with different thresholds in Table 1. We consider four performance measures: expected delay, frequency of delay over the threshold,

	TED	**L-DUM** (τ_p, τ_d)[1]					
		(1.5, 1.5)	(2, 2)	(2.5, 2.5)	(3, 3)	(3.5, 3.5)	(4, 4)
Patient 1	0	0	0	0	0	0	0
Patient 2	1	1	1	1	1	1	1
Patient 3	5	3.37	3.37	3.18	2.94	2.74	2.72
Patient 4	9	5.79	5.77	5.68	5.76	5.83	5.57
Patient 5	10	8.38	8.38	8.32	8.17	8.01	8.09
Patient 6	14	10.88	10.88	10.84	10.86	10.88	10.82
Patient 7	15	13.47	13.47	13.45	13.34	13.23	14.82

[1] τ_p: patients' delay threshold; τ_d: doctor's delay threshold.

Table 1 Patients' optimal appointment time under two scheduling methods.

standard deviation of delay, and expected delay over the threshold.

Table 2 summarizes the delay performance of the worst-off participants (including all patients and the doctor). Since the findings are similar, for convenience and clarity, we report the numerical performance for the case with patients' and doctor's threshold taking the value of two. In terms of

	Delay performance of the worst-off participants				**Total expected delays**
	Expected delay	Frequency of delay over the threshold[1]	Standard deviation of delay	Expected delay over the threshold[2]	
L-DUM(1.5,1.5)[3]	1.24	56%	1.74	0.57	8.43
TED	2.40	61%	2.26	1.48	6.74
L-DUM(2,2)	1.25	33%	1.73	0.44	8.44
TED	2.40	61%	2.26	1.17	6.74
L-DUM(2.5,2.5)	1.34	33%	1.72	0.32	8.57
TED	2.40	61%	2.26	0.86	6.74
L-DUM(3,3)	1.48	26%	1.71	0.24	8.65
TED	2.40	17%	2.26	0.56	6.74
L-DUM(3.5,3.5)	1.59	11%	1.74	0.20	8.74
TED	2.40	17%	2.26	0.47	6.74
L-DUM(4,4)	1.60	11%	1.81	0.14	8.36
TED	2.40	17%	2.26	0.39	6.74

[1] Frequency of delay over the threshold: $\mathbb{P}(\tilde{w} > \tau)$;
[2] Expected delay over the threshold: $\mathrm{E}_{\mathbb{P}}\left((\tilde{w}-\tau)^{+}\right)$;
[3] L-DUM(τ_p, τ_d).

Table 2 Delay performance under two scheduling methods (two-point).

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. The maximal average delay reduces from 2.40 to 1.25, and the frequency of delay over the threshold improves from 61% to 33%.

Thenceforth, we study the average performance of 100 randomly generated instances. The parameters determining the two-point distribution $\tilde{s}$ are specified as $\underline{s} = 3\varphi_1$, $\overline{s} = 3 + 5\varphi_2$, and $\mathbb{P}(\tilde{s} = \overline{s}) = 0.5\varphi_3$, where $\varphi_1, \varphi_2, \varphi_3$ are iid uniformly distributed, $U(0,1)$. The average service time, μ is therefore determined. Total session length is $L = 6\mu + \overline{s}$. The delay thresholds are set to three levels, namely, low, medium, and high, where $\tau_d(\text{low}) = \tau_p(\text{low}) = \underline{s}$, $\tau_d(\text{medium}) = \tau_p(\text{medium}) = \mu$, and $\tau_d(\text{high}) = \tau_p(\text{high}) = \overline{s}$. For each instance, we calculate the delay performance of the worst-off participants under the L-DUM model, and normalize it by the corresponding performance in the TED model. We summarize the average ratio in Table 3. The values less than one favor L-DUM model.

Threshold level	**Delay performance of the worst-off participants**				**Total expected delays**
	Expected delay	Frequency of delay over the threshold	Standard deviation of delay	Expected delay over the threshold	
Low	0.6813	0.8162	0.8494	0.4794	1.3134
Medium	0.6352	0.6185	0.8464	0.2892	1.31
High	0.7753	0.1886	0.8676	0.0867	1.2956

Table 3 Average performance analysis of two scheduling methods among 100 instances.

We also test our model using the empirical consultation data collected from the clinics in a local hospital in Singapore from March to May, 2012. The historical data during March and April (802 samples) is considered as the information to make scheduling decisions, while data in May (435 samples) is used for performance testing. The statistics of consultation time are summarized in Table 4.

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

Table 4 Statistics of consultation time from empirical data.

Our appointment design problem is to schedule ten patients within 150 minutes session length. The performance derived with similar procedures is listed in Table 5, which also manifests our conclusions for two-point distributions.

	Delay performance of the worst-off participants				**Total expected delays**
	Expected delay	Frequency of delay over the threshold	Standard deviation of delay	Expected delay over the threshold	
L-DUM(15,15)	13.37	37%	17.33	4.61	94.88
TED	24.12	63%	18.57	11.21	66.65
L-DUM(25,25)	14.45	16%	17.29	2.95	98.60
TED	24.12	35%	18.57	6.51	66.65
L-DUM(35,35)	15.07	9%	17.25	1.81	107.09
TED	24.12	19%	18.57	3.60	66.65

Table 5 Delay performance under two scheduling decisions (empirical data).

In general, compared with the TED method, the L-DUM model 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 ones under three scheduling decisions. The first two are derived by both stochastic optimization approach and distributionally robust optimization approach in the L-DUM model. Sampling average approximation is employed for stochastic optimization approach, and the information of bound support, mean, and mean absolute deviation for robust optimization approach is calculated accordingly. The third scheduling decision is derived from the TED method by using sampling average approximation scheme. Total session length is 7. We consider two types of distributions: uniform distribution $U(0,2)$ and beta distribution $3 \times \text{Beta}(2,4)$. Sample size for the L-DUM model and the TED model is 500 and 2000, respectively. The delay performance is listed in Table 6.

Distribution	**Approach**	**Delay performance of the worst-off participants**				**Total expected delays**
		Expected delay	Frequency of delay over the threshold	Standard deviation of delay	Expected delay over the threshold	
Uniform	L-DUMs(1.2,1.2)[1]	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	1.54	64%	0.82	0.52	3.46
	L-DUMs(1.4,1.4)	0.99	29%	0.87	0.16	5.84
	L-DUMr(1.4,1.4)	1.02	31%	0.89	0.19	6.26
	TED	1.55	55%	0.83	0.41	3.46
	L-DUMs(1.6,1.6)	0.95	21%	0.86	0.12	5.75
	L-DUMr(1.6,1.6)	1.12	28%	0.91	0.17	6.53
	TED	1.54	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	1.47	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	1.46	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	1.46	39%	0.79	0.26	3.18

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

Table 6 Delay performance under three scheduling decisions.

We observe the performance between stochastic optimization approach and robust optimization approach in the L-DUM model is 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 worth mentioning that the computation time for distributional robust optimization approach is relatively short. To solve each minimization problem, stochastic optimization approach requires 44 seconds, while distributional robust optimization approach only requires 8 seconds.

6.3. A sequencing and scheduling example

We also investigate the sequencing and scheduling problem with heterogeneous patients. Our program could easily solve a 10 patients' sequencing and scheduling problem. 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 repeated patients. Their demographics are shown in Table 7, and the information of mean absolute deviation is given as, for $i < k, i, k \in \{1, \ldots, N\}$,

$$\epsilon_{ik} = \begin{cases} 1.4, \forall\, i = k-1, \\ 1.7, \forall\, i = k-2, \\ 2.0, \forall\, i = k-3. \end{cases}$$

The sequencing and scheduling decisions are illustrated in Figure 6.3. For decades, researchers have debated whether to first schedule repeated patients (smallest variance), or new ones (largest

Type	N_j	μ_j	σ_j	$[\underline{z}, \overline{z}]$
New patient ($j=1$)	1	2.5	2.4	[-0.75,1.25]
Repeated patient ($j=2$)	3	1	0.8	[-0.75,1.25]

Table 7 Characterization of heterogeneous patients.

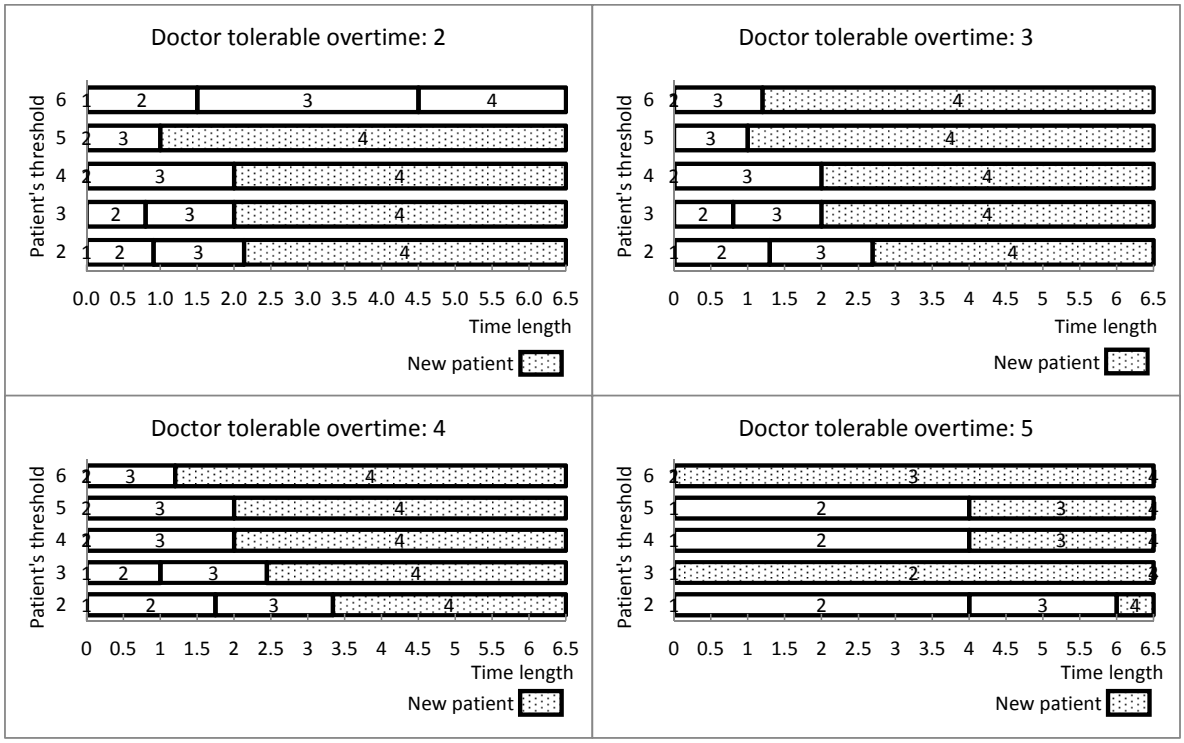

Figure 1 Sequencing and scheduling decisions with various tolerances.

variance). Our computational study actually suggests such universal rule may not be optimal, and the decisions may differ as participants' tolerable thresholds vary. For instance, as shown in the first graph of Figure 6.3, we generally observe that if the doctor's tolerance threshold is low, his/her delay can better be mitigated under L-DUM model if new patient, who may have longer and more uncertain consultation times, is scheduled first. On the other hand, if patients' waiting tolerance is low, for example, in Pediatrics clinic, the L-DUM method will arrange the new patient to arrive at the last position, such that his/her uncertain consultation time will not influence other patients' waiting as they are scheduled to arrive earlier.

7. Conclusions

This paper proposes a new quality measure named Delay Unpleasantness Measure (DUM) to describe individual's unsatisfaction attitude towards a waiting process, and then lexicographically minimizes the worst DUM to mitigate the delay and unfairness in the appointment system. The contributions of our paper stem from three key aspects:

Firstly, we develop the quality measure DUM to describe individual participant's behavior towards delay process. By taking each participant's tolerance threshold as an exogenous factor, DUM could not only provide an upper bound for the frequency of delay over a threshold, but also account for its intensity.

Secondly, we introduce lexicographic min-max concept to address the issue of fairness in the appointment system. As far as we are aware, this paper is the first analytical paper taking the fairness subject as the principle aim. Our model allows the decision maker of the appointment system to adjust participants' thresholds based on their needs and in accordance to their service times.

Thirdly, we provide formulation and solution techniques to encompass different information of uncertain service times. When the distributional information is completely known or with historical data, stochastic optimization approach is suggested for solving the problem. In our distributional uncertainty set, apart from support, and mean, we suggest using mean absolute deviation as descriptive statistics, which could capture the correlation and retain linearity of the nominal problem. The computational study suggests that even if distributions are known, the robust formulations, which are computationally more efficient, can be calibrated to provide competitive solutions to the stochastic programming problem.

Endnotes

1. The survey is crafted through Qualtrics and asks for respondents' preference. Option A: Waiting time of 3 hours (100% chance). Option B: Waiting time of 1 min (50% chance), 2 mins (25%), 4 mins (12.5%) ... and 2^{n-1} mins ($1/2^n$ chance).
2. Expected waiting time for Option A is 3 hours, and expected waiting time for Option B is infinity.
3. In the context of earlier example provided by Klassen and Rohleder (1996), if each patient's tolerable threshold is 3 minutes, the number of patients whose DUMs equal to 1 is 20 to the strategy that keeps only one patient waiting for 40 minutes, while that to the other strategy is 0.
4. According to data collection of Harper and Gamlin (2003) and Zhu et al. (2011), majority of patients arrive earlier than they are expected. This assumption avoids the complexity of modeling due to potential change in sequence.

Acknowledgments

The authors would like to acknowledge the assistance of Chew San Lee, Natalie from National University of Singapore for conducting survey research and provide suggestions for this paper. The authors would also like to thank Joe Sim (CEO, National University of Hospital, Singapore) for the support of this project and Heidi RAFMAN (Deputy Director, NUHS Way & Service Culture Unit) for providing the data.

References

Bailey NTJ (1952) A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Royal Statistical Society* 14:185–199.

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

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

Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. *Productions and Operations Management* 12:519–549.

Chen W, Sim M (2009) GoalDriven Optimization. *Operations Research* 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 Technology and Quantitative Management* 7:321–336.

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

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

Green L, Savin S (2008) Reducing delays for medical appointments: a queuing approach. *Operations Research* 56:1526–1538.

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

Gupta D, Denton B (2008) Appointment scheduling in health care: challenges and opportunities. *IIE Transactions* 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 Science* 54:565–572.

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

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

Isii K (1963) On the sharpness of Chebyshev-type inequalities. *Annals of the Institute of Statistical Mathematics* 12:185–197.

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

Kong QX, Lee CY, Teo CP, Zheng ZC (2012) Scheduling arrivals to stochastic service delivery system using copositive cones. *To appear in Operations Research*.

Mak HY, Rong Y, Zhang J (2012) Appointment ccheduling with limited distributional information. *Working paper.*

Mittal S, Stiller S (2011) Robust appointment scheduling. *Working paper.*

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

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, UK. 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 Journal on Optimization* 17:969–996.

Ogryczak W, Pióro M, Tomaszewski, A (2005) Telecomminications network design and max-min optimization problem. *Journal of Telecommunications and Information Technology* 3:43–56.

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

Rockafellar RT (2007) Coherent approaches to risk in optimization under uncertainty. *Tutorials in Operations Research, INFORMS.*

Rockafellar RT, Uryasev SP (2000) Optimization of Conditional Value-at-risk. *The Journal of Risk* 2:21–41.

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

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

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

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

Wang PP (1999) Sequencing and scheduling N customers for a stochastic server. *European Journal of Operational Research* 119:729–738.

Weiss EN (1990) Models for determining estimated start times and case orderings in hospital operation rooms. *IIE Transactions* 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. *Operations Research* 57:1155–1168.

Zhu ZC, Heng BH, Teow KL (2011) Reducing consultation waiting time and overtime in outpatient clinic: Challenges and solutions. *Chapter 11 in Management Engineering for Effective Healthcare Delivery: Principles and Applications* 229–245.