Reducing Disruptive Effects of Service Interruptions in Appointment Scheduling

Matthias Deceuninck[1], Stijn De Vuyst[1] and Dieter Fiems[2]
[1]*Department of Industrial Systems Engineering and Product Design, Ghent University, Technologiepark 903, Zwijnaarde, Belgium*
[2] *Department of Telecommunication and Information Processing, Ghent University, Gent, Belgium*
{*Matthias.Deceuninck, Stijn.DeVuyst, Dieter.Fiems*}*@ugent.be*

Keywords: OR in Health Services, Appointment Scheduling, Stochastic Programming

Abstract: This paper considers appointment scheduling for outpatient services when the service of scheduled patients can be interrupted by emergency arrivals. We consider a single doctor who consults K patients during a fixed-length session. Each patient has been given an appointment time during the session in advance. Our evaluation approach aims at obtaining accurate predictions at a very low computational cost for the waiting times of the patients and the idle time of the doctor. To this end, we investigate a modified Lindley recursion in a discrete-time framework. We assume general, possibly distinct, distributions for the patient's consultation times and allow for individual no-show probabilities. This fast evaluation method is then used in a local search algorithm to provide insights into scheduling with service interruptions. Numerical examples show that this method outperforms simulation optimization and naive approaches in terms of cost and running time.

1 INTRODUCTION

Due to the demographic development and increasing need of health care services, health care providers are faced with certain operational challenges. In order to give timely medical access to all patients while still maintaining a high level of service, hospitals need to improve the efficiency of their processes. One of the tools to achieve this is the design of the appointment system. A good and effective appointment scheduling system tries to balance two important factors: the waiting times experienced by the patients and the idleness experienced by the service provider. Scheduling appointments closely together leads to longer waiting times but less risk of idleness for the doctor. On the other hand, spacing appointments far apart reduces the waiting times at the expense of increased idleness of the doctor. This dilemma becomes even more complex when we also consider emergency arrivals which have to be served as soon as possible.

In this paper, we investigate the optimization of appointment schedules with heterogeneous patients in the presence of no-shows and service interruptions. We primarily focus on service interruptions that are caused by emergency arrivals and require non-preemptive priority. This contribution builds upon the fast procedures to evaluate patient schedules under uncertainty introduced in Lau and Lau (2000) and De Vuyst et al. (2014). We then include this fast evaluation method in a local search algorithm and compare our results with simulation optimization and naive methods.

The outline of this paper is as follows. In the next section, we review the relevant literature. In Section 3 we introduce our mathematical model. The calculations of the performance measures are presented in Section 4. Section 5 demonstrates our approach and presents some numerical results. Section 6 concludes and suggests ideas for future work.

2 Literature review

An overview of the literature on appointment scheduling can be found in the survey papers Cayirli and Veral (2003) and Gupta and Denton (2008). While being prevalent in many service systems, limited attention has been given to service interruptions. In what follows we discuss contributions on service interruptions and emergency arrivals.

Fiems et al. (2007) developed a discrete-time queueing model with preemptive service of emergency patients and loss of work. Service times are assumed to be deterministic and steady-state analysis is carried out to investigate the impact on the waiting time of regularly scheduled patients in a radiology department. In Begen and Queyranne (2011), a non-preemptive approach is discussed in which emergency jobs may arrive during the processing of another job. The approach considered in the paper falls short in taking into account emergency jobs that arrive during idle time. This can be a restriction if the service times of emergency patients are longer than those of scheduled patients. In addition, there are also limitations on the number of emergency jobs that can arrive during the processing of a job.

Luo et al. (2012) proposed a model where service interruptions have an exponentially distributed duration and occur according to a, possibly non-homogeneous, Poisson process. Additionally, the service times of scheduled patients are assumed to be identically distributed according to an exponential distribution. Their results indicate that significant savings can be made by including interruptions in the evaluation and optimization model. They also report that when the interruption rate is high the optimal policy has a monotone structure rather than a “dome-shape”. Klassen and Yoogalingam (2013) used a simulation optimization approach to study the effects of service interruptions on outpatient appointment scheduling. They report that a “plateau-dome” scheduling rule is robust for low interruption rates. The present study most closely relates to Koeleman and Koole (2012), where the scheduling problem is studied for homogeneous patients.

Furthermore, the problem of service failures and service vacations is also studied in the traditional queueing literature. The vast majority of these papers however conduct a steady-state analysis, which does not really fit for the appointment scheduling problem where only a limited number of services are performed. For example, Fiems et al. (2004) considers a discrete-time queueing model in which the service process is subject to interruptions which are modelled as an on–off-process with geometrically distributed on-times and generally distributed off-times.

Finally, mixed arrival processes are also studied in Kolisch and Sickinger (2008) and Sickinger and Kolisch (2009). Besides regularly scheduled patients and emergency patients, these studies also consider unscheduled inpatients who are available for treatment at any time during the day. Kortbeek et al. (2014) considers a non-stationary stream of unscheduled patients without priority (walk-ins). Their goal is to balance the access time of scheduled patients and the waiting time on the day of service.

3 Model description

In this section we briefly describe the methodology used in this paper. We adopt the notation of De Vuyst et al. (2014), which provides an evaluation method for the appointment scheduling problem under the implicit assumption that there are no interruptions.

3.1 Mathematical model

We consider a consultation session of a single doctor, which is divided into T slots of equal length Δ. The session spans a time period of $[0, t_{\max}]$. Prior to this session, a practitioner has to choose K, the number of patients to be scheduled in this session and subsequently needs to allocate appointment times to each of these K patients. Let τ_k denote the slot that is assigned to the appointment of the kth patient. A schedule is then fully defined by the vector $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_K)$. We assume that all patients either arrive punctually at their appointed time or do not arrive at all (no-show). Let p_k denote the probability that the kth patient does not show up. We assume that the consultation times form a sequence of independent random variables. Let $s_k(n) = \Pr[S_k = n]$ denote the probability mass function of the consultation time S_k of the kth patient.

Emergency arrivals are modelled by a sequence of independent Bernoulli random variables $\{N_t\}$, $t = 0, \ldots, T-1$ with constant event probability α, $N_t = 0$ if no emergency arrived at slot t. Here, we assume that whenever an emergency patient arrives, he gets non-preemptive priority over the regularly scheduled patients. That is, once started, the service of a patient needs to be carried out till completion. If there are multiple emergencies, they are served in order. The inter-arrival times of emergencies thus constitute a series of geometrically distributed random variables. Finally, the consultation times of emergencies are modelled as a series of i.i.d. positive random variables with common probability mass function $s_e(n) = \Pr[S_e = n]$.

The fact that each patient can have an individual service time distribution and no-show probability allows us to take prior knowledge about the patients into account. For example, for each appointment request, the scheduler can estimate the service time distribution based on the patient’s characteristics like age and medical record. Similarly, no-show probabilities can

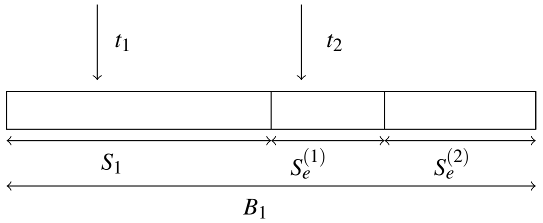

Figure 1: Illustration of the effective service time approach when there are two emergency arrivals, one at time t_1 and one at time t_2.

be estimated based on the type of service, appointment lead time and past no-show record.

Effective service times When emergencies occur during the service time of a patient, the waiting time of the next patient can be calculated by means of an effective service times approach. Such an approach replaces the service time of a patient by an effective service time which not only includes the patient's own service time, but also all time dedicated to emergency patients that arrived while the patient was being treated, as well as all service of emergency patients that arrived while an emergency patient was being treated, etc. Formally, the effective service time B_k starts when the patient's service time starts, and ends when the doctor becomes available for the next scheduled patient.

Let $S_k(z)$ be the probability generating function of the (discretised) service time S_k of the kth patient and $S_e(z)$ that of the consultation time S_e of an emergency patient, then the generating function of the effective service time B_k of the kth patient is

$$B_k(z) = S_k((\alpha B(z) + 1 - \alpha)z)$$

with $B(z)$ the probability generating function of the time to process a single emergency as well as all emergencies that arrived while processing this emergency, etc.

Because in every slot of the service time of the emergency there is a probability α to have a new emergency which needs to be processed, we have the following functional equation for the generating function $B(z)$,

$$B(z) = S_e((\alpha B(z) + 1 - \alpha)z) \tag{1}$$

We provide the derivation of (1) in Appendix. We thus have,

$$\mathrm{E}[B] = \frac{\mathrm{E}[S_e]}{1 - \alpha \mathrm{E}[S_e]}\,, \tag{2}$$

$$\mathrm{E}[B^2] = \frac{\mathrm{E}[S_e^2](1 - \alpha^2 \mathrm{E}[S_e])}{(1 - \alpha \mathrm{E}[S_e])^3} \tag{3}$$

and

$$\mathrm{E}[B_k] = \mathrm{E}[S_k](1 + \alpha \mathrm{E}[B])\,, \tag{4}$$

$$\mathrm{E}[B_k^2] = \mathrm{E}[S_k^2](1 + \alpha \mathrm{E}[B])^2 - \mathrm{E}[S_k](\alpha^2 \mathrm{E}^2[B] - \alpha E[B^2]) \tag{5}$$

We need the probabilities $b_k(n)$ corresponding to the generating function $B_k(z)$ as well as the corresponding moments. To obtain these, we first have to solve for the probabilities $b(n)$ corresponding to generating function $B(z)$. Since emergencies are independent of each other, the probabilities can be found by applying the property of composite generating functions:

$$\begin{aligned} B(z) &= S_e((\alpha B(z) + 1 - \alpha)z) \\ &= \sum_{m=0}^{\infty} s_e(m)\ ((\alpha B(z) + 1 - \alpha)z)^m \end{aligned}$$

The nth coefficient in the series expansion of $B(z)$ is the probability $b(n)$ equal to

$$\begin{aligned} b(n) &= \sum_{m=0}^{n} s_e(m)\ \ \Pr[\sum_{k=0}^{m} B_k A_k = n - m] \\ &= \sum_{m=0}^{n} s_e(m)\ \ x_m^*(n-m) \end{aligned}$$

with $x_m^*(n)$ recursively defined by

$$x_m^*(n) = \begin{cases} \mathbb{1}_{\{n=0\}} & \text{if } m = 0\,, \\ \sum_{\ell=0}^{n} x(n-\ell)\ \ x_{m-1}^*(\ell) & \text{otherwise.} \end{cases}$$

$$x(n) = \begin{cases} (1-\alpha) + \alpha s_e(0) & \text{if } n = 0\,, \\ b(n)\ \ \alpha & \text{otherwise.} \end{cases}$$

Analogously, for the nth probability of the effective consultation time we find

$$b_k(n) = \sum_{m=0}^{n} x_m^*(n-m)\ \ s_k(m)$$

Note that we need the emergencies to be independent of the scheduled patients as well as independent of each other.

Now, consider the situation where the doctor has finished the consultation of patient k and that there are no patients left in the waiting room. Without the possibility of emergencies arriving to the system, we know that the next scheduled patient, patient $k+1$, will experience zero waiting time. This is no longer true with emergencies since an emergency may arrive prior to the arrival of patient $k+1$. Since the idle times are bounded by the inter-arrival time, we can calculate the distribution of the waiting time of patient $k+1$. Let $g(n|i)$ denote this distribution, given that the idle time has length i, then

$$g(n|i) = \alpha\, b(n+i-1) + \alpha \sum_{\ell=0}^{i-1} b(\ell)\; g(n|i-\ell-1) \\ + (1-\alpha)\; g(n|i-1)$$

where the first term corresponds to the case an emergency period starts and ends after the idle time, the second to the case where an emergency period starts and ends before the end of the idle time and the last term corresponds to the case where no emergency arrives during the idle period. We are able to do the same for the moments of the waiting time, after an idle time. Let $\mathrm{E}[G_i^q]$ be the qth moment of the waiting experienced after an idle time of length i, then

$$\mathrm{E}[G_i^q] = \alpha \mathrm{E}[B^q \mathbb{1}_{\{B>i\}}] + \alpha \sum_{\ell=0}^{i-1} b(\ell)\, \mathrm{E}[G_{i-\ell-1}^q] \\ + (1-\alpha)\; \mathrm{E}[G_{i-1}^q]\,,$$

where the first expectation can easily be rewritten in terms of $\mathrm{E}[B]$ and the probabilities $b(n)$.

4 Performance measures

In this section we show how to evaluate the performance of a given schedule, i.e. assuming the appointment times $\boldsymbol{\tau}$ are fixed. We consider the following measures: the patient waiting times, the doctor's idle time and the session overtime.

First of all, we introduce a virtual arrival instant τ_{K+1} at the end of the session. This will enable us to calculate the overtime of the service provider (see Section 4.3). Furthermore, we introduce a notation for the time between consecutive appointment times: $a_k = \tau_{k+1} - \tau_k$, $k = 1, \ldots, K$ with $a_0 = \tau_1$. Note that, in accordance with this definition, a_K denotes the time between the appointment time of the last patient and the end of the session.

4.1 Waiting times

Let the waiting time W_k of the kth patient be the number of slots between the arrival of this patient and the start of his consultation. Consecutive waiting times then relate as

$$\mathrm{E}[W_{k+1}^q] = -\ell_k^{(q)} + \sum_{r=0}^{q} \sum_{m=0}^{q-r} \binom{q}{r} \binom{q-r}{m} \mathrm{E}[B_k^m] \\ \times (-a_k)^{q-r-m}\, \mathrm{E}[W_k^{(r)}] + \sum_{i=1}^{a_k} d_k(i)\, \mathrm{E}[G_i^q]\,,$$

with

$$\ell_k^{(q)} = \sum_{r=0}^{a_k-1} \sum_{m=0}^{r} b_k(r-m)\; w_k(m)\; (r-a_k)^q\,,$$

and with $d_k(i)$, the probability of an idle time of length i,

$$d_{k+1}(i) = \sum_{r=0}^{a_k-i} b_k(a_k-i-r)\; w_k(r)\,.$$

The probabilities of the waiting times are denoted by $w_k(n) = \Pr[W_k = n]$ and relate as

$$w_{k+1}(n) = \sum_{m=0}^{n+a_k} b_k(n+a_k-m)\; w_k(m) \\ + \sum_{m=0}^{a_k} \sum_{\ell=0}^{a_k-m} b_k(\ell)\; w_k(m)\; g(n|a_k-\ell-m)\,,$$

if $n > 0$ and

$$w_{k+1}(0) = \sum_{m=0}^{a_k} \sum_{\ell=0}^{a_k-m} b_k(\ell)\; w_k(m)\; g(0|a_k-\ell-m)\,.$$

The moments and probabilities of the first patients are treated separately. Note that it is possible that between the start of the session and the service of the first scheduled patient, emergency patients arrive and thus extend the waiting time of the first patient

$$\mathrm{E}[W_1^q] = \mathrm{E}[G_{\tau_1}^q]\,, \\ w_1(n) = g(n|\tau_1)\,.$$

Note that the calculations are also valid for a preemptive interruption if we assume that the waiting time is defined as the time the patient has to wait until the treatment starts and if there is no loss of work. That is, if we exclude any intermediate waiting time of a preemptive treatment of emergency patients.

4.2 Effective idle times

We define the idle time I_k of the kth patient as the time the doctor has to wait between the end of the service of the kth patient and the start of the service of the $(k+1)$th patient excluding any service of emergencies during this period of time. The qth moment of this effective idle time is then equal to

$$\mathrm{E}[I_k^q] = \sum_{r=0}^{a_k-1} \sum_{m=0}^{r} b_k(r-m)\; w_k(m)\; Z_k^q(a_k-r)\,,$$

with $Z_k(i)$ denoting the expected effective idle time given an idle time of length i,

$$Z_k^q(i) = \sum_{n=1}^{i} z_k(n,i) n^q$$

and,

$$\begin{aligned} z_k(n,i) = (1-\alpha)\, z_k(n-1,i-1) \\ + \alpha \sum_{m=1}^{i-1} z_k(n,i-m)\, b(m) \\ + \alpha\, z_k(n-1,0)\, \big(1 - \sum_{m=1}^{l-1} b(m)\big)\,. \end{aligned}$$

4.3 Overtime

The overtime O, which is the amount of time that the service provider works beyond the previsioned session length, can be calculated as the waiting time of the virtual patient. Indeed, if a patient were to be scheduled at the end of the session, this patient must wait till the overtime is completed. Hence, we find,

$$\mathrm{E}[O^q] = \mathrm{E}[W^q_{K+1}]\,.$$

4.4 Local Search Algorithm

Because of the sheer number of possible schedules, a heuristic method is required to find a good solution in a reasonable amount of time. Other studies in the literature have shown that local search procedures perform well for this type of problem (Kaandorp and Koole, 2007; Koeleman and Koole, 2012). The main idea of local search algorithms is to perform an iterative search throughout the solution space, by continuously evaluating and making small adjustments to a solution. The local search algorithm used in this study uses tabu search as a secondary heuristic and uses the search neighborhood $\mathcal{N}$ which is defined as,

$$\mathcal{N}(\boldsymbol{\tau}) = \{\boldsymbol{\tau}' : (\exists! k : \tau'_k = \tau_k \pm 1, \tau'_\ell = \tau_\ell, \ell \neq k)\}\,.$$

The algorithm is initialized with the best candidate solution from a reference set containing diverse solutions. The goal of the local search algorithm is to determine the vector of appointment times $\boldsymbol{\tau}$ which minimizes a certain objective function. For simplicity, we choose an objective function which only includes the first moments of the performance measures:

$$\mathrm{TC}(\boldsymbol{\tau}) = c_W\, \mathrm{E}[\sum_k W_k] + c_I\, \mathrm{E}[\sum_k I_k] + c_O\, \mathrm{E}[O]\,, \quad (6)$$

where c_W, c_I and c_O respectively denote the waiting, idle and overtime cost per time unit (e.g. dollars per time unit). Note that the relative importance of each term greatly depends on the type of service and organisation. The overtime cost c_O for example depends on the equipment and the number of assistants that are needed. In most environments, a greater weight will be assigned to idle time and overtime since the doctor's time is typically valued higher than the patient's time.

5 Numerical Results

In this section, we report the results of our numerical study. In particular, we focus on studying the effects of service interruptions and emergency arrivals on patient scheduling and the performance of our heuristic compared to simulation optimization.

5.1 Base case scenario

The parameters for our base case scenario are based on empirical results and assumptions made in prior studies. The parameters are given in Table 1. We use a time granularity of $\Delta = 1$ minute to make a reasonable trade-off between precision and computation time. In practice, data about service distributions will often be available as discrete data (a histogram). If this is not the case, discrete approximations can be obtained from the corresponding continuous distribution $\hat{S}$ as

$$s(n) = \Pr[\hat{S} < (n+\tfrac{1}{2})\Delta] - \Pr[\hat{S} < (n-\tfrac{1}{2})\Delta], \quad n \in \mathbb{N}.$$

Table 1: Parameters base case scenario

K	=	10 (number of patients)
$t_{\max}$	=	240 min (session length)
c_I	=	2 (idle time cost per time slot)
c_O	=	3 (overtime cost per time slot)
p_k	=	20% $\forall k$ (no-show probability)
α	=	0.5% (probability emergency arrival at slot)
$\hat{S}_k$	$\sim$	LogN($\mu = 25, \sigma$=15) $\forall k$
$\hat{S}_e$	$\sim$	Exp(μ=40)
Δ	=	1 min (slot length)

5.2 Comparison to simulation

First of all, to illustrate the usefulness of our exact evaluation algorithm, we compare its performance with simulation in terms of precision and running times. Most studies in the literature rely entirely on brute-force simulation to evaluate and optimize schedules. For example, Klassen and Yoogalingam (2014) applied a simulation optimization approach, in which they replicated each solution 10 000 times. Table 2 compares the computed values with the corresponding confidence intervals for their estimation by simulation for two sample sizes, i.e. 10 000 and 100 000 replications. It can be seen that the width of the 95% confidence intervals of the total cost TC is about 6% and 2% of the exact solution for respectively SS=10 000 and SS=100 000. The experiment is executed in Java Eclipse 2.0 on a Dell laptop with an i7-4900MQ 2.8 GHz processor and we find that the

Table 2: Comparison of computed values with the 95% confidence intervals of their estimates by simulation, for some performance measures of the base case scenario with a_k=a=24. (SS = sample size, number of replications)

Measure	Values	Simulated values	
		SS=10 000	SS=100 000
$E[W_2]$	8.93	8.43-9.62	8.88-9.24
$E[I_2]$	8.17	8.00-8.36	8.13-8.25
$E[W]$	272	259-279	270-277
$E[I]$	40.5	39.6-42.2	40.0-40.8
$E[O]$	63.8	62.0-65.3	63.5-64.5
TC	544	524-559	540-552

simulation run with sample size SS=10 000 (or 100 000) requires 1.3 (or 11) times more CPU time than our exact evaluation procedure which took less than 0.1s when we omit the preprocessing calculations of E[B] and E[G].

Next, we included our evaluation method in a local search algorithm as described in Section 4.4 to compare its performance with a simulation optimization method. Numerous test instances were developed to capture a diverse set of environments. For each instance, we run the local search algorithm five times using simulation to estimate a schedule's performance. We then compared the average cost TC_{sim} over these five runs with the cost obtained by using our exact evaluation method TC^*. The gap between these costs is defined as,

$$\text{gap} = \frac{TC_{\text{sim}} - TC^*}{TC^*} 100\%$$

The following parameters were represented in the experiment:

- The *number of patients K* in the schedule is equal to 6, 8, 10 or 12 patients.
- *Service interruptions* occur with a probability of α= 0.005 for each time slot, and are exponentially distributed with mean 30 minutes.
- The *service times* follow, before discretisation, a lognormal distribution with a mean equal to $\frac{200}{K(1-p_k)}$, resulting in an average of 200 minutes of work. The standard deviations σ_k are calculated in order to get coefficients of variation equal to one of the following values: $\{0.2, 0.4, 0.6\}$.
- The *no-show probability* p_k of a patient was selected from the set $\{0, 0.1, 0.2\}$.

This represents a total of 36 different environments. From a practitioners point of view, the choice of cost function is of great importance as well. To this end, we consider four different cost functions for which the c_I/c_O ratio is fixed at 1.5. The c_I level was then selected from the set $\{1, 2, 5, 10\}$. This adds up to a total of 144 test instances. These values reflect environments where patients' waiting times are highly valued as well as environments with high fixed costs for the service provider.

Table 3: Comparison between our approach and simulation optimization: the running time and gap between TC^* and TC_{sim} for sample sizes 10 000 and 100 000. The values in the table are the averages over the different scenarios for the given number of patients K.

	SS=10 000		SS=100 000		Exact
K	Gap (%)	Time (s)	Gap (%)	Time (s)	Time (s)
6	3.7	2.6	0.8	40.9	6.6
8	5.6	3.8	1.1	70.0	8.8
10	7.8	4.8	1.8	104.0	11.0
12	10.2	5.9	2.8	139.1	12.4

From Table 3 we can see that the exact evaluation method outperforms the simulation heuristics and significant cost reductions are obtained in about 10 seconds. Clearly, the variance on the simulated values has a big impact on the performance of the local search algorithm. The heuristic is often stuck in a suboptimal point after it underestimated the cost of a certain schedule.

Finally, we look at the performance of Bailey's rule in these environments. In Sickinger and Kolisch (2009), it is shown that Bailey's rule performs very well over a wide range of problem parameters if the cost of waiting is relatively low. Bailey's rule schedules two patients in the first slot, i.e. $\tau_1 = \tau_2 = 0$, while for the other patients the appointment time τ_k is equal to $\tau_{k-1} + E[S_{k-1}]$.

Figure 2 compares the heuristic solutions with Bailey's rule for the four different cost structures. It can be seen that the cost structure has a big impact on Bailey's performance.

5.3 Service time distribution emergency

In this section we look at the impact of the service time distribution of the emergencies S_e. We consider three different scenarios for the scheduled patients. For Scenario 1, we assume deterministic service times of 20 minutes and set $p = 0$. For Scenario 2, we assume $S_k \sim$ logN(20,4) with $p_k = 0$ while for Scenario 3 we set $p_k = 0.2$ and $\hat{S}_k \sim$ logN(25,12). For each scenario, we set K=10, $t_{\max} = 240$, $c_I = 2$ and $c_O = 3$.

For each scenario, we assume that the service time

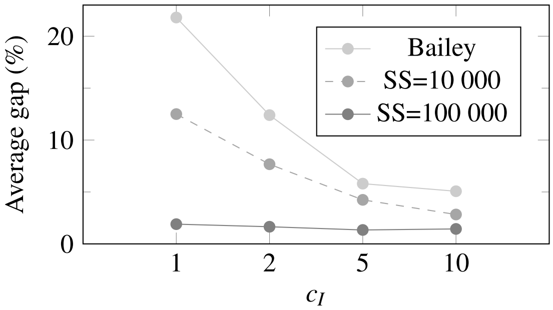

Figure 2: Effect of the cost structure on the performance of Bailey's rule and the heuristic solutions obtained with simulation. The gaps are averaged over all scenarios.

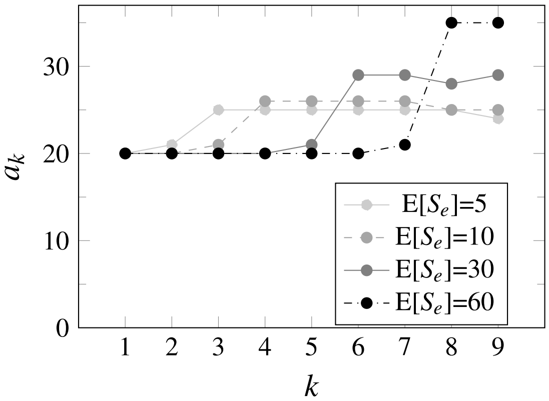

Figure 3: Heuristic solutions of inter-appointment times for Scenario 1.

of the emergency is exponentially distributed and vary its mean E[S_e], namely 5, 10, 30 and 60 minutes. The arrival rate of the emergencies, α, is chosen so that the expected effective service time is the same for each scenario (E[B_k]=22.22). For each instance of the problem, we first determine the local search solution by taking emergency arrivals into account. We denote the total cost of this heuristic solution as TC^*. In addition, we also determine the performance of the following policies: the policy that ignores emergencies and the policy that considers emergencies approximately by assuming that service times follow the corresponding distribution where the mean and variance are adjusted to its respective values of the effective service time given in equations 4 and 5. In the following, we use TC_{nointer} and TC_{approx} to denote the value of the total cost under these policies respectively.

Figure 3 depicts the heuristic solutions for Scenario 1 for different values of E[S_e]. Clearly, for this scenario, the best found inter-appointment times greatly depend on S_e and the interruption rate α. When the expected length of a service interruption is small and α is high, we find a dome-shaped pattern for the inter-appointment times.

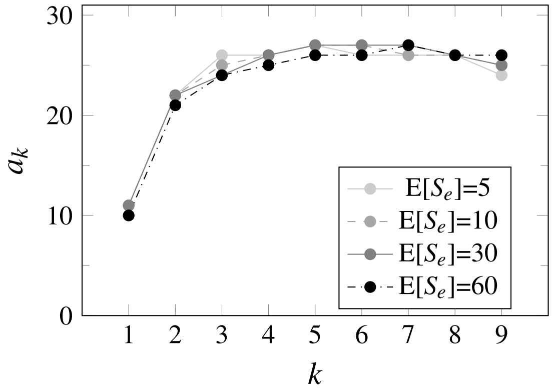

Figure 4: Heuristic solutions of inter-appointment times for Scenario 3.

Table 4: Numerical results for different service time distributions.

	E[S_e]	TC^*	TC_{nointer}	TC_{approx}
Scenario 1	5	121.2	190.8	122.4
	10	161.8	208.8	166.2
	30	234.9	248.9	248.4
	60	276.3	279.1	298.4
Scenario 2	5	152.9	157.6	153.2
	10	188.3	194.1	189.0
	30	259.0	265.8	263.9
	60	301.4	308.2	313.4
Scenario 3	5	360.4	362.0	360.5
	10	373.5	375.4	373.6
	30	416.5	418.3	416.7
	60	449.3	451.1	450.1

Figure 4 depicts the heuristic solutions for Scenario 3. It can be seen that in a more stochastic environment, the service time distribution of the emergencies S_e has a much smaller impact on the solution.

From Table 4, we can see that capturing the interruptions approximately by adjusting the service time distribution seems to work reasonably well for short and common service interruptions (low E[S_e], high α). However, when there are few other sources of variability (Scenario 1 and 2), the difference between TC^* and TC_{approx} is significantly greater for long and uncommon interruptions.

6 CONCLUSIONS

This paper presents a method to assess the moments of the waiting times of patients as well as the idle times and overtime of the doctor in a setting with emergency arrivals. The method allows patients to have general, distinct service time distributions and can handle no-shows. The algorithmic approach advocated here is fast in comparison with simulation and was included in a local search algorithm. Some numerical examples are presented in which we focus on the effects of emergency arrivals and service interruptions on patient scheduling. A possible direction for future research could be to investigate non-stationary arrival processes for the emergencies.

REFERENCES

Begen, M. A. and Queyranne, M. (2011). Appointment scheduling with discrete random durations. *Mathematics of Operations Research*, 36(2):240–257.

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

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

Fiems, D., Koole, G., and Nain, P. (2007). Waiting times of scheduled patients in the presence of emergency requests. *Technisch rapport. URL http://www. math. vu. nl/koole/articles/report05a/art. pdf,(Accessed on 18/12/2012)*, pages 1–19.

Fiems, D., Steyaert, B., and Bruneel, H. (2004). Discrete-time queues with generally distributed service times and renewal-type server interruptions. *Performance Evaluation*, 55(3):277–298.

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

Kaandorp, G. C. and Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science*, 10(3):217–229.

Klassen, K. J. and Yoogalingam, R. (2013). Appointment system design with interruptions and physician lateness. *International Journal of Operations & Production Management*, 33(4):394–414.

Klassen, K. J. and Yoogalingam, R. (2014). Strategies for appointment policy design with patient unpunctuality. *Decision Sciences*, 45(5):881–911.

Koeleman, P. M. and Koole, G. M. (2012). Optimal outpatient appointment scheduling with emergency arrivals and general service times. *IIE Transactions on Healthcare Systems Engineering*, 2(1):14–30.

Kolisch, R. and 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., Boucherie, R. J., Litvak, N., and Hans, E. W. (2014). Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Performance Evaluation*, 80:5–26.

Lau, H.-S. and Lau, A. H.-L. (2000). A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. *Iie Transactions*, 32(9):833–839.

Luo, J., Kulkarni, V. G., and Ziya, S. (2012). Appointment scheduling under patient no-shows and service interruptions. *Manufacturing & Service Operations Management*, 14(4):670–684.

Sickinger, S. and Kolisch, R. (2009). The performance of a generalized bailey–welch rule for outpatient appointment scheduling under inpatient and emergency demand. *Health care management science*, 12(4):408.

APPENDIX

Functional relation for the generating function of the unavailability time due to an emergency

Let B denote the time that the service provider is unavailable for scheduled patients due to the service of an emergency. This time equals the time that is needed to serve this initiating emergency as well as all other emergencies that arrived while serving these emergencies (Fiems et al., 2007):

$$B = S_e + \sum_{j=1}^{G_S} B^{(j)}$$

where G_S denotes the number of emergency arrivals during the service of the initiating emergency and where $B^{(j)}$ denotes the unavailable period corresponding to the jth emergency arrival during the service of the initiating emergency. In contrast to Fiems et al. (2007), service times are stochastic now. Due to the Bernoulli nature of the emergency arrival process, one easily verifies that the random variables $B^{(j)}$ are mutually independent and have the same distribution as B. This expression then translates into the following functional equation for the probability generating function $B(z)$ of the unavailable periods

$$\begin{aligned} B(z) &= E[z^{S_e+\sum_{j=1}^{G_S} B^{(j)}}] = E_S[E[z^{n+\sum_{j=1}^{G_n} B^{(j)}} | S_e = n]] \\ &= \sum_{n\geq 1} s_e(n)\; z^n E[z^{\sum_{j=1}^{G_s} B^{(j)}}] \\ &= \sum_{n\geq 1} s_e(n)\; z^n (1-\alpha+\alpha B(z))^n \\ &= \sum_{n\geq 1} s_e(n)\; [z(1-\alpha+\alpha B(z))]^n \\ &= S_e((\alpha B(z)+1-\alpha)z)\,. \end{aligned}$$