

Appointment scheduling with unscheduled arrivals and reprioritization

Nardo J. Borgman¹ · Ingrid M. H. Vliegen² ·
Richard J. Boucherie¹ · Erwin W. Hans¹

Published online: 20 January 2017

© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract Inspired by the real life problem of a radiology department in a Dutch hospital, we study the problem of scheduling appointments, taking into account unscheduled arrivals and reprioritization. The radiology department offers CT diagnostics to both scheduled and unscheduled patients. Of these unscheduled patients, some must be seen immediately, while others may wait for some time. Herein a trade-off is sought between acceptable waiting times for appointment patients and unscheduled patients' lateness. In this paper we use a discrete event simulation model to determine the performance of a given appointment schedule in terms of waiting time and lateness. Also we propose a constructive and local search heuristic that embeds this model and optimizes the schedule. For smaller instances, we verify the simulation model as well as compare our search heuristics' performance with optimal schedules obtained using a Markov reward process. In addition we present computational results from the case study in the Dutch hospital. These results show that a considerable decrease of waiting time is possible for scheduled patients, while still treating unscheduled patients on time.

Keywords Appointment scheduling · Radiology · Service operations · Heuristics

✉ Nardo J. Borgman
n.j.borgman@utwente.nl

¹ Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

² Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

1 Introduction

We study the optimization of an appointment scheduling problem with unscheduled arrivals and reprioritization. This research was inspired by the HagaZiekenhuis, a Dutch hospital where such situations are encountered in the radiology department. Radiology departments offer diagnostic services to other hospital departments, and outside health care providers. In addition to outpatients who receive an appointment, diagnostics requests are also received from the emergency department and the wards. Of these requests, some patients require immediate attention, and should be diagnosed as soon as possible, while others are urgent but may wait for some time. These patients however must be seen within a given time frame. We shall refer to these patients as semi-urgent.

The problem we study can also be found in outpatient departments that are faced with patients with appointments, as well as unscheduled arrivals with varying degrees of urgency (Gupta and Denton 2008). Our contribution is that we build upon work in appointment scheduling by incorporating this reprioritization effect, and apply our approach to both theoretical instances and a case study. To this end, we use discrete event simulation (DES) as well as a constructive and local search heuristic to find appointment schedules that minimize waiting times. In addition, we compare the simulation model outcomes with results obtained from an exact model, and test our search heuristic, before applying it to a case study.

The structure of this paper is as follows. Section 2 gives a problem description, and in Sect. 3 we review the literature. Section 4 describes our approach, and Sect. 5 details the simulation model and heuristics used to find and evaluate appointment schedules. In Sect. 6 we provide results of this research, using both theoretical instances and a case study from HagaZiekenhuis. Finally, Sect. 7 provides a discussion and conclusion.

2 Problem description

In this paper, we aim to evaluate and optimize an appointment schedule for a given day where part of the patients arrive via appointment, and others arrive without prior notice. This problem is studied on both a tactical, as well as an operational level. On an operational level, models are used to find policies that specify which patients to prioritize in an upcoming time period (e.g., time slot or day), given the current state (e.g., number and urgency of waiting patients) and associated costs. In contrast, on a tactical level, the prioritization policies are assumed fixed, and the aim is to create appointment schedules for appointment patients such that, given a fixed prioritization policy, performance such as waiting times are minimized. Our problem falls within the second category, we assume the prioritization of patients is fixed, and we aim to create an appointment schedule that minimizes waiting times. While it may be beneficial to create situation dependent policies, in practice such policies may be difficult to implement, while it is simple to only schedule appointments given a new appointment schedule.

An appointment schedule consists of several time slots during the day where appointment patients are scheduled. Besides these known arrivals, there are time dependent unscheduled arrivals during the day. These unscheduled arrivals may vary in priority, some must be scheduled as soon as possible, while others have a set due date (e.g., must be diagnosed within 2 h). However, if unscheduled patients are not seen, their due date comes closer and by this their urgency increases. When diagnostics become available, the next patient is selected based on his/her urgency. As such, patients are prioritized by their time remaining until due date, with the patient(s) with the least time until due date selected first (i.e., remaining slack time). Specifically, the prioritization is as follows:

1. Unscheduled patients at their due date, ordered by waiting time (longest waiting time first)
2. Scheduled patients that arrived via appointment
3. Unscheduled patients not at their due date, based on time until due date (closest first), and (in case of equal time until due date) on waiting time (longest waiting first).

3 Literature

Given the common use of appointment systems by healthcare providers it is not surprising that outpatient scheduling is a topic of interest and has been studied for a long time, starting with Bailey and Welch (1952). Those unfamiliar with the vast amount of earlier work on this topic we refer to Cayirli (2003) who has provided an extensive literature review. As noted in Sect. 2, we consider outpatient scheduling on a tactical level, taking into account (time dependent) patient arrivals of multiple urgency types, and reprioritizations that take place when unscheduled patients are left waiting. Therefore, we look at recent work that incorporates multiple urgency types, as well as papers including both scheduled and unscheduled arrivals.

Recent work is done by Patrick and Puterman (2007). Herein both high priority inpatients, as well as lower priority outpatients must be scheduled for a CT scan. A policy is provided where capacity is reserved for each priority level, allowing for carrying over a portion of the unscheduled demand to the next day. In their case however, no reprioritization takes place. Patrick et al. (2008) model a diagnostic resource where patients of multiple priority classes may request diagnostics with the aim to allocate capacity (daily) among the different classes such that the number of patients exceeding their waiting time is minimized. In their case however, a scheduling policy is sought that governs how many patients (per class) to schedule per day. This differs from our problem in that the number of appointment patients is fixed, and we must distribute them throughout the day such that waiting times are minimized.

Similar to our problem, Kolisch and Sickinger (2008) model a radiology department with both scheduled and unscheduled patients of multiple priorities. Modeling the problem as a Markov decision process, they dynamically allocate

available capacity to the patients such that a given cost function is minimized, thus studying the problem on an operational level. A later paper by Sickinger and Kolisch (2009) evaluates and searches for appointment schedules that minimize a cost function. Also in this work, both scheduled and unscheduled patients are taken into account, however the priority of patients herein is fixed.

Koeleman and Koole (2012a) evaluate and search for appointment schedules while considering emergency arrivals that have priority over appointment patients. These emergency arrivals are taken into account when constructing appointment schedules. Using a local search algorithm they find the optimal solution minimizing the weighted sum of overtime, idle time and waiting times. They expand upon this including late and early arrivals of patients (Koeleman and Koole 2012b), however in both papers there are only two patient classes, and there is no reprioritization of untreated patients.

Cayirli et al. (2006) evaluate different appointment schedules from literature using computer simulation. Herein they take into account no-shows, as well as walk-ins, which are given a lower priority than appointment patients. They extend upon this study with flexible appointment intervals (based on patient type) (Cayirli et al. 2008) and again evaluate several appointment schedules. In our case however, we aim to find the best possible appointment schedule, and not evaluate several schedule possibilities. In addition, in both papers no reprioritization of patients is taken into account.

Kortbeek et al. (2014) present an approach for optimizing appointment schedules for outpatient clinics with both scheduled and unscheduled arrivals. Herein they use two models to determine both the number of appointments to be offered over a planning horizon, and the times during the day these appointments should be offered. The latter is similar to our problem, but they model only one unscheduled patient type, and ignore the reprioritization of patients. Using a local search heuristic they iteratively improve the quality of the appointment schedule.

Our contribution is threefold. First we build on recent work in appointment scheduling in health care by incorporating different patient types (both scheduled and unscheduled), as well as urgency levels. Herein we consider the reprioritization of unscheduled patients, reflecting that if lower urgency unscheduled patients wait too long, they will be prioritized over appointment patients. To our knowledge, this reprioritization has so far not been addressed in a similar problem setting, while often seen in practice. Second, we use a generic simulation model and generic search heuristics to systematically evaluate appointment schedules and search for good appointment schedules. Both the model and the heuristics are easily adapted to include other scenario specific aspects (e.g., no shows) and can thus be applied to other (health care) settings. In addition we validate our approach with a Markov reward process, which may be applied itself to smaller instances. Finally, our approach enables health care providers to create more balanced appointment schedules in a timely manner, taking into account waiting time for both elective and urgent patients, while taking into account the reprioritizations that take place in health care settings.

4 Assumptions and approach

In this section we detail the assumptions made in modeling the radiology department, as well as the approach taken to evaluate and optimize appointment schedules.

Our approach builds upon the approach by Kortbeek et al. (2014), wherein we (1) need a method to evaluate the performance of an appointment schedule, and (2) a local search approach which incorporates the aforementioned method and iteratively optimizes the schedule. The underlying assumptions of this approach are as follows. We divide a day into T time slots of equal length h , with C servers (e.g., CT scanners) available every day. The division of the day into slots of fixed length is motivated by the fact that, in practice, radiology appointment lengths are reduced in variability, as preparatory steps, such as administering contrast fluid (Elkhuizen et al. 2007), are externalized, and better protocols and reconfiguration times are established.

We assume scheduled patients arrive on time for their appointment, and that all diagnostics (both scheduled and unscheduled) require one time slot. Unscheduled patients that arrive during the day may have different urgencies, reflected by a due date (i.e., time slot in which they ultimately must be seen). Let R be the number of time slots a patient of the lowest urgency may wait, then unscheduled patients have a time until due date of r , with $r = 0, \dots, R$. We assume unscheduled patients arrive via a non-stationary Poisson arrival process denoted by λ_{tr} , for time slot $t = 1, \dots, T$, and (time until) due date r .

Unscheduled patients whose due date has not passed wait for scheduled patients, and all other patients that arrived earlier, with an earlier or equal due date, as well as patients with an earlier due date that may arrive during their wait. Once an unscheduled patient is at their due date, their priority no longer increases, and they are served as soon as possible. This means that they only wait for current patients being diagnosed, other patients that reached their due date earlier, and in case of equal due date, patients that arrived earlier. Note that urgent patients may also arrive that should be served as soon as possible (i.e., their due date starts at 0). Finally, scheduled patients wait for all unscheduled patients at or past their due date, as well as other scheduled patients that arrived earlier. During the day, appointments are also scheduled in the time slots, and the number of patients scheduled for slot t is denoted by x_t . An appointment schedule for the day is then described by: $\mathbf{x} = (x_1, \dots, x_T)$. The notation introduced in this section is listed in Table 1.

We use discrete event simulation (DES) to evaluate the performance of an appointment schedule, paired with constructive and Tabu local search heuristics to search for appointment schedules that minimize waiting times. We discuss the simulation model as well as the heuristics in Sect. 5. To verify our DES, we also model the radiology department as a Markov reward process (MRP), wherein a time slot corresponds to a single diagnostic session (both appointment and unscheduled) during which patients are diagnosed. This MRP is detailed and discussed in the “Appendix”.

Table 1 Notation introduced in Sect. 4

Symbol	Description
T	Number of time slots during a day
t	Time slot index ($t = 1, \dots, T$)
h	Length of a time slot
C	Number of CT scanners (resources)
R	Highest patient due date (i.e., lowest urgency)
r	Due date index ($r = 0, \dots, R$)
λ_{tr}	Arrival rate of unscheduled patient with due date r during time slot $[t-1, t]$
x_t	Number of appointments patients that arrive for slot t

The use of simulation modeling is motivated by the fact that the state space quickly expands, and quickly becomes intractable using the MRP, making it impossible to evaluate a single appointment schedule for a realistic case instance, let alone search for the optimal schedule. When verifying our simulation model, we compare simulation outcomes of small test instances with the MRP model results. In addition, we use the MRP to find the optimal appointment schedules for the test instances. This is done by enumerating all possible schedules. We evaluate the effectiveness of our simulation model and heuristics by comparing results to those of the optimal schedules. Section 6.1 describes the used test instances, and details of the MRP are included in the [Appendix](#).

5 Simulation model and search heuristics

In this section we describe the simulation model (Sect. 5.1), as well as the performance criteria used when evaluating appointment schedules (Sect. 5.2). Following this we present the constructive and local search heuristics for optimizing the appointment schedule (Sect. 5.3).

5.1 Simulation model

Within the simulation model, a day is simulated by “jumping through” (the start of) the time slots (t) that make up a day. At the start of every time slot, scheduled patients arrive based on the appointment schedule x_t , and according to the unscheduled patient arrival rates λ_{tr} . Similar to the opening and closing of a radiology department the model stops when all patients that have arrived in regular time have been treated, and there are no more patients left. Events take place in the following order:

1. Patient arrivals (scheduled and unscheduled) are determined.
2. Patients are selected to be treated (following the prioritization detailed in Sect. 2).
3. Urgencies of patients not treated are updated.

A patient list contains the queue of patients currently present in the system with all relevant information (e.g., arrival time, initial urgency, etc.). When a patient is treated, the waiting time is recorded. When new patients arrive, they are placed at their proper place within the patient list.

To determine the number of simulation runs, we perform sufficient simulations, such that the specified precision of the time slot with the highest variability (of waiting time) in a schedule has at most a relative error of 5%, with a confidence level of 95% (Kelton and Law 2000). We initialize the number of simulation runs (i.e., days) to 20,000, and use common random numbers when evaluating different schedules. With the initial number of simulation runs we find that all time slots with considerable waiting times fall within the specified precision. For the sparse slots where the waiting time is close to 0 this is not the case. Since we assess the performance of the schedule based on the maximum waiting time during the day (Sect. 1), these sparse slots have no impact on the objective. Therefore we do not increase the number of simulation runs. In addition, we verify the simulation model results with those obtained by the MRP, and find that all MRP results are enclosed within the simulation model confidence intervals.

5.2 Performance criteria

To evaluate the performance of an appointment schedule we are interested in the waiting time of both scheduled and unscheduled patients. Specifically, we want to minimize, and distribute evenly, the waiting time for scheduled patients (caused by unscheduled arrivals). In addition, the appointment schedule should be such that most unscheduled patients are seen on time. Therefore, we also require that a percentage of unscheduled patients, specified by a pre-set norm, are seen before their stated due date. We denote this on time percentage as OTP. We formulate $\mathbb{E}[W^{t,a}]$ as the expected waiting time of a scheduled appointment (denoted by the superscript a) patient arriving at time slot t . In addition, we denote $V_{t,r}$ as the probability that an unscheduled patient arriving at time t , with initial time until the due date of r slots, is not treated on time. Our objective as follows:

$$\begin{aligned} \min \max_t \mathbb{E}[W^{t,a}] \\ \text{s.t.} \\ V_{t,r} < 1 - \text{OTP} \quad \forall t, r \end{aligned}$$

5.3 Constructive and local search heuristics

To optimize the appointment schedule, a constructive heuristic generates an initial appointment schedule, after which a local search heuristic improves (upon) it. Given the goal of minimizing the maximum expected waiting time during the day, the constructive heuristic starts with an empty appointment schedule $\mathbf{x}_t = 0$, $t \in \{1, \dots, T\}$, and iteratively adds one appointment to the time slot $t' := \arg \min_t \max \mathbb{E}[W^{t,a}]$ until all required appointment slots K have been assigned. In

other words, starting with an empty schedule, the effect of adding an appointment to a slot is evaluated for every possible time slot. Then the appointment is added to the time slot that results in the lowest maximum waiting time encountered during the day, and the next appointment is similarly added. As such, given a current schedule, the appointment is added to the best available slot.

As we aim to minimize the maximum expected waiting time, in the optimal schedule the waiting time (per time slot) is spread out evenly across the day. Therefore, it is likely that moving an appointment slot from a busy slot to a quiet slot improves the schedule's performance. We therefore perform a Tabu search as follows. We denote v "from slots" and w "to slots" ($v, w \in \mathbb{N}^+$), which are, respectively, the time slots with the highest and lowest expected waiting time ($E[W^{t,a}]$). Moving an appointment from a high to low waiting time slot is then a neighbor solution, with the neighborhood consisting of all possible moves, specifically $v \cdot w$ solutions. Our Tabu search then accepts the best neighbor solution that is not tabu, and adds it to the tabu list of size L^{size} . This is repeated until r iterations have been done, or no feasible solution is found. We experimented with several local search techniques, and Tabu search settings, and found this approach best performing with respect to computation time and outcomes.

6 Experiments and results

To evaluate the performance of our local search heuristic we first apply it to small test instances, and compare performance with the optimal solution obtained from the enumerated results of the MRP model. Using test instances with varying parameter settings not only allows us to compare simulation results with the MRP, but also evaluate the heuristics under more general settings. In addition, we apply our approach to a case study of a Dutch hospital where both appointment patients are scheduled, and urgent arrivals take place. This section first details the input parameters of the artificial test instances (Sect. 6.1), followed by numerical results of the test instances (Sect. 6.2). Following this, we present the case study (Sect. 6.3), and numerical results of applying the heuristic approach to the case study (Sect. 6.4). Results in this section are obtained using the simulation model from Sect. 5.1, and in case of the test instances, compared with the (optimal) results from the MRP. Programming was done using the Delphi programming language from CodeGear and all experiments were run on an Intel 2.4 GHz PC with 4 GB of RAM.

6.1 Input parameters test instances

In our test instances we consider a department with two resources ($C = 2$), and a day consisting of 8 time slots ($T = 8$). We vary the arrival patterns of unscheduled arrivals during the day, with two types of unscheduled arrivals. Specifically, $\lambda_{t,0}$ is the arrival rate of urgent patients (at time t), and $\lambda_{t,R}$ the arrival rate of unscheduled patients that may wait for R time periods (due date: $t + R$). In our test instances,

each time slot has similar arrival rates for the two unscheduled arrival types ($\lambda_{t,0} = \lambda_{t,R}, \forall t$). The considered arrival patterns are shown in Fig. 1.

Arrival pattern 1 resembles the arrival rates as seen in practice, albeit at a smaller time scale, where arrivals increase into the start of the afternoon, and then decrease back to just above the start-of-day arrival rates. In addition, we evaluate an arrival pattern (2) with two peaks, where walk-ins are more likely to arrive at the start or end of the day (e.g., walk-in blood donations before or after work). Another arrival pattern encountered in practice may be a high initial arrival rate which decreases during the day (pattern 5), resembling waiting patients that arrived overnight and are waiting to see a health care provider. Finally, we also evaluate an arrival pattern with ever increasing arrival rates (pattern 3) and a single (very) large peak (pattern 4) to evaluate heuristic performance under diverse arrival scenarios. We vary the urgency of the less urgent patient that may wait R slots, with $R \in \{1, 3\}$. In addition we vary the number of appointment patients K that should be scheduled, with $K \in \{5, 8\}$. When changing K , we correct our unscheduled arrivals accordingly, such that overall utilization is approximately 80%. Finally, we set the on time probability OTP at 0.75. Both fixed and varied inputs, as well as Tabu search settings are listed in Table 2. In total we evaluate 20 test instances.

6.2 Results test instances

In this section we discuss the results of the local search heuristic regarding the test instances. Table 3 contains the outcomes of the search heuristic per instance. Per test instance the maximum expected waiting time ($\mathbb{E}[W^{t,a}]$) for appointment patients arriving during the day is given for the schedule found by the constructive and local search. To evaluate the effectiveness of our approach we compare the outcomes with the optimal schedules. Also, we compare the found schedules with the performance of the scheduling policy used in practice. Currently, appointments are

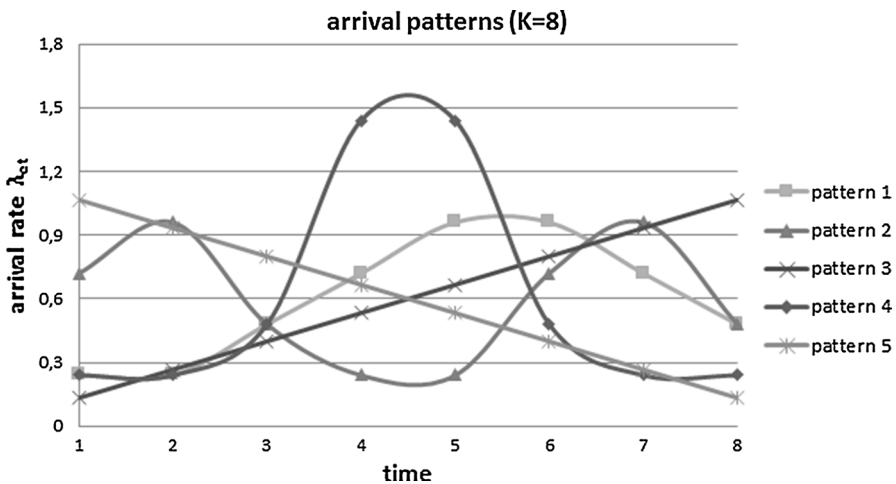


Fig. 1 Unscheduled patient arrival rates per time slot

Table 2 Fixed and varied input parameters

Parameters	Description	Value (s)
T	Time of day	8
S	Number of resources	2
OTP	% of unscheduled patients that must be seen on time	0.75
v	Number of from slots	3
w	Number of to slots	3
L^{size}	Tabu list size	10
r	Number of local search iterations	200
K	Number of appointments to schedule	{5, 8}
R	Highest due date (in slots) of unscheduled arrivals	{1, 3}
λ_{tr}	Arrival pattern of unscheduled patients	Pattern: {1, 2, 3, 4, 5}

booked every other time slot (e.g., 2, 0, 2, 0, ...). Finally, we show the additional improvement made by the local search heuristic over the constructive heuristic, as well as the runtime of the search heuristics.

Most schedules plan patients when the unscheduled arrival rate is low, such that the overall arrival rate of patients is leveled. The exceptions to this are the instances with arrival pattern 2. We illustrate this in Fig. 2, which concerns test instance 5. Here we see the arrival pattern, the number of appointments and the expected waiting time for scheduled patients. While this pattern has a peak at the start of the day, still patients are planned during these time slots. This makes sense as it takes time to have enough unscheduled patients to push back appointments. Also, we observe that if the maximum initial due date is high ($R = 3$), appointments are spread more over the day, as this allows unscheduled patients to wait longer and fill gaps in the schedule while still being treated on time.

We note that instance 13 has no feasible schedules. As this instance has a large peak in the middle of the day of unscheduled arrivals, all of which are (very) urgent, no schedule can guarantee that 75% of unscheduled patients are seen on time. In all other cases, the constructive and local search heuristic found the optimal schedule, or a good schedule in the case of instance 8. Here the found schedule plans one patient every slot, while the optimal schedule plans two patients in time slot four, and none in slot five (i.e., {1,1,1,2,0,1,1,1}). In Table 3 we also list the improvement (i.e., reduction of max. waiting time) of the schedules that is achieved by the local search heuristic over the initial constructive heuristic solution. We thus observe that the constructive heuristic is very effective for these small instances, in 11 out of 19 instances the optimal schedule is found. In addition, we see that our approach achieves considerable improvement over the current scheduling policy used in practice of booking appointments every other slot (e.g., 2,0,2,0, ...). Note that “inf” denotes instances where the base policy did not meet the on time probability constraint of unscheduled patients, and direct comparison is not possible. We conclude that the heuristic approach seems to find good schedules and apply it to our case study in the next section.

Table 3 Results test instances

No.	Pattern	Max. due date (R)	No. of apps. (K)	Found schedule	Max. waiting time ($\max \mathbb{E}[W^{i,d}]$)	Difference from opt (%)	Max. wait impr. over base pol. (%)	Impr. Local Search over constr. Heur. (%)	Runtime Sim (s)
1	1	1	5	{2,1,1,0,0,0,0,1}	0.323	0	Inf	0	295
2	1	1	8	{2,2,1,0,1,0,1,1}	0.214	0	-63.1%	-13.6%	346
3	1	3	5	{2,1,1,0,0,0,0,1}	0.128	0	Inf	0	276
4	1	3	8	{2,1,1,1,0,1,1,1}	0.109	0	-63.9%	0	363
5	2	1	5	{1,0,0,1,2,0,0,1}	0.363	0	Inf	0	351
6	2	1	8	{1,1,1,1,2,0,1,1}	0.217	0	-45.5%	-10.5%	406
7	2	3	5	{1,0,1,1,1,0,0,1}	0.165	0	-58.7%	0	357
8	2	3	8	{1,1,1,1,1,1,1,1}	0.128	0.005%	-55.1%	-10.1%	380
9	3	1	5	{2,2,0,1,0,0,0,0}	0.275	0	Inf	-26.4%	1158
10	3	1	8	{2,2,1,1,1,0,0,1}	0.188	0	-65.6%	-11.7%	332
11	3	3	5	{2,1,1,1,0,0,0,0}	0.133	0	-60.4%	0	244
12	3	3	8	{2,2,1,1,1,0,1,0}	0.105	0	-67.2%	-6.4%	312
13	4	1	5	No feasible schedule found	NA	NA	NA	NA	NA
14	4	1	8	{2,2,0,0,0,1,1,2}	0.206	0	-75.8%	-0.8%	249
15	4	3	5	{2,1,1,0,0,0,0,1}	0.166	0	-74.4%	0	264
16	4	3	8	{2,1,1,0,0,1,1,2}	0.122	0	-73.9%	331	0
17	5	1	5	{0,0,0,0,1,1,1,2}	0.335	0	-34.6%	0	798
18	5	1	8	{1,0,1,0,1,2,1,2}	0.219	0	-55.3%	-9.4%	906
19	5	3	5	{0,0,0,0,1,1,1,2}	0.182	0	-36.4%	0	903
20	5	3	8	{0,1,1,1,1,1,1,2}	0.111	0	-59.3%	0	1103

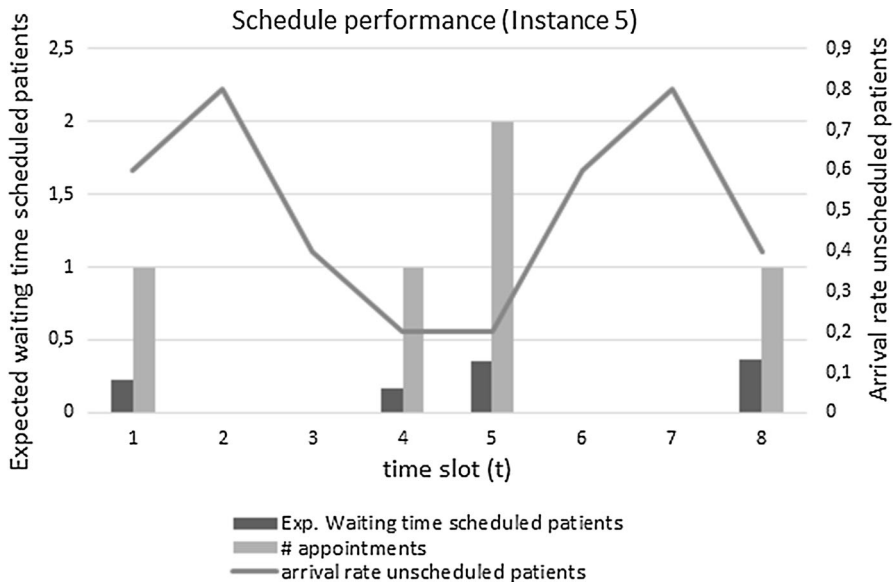


Fig. 2 Performance best found schedule for instance 5

6.3 Case study instance

HagaZiekenhuis has three CT scanners ($C = 3$), for which appointments are offered in 15 min slots. Regular opening hours of the department are from 8:00 to 16:30, resulting in 34 time slots per scanner ($T = 34$). Currently, every other time slot is reserved for scheduled patients. In practice, however, not all appointment slots are utilized, with the average number of appointments per day at $K = 36$. In addition there are unscheduled arrivals consisting of patients that must be seen as soon as possible, and patients that may wait for some time (2 h, $R = 8$). To estimate the unscheduled arrival rates, data of the total number of arrivals at the radiology department has been combined with arrival information from the Emergency Department. Figure 3 shows the arrival rates of unscheduled patients during the day. As available information was hourly based, the arrival rates per slot equal the hourly rates divided by four. The arrival pattern clearly follows (with a delay) the typical arrival pattern of the ED. In total, the utilization of the system is 77.2%. In addition, the required probability of unscheduled patients being treated on time is set at 90% ($OTP = 0.9$).

6.4 Results case study

This section presents the results of the case study. We first apply our approach to the current situation, and evaluate the found schedule. Also, we evaluate the effect of the appointment schedule on additional performance indicators. These are waiting time for unscheduled patients, utilization of CT scanners during the day, and overtime. Following the evaluation of the current situation, we then evaluate a

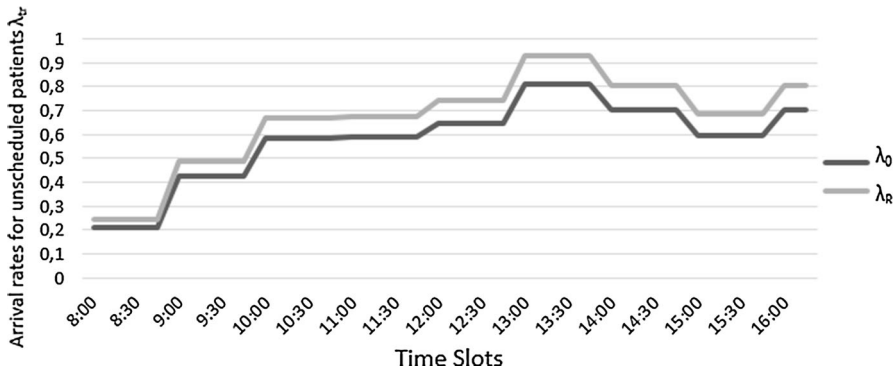


Fig. 3 Unscheduled patient arrival rates per time slot

what-if scenario with an increased utilization, where 20% more appointment patients are to be scheduled, to investigate performance with a potential increase of patient demand.

The runtime for the constructive and local search heuristic in the current scenario is 101.3 min, which is reasonable given that creating an appointment schedule only needs to be done every few months. Table 4 gives the performance of the current schedule, as well as the schedule found by the heuristic for the current situation. In the new schedule, more patients are scheduled at the start of the day. Following this, patients are scheduled evenly throughout the day, with a gap when there are many unscheduled arrivals expected. Similar to the test instances, appointments are spread out more over the available time slots, which reduces the peaks in scheduled patient waiting times. This makes sense, as for the time slots when three appointments are scheduled, only a single unscheduled (urgent) patient arriving causes one of the appointment patients to wait. The proposed schedule shows a reduction of the maximum (expected) waiting time during the day from 0.148 slots, to 0.046 slots (69% reduction), in comparison with the policy currently used in practice.

Figure 4 shows the waiting times specified throughout the day for appointment patients, for both the current and found schedule. In addition the arrival rates per time slot are displayed. Besides the reduction of the maximum expected waiting time, the even distribution also ensures a fairer distribution of waiting time across appointment slots. Under the current schedule, it is most beneficial to have an appointment just after the start of the day (around 9 AM), as the expected waiting time is considerably higher for appointments during the busiest moments in the afternoon. Conversely, this effect is lessened under the new schedule.

Table 4 Results case study (current situation)

Schedule	MaxE[W ^{1,σ}]	Schedule (x)
Current	0.148	{3,0,3,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0,2,0}
Heuristic	0.046	{2,2,2,2,1,2,1,2,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,1,1,1,1,1,1,1,1,1}

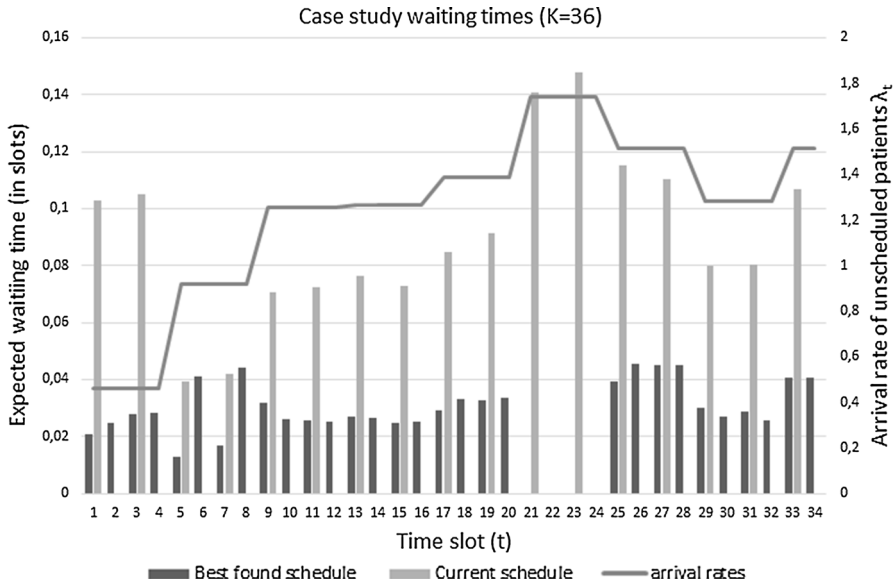


Fig. 4 Appointment patients expected waiting time per time slot (K = 36)

When evaluating the waiting times for urgent patients that should be seen as soon as possible we found that the differences between schedules were less prevalent, as these patients are always immediately prioritized over the lower urgency and appointment patients. We find that, for the current schedule, the fraction of unscheduled patients that are seen before their stated due date falls within the stated OTP norm (90%). Obviously, the results from the schedule resulting from the proposed algorithm also fall within the OTP norm of 90%.

We also evaluate the effect of the new schedule on waiting time for lower urgency unscheduled patients. Figure 5 shows the expected waiting time for the lower urgency patients, that should be seen within 2 h (i.e., 8 slots). We see that during the start of the day the expected waiting time is higher than under the current schedule. However, this stabilizes towards the afternoon. In contrast, later during the day the waiting time under the new schedule is at its lowest when no appointments are offered, and then remains below the waiting time of all time slots under the current schedule. From this we find that the new appointment schedule is not only beneficial for scheduled patients, it also reduces the average waiting time for unscheduled patients.

Besides the change in waiting times for appointment patients, the found appointment schedule also has an effect on utilization and overtime. As patients are scheduled more evenly across the day, not only are waiting times reduced, but also the utilization and workload during the day is more evenly spread. Comparing the schedules, we see that the under the current schedule the system is fully utilized when scheduled patients arrive, and during the slots with no scheduled arrivals the utilization steadily increases into the afternoon until it stabilizes. The new schedule however almost immediately stabilizes, and shows a drop in utilization when more

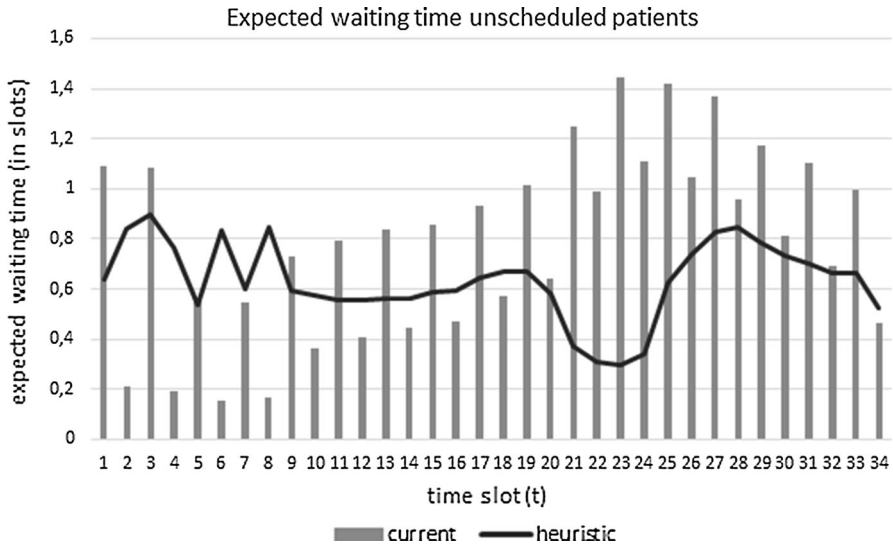


Fig. 5 Average waiting time of unscheduled patients that must be seen within 2 h ($r = 8, K = 36$)

unscheduled patients are expected, ensuring that unscheduled patients are seen on time.

Finally, Fig. 6 shows the overtime occurrence of both schedules. Specifically, it shows the fraction of times that the last patient is treated in the last time slot during regular time ($t = T$), and later time slots. We see that under the current schedule, 70% of the time the last patient is treated during regular time, and there is no overtime. With the new schedule this reduces to 65%. When patients are treated in overtime, this mostly extends to a single slot, which corresponds to an appointment of 15 min. Under both schedules, in less than 5% of the days the overtime is two slots or more. As the new schedule improves waiting times for both scheduled and unscheduled patients, this does come at a cost with regards to overtime.

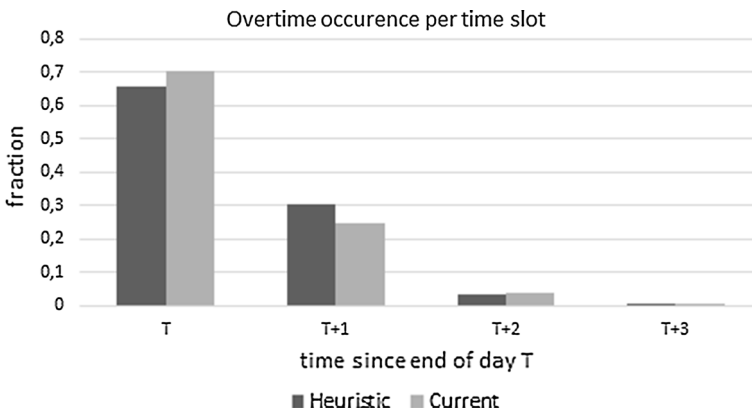


Fig. 6 Probabilities that the last patient treated falls outside regular hours ($K = 36$)

Table 5 Results case study (increased appointments)

Schedule	$\text{MaxE}[W^{t,a}]$	Schedule (x)
Current	0.262	{3,0,3,0,3,0,3,0,3,0,3,0,3,0,3,0,2,0,2,0,2,0,2,0,2,0}
Heuristic	0.081	{2,2,2,2,2,2,2,1,1,2,1,1,1,1,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1}

In addition to the current situation we also ran the heuristic for a scenario where eight more appointments should be scheduled during the day, resulting in a utilization of 85.5%. The runtime for this scenario was 115.7 min. Table 5 shows the outcome of the scenario using the current schedule and the heuristic. Similar to the current situation more patients are scheduled at the start of the day, and the remaining patients are scheduled evenly across the day. The maximum expected waiting time during the day is currently 0.148 slots, and 0.081 under the increased load and adapted schedule. Thus, using our approach, future increases in patient demand may be handled with the current capacity while still having acceptable waiting times. We conclude that our approach finds a good schedule within an acceptable amount of time for realistic instances, that takes into account the time dependent arrivals and reprioritization that take place in practice.

7 Discussion

In this paper we studied the optimization of an appointment schedule with time dependent unscheduled arrivals and this research was inspired by the HagaZiekenhuis hospital where such situations are encountered in the radiology department. To find appointment schedules that minimize patient waiting times in a timely manner, we use a discrete event simulation model in combination with constructive and local search heuristics. To validate this model, and evaluate the performance of the heuristics, we also evaluate multiple test instances using a Markov reward process model and enumerate all possible appointment schedules to find the optimal solution. From this comparison we find that our simulation and heuristics approach is able to find the optimal appointment schedule in most test scenarios, and finds good schedules in the remaining scenarios. We also apply our approach to a case study and find that significant waiting time reductions may be possible by adapting the slots available for appointments to the arrival rates of unscheduled patients.

Besides the reduction in waiting time for scheduled patients, the new schedule also shows a reduction in waiting time for lower urgency unscheduled patients, while still treating high urgency patients in time. Also, the utilization of CT scanners is spread evenly during the day, resulting in a more leveled workload. In addition to the even inflow of patients into the radiology department, the new schedule also allows for a more stable outflow of unscheduled patients back to wards and the Emergency Department. By spreading appointments during the day, also a more equitable distribution of the waiting time is obtained, so that patients having an appointment in the afternoon do not wait much longer than those arriving in the morning.

When looking at the found appointment schedules for both the current and busier scenario, we see that at no point three appointments are scheduled, as then only a single high urgency arrival means that an appointment patient is pushed back and must wait. We also observe in all experiments that the constructive heuristic by itself is already very effective in finding good solutions, as it iteratively adds appointment patients to the next best available time slot. In practice, our model may be used to periodically evaluate the offered appointment schedule, and re-optimize if either scheduled or unscheduled patient arrivals change.

For future work it may be interesting to evaluate the stochasticity of service times in practice. While in our case study, using the fixed service time assumption allows us to also enumerate exact results in order to compare our heuristics, incorporating these stochastic service times in follow up research can be interesting to see if a similar approach works. In addition, there can be seasonality effects in the arrival rates of unscheduled patients, or day-to-day effects regarding arrivals from the ED. In this research we use a single (averaged) day of arrival rates from the ED as little data on arrival numbers was available. However, our model and approach may be run for individual days with their unique arrival patterns, in order to construct appointment schedules for specific days. Regarding the unscheduled patient arrivals from the wards, while these are currently assumed as given, it may also be interesting to investigate the effect on the radiology department of changing the patient rounds on the wards, and thus the arrival rates of patients from the wards. Another example of future research may be the inclusion of no-shows of patients. By including this into the model the appointment schedule may also anticipate for this effect. Furthermore, different patient types or priority rules could also be incorporated, investigating the effect of differently prioritizing patients on the appointment schedule.

Concluding, our model generates appointment schedules, taking into account both elective and unexpected arrivals of patients, and the possible reprioritization that takes place between different patient types, during which patients may overtake each other and offers a practical and usable solution for practitioners. Applying our approach to a case study, the found appointment schedule considerably reduces expected waiting time for scheduled patients, while still ensuring unscheduled patients are seen on time. Additionally, the appointment schedule more fairly spreads the waiting time over patients, and results in a more even utilization, and thus workload, across the day.

Acknowledgements The authors thank the HagaZiekenhuis who inspired this study, provided data for our case study, and funded this project.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

In this appendix we describe the Markov reward process (MRP) used to explicitly evaluate the day process and its performance. By enumerating all possible appointment schedules and evaluating them with the MRP, the optimal schedule may be found for the test instances described in Sect. 6.1. Using these optimal schedules we compare the effectiveness of the simulation model and heuristics approach in Sect. 6.2. As noted, the radiology department is modeled as a Markov reward process. In the MRP a time slot corresponds to a single diagnostic session (both appointment and unscheduled) during which patients are diagnosed. To formulate our model we first give the state formulation followed by transition probabilities.

Model

State formulation: To formulate our model, we use the notation introduced in Table 1, introduced in Sect. 4. In addition, we denote u_r as the number of unscheduled patients with r remaining time slots until due date. Also, R denotes the highest time until due date an arriving unscheduled patient can have. Following this, $\mathbf{u} = (u_0, \dots, u_R)$ denotes all waiting unscheduled patients with different possible due dates. The state of the system is then denoted by the tuple (t, a, \mathbf{u}) , with a the number of scheduled (appointment) patients, and \mathbf{u} , the vector of unscheduled patients with respectively 0 to R time slots remaining before their due date, at the beginning of time slot t . Specifically, it is the number of patients in the system after patient arrivals at time t , and prior to patient selection for treatment in time t . For example, $(3, 1, (1, 0, 2))$, indicates that, at the start of time slot 3, there are four patients in the system. These are, respectively, one scheduled patient, one unscheduled patient at (or past) their due date, and two patients that may wait two more time slots. Notation introduced in this appendix is given in Table 6.

Table 6 Notation introduced in Sect. 7

Symbol	Description
a	Number of appointment patients present in the system at the beginning of a time slot.
u_r	Number of unscheduled patients with remaining due date of r in the system at the beginning of a time slot
\mathbf{u}	Number of unscheduled patients in the system at the beginning of a time slot, $\mathbf{u} = (u_0, \dots, u_R)$
(t, a, \mathbf{u})	State of the system: at the start of time slot t , there are a appointment and \mathbf{u} unscheduled patients present
$p_i^j(n)$	P [Number of arrivals of type i , ($i = a, u_0, \dots, u_R$), during slot $[t-1, t]$ is n]
$P[(b, \mathbf{f})_{t+1} (a, \mathbf{u})_t]$	P [Transition probability from state (t, a, \mathbf{u}) to state $(t + 1, b, \mathbf{f})$]
$n^a; n^u$	Number of treated scheduled and unscheduled patients respectively

Transition probabilities: Let the probability of going from state (t, a, \mathbf{u}) to $(t + 1, b, \mathbf{f})$ be denoted by $P[(b, \mathbf{f})_{t+1} | (a, \mathbf{u})_t]$, and $p_t^i(n)$ the probability that for time slot t , n patients arrive of type i , with $i = (a, u_0, \dots, u_R)$. Note that for scheduled patients $p_t^a(x_t) = 1$ (i.e., scheduled arrivals follow the appointment schedule). With an empty system, this gives:

$$P[(b, \mathbf{f})_{t+1} | (a, \mathbf{u})_t] = p_{t+1}^a(b) \cdot p_{t+1}^{u_0}(f_0) \cdot p_{t+1}^{u_1}(f_1) \cdot \dots \cdot p_{t+1}^{u_R}(f_R)$$

To formulate the transition probabilities for a non-empty system we introduce n^a , and n^{u_r} (for $r = 0, \dots, R$), which denote the number of treated scheduled and unscheduled patients with time until due date remaining of r respectively. Using this, we are able to define the state of the system after starting treatment of patients, but before new arrivals come in. Given the state of the system and prioritization, the number of treated patients per type is determined as follows:

$$\begin{aligned} n^{u_0} &= \min\{C; u_0\} \\ n^a &= \min\{\max\{C - u_0; 0\}; a\} \\ n^{u_r} &= \min\{C - n^a - \sum_{i=0}^{r-1} n_i; n_r\}, \quad \text{for } r = 1, \dots, R \end{aligned}$$

Continuing the previous example, suppose the state of the system is $(3, 1, 1, 0, 2)$ and there are two CT machines ($C = 2$), then the unscheduled patient at their due date, as well as the appointment patient are treated (e.g., $n^{u_0} = n^a = 1$). Following this, the state of the system after starting treatment of patients, but before new arrivals, is $(4, 0, 0, 2, 0)$. Note that the two unscheduled patients are one time slot closer to their due date. Using the number of treated patients and state of the system at time t , the partial transition probabilities per patient type are as follows:

$$\begin{aligned} & \text{(scheduled patients)} \\ & P[(b)_{t+1} | (a)_t] = p_{t+1}^a(b + n_a - a) \\ & \text{(unscheduled patients)} \\ & P[(f_r)_{t+1} | (\mathbf{u})_t] = p_{t+1}^{u_r}(f_r - (u_{r+1} - n^{u_{r+1}} + u_r - n^{u_r})), \text{ for } r = 0 \\ & P[(f_r)_{t+1} | (\mathbf{u})_t] = p_{t+1}^{u_r}(f_r - (u_{r+1} - n^{u_{r+1}})), \text{ for } r = 1, \dots, R - 1 \\ & P[(f_r)_{t+1} | (\mathbf{u})_t] = p_{t+1}^{u_r}(f_r), \text{ for } r = R \end{aligned}$$

The transition probability $P[(b, f)_{t+1} | (a, u)_t]$ can then be constructed by multiplying the partial transition probabilities:

$$P[(b, f)_{t+1} | (a, u)_t] = P[(b)_{t+1} | (a)_t] \cdot P[(f_0)_{t+1} | (u_0)_t] \cdot \dots \cdot P[(f_r)_{t+1} | (u_r)_t]$$

System performance

To evaluate the performance of an appointment schedule we are interested in the waiting time of both scheduled and unscheduled patients. Specifically, we want to minimize, and distribute evenly, the waiting time for scheduled patients (caused by

unscheduled arrivals), while a percentage, specified by a pre-set norm, of unscheduled patients are seen before their stated due date.

The waiting time of a patient depends on the state of the system at their time of arrival (i.e., the number of patients already present and of higher priority), as well as the number of patients that arrive later, but are still of higher priority. In the remainder of this subsection we first determine the arrival probabilities (i.e., the probability that an arriving patient sees a certain system state), and then the waiting time distributions conditioned on these encountered system states. Combining these two gives the complete waiting time distribution of a patient arriving at a certain time slot. The notation introduced in Sect. 1 is listed in Table 7.

Arrival probabilities: To determine the waiting time for patients we need the state probabilities seen by patients as they arrive. Suppose that just after starting patient diagnostics there are k patients of type i ($i = a, u_0, \dots, u_R$). Then an arriving patient of type i only sees k patients if he is the only arrival, or is the first of multiple arrivals. The probability that an arriving patient sees a certain state can be calculated by conditioning on the number of arrivals during the slot.

We denote $O_t(a, \mathbf{u})$ as the probability that at the start of time slot t , there are (a, \mathbf{u}) patients present. Additionally, we denote $O'_t(a, \mathbf{u})$ as the probability that at the start of slot t , after starting diagnostics and before new arrivals, there are (still) (a, \mathbf{u}) patients present. Using $O_t(a, \mathbf{u})$ and $O'_t(a, \mathbf{u})$ we can determine the arrival probability $Q^i_t(a, \mathbf{u})$, defined as the probability that a patient of type i , arriving for slot t , finds state (a, \mathbf{u}) . $O_t(a, \mathbf{u})$ and $O'_t(a, \mathbf{u})$ are calculated as follows:

$$\begin{aligned}
 O_1(b, \mathbf{f}) &= p_1^a(b) \cdot p_1^{u_0}(f_0) \cdot p_1^{u_1}(f_1) \cdot \dots \cdot p_1^{u_R}(f_R) \\
 O_{t+1}(b, \mathbf{f}) &= \sum_a \sum_{u_0} \dots \sum_{u_R} (O_t(a, \mathbf{u}) \cdot P[(b, \mathbf{f})_{t+1} | (a, \mathbf{u})_t]), \\
 &\text{for } t = 2, \dots, T
 \end{aligned}$$

Let n^u be the number of treated unscheduled patients, $n^u = (n^{u_0}, \dots, n^{u_R})$, then

Table 7 Notation introduced in Sect. 1

Symbol	Description
$O_t(a, \mathbf{u})$	P [Number of scheduled and unscheduled patients in slot t is (a, \mathbf{u})]
$O'_t(a, \mathbf{u})$	P [Number of scheduled and unscheduled patients in slot t , just after starting diagnostics, is (a, \mathbf{u})]
$Q^i_t(a, \mathbf{u})$	P [Patient of type i encounters state (a, \mathbf{u}) when arriving in slot $[t-1, t]$ ($i = a, u_0, \dots, u_R$)]
$O^{\tau,c,i}_t(a, \mathbf{u})$	P [Patient of type i , ($i = a, u_0, \dots, u_R$), arriving in state c ($c = b, \mathbf{f}$) in slot t , is in state (a, \mathbf{u}) after τ time slots]
$W^{\tau,i}_t(b, \mathbf{f})$	P [Patient of type i , ($i = a, u_0, \dots, u_R$), that arrives in state (b, \mathbf{f}) in slot t waits for τ time slots]
$W^{\tau,i}_t$	P [Patient of type i , ($i = a, u_0, \dots, u_R$), arriving at time t waits for τ time slots]
OTP	(Minimum) probability constraint that an unscheduled patient is treated on time

$$O'_{t+1}(b, \mathbf{f}) = \sum_{(a, \mathbf{u}) | [(a, \mathbf{u}) - (n^a, n^u) = (b, \mathbf{f})]} O_{t+1}(a, \mathbf{u}),$$

for $t = 2, \dots, T$

The probability that an arriving patient sees a certain state can be calculated by conditioning on the number of arrivals during the slot. Suppose the patient type under consideration are unscheduled patients with due date 0 (u_0) patients, then:

$$Q_1^{e_0}(b, \mathbf{f}) = p_1^a(b) \cdot \frac{1}{\lambda_{1,u}} \sum_{n=f_0+1}^{\infty} [p_1^{u_0}(n)] \cdot p_1^{u_1}(f_1) \cdot \dots$$

$$\cdot p_1^{u_R}(f_R)$$

$$Q_{t+1}^{u_0}(b, \mathbf{f}) = \sum_{(a, \mathbf{u})} [O'_t(a, \mathbf{u}) \cdot p_{t+1}^a(b - a)] \cdot \frac{1}{\lambda_{t+1,u}}$$

$$\sum_{n=f_0+1}^{\infty} [p_{t+1}^{u_0}(n)] \cdot p_{t+1}^{u_1}(f_1 - u_1) \cdot \dots$$

$$\cdot p_{t+1}^{u_R}(f_R - u_R), \text{ for } t = 2, \dots, T$$

Note that, $\frac{1}{\lambda_{1,u}}$ denotes the average size of an arriving 'group' of patients during a time slot. Similarly these probabilities may be constructed for the other patient types. Using this we can determine the waiting time distributions for scheduled and unscheduled patients.

Conditional waiting time of scheduled patients: As mentioned, the waiting time of an appointment patient depends on the number of appointment patients already waiting in the system, the patients with a higher urgency in the system, as well as patients that may (arrive and) become higher urgency during their waiting time. We derive the conditional waiting time distribution of a patient, by conditioning on the state of the system when the appointment patient enters (Litjens and Boucherie 2002). By setting the arrival rates of scheduled patients to 0 ($p^a(0) = 1$) for all future time periods, and defining all states where $(u_0 + a) < C$ as absorbing states. Then, the time it takes to get to an absorbing state, is the waiting time of the scheduled patient in the original system.

For notational simplicity, we denote $\mathbf{c} = (b, \mathbf{f})$, as the state encountered by an arriving scheduled patient, and the state conditioned on. Following this we can calculate the time until absorption, as once an absorbing state is reached, the tagged (under evaluation) patient is treated. We denote $O_{\tau}^{t,c,a}(a, \mathbf{u})$ as the probability that τ time slots after the tagged appointment patient's arrival in the conditioned state \mathbf{c} and time t , the system is in state (a, \mathbf{u}) . We then calculate $O_{\tau}^{t,c,a}(a, \mathbf{u})$ as follows:

$$O_0^{t,c,a}(\mathbf{c}) = 1$$

$$O_{\tau+1}^{t,c,a}(b, \mathbf{f}) = \sum_{(a, \mathbf{u}) | (u_0+a) \geq C} O_{\tau}^{t,c,a}(a, \mathbf{u}) \cdot P[(b, \mathbf{f})_{\tau+1} | (a, \mathbf{u})_{\tau}], \text{ for } \tau \geq 1$$

Note that $P[(b, \mathbf{f})_{\tau+1} | (a, \mathbf{u})_{\tau}]$ are the updated transition probabilities, where appointment patient arrivals are set to 0, starting from the initial time of arrival t . Using this approach we can determine the waiting time distribution for a given arrival state. Let $W_{\tau}^{t,a}(b, \mathbf{f})$ be the probability that a scheduled patient that arrives in state $\mathbf{c} (= (b, \mathbf{f}))$ at time t waits for τ time periods. This can be determined by:

$$W_{\tau}^{t,a}(b, \mathbf{f}) = \sum_{(a, \mathbf{u}) | (u_0+a) < C} O_{\tau}^{t, \mathbf{c}, a}(a, \mathbf{u}), \text{ where } \mathbf{c} = (b, \mathbf{f}) \tag{1}$$

Combining (1) with the arrival probabilities allows for the formulation of $W_{\tau}^{t,a}$, the probability that an arriving appointment patient at time t , regardless of arrival state, waits for τ time periods. Also, we formulate $\mathbb{E}[W^{t,a}]$, the expected waiting time for an appointment patient arriving at time t :

$$W_{\tau}^{t,a} = \sum_{(b, \mathbf{f})} Q_t^a(b, \mathbf{f}) \cdot W_{\tau}^{t,a}(b, \mathbf{f})$$

$$\mathbb{E}[W^{t,a}] = \sum_{\tau=0}^{\infty} \tau \cdot W_{\tau}^{t,a}$$

Lateness of unscheduled patients: Similar to calculating the waiting time of scheduled patients, we can determine the waiting time for unscheduled patients, and subsequently the lateness (waiting time past due date). Again we consider a system where the arrival rate of the unscheduled patient’s type, as well as all lower urgency patient types, is set to 0. In addition, as every unscheduled patient type that is not treated will be a higher urgency in the next time slot, we then also set the arrival rate of the tagged patient’s new type to 0, up to u_0 .

Finally, as the arriving patient is prioritized over all lower urgency patients, from the patient’s perspective these may be ignored (i.e., set to 0). Suppose the unscheduled patient’s due date is m slots from now, and arrives in state (a, \mathbf{u}) , then the conditioned on state can be formulated by $c = (b, \mathbf{f})$, with

$$\begin{cases} b = a, & \text{if } m \geq 0 \\ b = 0, & \text{otherwise} \end{cases}, \text{ and } \begin{cases} f_r = u_r, & \text{if } r < m \\ f_r = u_r - 1, & \text{if } r = m \\ f_r = 0, & \text{if } r > m \end{cases}$$

We denote $O_{\tau}^{t, \mathbf{c}, m}(a, \mathbf{u})$ as the probability that τ time slots after the tagged unscheduled patient’s arrival in the conditioned state \mathbf{c} , at time t and m slots until due date, the system is in state (a, \mathbf{u}) . We then calculate $O_{\tau}^{t, \mathbf{c}, m}(a, \mathbf{u})$ as follows:

$$O_0^{t, \mathbf{c}, m}(c) = 1$$

$$O_{\tau+1}^{t, \mathbf{c}, m}(b, \mathbf{f}) = \sum_{(a, \mathbf{u}) | k_{\tau}^m(a, \mathbf{u}) \geq C} (O_{\tau}^{t, \mathbf{c}, m}(a, \mathbf{u}) \cdot P[(b, \mathbf{f})_{\tau+1} | (a, \mathbf{u})_{\tau}]), \text{ for } \tau \geq 1$$

with, $k_{\tau}^m(a, \mathbf{u}) = a \cdot \mathbf{1}_{\{m-\tau > 0\}} + \sum_{r=0}^{\max\{0, m-\tau\}} (u_r)$, the sum of all patient types prioritized over the tagged patient with time until due date of m , at τ time slots after

arrival. Note that $\mathbf{1}_{\{x\}}$ is set to 1 if x evaluates true, and 0 otherwise, and $P[(b, \mathbf{f})_{\tau+1} | (a, \mathbf{u})_{\tau}]$ are the updated transition probabilities. Similar to the waiting time for scheduled patients, $W_{\tau}^{t,m}(b, \mathbf{f})$ denotes the probability that an unscheduled patient with time until due date of m , that arrives in state (b, \mathbf{f}) at time t , waits for τ time periods can be calculated by:

$$W_{\tau}^{t,m}(b, \mathbf{f}) = \sum_{(a, \mathbf{u}) | k_{\tau}^m(a, \mathbf{u}) < C} O_{\tau}^{t, \mathbf{c}, m}(a, \mathbf{u}), \text{ where } \mathbf{c} = (b, \mathbf{f})$$

To formulate the performance constraint of unscheduled patients, we denote OTP as the probability constraint that unscheduled patients must be treated before their due date. Second, $W_{\tau}^{t,m}$ denotes the probability that an arriving unscheduled patient in slot t , with time until due date m , regardless of arrival state, waits for τ slots. The performance can then be calculated by summing the probabilities of waiting past due date m , these are given respectively by:

$$W_{\tau}^{t,m} = \sum_{(b, \mathbf{f})} Q_t^m(b, \mathbf{f}) \cdot W_{\tau}^{t,m}(b, \mathbf{f})$$

$$\sum_{\tau=m+1}^{\infty} W_{\tau}^{t,m} < 1 - \text{OTP} \quad \forall t, m$$

As we aim to minimize and evenly spread the waiting time of scheduled patients during the day, while ensuring a pre-specified percentage of unscheduled arrivals is seen before their due date, we formulate our objective as follows:

$$\text{minimize } \max_t \mathbb{E}[W^{t,a}]$$

s.t.

$$\sum_{\tau=m+1}^{\infty} W_{\tau}^{t,m} < 1 - \text{OTP} \quad \forall t, m$$

References

- Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. *Prod Oper Manag* 12(4):519–549
- Cayirli T, Veral E, Rosen H (2006) Designing appointment scheduling systems for ambulatory care services. *Health Care Manag Sci* 9(1):47–58
- Cayirli T, Veral E, Rosen H (2008) Assessment of patient classification in appointment system design. *Prod Oper Manag* 17(3):338–353
- Elkhuizen SG, van Sambeek JR, Hans EW, Krabbendam KJ, Bakker PJ (2007) Applying the variety reduction principle to management of ancillary services. *Health Care Manage Rev* 32(1):37–45
- Gupta D, Denton B (2008) Appointment scheduling in health care: challenges and opportunities. *IIE Trans* 40(9):800–819
- Kelton WD, Law AM (2000) *Simulation modeling and analysis*. McGraw Hill, Boston
- Kooleman PM, Koole GM (2012a) Appointment scheduling using optimisation via simulation. In: *Proceedings of the winter simulation conference, winter simulation conference*

- Koeleman PM, Koole GM (2012b) Optimal outpatient appointment scheduling with emergency arrivals and general service times. *IIE Trans Healthcare Syst Eng* 2(1):14–30
- Kolisch R, Sickinger S (2008) Providing radiology health care services to stochastic demand of different customer classes. *OR Spectr* 30(2):375–395
- Kortbeek N, Zonderland ME, Braaksmā A, Vliegen IM, Boucherie RJ, Litvak N, Hans EW (2014) Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Perform Eval* 80:5–26
- Litjens R, Boucherie RJ (2002) Performance analysis of fair channel sharing policies in an integrated cellular voice/data network. *Telecommun Syst* 19(2):147–186
- Patrick J, Puterman ML (2007) Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization. *J Oper Res Soc* 58(2):235–245
- Patrick J, Puterman ML, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagnostic resource. *Oper Res* 56(6):1507–1525
- Sickinger S, Kolisch R (2009) The performance of a generalized bailey-welch rule for outpatient appointment scheduling under inpatient and emergency demand. *Health Care Manag Sci* 12(4):408–419
- Welch J, Bailey NJ (1952) Appointment systems in hospital outpatient departments. *Lancet* 259(6718):1105–1108

Nardo J. Borgman received his M.Sc. degree in 2012 in Industrial Engineering and Management from the University of Twente. He is currently a Ph.D. candidate at the HagaZiekenhuis hospital in The Hague, and at the University of Twente Center for Healthcare Operations Improvement and Research (CHOIR). His research is focused on the modeling and optimization of (acute) health care processes.

Ingrid M. H. Vliegen received her M.Sc. degree (2005) and Ph.D. degree (2009) in Industrial Engineering from Eindhoven University of Technology. From 2010 until 2016, she has been working as an Assistant Professor within the CHOIR research center and in the Department of Industrial Engineering and Business Information Systems of the University of Twente. Since 2016, she is working as project officer at the Eindhoven University of Technology, focusing on logistics and its interfaces.

Richard J. Boucherie received M.Sc. degrees in mathematics and theoretical physics from the Universiteit Leiden in 1988, and received the Ph.D. degree in econometrics in 1992 from the Vrije Universiteit, Amsterdam. Since 2000 he is with the department of Applied Mathematics of the University of Twente, where he was appointed in 2003 as Full Professor of Stochastic Operations Research. He is co-founder of the University of Twente CHOIR research center in the area of healthcare logistics, and chair of the Postdoctorate program in healthcare logistics. In 2014 he co-founded the spin-off company Rhythm, that carries out actual implementations of healthcare logistics solutions in healthcare organizations. His research interests are in queueing theory, Petri nets and random walks with application areas including wireless and sensor networks, healthcare, road traffic, and network intrusion detection and prevention.

Erwin W. Hans received his M.Sc. degree (1996) and Ph.D. degree (2001) in Applied Mathematics from the University of Twente. Since 2001 he has continued his academic career within Department of Industrial Engineering and Business Information Systems of the University of Twente, where he was appointed as Full Professor of Operations Management in Healthcare in 2013. He is Program Director of the Industrial Engineering and Management undergraduate and graduate programs. He is co-founder and co-chair of the University of Twente CHOIR research center and spin-off company Rhythm. His research interests are in the application of Operations Management/Operations Research in the healthcare domain.