Appointment games with unobservable and observable schedules

Matthias Deceuninck Stijn De Vuyst Dieter Claeys Dieter Fiems*

Abstract

Appointment scheduling assigns start times in a session or consultation block to a set of tasks that share a common resource. For a generic repeated appointment scheduling problem, we study the trade-off between waiting for an appointment and waiting at the appointed time. Assuming that being scheduled later during a session implies that one has to wait longer for service (on average), it is often beneficial to choose a consultation block further away. We study this trade-off both when the patients have no information on how many patients are already scheduled in future consultation blocks and when they can observe the future block schedules. By some numerical examples, we find that in both cases the rational choice considerably changes between consecutive appointment blocks, patients favouring later blocks when the next appointment block starts in the near future. We also compare the rational choice with the socially optimal schedule and find that socially optimal scheduling can significantly reduce the waiting cost.

Keywords: Appointment scheduling; Game theory; Wardrop equilibrium.

1 Introduction

Many primary care practices are struggling to maintain timely access to medical care. Patients often need to schedule routine appointments many months in advance due to overfilled appointment books [27] or experience long delays in the waiting room on the day of their appointment [13]. In response to these delays, the so-called open-access (OA) policy was introduced as a more patient-centred method [23]. By offering same-day appointments to patients, OA scheduling promotes timely access to care and often improves the clinic's operational efficiency since no-show rates likely decrease when the time till the appointment becomes shorter [25]. Clinics no longer have to use strategies such as appointment overbooking which mitigate the negative effects of no-shows but in turn increase the in-clinic waiting time of patients [28].

Despite these clear benefits, empirical evidence suggests however that implementing OA scheduling does not always lead to higher patient satisfaction levels [22, 29, 30]. Some patients favour scheduling an appointment at a convenient time over a same-day appointment and taking such preferences into account is an important feature of a good appointment system. Recently, Liu et al. [19] conducted four discrete choice experiments to better understand how patients value 'operational' attributes such

*email: Dieter.Fiems@UGent.be

as appointment delay and flexibility in the primary care set. They found that the average utility loss of a 3-day delay is comparable with 30 minutes of physical waiting in the clinic or alternatively, with giving patients more choice. Patients thus take a variety of factors into account when scheduling an appointment.

In this paper, we study the overall consequences of providing more flexibility to patients regarding their appointment time. It is easy to imagine that giving patients more freedom to choose results in a suboptimal allocation. In particular, we consider a setting that operates under appointment scheduling. During the day, patients call in to schedule an appointment in one of the future *consultation blocks*. We assume that patients make a rational choice by comparing two costs. First, there is a cost proportional to the *appointment delay*. This is the time between making the appointment and the start of the chosen consultation block. This cost is also called the indirect waiting time or access time [37]. Secondly, there is a cost proportional to the number of available appointment slots in the consultation block, which we will refer to as the *appointment flexibility*. When more appointment slots have been booked in a consultation block, there are less desirable visit times left for a patient to choose from. Hence, there indeed is a trade-off: by choosing a later block, one experiences a larger appointment delay but one can be scheduled at a more convenient time within the consultation block.

1.1 Contribution

This paper provides a framework for studying appointment scheduling in a game theoretical framework where patients rationally choose an appointment time in one of the future appointment blocks. We consider two different settings: an unobservable game setting where patients do not know how many patients already have been scheduled in the future consultation blocks and an observable game setting, where patients do have full information on the future consultation blocks. For both settings, we study the time-dependent Wardop equilibrium and compare the cost of this rational Wardrop strategy with the cost of the socially optimal allocation which minimises the total cost of all patients in a consultation block.

1.2 Related work

Although the literature on appointment scheduling is vast, only a few papers explicitly take into account patient choice behaviour. As Gupta and Denton [11] indicate, most of the existing works focus solely on the in-clinic waiting times to measure patient satisfaction. Exceptions include the contributions [7, 12, 33, 35]. These studies consider the effect of patient-choice on a clinic capacity management problem. The objective is to offer appointment slots in such a way that the revenue is maximised, given that patients have preferences for certain slots. The trade-off between choosing the appointment on two time scales (day vs. time) has also been considered in other works. Patrick et al. [26] study the problem of dynamically allocating available capacity to incoming demand in a cost-effective manner while maintaining appointment delay targets. Luo et al. [20] developed a queueing model for an appointment-based service system that consists of two queues in tandem: an appointment queue to model the appointment delay and a service queue to model the in-clinic waiting time. In Zacharias and Armony [36], an analytical model is proposed to solve the joint problem of appointment scheduling and determining the panel size.

The appointment game at hand also somewhat relates to sequencing games [4]. A sequencing game

is a cooperative game in which the players aim at finding the optimal order in which they are served, assuming they have all arrived at the start of the game. In Curiel et al. [4], the players have a fixed service time and each incur a cost linear in their queueing delay. In case of identical service times, fairness properties of splitting the cooperation gains according to the Shapley value are discussed by Maniquet et al. [21], while for example Chun et al. [2] study the fairness of more general rules for allocating money to the players. In the 'concert' game of Jain et al. [16], players have independent and identically distributed service times and can choose their own arrival time distribution so as to be served the soonest and with minimal waiting time. The equilibrium arrival profile and Price of Anarchy (PoA) are obtained in the fluid limit for the number of players, both in case of identical players and for the multi-class case where players have different costs. Some authors have studied appointment problems apart from sequencing games. Chun et al. [3] for example use concepts from cooperative game theory to divide the travelling cost of a salesman among his appointments.

In contrast to sequencing games where a finite number of players form a queue, queueing games consider optimal strategies for infinite populations of customers arriving to a queueing system in steady-state [14]. In the seminal model of Naor [24], the customers receive a reward for joining an M/M/1 queue but incur a loss proportional to their queueing delay. Their possible strategies are then to join the queue or not. This is extended by Wang and Zhang [34] to a queue with server breakdowns and delayed repair. Haviv and Roughgarden [15] obtain bounds on the PoA in an exponential multiserver system where the customers' strategies are to choose one of the differently-rated servers without observing the queue lengths. Not only customers are players in queueing games. Instead of games between customers, the queueing game can also be played between customers and the service provider, such as on a market between sellers and buyers [6, 31]. Finally, the time-dependent Wardrop equilibrium at hand is also studied for some queueing systems. In [8, 9] a network of processor sharing queues is studied in the context of a rush-hour traffic problem in the fluid limit regime.

1.3 Outline of the paper

The remainder of the paper is organised as follows. The mean waiting times of the consecutive patients in a consultation block are studied in the next section. Given these mean waiting times, section 3 introduces the unobservable appointment game and its solution. In section 4, we then compare the rational solution with the socially optimal solution by some numerical examples. In addition, section 5 studies an alternative appointment game where patients can observe the number of patients that have already been scheduled in the different future appointment blocks. Finally, we summarise our results, discuss some extensions and draw conclusions in section 6.

2 Mathematical model

We consider a sequence of equally spaced consultation blocks in time, let T denote the time between the start of consecutive blocks. In other words, the schedule is assumed to be cyclic [17]. Patients call in accordance with a Poisson process with fixed rate λ to schedule an appointment in one of the P future consultation blocks. We hereby assume that patients always choose the appointment block that minimises the combined cost of appointment delay and flexibility. Furthermore, we initially assume that patients cannot observe how many patients are already scheduled in any of the future consultation blocks and

thus first choose their consultation block, e.g. Monday afternoon or Tuesday morning, before they are assigned to an appointment time during that block. In section 5, the converse case where calling patients observe the appointment blocks is considered.

The cost of the appointment flexibility can incorporate various factors such as desirable visit times and expected waiting times, as long as its expectation only depends on the number of already scheduled patients and is increasing with an increasing number of occupied slots. To ease the exposition however, we will make the assumption that the cost of appointment flexibility only consists of the expected waiting times at the appointment time. We further assume that patients are scheduled at equidistant points in time with distance Δ and that patient service times constitute a sequence of independent and identically distributed random variables. These assumptions guarantee that the expected waiting time of the patients within each consultation block increases with the position of the patient in the block. The cost function is thus strictly increasing and patients will always choose the earliest available appointment time within each consultation block.

2.1 Waiting time analysis

We consider a generic consultation block. Patients arrive at equidistant points in time and are served in order of arrival. Without loss of generality, we assume that the first patient's appointment is at time 0. The kth patient is then scheduled at time $(k-1)\Delta$, where Δ denotes the time between the appointment times of consecutive patients. The service times of the consecutive patients constitute a sequence of independent and identically distributed random variables. Let S_k denote the service time of the kth patient, and let $s(t) = P[S_k \leq t]$ and $\bar{s}$ denote its distribution function and mean value, respectively.

We now calculate the mean waiting times for the consecutive patients; let W_k denote the waiting time of the kth patient, and let $w_k(t) = P[W_k \leq t]$ and $\bar{w}_k$ be its distribution function and expected value, respectively. We then have $W_1 = 0$ and the following Lindley equation,

$$W_{k+1} = (W_k + S_k - \Delta)^+ .$$

We can then express the distribution of the $k+1$st waiting time in terms of the distribution of the kth waiting time as follows,

$$\begin{aligned} w_{k+1}(t) &= P[(W_k + S_k - \Delta)^+ \leq t] \\ &= \int_0^\infty P[(W_k + u - \Delta)^+ \leq t] ds(u) \\ &= \int_0^{t+\Delta} w_k(t + \Delta - u)\, ds(u) , \end{aligned}$$

with $w_1(t) = 1$ for $t \geq 0$, and zero elsewhere. The mean value of the waiting time of the $k+1$st patient equals,

$$\bar{w}_{k+1} = \int_0^\infty (1 - w_{k+1}(t)) dt ,$$

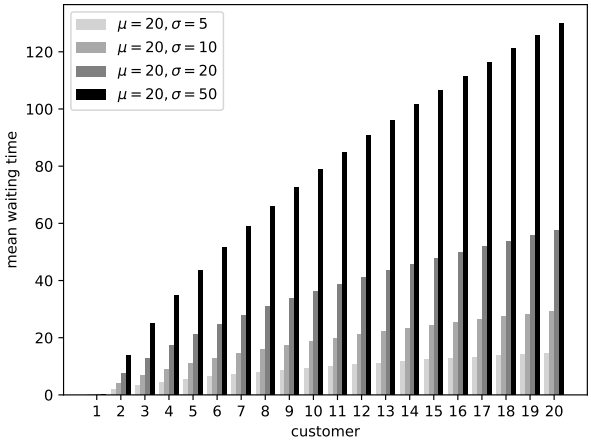


Figure 1: Mean patient waiting time for service times with mean $\mu = 20$, and standard deviation σ as indicated.

or, equivalently,

$$\begin{aligned}
\bar{w}_{k+1} &= \mathsf{E}[(W_k + S_k - \Delta)^+] \\
&= \mathsf{E}[W_k + S_k - \Delta] + \mathsf{E}[(W_k + S_k - \Delta)1_{\{W_k + S_k - \Delta < 0\}}] \\
&= \bar{w}_k + \bar{s} - \Delta - \int_0^{\Delta} \int_0^{\Delta - u} (u + v - \Delta) \mathrm{d}s(v) \mathrm{d}w_k(u) \\
&= \bar{w}_k + \bar{s} - \Delta + \int_0^{\Delta} \mathrm{d}u \int_0^{u} w_k(u - v) \mathrm{d}s(v) \, .
\end{aligned}$$

Here $1_{\{\cdot\}}$ is the indicator function which evaluates to 1 if its argument is true and to 0 if this is not the case. Note that the latter expressions are more convenient for numerically calculating the mean waiting times, as one only needs to determine the waiting time distribution $w_k(t)$ of the preceding patient for $t \leq \Delta$.

Remark 1. One can easily avoid dealing with the integral expressions for the calculation of the mean waiting times by discretising the patient service times. Let $\delta = \Delta/N$ for some $N \in \mathbb{N}$. Now consider random variables $\tilde{S}_n = \lfloor S_n/\delta \rfloor \delta$ and $\hat{S}_n = \lceil S_n/\delta \rceil \delta$ taking values in $\{n\delta, n \in \mathbb{N}\}$, such that $\tilde{S}_n \leq S_n \leq \hat{S}_n$. By Lindley's recursion, the waiting times are increasing functions of the patient service times. Hence if one replaces the distribution of S_n by the distribution of $\tilde{S}_n$ (for $\hat{S}_n$) in the calculations above, one obtains a lower (an upper) bound for the mean waiting time. Moreover, one only needs to calculate the probability mass function of the patient waiting times for a finite number of discrete values, and the integrals simplify to finite sums. See [5] for details on the discretised algorithm and its numerical complexity.

In practice, it is not uncommon to choose the slot length equal to the mean service time, see for example [1]. This choice is illustrated in Figure 1 which depicts the mean waiting times for the consecutive patients, assuming that the service times are gamma distributed random variables. The mean service time and slot length equal $\mu = 20$, while different values of the standard deviation σ of the service

σ	$\mathbf{a}$	$\mathbf{b}$	abs. error	rel. error
5	[37.63, 4.34, 1.24]	[0.015, 0.171, 0.901]	0.077	0.35%
10	[75.56, 8.66, 2.47]	[0.015, 0.169, 0.878]	0.147	0.39%
20	[154.86, 17.48, 4.74]	[0.014, 0.158, 0.806]	0.208	0.40%
50	[367.30, 35.34, 5.54]	[0.015, 0.164, 0.729]	1.01	0.43%

Table 1: Coefficients and accuracy of the 3-term approximation of the waiting times

times are assumed as depicted. It is readily seen that the mean waiting times of the consecutive patients form an increasing sequence. This is in line with queueing theory: one knows that for $\bar{s} < \Delta$ and for finite service time variance, the mean waiting times of the consecutive patients constitute an increasing sequence, converging to a finite limiting value. In contrast, for $\bar{s} \geq \Delta$, the mean waiting times constitute an increasing unbounded sequence. In addition, an increase of the standard deviation σ leads to longer waiting times.

For the remainder, it is convenient to be able to express $\bar{w}_n$ explicitly in terms of n. To this end, we introduce the following representation

$$\hat{w}_n \doteq \sum_{m=1}^{M} \left(a_m - a_m \exp(-b_m(n-1))\right), \tag{1}$$

for some constant M, and where $\mathbf{a} = [a_1, \ldots, a_M]$ and $\mathbf{b} = [b_1, \ldots, b_M]$ are known vectors. It will become evident later that this exact form greatly simplifies the cost calculations in the game below. Moreover, some experimentation — we consider the mean waiting times of Figure 1, for the different σ — shows that this alternative characterisation hardly introduces errors. In table 1, we list the vectors $\mathbf{a}$ and $\mathbf{b}$ that minimise the Euclidean distance between $[\bar{w}_n]_{n=1}^{N}$ and $[\hat{w}_n]_{n=1}^{N}$ for $N = 50$ patients, assuming $M = 3$ terms in the approximation. We relied on the basin-hopping algorithm to find the optimal parameters [32]. The reported absolute and relative errors are defined as $\sup_{1 \leq n \leq N} |\bar{w}_n - \hat{w}_n|$ and $\sup_{1 < n \leq N} |1 - \hat{w}_n / \bar{w}_n|$, respectively.

Remark 2. In fact, the representation (1) allows for arbitrarily small absolute errors for the first N patients, by including more terms in the representation. To this end, consider a continuous function $\phi(x)$ such that $\phi(n) = w_{n+1}$ for $n \in \{0, 1, \ldots, N-1\}$. From Theorem 1 of [10], this function can be approximated arbitrarily well in the interval $[0, N-1]$ by a sum of exponential functions. In particular, for every $\epsilon > 0$, there exist M, a_m and b_m such that

$$\left| \phi(x) - \sum_{m=1}^{M} a_m \exp(-b_n x) \right| < \epsilon \,,$$

for $x \in [0, N-1]$. We therefore have

$$\left| w_n - \sum_{m=1}^{M} a_m \exp(-b_n(n-1)) \right| < \epsilon \,.$$

Moreover, as $w_0 = 0$, we have

$$|w_n - \hat{w}_n| = \left| w_n - w_0 - \sum_{m=1}^{M} a_m \exp(-b_n(n-1)) + \sum_{m=1}^{M} a_m \right|$$
$$\leq \left| w_n - \sum_{m=1}^{M} a_m \exp(-b_n(n-1)) \right| + \left| w_0 - \sum_{m=1}^{M} a_m \right| \leq 2\epsilon\,.$$

As ϵ can be chosen freely, the approximation error can be made arbitrarily small.

3 Appointment game

We now use the waiting time calculations of the preceding section to find the optimal scheduling policy. We here assume that new appointments are made in accordance with a Poisson process with rate λ in any of the first P future appointment blocks. We refer to P as the scheduling horizon.

Recall that T denotes the time between consecutive appointment blocks. It suffices to consider the arrival process in an interval of length T, prior to an appointment block. Let $\lambda_i(t), t \in (0, T]$ denote the arrival rate of patients that choose the ith future consultation block. We obviously have

$$\sum_{i=1}^{P} \lambda_i(t) = \lambda\,, \tag{2}$$

as patients will always make an appointment. Clearly, assuming that the $\lambda_i(t)$'s are fixed, the arrival stream to a single consultation block (say, starting at $t = 0$) is a non-homogeneous Poisson process with rate,

$$\alpha(t) = \lambda_i(t + iT) \qquad \text{for } t \in (-iT, -(i-1)T] \text{ and } i \in \{1, 2, \ldots, P\},$$

and $\alpha(t) = 0$ for $t < -PT$. Hence, at time t prior to the consultation block at time 0, the number of patients that have joined this block is Poisson distributed with mean,

$$\beta(t) = \int_{-PT}^{t} \alpha(u)du\,, \tag{3}$$

for $-PT < t \leq 0$. In other words, the chance that n patients have joined the consultation block prior to time t is,

$$\gamma_n(t) = \frac{\beta(t)^n}{n!} e^{-\beta(t)}\,,$$

for $-PT < t \leq 0$.

We are now ready to express the total cost of joining this consultation block at time t prior to the block. First there is the cost of having a future appointment:

$$C_1(t) = -c_1 t \tag{4}$$

for some constant $c_1 > 0$ where $-t$ is the time till the consultation block at time 0. Secondly, there is the cost of waiting during the appointment. At time t, there are already n patients scheduled with probability $\gamma_n(t)$. Hence we have,

$$C_2(t) = \sum_{n=0}^{\infty} \gamma_n(t) \bar{w}_{n+1}\,, \tag{5}$$

where $\bar{w}_n$ is the mean waiting time of the nth patient in the schedule, as calculated in the preceding section. Replacing $\bar{w}_n$ by its approximation $\hat{w}_n$, allows to further simplify the expression above,

$$\begin{aligned} C_2(t) &= \sum_{n=0}^{\infty} \frac{\beta(t)^n}{n!} e^{-\beta(t)} \sum_{m=1}^{M} (a_m - a_m \exp(-b_m n)) \\ &= \sum_{m=1}^{M} a_m \left(1 - \exp\left(-\beta(t)\left(1 - e^{-b_m}\right)\right)\right) . \end{aligned} \tag{6}$$

Now the simplification by the proposed approximation is clear. Indeed, by the approximation there is no need to calculate (and truncate) the infinite sum in (5).

For the formulation of the game equilibrium, it is convenient to split up the total cost function $C(t) = C_1(t) + C_2(t)$ in intervals:

$$\phi_i(t) = C(t - iT) , \tag{7}$$

for $t \in (0, T]$ and $i \in \{1, 2, \ldots, P\}$. Clearly $\phi_i(t)$ depends on $\alpha(t)$ and therefore on $\lambda_i(t)$. To find a rational schedule, patients should choose between the best options only, and never choose inferior options. This means that there exists a function $f(t)$ such that for $i \in \{1, 2, \ldots, P\}$,

$$\begin{aligned} &\phi_i(t) = f(t) \quad \text{for } \lambda_i(t) > 0 ; \\ &\phi_i(t) \geq f(t) \quad \text{for } \lambda_i(t) = 0 . \end{aligned} \tag{8}$$

The first equation corresponds to the notion that for all blocks where patients join with a positive probability, the cost should be equal. The second equation says that one cannot reduce the cost by scheduling to another consultation block. The notion of equilibrium above is similar to that of a Wardrop equilibrium, but explicitly adds time-dependence. One however cannot study this type of equilibrium at a single time-point as past decisions affect the cost at each point in time.

The formulation (8) of the equilibrium above, does not immediately indicate how the equilibrium can be found. To find the equilibrium in practice, we rely on the following iterative approach. For L sufficient large, we consider the set of points $\mathcal{L} = \{\ell T/L : \ell = 1, \ldots, L\} \subset (0, T]$. We initialise the rates $\lambda_i(t) = \lambda/P$, for $i \in \{1, 2, \ldots, P\}$ and $t \in \mathcal{L}$, and then update the values according to the following iterative procedure.

1. Calculate the average cost for $t \in \mathcal{L}$,

$$\nu(t) = \frac{1}{P} \sum_{i=1}^{P} \phi_i(t)$$

2. For all $i \in \{1, 2, \ldots, P\}$ and $t \in \mathcal{L}$, update the rates according to,

$$\tilde{\lambda}_i(t) \leftarrow \lambda_i(t) \exp(-\theta(\phi_i(t) - \nu(t))) .$$

   Hence, the rate to appointment blocks with a higher than average cost is decreased, while the rate to appointment blocks with a lower than average cost is increased.

3. For all $i \in \{1, 2, \ldots, P\}$ and $t \in \mathcal{L}$, normalise the rates,

$$\hat{\lambda}_i(t) \leftarrow \lambda \frac{\tilde{\lambda}_i(t)}{\sum_{k=1}^{P} \tilde{\lambda}_k(t)} .$$

   This step ensures that the sum of the rates to the different blocks is equal to λ.

4. If $\sup_{t\in\mathcal{L}} |\lambda_i(t) - \hat{\lambda}_i(t)| > \epsilon$, set $\lambda_i(t) \leftarrow \hat{\lambda}_i(t)$ and go back to step 1. If not, return the solution $\hat{\lambda}_i(t)$.

The algorithm depends on 2 parameters: θ and ϵ. The value of θ determines how fast the rate is adapted in accordance to the cost difference. Convergence can speed up by increasing this value, but the algorithm may not converge at all if this value is too large. The value of ϵ determines when we are sufficiently close to convergence. In the algorithm, equations (4), (6) and (7) are used to calculate $\phi_i(t)$ for each $t \in \mathcal{L}$. The function $\beta(t)$ in (6) is approximated by the following sum,

$$\beta(t) \approx \frac{T}{L} \sum_{\ell=-PL}^{t/T-1} \alpha(\ell T/L)$$

which only requires one to evaluate $\lambda_i(t)$ for $t \in \mathcal{L}$.

To illustrate our approach, Figure 2 shows the outcome of the game for different values of the standard deviation σ of the service times. The slot length within an appointment block is $\Delta = 20$, and the mean service time is $\mu = 20$ as well. The time between appointment blocks is $T = 1$ (this is e.g. natural if T is expressed in days), and the arrival rate is $\lambda = 20$, such that on average $\lambda T = 20$ patients are served in an appointment block. Moreover, the patients can make appointments in one out of $P = 5$ future blocks. The cost $c_1 = 14$ is chosen such that for $\sigma \in \{5, 10, 20\}$, the limit P does not influence the outcome of the game, while it does for $\sigma = 50$. The figures depict the arrival rate $\alpha(t)$ to the block at time 0, and the corresponding costs $C_1(t)$, $C_2(t)$ and $C(t)$.

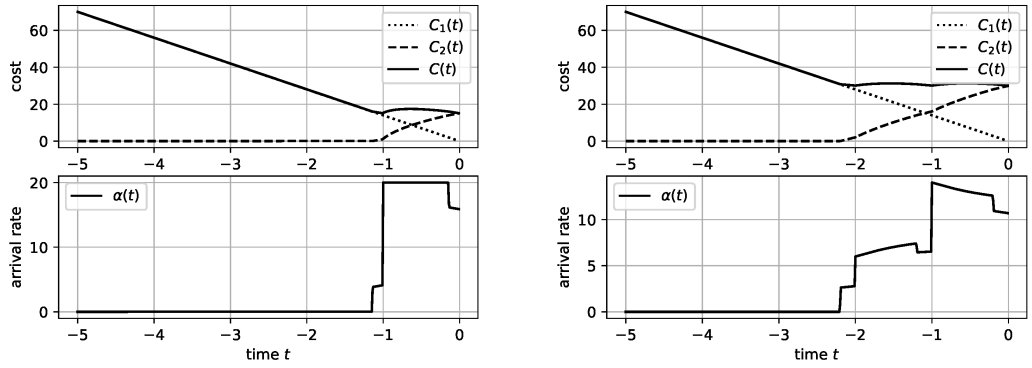
Foremost, it is easily verified from the numerical results that the solution indeed satisfies the constraints (8). These constraints also explain the specific shape of the $\alpha(t)$ curves. For a certain value t, the cost $C_1(t)$ becomes too large and patients are not sent to the block at time 0 anymore. However, this implies that more patients are sent to this block at times $t + T$, $t + 2T$, etc. Similarly, the chance to select a block may either increase or decrease between two consultation blocks, as an increase for one block implies a decrease for another.

We now compare the curves for different σ. As can be seen from Figure 1, larger σ-values translate into larger mean waiting times, hence it is more beneficial to be scheduled at the start of the appointment block. This also means that it is more beneficial to choose an appointment block further in the future. This is indeed observed. For $\sigma = 5$, most patients choose the next block. Only if the next appointment block is about to start, some patients opt for the second block as already many patients will have been scheduled in the first. For $\sigma = 10$, patient mostly choose for the first two blocks, and only some choose for the third block, again when the next block is near. Similar observations hold for $\sigma = 20$. However, for $\sigma = 50$ the limit on the number of future appointment blocks $P = 5$ comes into play. Now the cost of waiting at the appointment is so high, that it is beneficial to schedule to the last possible block, in order to reduce the waiting time at the appointment.

4 Rational vs. social optimum

In the section above, patients could freely choose. It may however be more beneficial if there is a central control. We now compare the rational choice of the preceding section with the socially optimal choice. For a given $\alpha(t)$ (or equivalently, for given $\lambda_i(t)$), the total cost of all patients in a session equals,

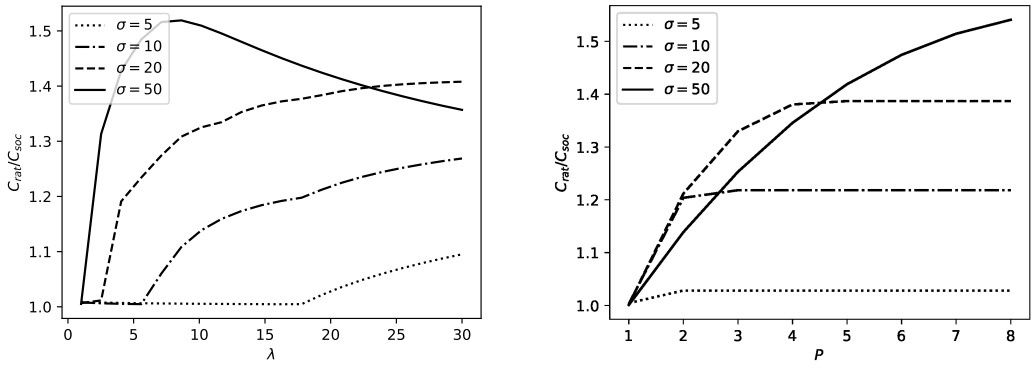
$$C = -\int_{-PT}^{0} \alpha(t) c_1 t dt + \sum_{n=1}^{\infty} \frac{\beta(0)^n}{n!} e^{-\beta(0)} \sum_{k=1}^{n} \bar{w}_k \,, \tag{9}$$



(a) $\sigma = 5$ (b) $\sigma = 10$

(c) $\sigma = 20$ (d) $\sigma = 50$

Figure 2: Arrival rate $\alpha(t)$ and corresponding costs for different values of the service time standard deviation σ as indicated.



(a) effect of λ

(b) effect of P

Figure 3: Price of Anarchy vs. the arrival rate λ (a) and the scheduling horizon P (b).

with $\beta(0) = \int_{-PT}^{0} \alpha(t)dt = \lambda T$. Evaluating this expression for the solution $\alpha(t)$ of the game gives the rational cost C_{rat}.

As the second term in (9) does not depend on $\alpha(t)$, the socially optimal cost can be determined by minimising the first term over the set of admissible $\alpha(t)$. This is the set of non-negative functions that adhere the constraints (2). As the cost grows with $-t$, one immediately finds that making appointments in the first block is socially optimal. The corresponding cost equals,

$$C_{soc} = c_1 \lambda \frac{T^2}{2} + \sum_{n=0}^{\infty} \frac{(\lambda T)^n}{n!} e^{-\lambda T} \sum_{k=1}^{n} \bar{w}_k \,.$$

A measure of inefficiency of individual decision making is the Price of Anarchy [18] which is the ratio of the cost at the game equilibrium to the optimal social cost, $PoA = C_{rat}/C_{soc}$. The Price of Anarchy (PoA) is depicted in figures 3(a) and 3(b) as a function of the arrival rate λ and the scheduling horizon P, respectively. We assume $\mu = \Delta = 20$ as before, and again consider different σ as indicated. For Figure 3(a) , we have $P = 5$, while for Figure 3(b), we have $\lambda = 20$. Both increasing λ and P has a negative effect on the Price of Anarchy, the effect being more pronounced for high σ. For small λ, both the rational and social optimum always schedule to the first available appointment block, such that the PoA is 1. When the rate increases further, individual patients have an incentive to schedule to blocks further way, and the PoA increases. From Figure 3(b), the PoA is 1 for $P = 1$ as all patients then have to make appointments in the first block. Once P increases, individuals have an incentive to make schedules in later appointment blocks and the PoA again increases. There is one notable exception though. For increasing λ, the PoA for $\sigma = 50$ first increases and then decreases again. In this particular case, the waiting times $\bar{w}_k$ are considerable and the scheduling horizon $P = 5$ comes into play (see also fig. 2(d)). The scheduling horizon ensures that the cost C_1 does not increase any further, while the waiting time cost C_2 does increase as there are more patients in every session, leading to a decrease of the PoA.

5 Observable appointment game

In contrast to the preceding sections, we now assume that patients can observe the number of patients in the different blocks upon arrival. Therefore, they can select the block which induces the least cost. In the remainder, to simplify notation, we assume that the kth block starts at time kT, for $k \in \mathbb{N}$. As in the preceding sections, a patient can select any of the P future blocks. We use the following notation to describe the evolution of the appointments: $A(t)$ denotes the number of arrivals up to time t and $N_k(t)$ denotes the number of patients assigned to block k at time t.

For an arrival at time t, with $kT \leq t < (k+1)T$, the cost to schedule to the ℓth future block $k+\ell$ is,

$$\hat{C}_\ell(t) = c_1((k+\ell)T - t) + \bar{w}_{N_{k+\ell}(t)+1}$$

where $N_{k+\ell}(t)$ can be expressed in terms of $N_{k+\ell}(kT)$ as follows,

$$N_{k+\ell}(t) = N_{k+\ell}(kT) + \int_{kT}^{t} 1_{\{\arg\min_h \hat{C}_\ell(t)=\ell\}} dA(t)\,.$$

The second term counts the number of arrivals in $(kT, t]$ which choose block $k+\ell$. These expressions in particular show that the sequence

$$\{(N_k(kT), \ldots, N_{k+P-1}(kT)), k \in \mathbb{N}\}$$

constitutes a discrete-time Markov process. As each appointment block can hold any number of patients, the state space of the Markov process is $\mathbb{N}^P$. One can argue that in practice there is an upper bound K on the number of patients that can be assigned to the same block. However, even with this assumption, the size of the state space $(K+1)^P$ is still prohibitively large to allow for its numerical computation in reasonable time. Not only is the size of the state space very large, even for moderate P, the transition matrix is typically not sparse, and calculating the transition probabilities also requires considerable numerical effort. Therefore, we rely on simulation to assess the observable appointment scheduling game.

To simulate the appointment system, we calculate the state of the appointment blocks as seen by consecutive patients. In particular, the simulation calculates a sequence $\{(t_k, n_k), k = 1, \ldots, K\}$ where t_k denotes the waiting time till the appointment block of the kth patient, and n_k denotes the position of the kth patient in its appointment block. The simulation logic is summarised as follows.

1. Initialise:
   (a) Initialise the number of patients in the P future appointment blocks: $N_i \leftarrow 0$ for $i = 1, \ldots, P$; N_i is the number of patients in the ith future appointment block as seen by patients.
   (b) Initialise time: $t \leftarrow 0$ denotes the time since the last appointment block;
   (c) $k \leftarrow 1$ tracks the number of patients that have been scheduled
2. Draw an exponentially distributed inter-arrival time a and calculate the next arrival instant: $t \leftarrow t + a$;
3. While $t > T$:
   (a) shift the appointment blocks $N_i \rightarrow N_{i+1}$ for $i = 1, \ldots, P-1$ and set $N_P = 0$
   (b) $t \leftarrow t - T$

4. Select the appointment block with minimal cost $b \leftarrow \arg\min_{i \le P} c_1(iT - t) + \bar{w}_{N_i+1}$;
5. Add a patient to appointment block b: $N_b \rightarrow N_b + 1$;
6. Store the time till the selected block, $t_k \leftarrow bT - t$, and the position of the patient in the selected block $n_k \leftarrow N_b$;
7. Set $k \leftarrow k + 1$
8. If $k < K$ return to 2.

Having obtained the sequence (t_k, n_k) by simulation, we now calculate the performance measures. Recall that for $t \in [-PT, 0]$, $\alpha(t)$ denotes the fraction of patients that choose the appointment block at time 0 if they arrive at time t. For the observable game, we can approximate $\alpha(t)$ by,

$$\alpha(t) \approx \alpha^\epsilon(t) = \frac{1}{\epsilon} \int_t^{t+\epsilon} \alpha(t) dt$$

and the corresponding cost by,

$$C_2(t) \approx C_2^\epsilon(t) = \frac{\int_t^{t+\epsilon} C_2(t) dt}{\int_t^{t+\epsilon} \alpha(t) dt}.$$

The former approximations replace the fraction of patients at time t that opt for the appointment at time 0 (and the corresponding cost) by the average fraction over a small interval after t, and the corresponding cost averaged over this interval. As the sequence $\{(t_k, n_k), k = 1, \ldots, K\}$ is asymptotically stationary ergodic, we can approximate $\alpha^\epsilon(t)$ and $C_2^\epsilon(t)$ by,

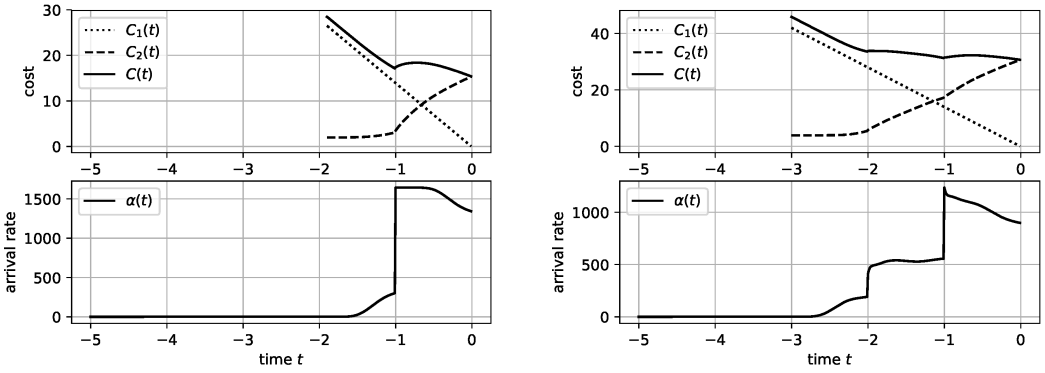
$$\alpha^\epsilon(t) \approx \frac{1}{\epsilon K} \sum_{k=1}^{K} 1_{\{-t < t_k \le -t+\epsilon\}}$$

and,

$$C_2^\epsilon(t) \approx \frac{\sum_{k=1}^{K} 1_{\{-t < t_k \le -t+\epsilon\}} (c\, t_k + \bar{w}_{n_k})}{\sum_{k=1}^{K} 1_{\{-t < t_k \le -t+\epsilon\}}}.$$

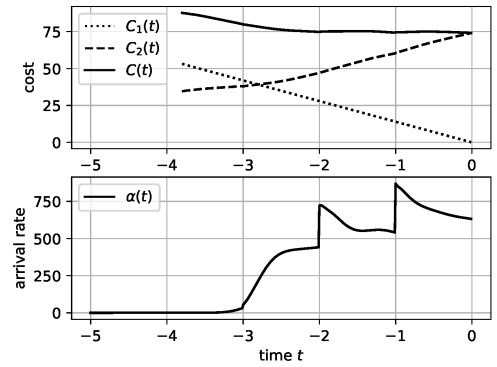
To illustrate our approach, Figure 4 shows the outcome of the game for different values of the standard deviation σ of the service times. We retain the parameters of the unobservable game. That is, the slot length within an appointment block is $\Delta = 20$, and the mean service time is $\mu = 20$ as well, while $T = 1$. We only depict the costs $C_1(t)$, $C_2(t)$ and $C(t)$ in the range where $\alpha(t)$ is sufficiently large. Outside that range, there are but a few observations of the cost, and it is not possible to accurately assess the cost by simulation.

For the interval $[-1, 0]$, the chance to select the block at time 0 decreases with time. The closer to 0, the more likely it is that there are already several patients scheduled, hence it is beneficial to select a later block. For $\sigma = 20$ and $\sigma = 50$ (Figs. 4(c) and 4(d)), the same behaviour is observed for the interval $[-2, -1]$, and can be explained by similar arguments. Finally, for $t < -1$ in Figs. 4(a) and 4(b) and for $t < -2$ in Figs. 4(c) and 4(d), we observe an increase. As for the unobservable game, this can again be explained by the constraints (8). I.e., as the arrival rate is constant, a decrease in $[-2, 0]$ or $[-1, 0]$, must be compensated by an increase in another interval.

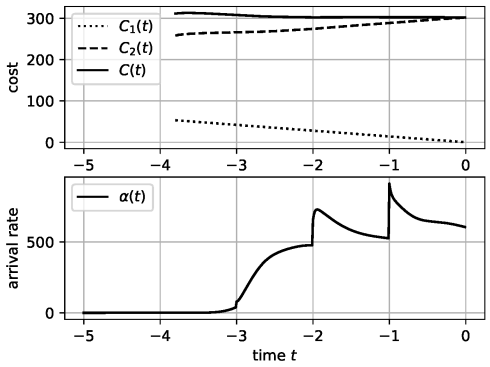


(a) $\sigma = 5$

(b) $\sigma = 10$

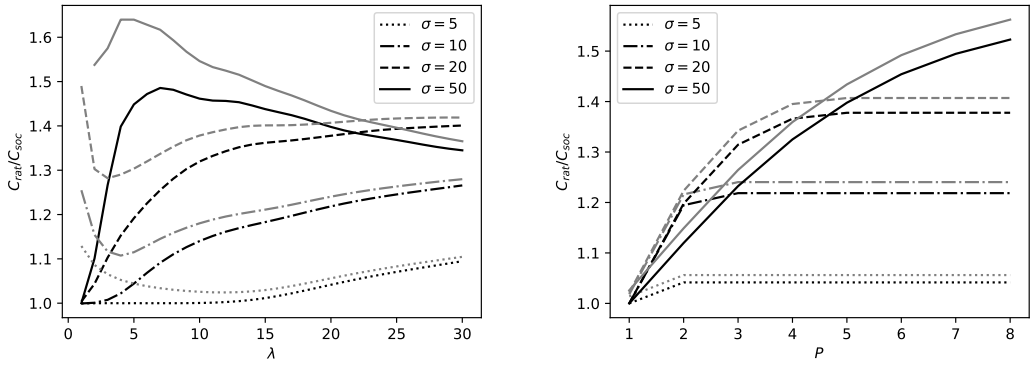


(c) $\sigma = 20$



(d) $\sigma = 50$

Figure 4: Arrival rate $\alpha(t)$ and corresponding costs for different values of the service time standard deviation σ as indicated.



(a) effect of λ

(b) effect of P

Figure 5: Upper (gray) and lower (black) bounds for the Price of Anarchy vs. the arrival rate λ (a) and the scheduling horizon P (b).

A comparison the unobservable and observable games, shows that the outcomes of the games are comparable. While there are a distinct points in time where the patient strategies change for the unobservable game, this is not the case for the observable game. In the latter game, the patients possess additional knowledge, leading to a more refined choice.

Finally Figure 5(a) and 5(b) study the Price of Anarchy for the observable game. Given the sequence (t_k, n_k), we can directly calculate the expected cost of patient k. Averaging over all patients yields the average cost per patient, and multiplying with the average number λT of patients per session yields the average cost per session. Calculating the socially optimal strategy is however not trivial. The socially optimal strategy is the strategy which assigns patients to blocks based on the available information such that the cost per session (or per patient) is minimal. This is a non-trivial control problem. We can however easily retrieve an upper and lower bound for the cost in the socially optimal strategy. Noting that any particular strategy is an upper bound, the cost of assigning to the first block provides an obvious upper bound. For a lower bound, we note that the minimal waiting cost within a session is attained when each session has the same deterministic number of patients. As the mean number of patients per session equals the mean number of arrivals between sessions λT, we find the following lower bound,

$$C_{soc}^{LB} = c_1 \lambda \frac{T^2}{2} + \sum_{k=1}^{\lfloor \lambda T \rfloor} \bar{w}_k \,.$$

The first term corresponds to the cost of waiting till the next session, and the second term is the total waiting time within a session with $\lfloor \lambda T \rfloor$ patients.

The lower and upper bounds of the socially optimal cost translate into upper and lower bounds for the PoA. Both bounds are depicted in Figures 5(a) and 5(b) which study the influence of λ and P on the Price of Anarchy, respectively. From the figures, we immediately observe that the bounds are fairly close for $\lambda > 10$, which corresponds to sessions with 10 patients on average. Moreover, the PoA for the observable and unobservable game is comparable, and therefore the discussion of the effect of λ and P on the PoA is identical to the observable setting.

6 Conclusions

We have studied appointment scheduling when the order of the appointments within an appointment block is first-come-first-served. As during an appointment block, it is beneficial to be scheduled early, the patients have an incentive not to choose the first appointment block. Accounting for the waiting time till the appointment block as well, this leads to the interesting trade-off between waiting till the appointment block and waiting at the appointment. To study this trade-off, we have used game-theoretic concepts to study individually optimal decision making and compare with the socially optimal schedule. We have studied both the case where patients do and do not know how many patients are already scheduled in the future appointment blocks.

The methodology developed in this paper can be easily extended to include more realistic features of appointment scheduling problems. Foremost, while the cost at the appointed time now only includes the waiting cost, a more general cost function can be used and the patients do not have to be seen in order of arrival. For example, in the unobservable game, some later slots may be generally preferred over earlier slots (for example slots after working hours) and patients can be assigned to these slots first. The simulation methodology for the observable case is even easier to extend. For example, patients can have preferred slots and the assignment of a slot can be based on a combination of the cost or utility of these preferences and the expected waiting times.

Secondly, the assumptions on the arrivals and the placement of the appointment blocks can be relaxed. There is no need to have equidistantly spaced consultation blocks. One can e.g. consider cycles of a week with unevenly spaced consultation blocks. Moreover, within such a (weekly) cycle, there is no need to have time-homogeneous Poisson arrivals; the arrival rate can e.g. fluctuate over the days of the week.

Finally, our approach can support the design of specific appointment systems that e.g. aim for evenly spreading the load over the different appointment blocks, or that aim at minimising the appointment delay to reduce the number of no-shows per appointment block. Our game-theoretic findings can then be part of a larger optimisation problem with one of the aforementioned objectives. By predicting the behaviour of rational patients, the optimisation problem can include the reaction of the patients on the particular design, which in turn allows for a more accurate assessment of the performance of the appointment system.

References

[1] T. Cayirli and E. Veral. Outpatient scheduling in health care: a review of literature. *Production and operations management*, 12(4):519–549, 2003.

[2] Y. Chun, M. Mitra, and S. Mutuswami. Egalitarianism in the queueing problem. *Journal of Mathematical Economics*, 81:48–56, 2019.

[3] Y. Chun, N. Park, and D. Yengin. Coincidence of cooperative game theoretic solutions in the appointment problem. *International Journal of Game Theory*, 45(3):699–708, 2016.

[4] I. Curiel, G. Pederzoli, and S. Tijs. Sequencing games. *European Journal of Operational Research*, 40(3):344–351, 1989.

[5] S. De Vuyst, H. Bruneel, and D. Fiems. Computationally efficient evaluation of appointment schedules in health care. *European Journal of Operational Research*, 237(3):1142–1154, 2014.

[6] L.G. Debo, C. Parlour, and U. Rajan. Signaling quality via queues. *Management Science*, 58(5):876–891, May 2012.

[7] J. Feldman, N. Liu, H. Topaloglu, and S. Ziya. Appointment scheduling under patient preference and no-show behavior. *Operations Research*, 62(4):794–811, 2014.

[8] D. Fiems and B. Prabhu. Macroscopic modelling and analysis of rush-hour congestion. In *Proceedings of the 13th EAI International Conference on Performance Evaluation Methodologies and Tools*, 2020.

[9] D. Fiems, B. Prabhu, and K. De Turck. Travel times, rational queueing and the macroscopic fundamental diagram of traffic flow. *Physica A*, 524:412–421, 2019.

[10] M. v. Golitschek. Linear approximation by exponential sums on finite intervals. *Bulletin of the American Mathematical Society*, 81(2), 1975.

[11] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819, 2008.

[12] D. Gupta and L. Wang. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, 56(3):576–592, 2008.

[13] P.R. Harper and H.M. Gamlin. Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *Or Spectrum*, 25(2):207–222, 2003.

[14] R. Hassin and M. Haviv. *To queue or not to queue*. Kluwer, 2009.

[15] M. Haviv and T. Roughgarden. The price of anarchy in an exponential multi-server. *Operations Research Letters*, 35(4):421–426, 2007.

[16] R. Jain, S. Juneja, and N. Shimkin. The concert queueing game: to wait or to be late. *Discrete Event Dynamical Systems*, 21:103–138, 2011.

[17] N. Kortbeek, M. E. Zonderland, A. Braaksma, I. M. H. Vliegen, R. J. Boucherie, N. Litvak, and E. W. Hans. Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Performance Evaluation*, 80:5–26, 2014.

[18] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 404–413. Springer, 1999.

[19] N. Liu, S.R. Finkelstein, M.E. Kruk, and D. Rosenthal. When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Science*, 64(5):1975–1996, 2017.

[20] J. Luo, V.G. Kulkarni, and S. Ziya. A tandem queueing model for an appointment-based service system. *Queueing Systems*, 79(1):53–85, 2015.

[21] F. Maniquet. A characterization of the Shapley value in queueing problems. *Journal of Economic Theory*, 109:90–103, 2003.

[22] A. Mehrotra, L. Keehl-Markowitz, and J.Z. Ayanian. Implementing open-access scheduling of visits in primary care practices: a cautionary tale. *Annals of internal medicine*, 148(12):915–922, 2008.

[23] M. Murray and D.M. Berwick. Advanced access: reducing waiting and delays in primary care. *Jama*, 289(8):1035–1040, 2003.

[24] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, January 1969.

[25] N. Osadchiy and D. KC. Are patients patient? the role of time to appointment in patient flow. *Production and Operations Management*, 26(3):469–490, 2017.

[26] J. Patrick, M.L. Puterman, and M. Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, 56(6):1507–1525, 2008.

[27] D. Polsky, M. Richards, S. Basseyn, D. Wissoker, G.M. Kenney, S. Zuckerman, and K.V. Rhodes. Appointment availability after increases in medicaid payments for primary care. *New England Journal of Medicine*, 372(6):537–545, 2015.

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

[29] K.D. Rose, J.S. Ross, and L.I. Horwitz. Advanced access scheduling outcomes: a systematic review. *Archives of internal medicine*, 171(13):1150–1159, 2011.

[30] F. Sampson, M. Pickin, A. O'Cathain, S. Goodall, and C. Salisbury. Impact of same-day appointments on patient satisfaction with general practice appointment systems. *Br J Gen Pract*, 58(554):641–643, 2008.

[31] D.K. Sundar and K. Ravikumar. An actor-critic algorithm for multi-agent learning in queue-based stochastic games. *Neurocomputing*, 127:258–265, 2014.

[32] D.J. Wales and J.P.K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *Journal of Physical Chemistry A*, 101:5111–5116, 1997.

[33] J. Wang and R.Y.K. Fung. Dynamic appointment scheduling with patient preferences and choices. *Industrial Management & Data Systems*, 115(4):700–717, 2015.

[34] J. Wang and F. Zhang. Equilibrium analysis of the observable queues with balking and delayed repairs. *Applied Mathematics and Computation*, 2018:2716–2729, 2011.

[35] W.-Y. Wang and D. Gupta. Adaptive appointment systems with patient preferences. *Manufacturing & Service Operations Management*, 13(3):373–389, 2011.

[36] C. Zacharias and M. Armony. Joint panel sizing and appointment scheduling in outpatient care. *Management Science*, 63(11):3978–3997, 2016.

[37] A Zander. Modeling indirect waiting times with an M/D/1/K/N queue. In *Proceedings of the Second KSS Research Workshop*, Karlsruhe, Germany, 2016.