HEALTH SYSTEMS, 2017
https://doi.org/10.1080/20476965.2017.1403675

THE OPERATIONAL RESEARCH SOCIETY
Taylor & Francis
Taylor & Francis Group

Static and dynamic appointment scheduling to improve patient access time

Corine Laan[a], Maartje van de Vrugt[a,b,§], Jan Olsman[c], Richard J. Boucherie[a]

[a]Centre for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Enschede, The Netherlands; [b]Healthcare Innovations Programme, Leiden University Medical Centre, Leiden, The Netherlands; [c]Department of Surgery, Jeroen Bosch Hospital, 's-Hertogenbosch, The Netherlands

ABSTRACT

Appointment schedules for outpatient clinics have great influence on efficiency and timely access to health care services. The number of new patients per week fluctuates, and capacity at the clinic varies because physicians have other obligations. However, most outpatient clinics use static appointment schedules, which reserve capacity for each patient type. In this paper, we aim to optimise appointment scheduling with respect to access time, taking fluctuating patient arrivals and unavailabilities of physicians into account. To this end, we formulate a stochastic mixed integer programming problem, and approximate its solution invoking two different approaches: (1) a mixed integer programming approach that results in a static appointment schedule, and (2) Markov decision theory, which results in a dynamic scheduling strategy. We apply the methodologies to a case study of the surgical outpatient clinic of the Jeroen Bosch Hospital. We evaluate the effectiveness and limitations of both approaches by discrete event simulation; it appears that allocating only 2% of the capacity flexibly already increases the performance of the clinic significantly.

ARTICLE HISTORY
Received 20 March 2017
Accepted 4 November 2017

KEYWORDS
Queueing; mathematical programming; decision process; simulation; capacity analysis

1. Introduction

There is a growing need to improve efficiency as the expenditures of healthcare, one of the largest industries in the developed world, are rapidly rising. Appointment schedules for outpatient clinics have great influence on efficiency and patient access times to health care services, which is important for both medical outcomes and patient satisfaction (Gupta & Denton, 2008). To this end, in the Netherlands all outpatient clinics have to comply with nationally set upper bounds for patient access time (van Boven (RIVM), 2007), which impose a maximum on the time between requesting an appointment and the appointment itself for every patient (which is generally four weeks).

Many outpatient clinics use tactical appointment schedules in which the available capacity is divided among all patient types (Hulshof, Kortbeek, Boucherie, Hans, & Bakker, 2012). This tactical schedule determines to a large extent patients' access times, as patients often get the first available slot for their type when they request an appointment. However, the number of arriving patients fluctuates over the year due to, for example, the weather conditions (more bone fractures in icy weather) and the day of the week during which the arrival takes place (more patients with sports injuries arrive on Mondays). Moreover, the availability of the physicians is not constant, due to other obligations and holidays. As a consequence, overbooking is often unavoidable to provide patients with reasonable access times. Overbooking may result in physicians working overtime, and in large waiting times for patients.

Appointment scheduling in healthcare has received considerable attention from academics in the past years. In this paper, we summarise the most related results, and refer the reader to the reviews (Cayirli & Veral, 2003; Gupta & Denton, 2008; Hulshof et al., 2012) for a broader view of this field. In this paper, we optimise the tactical appointment schedule with respect to access time and idle time, in such a way that the performance is robust against varying demand and supply. We have found three papers studying access time improvements at outpatient clinics at the tactical planning level. Joustra et al. (2010) determine an appointment schedule for one week using linear programming, and use simulation to evaluate patients' access times and the number of overbooked appointments for that schedule. If it appears that not all access time upper bounds are met, the linear program is run again with input based on the simulation results, and these iterative steps are repeated if necessary. Elkhuizen et al. (2007) obtain global insights into the required capacity to meet the access time upper bounds by means of a queuing model, and evaluate the performance of the clinic in more detail by means of a simulation model. Creemers et al. (2012) invoke

CONTACT Maartje van de Vrugt ✉ n.m.vandevrugt@utwente.nl

[§]During this project, the second author was working at the Jeroen Bosch Hospital.

a bulk-server queuing model to evaluate access times, and use complete enumeration to determine the best assignment of the capacity to different patient types.

Dynamic capacity allocation typically includes reserving capacity for urgent or walk-in patients, or obtaining policies that describe how many patients of each diagnostic group should be admitted into the hospital from a waiting list to optimise utilisation and access times (cf. Hulshof, Boucherie, Hans, & Hurink, 2013; Kolisch & Sickinger, 2008). Several studies evaluate specific scheduling policies for reserving capacity for different patient types by means of simulation (cf. Klassen & Rohleder, 1996; Vermeulen et al., 2009).

A different, but related topic is the master surgery scheduling problem, in which tactical schedules assign operating rooms and time slots to all (sub-)specialties in a hospital (van Oostrum et al., 2008). This field of literature typically studies the objective of balancing the occupancy of post-operative wards, which differs from our objective. Moreover, only a few papers incorporate uncertain demand; for example Holte and Mannino (2013) minimise the (weighted) queue lengths of different patient types for a cyclical schedule with stochastic demand.

In this paper, we present a new stochastic mixed integer programming model (SMIP) with both stochastic patient demand and physician capacity (the number of available consultation hours). The objective of the SMIP is to obtain a tactical appointment schedule with minimal access times for a reasonable idle time. This objective is tailored to the practical problems encountered in hospitals. In particular, at the Jeroen Bosch Hospital (JBH), a large Dutch teaching hospital, the surgical outpatient clinic struggles with the questions: what is the minimally required capacity to comply with the access time upper bound, and how could the outpatient clinic's appointment schedule be made more robust with respect to the variance in demand and capacity? To address these questions, we approximate the SMIP with two different approaches. We use discrete event simulation to evaluate the performance of the approaches, and this numerical analysis helped the JBH assess the benefits of the different schedules. Currently, the hospital is investigating the possibilities and requirements for implementing the results. The model developed in this paper can readily be adapted to other outpatient clinics and to appointment systems in other applications than health care.

2. Mathematical model

The model presented in this paper mimics a typical Dutch outpatient clinic that uses a tactical appointment schedule in which a fixed capacity (number of appointment slots) is reserved for each patient type each week. For simplicity, we assume that each arriving patient gets the first available appointment slot that is reserved for that patient's type.

The clinic works with a block schedule, in which each morning or afternoon block (in practice often called "outpatient clinic session") consists of a specific number of time slots that are reserved for certain patient types; for example, the first five slots are reserved for new patients and after that 10 slots are reserved for follow-up patients. Once set, these compositions of morning or afternoon blocks are not allowed to change often, due to physician preferences and technical limitations. If a physician is not available, for example due to a conference or holiday, typically an entire block is cancelled. If access times exceed the upper bound, physicians work overtime or add an entire block to the schedule to ensure timely access for patients. Therefore, the composition of the blocks, i.e., how many appointment slots for each patient type, plays a significant role in patients' access times. We do not incorporate patient no-shows; the capacity is lost when patients do not show up. The term "cancellation" is used for blocks being cancelled due to the occasional unavailability of the physician. These blocks are always cancelled sufficient time in advance to avoid that patients need to be rescheduled, which is common practice at Dutch outpatient clinics.

As overbooking a clinic block involves much additional scheduling efforts and often increases patient waiting times and/or physicians working overtime, we determine the minimally required number of scheduled blocks and each block type's composition without surgeons working overtime. To this end, we present a stochastic mixed integer programming model (SMIP) that incorporates stochastic arrivals and capacity. A direct solution of the SMIP cannot be obtained, due to the stochasticity involved. Therefore, we approximate the SMIP's solution in two ways: (1) a MIP in which we incorporate stochastic arrivals but assume that the capacity is fixed, which results in the optimal static schedule, and, based on this static schedule, (2) a Markov decision model, which results in the optimal flexible scheduling policy. The SMIP and both approximation approaches are developed in this section in a general form; all input details are given in Section 3.

2.1. The stochastic scheduling problem

The JBH uses a cyclic tactical schedule that distributes the capacity among the different patient types. We develop a SMIP to optimise both the number of scheduled blocks and the number of appointment slots that are planned for each patient type during each block. The objective of the SMIP is to balance patient access time and physician idle time.

In the following, we define the sets, parameters, variables and constraints of the SMIP. Let $p \in \mathcal{P} = \{1, \ldots, P\}$ denote the patient types, $d \in \mathcal{D} = \{1, \ldots, D\}$ the days, and $b \in$ {morning, afternoon} the block types. Note that in the SMIP it is possible to schedule more days than the number of clinic days in the cycle, because there are multiple surgeons who may work in parallel as long as there are enough consultation rooms. The parameters and variables are denoted as follows, with the indices subscripted and the "type" of parameter or variable superscripted.

L_p appointment length for type p (in number of time slots)

N_b capacity of block b (number of time slots)

M maximum number of blocks during one cycle

C^a "cost" of patient access time

C^e "cost" of an empty time slot

X_{pdb} number of type p patients scheduled on day d, block b

Y_{db} indicator that equals 1 if at least one patient is scheduled on day d, block b

T_p total number of appointment slots scheduled for type p

The maximum number of blocks during one cycle, M, is used to limit the solution space of the SMIP, and is set sufficiently large to accommodate all demand. The cost parameters C^a and C^e can be used to give priority to minimising either the access time or the idle slots, and can be chosen according to the preferences of the hospital. Note that the parameters and variables of the SMIP are not stochastic. As in practice not the schedule itself, but the realisation of the schedule and the patient arrivals are stochastic, we formulate the SMIP such that the stochasticity is incorporated in the objective function only.

The following constraints hold for the parameters and variables.

$$T_p = \sum_{d,b} X_{pdb} \quad \forall p \tag{1}$$

$$\sum_p X_{pdb} \cdot L_p = N_b \quad \forall d, b \tag{2}$$

$$\sum_{d,b} Y_{db} \leq M \tag{3}$$

$$X_{pdb} \leq Y_{db} \cdot N_b \quad \forall p, d, b \tag{4}$$

$$X_{pd'b} \leq X_{pdb} + 1 \quad \forall p, d, d', b \tag{5}$$

$$X_{pd'b} \geq X_{pdb} - 1 \quad \forall p, d, d', b \tag{6}$$

$$\sum_d Y_{db} \leq \sum_d Y_{db'} + 1 \quad \forall b, b' \tag{7}$$

$$\sum_d Y_{db} \geq \sum_d Y_{db'} - 1 \quad \forall b, b' \tag{8}$$

$$T_p, X_{pdb} \in \mathbb{N} \quad \forall p, d, b \tag{9}$$

$$Y_{db} \in \{0, 1\} \quad \forall d, b \tag{10}$$

The first constraint (1) sets T_p, which is used to calculate the average access time. We assume that the total number of appointment slots is distributed evenly over the days in the cycle, but the model is readily adapted to other arrival patterns. Constraints (2) and (3) ensure that all time slots of each block are used in the schedule, and the available capacity is not exceeded. Constraint (4) makes sure that a surgeon is scheduled when patients are scheduled on that day. In practice, the capacity of block b may also depend on the day of the week; this could be incorporated in the model and would only affect Constraints (3) and (4). Constraints (5)–(8) are not necessary for obtaining a feasible schedule, but reflect preferences of hospital management and physicians. Constraints (5) and (6) balance the workload over the surgeons by requiring that the number of scheduled patients of each type on each block differs at most one. This constraint should be adapted when physicians are not all allowed to treat all patient types; in such clinics the constraint can better limit the differences in the number of slots per block. Constraints (7) and (8) ensure that the number of morning and afternoon blocks are balanced, such that on most days both a morning and afternoon block is scheduled. The last constraints (9) and (10) define the variable types.

The SMIP formulation above does not include stochastic parameters or variables, and is not related to the actual demand of the clinic. For simplicity we assume that each block has a cancellation probability $u \in [0, 1]$ and patient arrivals follow a certain probability distribution with rate λ_p for type p patients, which both may depend on for example the weekday. The stochastic capacity and demand is incorporated in the objective function, which is represented by the general notation:

$$\min \quad C^a \sum_p f^a\left(T_p, \lambda_p, u\right) + C^e \sum_p f^e\left(T_p, \lambda_p, u\right).$$

The functions f^a and f^e may denote any function depending on T_p, λ_p and u. We choose to let these function relate the stochastic total realised number of appointment slots and arrival distribution for each patient type, to the expected access time and expected number of empty slots, respectively. We clarify f^a (f^e) below, after we derive a relation between the access (idle) time, the arrival distribution and the realised number of appointment slots. As both the demand and capacity are stochastic variables, $f^a\left(T_p, \lambda_p, u\right)$ and $f^e\left(T_p, \lambda_p, u\right)$ are stochastic variables. Furthermore, this objective allows to set weights for each patient type individually, if necessary.

In this SMIP, both the demand and the capacity are stochastic. There exist several approaches for solving SMIPs, for example decomposition algorithms, cf. (Sen, 2005), and robust optimisation approaches (cf. Beyer & Sendhoff, 2007). In this paper, we incorporate the stochasticity in the patient arrivals in the SMIP by means of a queuing model, and solve the SMIP with a fixed and a flexible approximation approach. For the first approach we assume that the capacity is deterministic. For the second approach, we make the static schedule more flexible using Markov decision theory. We describe the queuing model in the remainder of this subsection, and the two approaches in the following two subsections.

We incorporate the stochasticity in the patient arrivals in the SMIP by means of a discrete time queueing model. The model relates the realised capacity to the expected access time (f^a) and expected number of idle slots (f^e) for each patient type. The discrete time queuing model is presented in analogy by the ones presented in Masselink, van der Mijden, Litvak, and Vanberkel (2012), Kortbeek et al. (2014) and van de Vrugt, Boucherie, Smilde, de Jong, and Bessems (2017). The state of the model is the size of the backlog of a specific patient type p at the end of day d, B_{dp}. Every day a number of patients, at most equal to the capacity S_{dp} for this type on this day, is removed from the backlog, and a random number of new patients, A_{dp}, is added to the backlog. We do not allow arriving patients to be scheduled on the same day, which is why a regular $M/D/c$ queue cannot be used in this analysis even if the capacity S_{dp} would be the same for all days d. For the backlog on day d, we have:

$$B_{dp} = \left(B_{d-1,p} - S_{dp}\right)^+ + A_{dp}, \qquad (11)$$

in which $x^+ = x$ if $x > 0$ and $x^+ = 0$ otherwise. Eq. (11) is known as Lindley's recursion (cf. Cohen, 1982). By means of the queueing model we derive a relation between the daily capacity and expected access and idle times by calculating the stationary distribution of B_{dp}, if we assume that S_{dp} is not stochastic. Appendix 1 presents more details on the queuing model; the definition of f^a is given by (A2), and for f^e by (A1). Note that these functions depend on T_p, λ_p and u in the different calculation steps. These expressions allow us to store all possible values for f^a and f^e before solving the SMIP, which is clarified in the next subsection. Note that the model formulation of both the queuing model and the SMIP allow to incorporate other performance measures than the expected values, for example the probability that the access time upper bound is met for 90% of the patients.

The two approximation methods are both practically relevant and offer the JBH two distinct alternatives with good performance. Several modelling assumptions are based on the JBH case study, but we indicated how the SMIP can be adjusted to other outpatient clinics. Additionally, we assume that capacity is distributed evenly over the week as we cannot incorporate time-dependent arrival rates due to lack of data. The JBH requires that the block-sizes, measured both in number of appointment slots for each patient type, are equal for each day and that the number of morning and afternoon blocks are balanced.

2.2. Static scheduling

We cannot solve the SMIP as presented above due to the stochasticity in the capacity. We first approximate the SMIP by a mixed integer program (MIP) in which we assume that exactly a fraction u of all blocks is canceled, so the realised capacity is $(1-u)T_p$ for each patient type. The MIP provides both the number of blocks that should be scheduled each week, and the number of appointment slots in each block-type for each patient type. This schedule is the base scenario for our numerical study. If the variance in the capacity of the clinic is relatively low, the schedule obtained with the MIP should result in acceptable performance in practice. In order to accommodate for higher variability, in the following subsection we present a flexible scheduling approach, which is based on the MIP schedule.

As stated before, we assume that a fraction u of the blocks is cancelled (opposed to each block being canceled with probability u). Therefore, the conversion from T_p to S_{dp} is $S_{dp} = \lfloor (1-u)T_p/D \rfloor$, with $\lfloor x \rfloor$ denoting rounding down to the nearest integer, and the remaining $(1-u)T_p - D \cdot \lfloor (1-u)T_p/D \rfloor$ days added to days with the lowest index and enough surgeons available. We assume that the total realised capacity is divided equally over the weekdays, but the queuing model is readily adapted for different capacity distributions. Note that the assumption that capacity is divided equally over the weekdays represents the best-case scenario because we also assume a constant arrival rate.

To be able to incorporate functions f^a and f^e in the objective of the SMIP, we introduce parameters m_p and M_p denoting the minimum and maximum number of appointment slots scheduled for type p, respectively. We create an array with $f^a\left(T_p, \lambda_p, u\right)$ and $f^e\left(T_p, \lambda_p, u\right)$, for a fixed λ and u in each scenario, and calculate these functions for each patient type and all possible $(1-u)T_p$ between m_p and M_p. When the solver solves the MIP, it can access the array to obtain the objective value corresponding to the current value of the realised capacity, $(1-u)T_p$. We assume parameter m_p is larger than the average arrival rate of type p patients, otherwise the queuing model is not solvable. Furthermore, M_p is set sufficiently large to accommodate all arrivals, and we introduce additional constraints to ensure that $m_p \leq (1-u)T_p \leq M_p$. Using this array, we can approximate the SMIP by a MIP.

2.3. Dynamic scheduling

Since both the capacity and the number of patients requesting an appointment fluctuate during the year, the performance of the clinic will improve by allocating capacity dynamically. However, physicians have the right to know their working hours several weeks in advance, so the flexibility of the schedule is limited. Therefore, the cyclical schedule as determined above is set as the basis of the flexible schedule, but the clinic may decide to add an extra block each cycle. The optimal policy for determining whether or not to schedule an additional block is determined by means of a Markov decision process (MDP), which is explained in the following.

The state of the Markov process is the number of patients in the queue for all patient types: $s = [q_1, ..., q_N] \in \mathcal{S}$, with N the number of patient types, and q_p is finite for all p. Transitions and decisions occur at the end of every cycle. Each decision epoch, the clinic either schedules an additional block at the end of the next cycle ($a = 1$), or not ($a = 0$). The composition of the additional block is the same each time it is added to the schedule, and is obtained with the MIP. Let $P(q_p'|q_p, a)$ denote the probability that the queue length of type p equals q_p' at the next decision epoch, given that the current queue length is q_p and decision a is taken. Furthermore, let Y^t denote the total number of blocks scheduled by the MIP in one cycle, and α_p the number of appointment slots per block for type p patients. Then, given that each patient type can only be assigned to appointment slots that are reserved for their type, the transition probabilities are:

$$P(s'|s, a) = P(q_1'|q_1, a) \cdot \ldots \cdot P(q_N'|q_N, a),$$

with

$$P(q_p'|q_p, a) =$$
$$\begin{cases} P(A_p = q_p') & \text{if } q_p \leq t_{pY^t}, \\ P(A_p = q_p') \sum_{i=0}^{j-1} P(T_p = t_{pi}) & \text{if } t_{pj} \leq q_p \leq t_{p,j-1}, \\ \quad + \sum_{i=j}^{Y^t} P(A_p = q_p' - q_p + t_{pi} + a \cdot \alpha_p) P(T_p = t_{pi}) & \text{for } j = 1, \ldots, Y^t, \\ \sum_{i=0}^{Y^t} P(A_p = q_p' - q_p + t_{pi} + a \cdot \alpha_p) P(T_p = t_{pi}) & \text{else.} \end{cases}$$

Here, A_p is the random number of arrivals for patient type p each cycle, and t_{pi} the capacity for patients of type p in case i blocks are cancelled. We assume that each block is cancelled with probability u, so $P(T_p = t_{pi})$ follows a binomial distribution with probability u.

To enhance readability, we assume that the cycle length equals the access time upper bound (the maximally allowed time between a patient's arrival and appointment) and this upper bound is the same for all patient types, but the model can be adapted to different values. Therefore, we incur costs in the MDP when the access time exceeds one cycle length. Additionally, costs are incurred in the MDP for idle appointment slots. We use the expected number of empty appointment slots and the expected queue length at the end of the cycle for the direct costs, resulting in:

$$C(s, a) = \sum_{i=0}^{Y^t} \Bigg[P(T_p = t_{pi}) \cdot \Bigg(C^a \sum_{p=1}^{N} (q_p + \mathbb{E}[A_p] \\ - t_{pi} + a \cdot \alpha_p)^+ + C^e \sum_{p=1}^{N} (t_{pi} + a \cdot \alpha_p \\ - q_p - \mathbb{E}[A_p])^+ \Bigg) \Bigg]$$

The total discounted value of the MDP, given policy π, satisfies the Bellman-equations:

$$V_\pi(s) = \max_a \left\{ C(s, a) + \beta \sum_{s' \in S} P(s|s', a) V_\pi(s') \right\},$$

with β the discount factor. We assume $\beta = 0.95$ to reflect that future access times are important to take into account with the current decision making.

Opposed to the queuing model, the MDP cannot be evaluated for each patient type independently, as we may only add an entire block and blocks consist of appointment slots for all patient types. The purpose of the MDP is only to determine whether or not to add a block and the actual scheduling is carried out following the results of Section 2.2. Therefore, we aggregate all patient types in the numerical analysis of the MDP model. Note that this aggregation also substantially reduces the size of the state space, which makes the MDP more tractable. This result can also be obtained if we assume that patients arrive and are treated in batches, but that would result in a more difficult and therefore less practical policy.

3. Application

In this section, we assess the quality of the schedules derived by the MIP and MDP approaches by means of the case study of the surgical outpatient clinic of the JBH, for the sub-specialty oncology. To this end, we compare the different appointment schedules on the following performance measures: average access time, the probability that the access time exceeds one week, and the utilisation of the appointment slots. The JBH aims for an access time of at most one week for as many patients as possible, which is a tighter upper bound than the national one.

3.1. *Case study input*

In this subsection, we provide all case study specific input and assumptions for the methods, how we test the methods, and which scenarios we use in the numerical analysis. In the objective of the (S)MIP we include the expected access and idle time instead of, for example, the probability that the access time upper bound is exceeded. The JBH prefers the average performance measures, as they do not have the capacity to accommodate rarely occurring high peaks in the number of arriving patients. In the (S)MIP, the access and idle time can be given weights to reflect that one is a more important performance measure than the other. We first use $C^a = C^e$, so equal weights for access and idle time, as the JBH indicated to value patient access time and physician idle time equally. Additionally, we investigate scenarios with $C^a \neq C^e$.

As in many hospitals, in the JBH it is not possible to use data on the times appointments were requested, as only the realised appointment times are registered. This also implies that we cannot incorporate appointments scheduled for patients who did not show up, and cannot derive the arrival distribution from the data. Therefore, as input for the models we use the data in Table 1, which is based on the realised patient appointments instead of the times the appointments are requested. Additionally, we assume that the arrival process is Poisson with a constant rate.

According to the preferences of the JBH we set the cycle length to one week. We assume that the probability that a block is cancelled, is independent for each block and equals 10%, so $u = 0.1$, which is desired by the JBH but currently the cancellation probability is slightly higher.

We solve the MIP using AIMMS (Bisschop, 2014) and use the policy iteration algorithm (Puterman, 2005), implemented and solved in Matlab (MATLAB, 2010) to obtain the optimal policy for the MDP. As stated before, we solve the aggregated version of the MDP. This because the state space explodes already if we include a few patient types and assume that access times are at most four cycle lengths (which is the national upper bound), as the number of appointment slots per patient type per cycle can be over 100.

We evaluate the schedules invoking discrete event simulation. The advantage of simulation is that we can mimic the processes at the JBH realistically, so we can incorporate random cancellation of blocks, fluctuating arrival rates and even patients who have preferences for a particular physician. In the discrete event simulation patients of different types arrive one-by-one according to a Poisson process and take the first available appointment slot that is reserved for their type. The search for an available slot starts from the next day, as patients are not scheduled on the same day when they request an appointment. Furthermore, all patients show up for their appointment, which is realistic for the JBH but can readily be adapted to other clinics. Each block has a certain probability – independent of the probabilities for the other blocks – of being cancelled by the physician in advance, so it is not necessary to reschedule patients. This cancellation probability is varied in the simulation. The simulation is programmed in C++. We use common random numbers (Law & Kelton, 2000) for a fair comparison between the schedules. We simulate 200 runs with 260 clinic days each, in order to obtain minimally 5% relative precision.

Table 1. Weekly demand and capacity for the case study.

Type p	L_p	$\mathbb{E}[A_p]$ (patients)	MIP capacity (appointment slots)
1	2	7.4	9
2	2	115.9	130
3	2	14.0	17
4	2	29.4	34
5	2	8.3	11
6	2	27.7	33
7	2	5.7	8
8	1	24.7	28

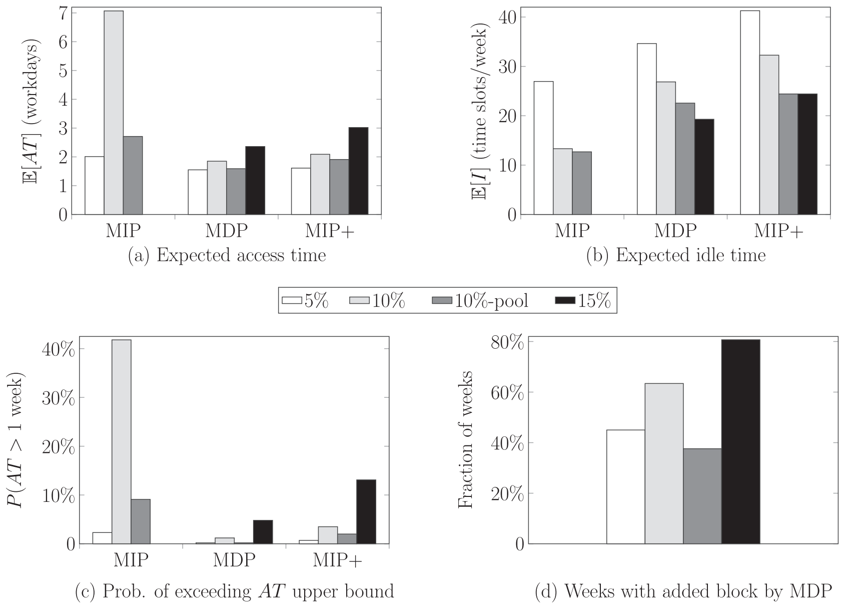
Next to the MIP and MDP schedule, we investigate scenarios in which the capacity is pooled for all patient types, denoted by suffix '-pool'. In these scenarios, we investigate to what extent the JBH could improve the accessibility of the clinic when there are no reserved appointment slots for each patient type, but each patient would be treated first come, first served.

The MDP schedule adds capacity to the MIP schedule in a significant number of weeks. In order assess the added value of dynamic scheduling compared to static scheduling we additionally investigate a MIP schedule with the same total capacity as the MDP schedule, which is labelled 'MIP+'; the additional capacity is added by adding (part of) a block to the MIP schedule each week such that the total capacity equals the capacity of the MDP schedule over the entire simulation run.

3.2. Case study results

The schedule created by the MIP for this scenario requires 512 time slots in 15 blocks each week, divided over 7 × 18 appointment slots in the morning and 8 × 18 appointment slots in the afternoon. Note that most patient types require two time slots per appointment slot, and we can schedule more than ten blocks per week (five working days × two blocks per day) because multiple surgeons can work in parallel. Figure 1 depicts simulation results for different cancellation probabilities, performance measures and scheduling policies. Recall that all schedules in this figure are obtained assuming that $u = 10\%$ in the MIP and MDP models, so the cancellation probability is only varied in the simulation. The simulated scenario with 15% cancellation probability has insufficient capacity in the MIP scenario to schedule all arriving patients and is therefore not depicted in Figure 1.

From the simulation results, it appears that the average access time and the access time upper bound compliance improve significantly in case of a lower cancellation probability, more capacity (MDP and MIP+ schedules), and when all capacity is pooled, which are



(a) Expected access time

(b) Expected idle time

(c) Prob. of exceeding AT upper bound

(d) Weeks with added block by MDP

Figure 1. Results static and dynamic scheduling with $u = 10\%$ for multiple cancellation probabilities.

intuitive results. With the current capacity and 10% cancellations 38.3% of the patients cannot be scheduled within one week, see Figure 1(c). Also in the other schedules there are patients for which the access times exceed one week, but these numbers of patients are acceptable for the clinic.

For the performance of the clinic it is important to minimise the cancellation probability, as the simulation results show that this probability significantly affects all performance measures when the load of the system is close to one. For the MDP and MIP+ schedules, the effect of different cancellation probabilities is relatively small. Aggregating the capacity for all patient types would significantly increase the clinic's performance, even resulting in better performance compared to the scenario with only 5% cancellation probability.

The MDP and MIP+ schedules have relatively many idle slots each week, the MDP slightly less than the MIP+ schedule. As each block consists of 32 (morning) or 36 (afternoon) time slots, on average one block is canceled each week, which is acceptable for the JBH; because the appointments are planned in advance the surgeons can use this time to schedule other activities. Figure 1(d) depicts how often the MDP policy prescribes to schedule an additional block; this policy implies that for the 10%-scenario on average 2.5% capacity is added to the schedule.

The decision rule of the MDP can affect at most one physician for one morning or afternoon per cycle, which is less than 7% of the capacity in the JBH case study. From Figure 1(a) and (c) it appears that the MDP policy outperforms the MIP+ policy with respect to both the access and idle time, i.e., dynamic scheduling outperforms static scheduling already for this small fraction of flexible capacity. This is an intuitive result, as both schedules contain the exact same capacity, but the dynamic schedule schedules the capacity only at times with many patients in the system. We conjecture that dynamic scheduling outperforms static scheduling for all realistic instances. However, in an (extremely unrealistic) artificial instance in which alternately $Y^t < x < 2Y^t$ patients and zero patients arrive each week, the MIP+ schedule will outperform the MDP schedule; the MDP policy will add capacity in quiet weeks and will not add capacity in busy weeks, thereby increasing the access times significantly.

Table 2. MDP policy queue length threshold values for different C^a and C^e.

C^e	C^a		
	1	2	5
1	22	20	1
2	32	22	9
5	41	35	22

Similar to the MIP, Figure 1 only depicts the results of the MDP with equal costs for access and idle time. With these parameters, it is optimal to add an extra block when the total queue length exceeds 22 patients. We investigated the effect of other values of C^a and C^e, see Table 2. As expected, the threshold to schedule an additional block decreases for relatively large access time costs, and vice versa. As the results of these schedules are obvious, they are omitted in this paper. Additionally, we investigated policies that depend on all patient types separately, assuming that the patient arrivals for type 2 patients arrive and are served in batches of size five. We do not present these results here, as the resulting optimal policies are complex and

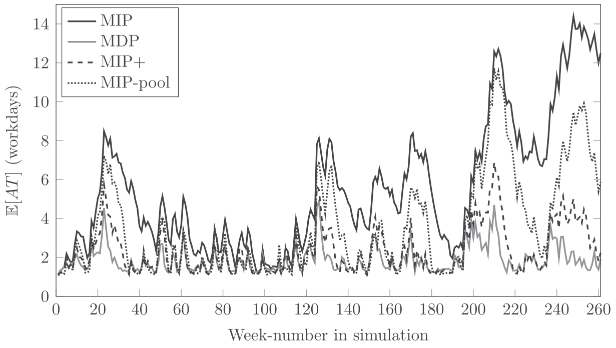


Figure 2. Week-dependent results for 10% cancellation.

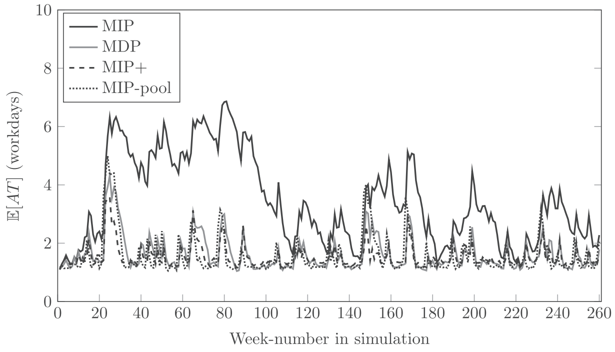


Figure 3. Week-dependent results for 5% cancellation.

Table 3. Simulation results for different values of u.

u	Schedule	$\mathbb{E}[AT]$	$P(AT > 5)$	$\mathbb{E}[I]$
5%	MIP	3.14	14.6%	22.32
5%	MDP	1.61	0.2%	32.41
5%	MIP+	1.69	1.0%	39.77
5%	MIP-pool	1.55	0.1%	23.31
10%	MIP	7.07	41.8%	13.32
10%	MDP	1.85	1.2%	26.86
10%	MIP+	2.09	3.5%	32.27
10%	MIP-pool	2.71	9.1%	12.69
15%	MIP	4.20	27.1%	17.03
15%	MDP	1.88	0.8%	29.99
15%	MIP+	2.30	4.4%	32.00
15%	MIP-pool	2.24	2.2%	19.45

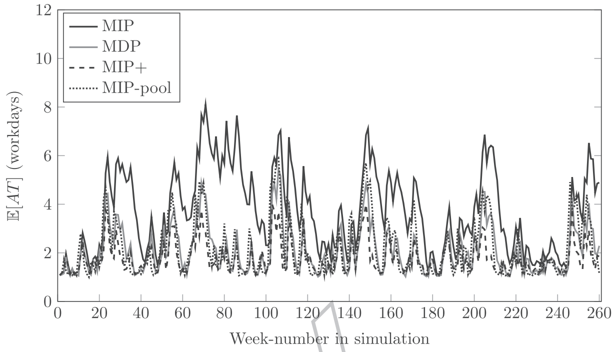


Figure 4. Week-dependent results for 15% cancellation.

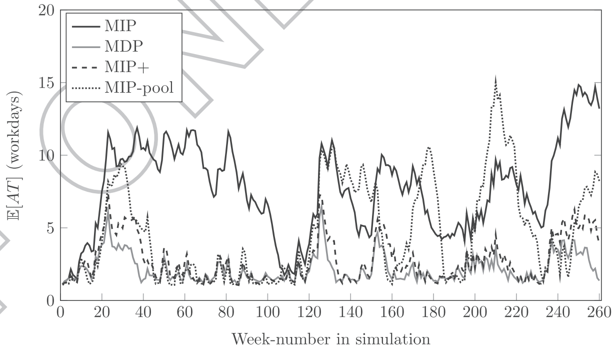


Figure 5. Results static and dynamic scheduling for fluctuating arrival rates for 10% cancellation.

these policies did not improve the results significantly. From a practical perspective, more complicated policies are of little use for the JBH.

In Figure 2 we depict simulation results of the conditional average access time of patients arriving in week x for several scheduling policies with 10% cancellation probability. From these results it is clear that pooling the capacity can already improve the performance significantly. The conditional average access time is always below the upper bound (five working days) with the dynamic schedule, while for the other schedules in some weeks the upper bound is exceeded.

In order to investigate to what extent the cancellation probability affects the schedules, we additionally obtain the fixed and flexible schedule for $u = 5\%$ and $u = 15\%$. In the MIP schedule for $u = 5\%$, eight morning blocks and seven afternoon blocks are scheduled, while for $u = 15\%$ the schedule contains eight morning and eight afternoon blocks. Note that each afternoon block consists of four time slots more than each morning block, so the MIP schedule for $u = 10\%$ contains more time slots than the MIP schedule for $u = 5\%$. We depict simulation results for $u = 5\%$ and $u = 15\%$ in Figures 3 and 4, respectively, and in Table 3. Note that the performance of both the 5% and 15% scenario is better than for $u = 10\%$; in these scenarios the MIP schedules relatively much overcapacity due to rounding of the number of blocks required, which appears to improve the performance on the access time, but the number of idle slots increases. We conclude that for this case study in most weeks the conditional average access time of a patient is less than five days, but the static schedule results in several time periods where the upper bound is exceeded, especially for higher cancellation probabilities.

In reality, patients could have a preference for one of the doctors when requesting an appointment. With the simulation model, we investigated scenarios in which 40% or 80% of the patients had a preference for a random physician; the JBH does not have data on patient preferences. These preferences, obviously, increased the average access time and number of idle slots. The probability that the access time exceeds one week approximately doubled compared to the scenarios where

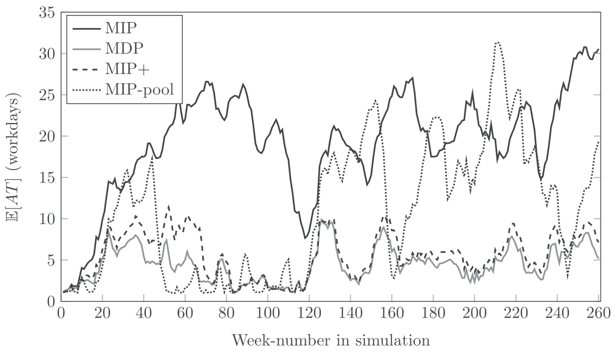


Figure 6. Results static and dynamic scheduling for strongly fluctuating arrival rates for 10% cancellation.

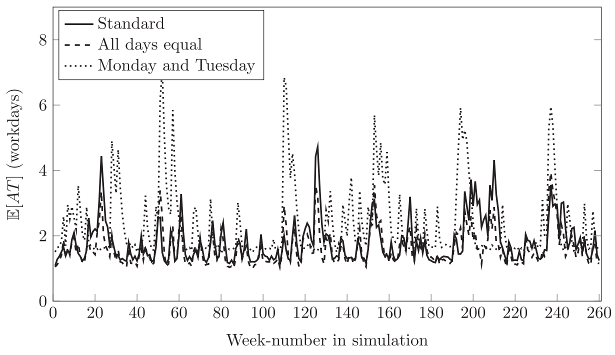


Figure 7. Flexible schedule simulation results of different weekly capacity distributions and $u = 10\%$.

patients do not have preferences. These effects should be taken into account when the results of the models would be implemented in practice, and imply that the performance as presented here will probably not be achieved by the JBH.

For this case study we do not have data on week- or season-dependent patient arrival rates. However, in practice the arrivals at the surgical outpatient clinic are significantly higher in certain weeks, for example during cold winter months with icy weather. Dynamic scheduling is promising for coping with these varying arrival rates, and to investigate to what extent we simulate scenarios in which the arrival rate changes every four weeks. When the arrival rate is changed, we multiply the rate by a (uniformly drawn) random number between 0.8 and 1.2 to obtain the new arrival rate. Additionally, we investigate a scenario where the random number is between 0.5 and 1.5, which reflects more strongly fluctuating arrival rates. The results for these simulations are depicteds in Figures 5 and 6.

With fluctuating arrival rates, the capacity at the outpatient clinic may be insufficient to cope with the arriving patients for several weeks in the simulation. As a consequence, patient access times increase, as Figures 5 and 6 depict, especially for the MIP schedule. In these scenarios, pooling the MIP capacity for all patient types cannot always avoid exceptionally long access times. Additionally, the numerical results in both scenarios indicate that the expected access time is at most equal, but more often lower with dynamic scheduling, compared to static scheduling. Especially in the first 70 weeks of the simulation the access times are significantly lower with dynamic scheduling, even for the scenario with strongly fluctuating arrival rates. In weeks 120–160 of the scenario with strongly fluctuating arrival rates, both methods seem to perform comparably bad; this may indicate that one block of flexible capacity per week is not enough to anticipate for all fluctuations in the arrival rate.

In the models, we assume that the capacity of the clinic is distributed evenly over all weekdays, which is often not common practice in hospitals. As we do not have information on weekday-dependent patient arrival rates, we assume that this rate is constant. Because of this assumption, distributing capacity evenly over the weekdays will represent a best-case scenario for the JBH. When a hospital knows that there are relatively many arrivals on Mondays, the capacity distribution over the weekdays can be adapted accordingly. Figure 7 depicts simulation results for a constant arrival rate, flexible schedule with $u = 10\%$, and two scenarios with a different capacity distribution: "All days equal" denotes that each day has the exact same number of appointments for each patient type, and "Monday and Tuesday" denotes that all capacity is allocated to Mondays and Tuesdays. The "Standard" scenario schedules two blocks each weekday, and adds an additional block on Monday, Wednesday and Friday, which is motivated by the current schedule at the JBH. From the simulation results it appears that the "Standard" schedule and "All days equal" schedule perform comparably well, which is intuitive as these schedules do not differ much. When the clinic is only open two days per week, Mondays and Tuesdays, conditional patient access times are higher on average, and the peaks of the average conditional access time exceeds the upper bound of five workdays quite often in the simulated days. Concluding, when we assume that patient arrival rates are equal for each weekday, the "Standard" scenario investigated in this paper is close to best-case performance.

The "Standard" scenario investigated in this paper is also best-case because we assume that patients always take the first available slot. In practice, this is not always the case as patients may for example have preferences for a different weekday. It would be interesting for further research to extend the models to incorporate such practical settings.

For this outpatient clinic, the JBH aims to schedule similar blocks for all weekdays and physicians because this ensures a balanced workload for all physicians. For other outpatient clinics with, for example, many sub-specialisations of the physicians or many specific appointments that only take place on certain weekdays, it would be interesting for further research to extend the models accordingly. Additionally, in some clinics the different patient types may have different access time upper bounds, physicians are allowed to work overtime up to a certain maximum per block or week, and patients may take an appointment slot that is reserved for a different patient type when the slot is expected to remain idle. Including these practical assumptions would be interesting for further research.

This research results in a structured decision rule for adding capacity in case patient access times are perceived as excessive, which is often done at hospitals in an unstructured fashion. For further research, we aim to investigate if we can add a block to the schedule based on actual patient access times. The question how to include access time information in the state space while avoiding state space explosion is intriguing. Additionally, future work will focus on extending the models to make longer-term decisions, so a block could be added for, for example, next month. The hospital in this case study uses the same access time upper bound for all patient types, but to generalise the model it would be interesting to investigate patient types having different access time upper bounds.

The two approximation approaches presented in this paper do not solve the SMIP to optimality, or guarantee a certain performance for the clinic. When all capacity is allocated flexibly in a pooled way, the clinic's performance will be close to optimal. However, physicians have the right to know their working schedule some time in advance, which restricts the flexibility of the schedule. To this end, we restricted the flexible capacity in this paper and focused on practically relevant approximation approaches. It would be interesting for further research to investigate different solution approaches.

4. Conclusion

In this paper, we have formulated a SMIP and developed two approximation approaches by which a practically relevant static and dynamic appointment schedule can be made for an outpatient clinic with time-varying demand and capacity. The schedule resulting from the MIP model can replace the current schedule at the JBH without further changes when the physicians agree on their new working hours. The MDP decision rule that extends the fixed schedule will further improve the clinic's performance with respect to patient access times and the number of idle slots. We have shown that optimally allocating only a small fraction of the capacity dynamically, already increases both the efficiency and performance of the hospital, thus improving the quality of care. Although this methodology is designed for a hospital case study, it can also be applied to different appointment systems with time-varying demand and/or capacity.

For the JBH, the insights in their currently available capacity and patient arrival rate were already very valuable. At first, the idea of flexible capacity did not appeal to the JBH, but the effect of allocating only 2% of the capacity dynamically opened up the discussion on how to implement flexible capacity at the clinic. However, the main discussion at the clinic is on the addition of 2.5% capacity. The required additional capacity is currently not available on paper, but is presumably already used because surgeons work overtime. Aggregating capacity does not seem a valid option for all patient types at the JBH, as there are some patient types with shorter access time upper bounds that arrive relatively less often. The JBH is eager to investigate more scenarios, mentioned in the end of the previous section, to tailor the models more to the JBH case study.

Acknowledgements

The authors are grateful for the cooperation, input and support of the employees of the surgical department of the Jeroen Bosch Hospital. Additionally, we thank several anonymous referees, as their valuable comments have helped to improve the paper significantly.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Beyer, H. G., & Sendhoff, B. (2007). Robust optimization - a comprehensive survey. *Computer Methods in Applied Mechanics and Engineering*, *196*(33–34), 3190–3218. ISSN 0045–7825

Bisschop, J. (2014). *AIMMS: Optimization modeling*. Haarlem: Paragon Decision Technology.

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, *12*(4), 519–549.

Cohen, J. W. (1982). *The single server queue, volume 8 of North-Holland series in applied mathematics and mechanics*. (2nd ed.). Amsterdam: North-Holland Publishing Co.

Creemers, S., Beliën, J., & Lambrecht, M. (2012). The optimal allocation of server time slots over different classes of patients. *European Journal of Operational Research*, *219*(3), 508–521.

Elkhuizen, S. G., Das, S. F., Bakker, P. J. M., & Hontelez, J. A. M. (2007). Using computer simulation to reduce access time for outpatient departments. *Quality and Safety in Health Care*, *16*(5), 382–386.

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions (Institute of Industrial Engineers)*, *40*(9), 800–819.

Holte, M., & Mannino, C. (2013). The implementor/adversary algorithm for the cyclic and robust scheduling problem in health-care. *European Journal of Operational Research*, *226*(3), 551–559.

Hulshof, P. J. H., Kortbeek, N., Boucherie, R. J., Hans, E. W., & Bakker, P. J. M. (2012). Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health systems*, *1*(2), 129–175.

Hulshof, P. J. H., Boucherie, R. J., Hans, E. W., & Hurink, J. L. (2013, June). Tactical resource allocation and elective patient admission planning in care processes. *Health Care Manag Science, 16*(2), 152–166.

Joustra, P. E., de Wit, J., Struben, V. M. D., Overbeek, B. J. H., Fockens, P., & Elkhuizen, S. G. (2010). Reducing access times for an endoscopy department by an iterative combination of computer simulation and linear programming. *Health Care Management Science, 13*(1), 17–26.

Klassen, K. J., & Rohleder, T. R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management, 14*(2), 83–101.

Kolisch, R., & Sickinger, S. (2008). Providing radiology health care services to stochastic demand of different customer classes. *OR Spectrum, 30*(2), 375–395.

Kortbeek, N., Zonderland, M. E., Braaksma, A., Vliegen, I. M. H., Boucherie, R. J., Litvak, N., Hans, E. W. (2014). Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Performance Evaluation*, *80*, 5–26. ISSN 1874–4850

Law, A. M., & Kelton, W. D. (2000). *Simulation modeling and analysis* (3rd ed.). New York, NY: McGrawHill.

Masselink, I. H. J., van der Mijden, T. L. C., Litvak, N., & Vanberkel, P. T. (2012). Preparation of chemotherapy drugs: Planning policy for reduced waiting times. *Omega*, *40*(2), 181–187. ISSN 0305–0483.

MATLAB. (2010). *Version 7.11.0 (R2010b)*. Natick, MA: MathWorks Inc.

Puterman, M. L. (2005). *Markov decision processes: Discrete stochastic dynamic programming*. Hoboken, NJ: Wiley.

Sen, S. (2005). Algorithms for stochastic mixed-integer programming models. In K. Aardal, G. L. Nemhauser, & R. Weismantel (Eds.), *Discrete optimization, volume 12 of handbooks in operations research and management science* (pp. 515–558). Elsevier. Retrieved from http://www.sciencedirect.com/science/handbooks/09270507/12

van Boven (RIVM), P. F. (2007). *Het treekoverleg: streefnormen wachttijden curatieve sector*. Volksgezondheid Toekomst Verkenning, Nationale Atlas Volksgezondheid. Bilthoven: RIVM. Retrieved from http://www.zorgatlas.nl/thema-s/wachtlijsten/wachtlijsten-ziekenhuiszorg/

van de Vrugt, N. M., Boucherie, R. J., Smilde, T. J., de Jong, M., & Bessems, M. (2017). Rapid diagnoses at the breast center of Jeroen Bosch Hospital: a case study invoking queueing theory and discrete event simulation. *Health Systems, 6*(1), 77–89.

van Oostrum, J. M., Houdenhoven, M., Hurink, J. L., Hans, E. W., Wullink, G. & Kazemier, G. (2008). A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum*, *30*(2), 355–374. ISSN 0171–6468.

Vermeulen, I. B., Bohte, S. M., Elkhuizen, S. G., Lameris, H., Bakker, P. J. M., & Poutré, H. L. (2009). Adaptive resource allocation for efficient patient scheduling. *Artificial Intelligence in Medicine, 46*(1), 67–80.

Appendix 1. Discrete time queueing model

This appendix provides the formulas relating the daily capacity to the expected access time and number of empty slots through the stationary distribution of B_{dp}, the backlog for patient type p on day d (which are patients that have requested an appointment, but have not yet been treated at the clinic). These results are similar to the ones presented in (Kortbeek et al., 2014), but we simplified the formulas where possible. Recall that the conversion from T_p to S_{dp} is $S_{dp} = \lfloor (1-u)T_p/D \rfloor$, with $\lfloor x \rfloor$ denoting rounding down to the nearest integer, and the remaining $(1-u)T_p - D \cdot \lfloor (1-u)T_p/D \rfloor$ days added to days with the lowest index and enough surgeons available.

Recall that $B_{dp} = \left(B_{d-1,p} - S_{dp}\right)^+ + A_{dp}$, with S_{dp} the daily capacity and A_{dp} the random number of new appointment requests on day d. The transition probabilities for this model are given by:

$$P\left(B_{dp} = q' \middle| B_{d-1,p} = q\right) = P(A_{dp} = q' - (q - S_{dp})^+).$$

We consider a cyclic schedule, so index $d := d \bmod D$, with D the cycle length. As the JBH does not have data on the arrival distribution, we assume Poisson arrivals, which implies:

$$P(A_{dp} = j) = \frac{\lambda_{dp}^{j}\, e^{\lambda_{dp}}}{j!},$$

with λ_{dp} the arrival rate on day d. The stationary distribution of B_{dp} is obtained by solving $\pi P = \pi$, with P the transition probability matrix.

Let π_{dpq} the stationary probability that the backlog on day d equals q for type p, and T_p the total number of slots scheduled for this patient type during one cycle. The expected number of empty slots reserved for type p patients per cycle, $\mathbb{E}[I_p]$, is given by:

$$\mathbb{E}[I_p] = \sum_{d=0}^{D-1} \sum_{q=0}^{S_{dp}-1} (S_{dp} - q)\pi_{dpq}. \tag{A1}$$

The expected access time is derived by conditioning on the backlog of patients on a day d:

$$\mathbb{E}[AT_p] = \sum_{d=0}^{D-1} \sum_{y=1}^{\infty} P(AT_{dp} > y | B_{dp} = q)\pi_{dpq} \cdot \frac{\mathbb{E}[A_{dp}]}{\sum_{j=0}^{D-1} \mathbb{E}[A_{jp}]}.$$

Here, $P(AT_{dp} > y | B_{dp} = q)$ is the probability that the access time of a type p patient arriving on day d exceeds y days, given that the backlog at the end of day d equals q. Define $\bar{S}_{dp}(y) := \sum_{i=1}^{y} S_{d+i,p}$ the sum of the capacity from day $d+1$ until day y. Then $P(AT_{dp} > y | B_{dp} = q) = 1$ if $\bar{S}_{dp}(y) \leq q$, and for $\bar{S}_{dp}(y) > q$ it holds:

$$P(AT_{dp} > y | B_{dp} = q) = \frac{\sum_{j=\bar{S}_{dp}(y)+1}^{\infty} (j - \bar{S}_{dp}(y))P(A_{dp} = j)}{\mathbb{E}[A_{dp}]}.$$

Concluding, the average access time is given by:

$$\mathbb{E}[AT_p] = \sum_{d=0}^{D-1} \sum_{y=1}^{\infty} \frac{\sum_{j=\bar{S}_{dp}(y)+1}^{\infty} (j - \bar{S}_{dp}(y))P(A_{dp} = j)}{\sum_{j=0}^{D-1} \mathbb{E}[A_{jp}]} \cdot \pi_{dpq}. \tag{A2}$$

Formula $f^e\left(T_p, \lambda_p, u\right)$ equals Equation (A1), for which we need to derive S_{dp} from T_p as explained above and in Section 2.2. In order to evaluate formula $f^a\left(T_p, \lambda_p, u\right)$, we determine $\bar{S}_{dp}(y)$ with y equal to five days in our case study, and evaluate Equation (A2).