European Journal of Operational Research 265 (2018) 195–207

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

# Outpatient scheduling with unpunctual patients and no-shows

Matthias Deceuninck[a,*], Dieter Fiems[b], Stijn De Vuyst[a]

[a] Department of Industrial Systems Engineering and Product Design, Ghent University, Technologiepark 903, 9052 Zwijnaarde, Belgium
[b] Department of Telecommunication and Information Processing, Ghent University, St-Pietersnieuwstraat 41, 9000 Gent, Belgium

ARTICLE INFO

*Article history:*
Received 13 January 2016
Accepted 4 July 2017
Available online 8 July 2017

*Keywords:*
OR in health services
Appointment scheduling
Stochastic programming
Unpunctuality

ABSTRACT

We assess appointment scheduling strategies for outpatient services. To this end, we consider a fixed-length consultation session in which $K$ patients have to be scheduled at predefined appointment times, but who may not be punctual. Their effective arrival time deviates from their appointment time by a stochastic unpunctuality time. We assume general, possibly distinct distributions for the patients' consultation times as well as for their unpunctuality. The heterogeneity of the consultation times is motivated by patient classification: the schedule can be adapted to the patients' characteristics. Our evaluation approach is based on a modified Lindley recursion in a discrete-time framework and obtains accurate predictions for the moments of the patient waiting times as well as the doctor's idle times and overtime. This evaluation method is then included in a local search algorithm to provide general insights into appointment scheduling under unpunctual patients. Our results suggest that the proposed method obtains substantial cost reductions when patient classification is correctly exploited. Finally, it is shown that our analysis can also be used to determine optimal sequencing rules for patients who arrive out of turn.

## 1. Introduction

As the need for health care services continues to grow, health care providers are faced with challenges that have increased the attention for the field of Operations Research in health care considerably (Brailsford & Vissers, 2011). Given the current emphasis on preventive medicine and shorter lengths of stay at the hospital, one process clearly affected by the increase in patients is outpatient scheduling. Outpatient services constitute those medical procedures that can be performed in a medical center or hospital without an overnight stay.

The seemingly simple task of scheduling patients into the consultation session of a doctor is key in ensuring cost-effective health care and patient satisfaction. Like any service providing industry, health care providers can only cope with fluctuations in demand by using appointment systems to manage the arrival process. In doing so, clinics avoid overcrowded waiting rooms during peak hours and idle doctors in between. A good appointment system therefore aims to balance the following two conflicting objectives. On the one hand, service providers want to avoid both valuable resources being idle and the additional cost of having to complete the scheduled services during overtime. On the other hand, customers seek to minimize the time they spend waiting for the service. Indeed, waiting time has shown to be an important factor affecting the satisfaction of patients with medical services (Bardayan, 2002). Waiting does not only lead to economic costs for the society (e.g. opportunity cost), it also implies costs of a psychological nature, which amount to the stress that is incurred during the waiting period (Kocas, 2015; Osuna, 1985) or perceived quality of service and hence reputability of the health care provider. Therefore, it is no surprise that appointment scheduling has been an area of intense inquiry in the fields of operations research over the last sixty years. However, despite this great interest, the impact on outpatient clinics has been very limited (Cayirli & Veral, 2003). Most studies lack general applicability or have a set of very restrictive assumptions (Cayirli & Veral, 2003; Kuiper, Kemper, Mandjes, 2015).

In this paper, a numerical methodology for the evaluation of outpatient schedules is introduced where no-shows and the unpunctuality of patients are explicitly taken into account. The aim of our evaluation approach is to obtain accurate performance predictions at a low computational cost in order to make it suitable for heuristic search methods. This is achieved by (1) using the transient solution of a modified Lindley recursion which allows explicit expressions of each performance measure and (2) a discrete-time (slotted) setting to make those expressions easy to compute. Our analysis then essentially comes down to a discrete-time queueing model with a two-dimensional state space containing

[*] Corresponding author.
*E-mail addresses:* matthias.deceuninck@ugent.be (M. Deceuninck), dieter.fiems@ugent.be (D. Fiems), stijn.devuyst@ugent.be (S. De Vuyst).

http://dx.doi.org/10.1016/j.ejor.2017.07.006
0377-2217/© 2017 Elsevier B.V. All rights reserved.

both the amount of work in the queue and the unpunctuality on effective patient arrival times. The approach requires no assumptions on the consultation times of the patients other than that their distributions are known and that they are independent. The fact that each patient can have a general, distinct consultation time distribution, an individual unpunctuality distribution as well as an individual no-show probability, allows for evaluating schedules containing heterogeneous patients. This is an important feature in health care settings, because data may be available (or collected) on consultation times, no-show rates and punctuality of different types of patients and treatments. For example, Samorani and LaGanga (2015) recently showed that using individual no-show predictions may significantly improve the performance of a schedule by strategically scheduling ‘expected’ no-shows.

Finally, patients can also be given general, distinct costs of waiting since our approach allows for explicit expressions and moments up to any order. This contribution builds upon the development of fast and versatile algorithms for the evaluation of outpatient schedules introduced in De Vuyst, Bruneel, and Fiems (2014) and Fiems and De Vuyst (2013). The method in Fiems and De Vuyst (2013) is extended here to include no-shows and late cancellations under more general conditions as well as doctor lateness. Furthermore, we generalized and improved the analysis and investigated the impact of unpunctuality on patient scheduling using a local search algorithm. Our results suggest that practitioners can benefit by incorporating knowledge about the patients’ unpunctuality especially when there are few other sources of uncertainty. Our proposed analytic method can obtain substantial cost reductions in environments with low no-show rates and quasi-deterministic consultation times. In case of patient heterogeneity, it is shown that the cost of a schedule could be dramatically reduced by effectively using patient classification. Finally, we show that the evaluation algorithm can be used to address the problem of sequencing patients who arrive out of order. Should an idle doctor still follow the appointment order or should she see an early patient instead? Our analytical approach shows that this decision depends on several environmental factors.

The remainder of the paper is structured as follows. Section 2 reviews the relevant literature. Sections 3 and 4 then introduce the mathematical model and evaluation method, respectively. Section 5 presents some numerical results and discusses some insights. In Section 6, we address the problem of sequencing patients who arrive out of order. Finally, conclusions are drawn in Section 7.

## 2. Brief literature review

From a methodological point of view, the literature on appointment scheduling can generally be broken down into the following four approaches: (i) queueing models; (ii) simulation methods; (iii) mathematical programming; and (iv) heuristics. For a thorough overview on these approaches we refer to our previous related work (De Vuyst et al., 2014) and the survey papers (Ahmadi-Javid, Jalali, & Klassen, 2017; Cayirli & Veral, 2003; Gupta & Denton, 2008). Noting that most of the studies on outpatient scheduling focus on punctual and identical patients, we discuss contributions on patient heterogeneity and unpunctuality below.

Cayirli et al. report that the clinic performance can be improved by using patient-related information either for sequencing alone (Cayirli, Veral, & Rosen, 2006) or together with interval adjustment (Cayirli, Veral, & Rosen, 2008). This builds upon the conclusions of the earlier studies by Klassen and Rohleder (1996) and Rohleder and Klassen (2000) who evaluate the use of the patient’s consultation time variance on scheduling. More recent studies also recognize the importance of assuming heterogeneous patients and indicate the relevance of sequencing patients (Chen & Robinson, 2014; Creemers, Beliën, & Lambrecht, 2012; Erdogan, Gose, & Denton, 2017; Salzarulo, Mahar, & Modi, 2016). In Salzarulo et al. (2016) it is demonstrated how information on individual patient characteristics can be effectively incorporated in scheduling decisions.

Unpunctuality complicates the analysis of appointment-driven scheduling even further. Empirical studies indicate that patients tend to arrive early on average, but the variance in arrival time is high (Fetter & Thompson, 1966; White & Pike, 1964). Cayirli et al. (2006) found empirically that the earliness of patients at a primary health care clinic had an average of 17 minutes and a standard deviation of 27 minutes. White and Pike (1964) investigate the effects on efficiency and proposed distinct appointment schemes for the punctual and unpunctual case. Other simulation-based studies have shown that patient tardiness increases delays and extends the duration of the sessions (Meza, 1998; Perros & Frier, 1996). Alexopoulos, Goldsman, Fontanesi, Kopald, and Wilson (2008) give an elaborate discussion of the arrival process and state that unpunctuality is best modeled by an asymmetric Johnson distribution. Tai and Williams (2012) propose a mixed distribution, referred to as ‘F3’, to model patient unpunctuality and to provide a richer representation of shape. The idea of using mixed distributions is further investigated in Cheong, Bitmead and Fontanesi,Cheong (2013). To obtain a better fit, they treat patient earliness and lateness separately and model unpunctuality by a mixture of two one-sided exponential distributions.

Koeleman and Koole (2012a) recently noted that most papers that obtain numerical or structural results assume punctuality, Jouini and Benjaafar (2009) and Creemers, Colen, and Lambrecht (2012) being notable exceptions. Jouini and Benjaafar (2009) developed an exact analytical approach under the condition that customers always arrive in the order of their appointments and that they have exponential consultation time distributions. Various performance measures related to patient waiting time were obtained and with numerical results they illustrated the impact of not accounting for unpunctuality and no-shows. Klassen and Yoogalingam (2014) used a simulation optimization framework to investigate the effects of patient unpunctuality. They report that dome-shaped scheduling rules, for which times between appointments first increase and then decrease, still perform reasonably well in a variety of clinic environments, but propose a variable-length and block scheduling policy which is better at mitigating the adverse effects of unpunctuality. The same authors also studied the effect of unpunctuality of the doctor, again using a simulation optimization framework. It was shown that doctor’s unpunctuality, measured as the lateness to the start of the session, plays an important role (Klassen & Yoogalingam, 2013). This was also concluded by Fetter and Thompson (1966), which reported that the doctor arrived more than 15 minutes late in 37 of the 61 observed cases.

Finally, Williams, Chambers, Dada, McLeod, and Ulatowski (2014) examine the impact of an intervention intended to motivate patients to arrive more punctually: tardy patients were not seen and had to be rescheduled. They observed that the patients’ unpunctuality distribution was changed significantly after 12 months. Reducing tardiness does not always improve clinic performance however, since the waiting times increase because of early arrivals and the patients may be upset about having to reschedule. This clearly conflicts with the aim of appointment systems to provide a good service towards the patients and reduces the utilization of the clinic. On the other hand, a less drastic measure may lose its effect. Gneezy, Meier, and Rey-Biel (2011) discuss when and why some incentives may not work if they are not large enough, with Gneezy and Rustichini (2000) providing a good example of late-coming parents in a daycare center. It may thus be better to incorporate unpunctuality in the model rather than trying to change the patients' behavior.

## 3. Mathematical model and assumptions

We consider a single server in an outpatient clinic over the course of a session. During this time period $[0, t_{\max}]$, $K$ patients need to be given a fixed appointment time for their consultation so that a certain objective function is minimized. Scheduling the patients' appointments can be seen as a two-fold decision: we need (1) to decide in which *order* the $K$ appointments will be scheduled, i.e. which patient will be seen first, which one second and so on; and (2) to decide on the appointment times themselves. We refer to the first decision as 'sequencing' the patients and we relabel the patients according to the outcome of that decision as patient 1, 2, up to $K$. The second decision is on the appointment times $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ of these ordered patients.

In our model, we make the following assumptions to account for the complex environmental characteristics:

[A1] *Unpuctuality* is modeled by a sequence of independent random variables $\{U_k\}$ that denote the difference between actual arrival times and appointed arrival times. The unpunctuality $U_k$ is positive if the $k$th patient is late, negative if this patient is early and equals 0 if the patient arrives exactly at the appointed time. Let $u_k(n) = \Pr[U_k = n]$ denote the probability mass function (pmf) of $U_k$. We assume $U_k$ to have a bounded support $[\underline{u}_k, \overline{u}_k]$, with $\underline{u}_k$ and $\overline{u}_k$ respectively denoting the lower and upper bound for the $k$th patient.

[A2] The *consultation times* constitute a sequence of independent random variables. Let $s_k(n) = \Pr[S_k = n]$ denote the pmf of the consultation time $S_k$ of the $k$th patient.

[A3] *Late cancellations* and *no-shows* are modeled by a sequence of independent Bernoulli random variables $\{B_k\}$, $B_k = 0$ if the $k$th patient fails to show up for his consultation and $B_k = 1$ otherwise. It is assumed that the doctor is either informed prior to the appointment but too late to schedule a new patient (late cancellation) or that she is not informed at all (no-show).

[A4] *Doctor lateness* is modeled by a discrete random variable $\Theta$. Let $\theta(n) = \Pr[\Theta = n]$ denote the pmf of $\Theta$, with $n \geq 0$. We assume that the doctor remains available until the last patient is served or until it is evident that the patient will not show up any more.

These assumptions support heterogeneity in patient characteristics which allows us to take prior knowledge about the patients into account. For example, for each appointment request, the scheduler can estimate the required consultation time distribution based on the required type of medical treatment and the person's characteristics like age, medical record and the number of previous visits.

Throughout the analysis, a discrete-time setting is assumed. That is, all time-related quantities in the model, including appointment, arrival, consultation and unpunctuality times, are expressed as an integer multiple of the slot length $\Delta$. As argued by De Vuyst et al. (2014), a suitable choice of $\Delta$ follows from a trade-off: whereas using small slots ensures a maximal accuracy of the performance predictions, choosing large slots results in a lower computational effort.

It should be noted that patient unpunctuality can lead to the effect of patients overtaking each other, which is not possible if patients arrive exactly at their appointment time. For example, in case $\tau_2 + \overline{u}_2 > \tau_3 + \underline{u}_3$, patient 3 may arrive earlier than patient 2, contrary to their appointment order. The doctor may follow different queueing disciplines if she becomes available between those arrivals: either strictly following Appointment Order (AO) and remain idle until either patient 2 arrives or $\tau_2 + \overline{u}_2$ expires, or instead follow a work-conserving discipline and start the consultation of patient 3 right away. We will denote the latter policy where the doctor always chooses the patient in the waiting room with lowest appointment number as Appointment Order Work Conserving (AOWC). An extreme choice of discipline is first-come first-served (FCFS), where the appointment order no longer matters and patients are served in their actual arrival order. However, using a FCFS discipline may conflict with the goal of appointment scheduling, i.e. create a smooth flow of arrivals. As FCFS potentially rewards patients for being unpunctual, it may enable a strategic arrival behavior on the part of patients to 'beat' the system. For tractability, the model at hand assumes that the doctor strictly follows AO. In Section 5.2, we look at what happens if the queueing discipline differs from AO while in Section 6 we show that the optimal policy may actually lie somewhere between AO and AOWC.

Finally, for AO, there is a difference between how no-shows and late cancellations are handled. If patient $k$ does not inform the doctor that he will not show up, then due to the unpunctuality it is not clear prior to slot $\tau_k + \overline{u}_k$ if patient $k$ is merely late or a no-show so the doctor has to wait until $\tau_k + \overline{u}_k$ before serving the next patient. On the other hand, if the doctor is aware of the cancellation she should not wait and serve the following patients right away.

## 4. Schedule evaluation

In this section we show how to evaluate the performance of a given schedule, i.e. assuming the patient sequence and appointment times $\boldsymbol{\tau}$ are fixed. We consider the following measures: the patient waiting times, the doctor's idle time and the session overtime. The objective function $\mathrm{TC}(\boldsymbol{\tau})$ is then defined as a polynomial function in the moments of these performance measures.

First of all, to ease the exposition of our analysis, we create a modified representation of the system in which an arrival time is associated with every patient regardless of whether or not that patient actually shows. We will refer to this as the virtual system with virtual patients and AO discipline. In this virtual system, waiting times and idle times are calculated for all patients, including no-shows and late cancellations. After these calculations, we translate the results back to the original problem.

### 4.1. Performance of virtual system

We start by introducing a dummy arrival instant $\tau_{K+1} = t_{\max}$ at the end of the session for further use. This dummy patient is assumed to be punctual: $\overline{u}_{K+1} = \underline{u}_{K+1} = 0$ and will be useful in the calculation of the overtime (see Section 4.2). Furthermore, we also introduce notation for the time between consecutive appointment times:

$$a_k = \tau_{k+1} - \tau_k,$$

for $k = 1, \ldots, K$ and with $a_0 = \tau_1$. This will be referred to as the inter-appointment time. 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.

Since an arrival time is associated with every virtual patient, the question now arises which unpunctuality has to be assigned to no-shows and late cancellations. As discussed in Section 3, from the vantage point of the doctor, no-shows cannot be separated from a late patient prior to $\tau_k + \overline{u}_k$. Therefore, we set $U_{\text{no-show}} = \overline{u}_k$ to reflect that the doctor waits until this last possible arrival moment. Likewise, the unpunctuality of a late cancellation has to correspond to the time that the doctor is aware of the non-attendance

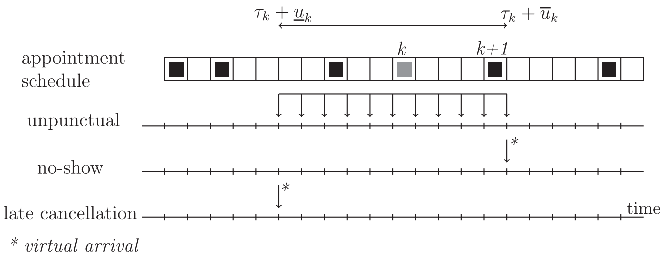

**Fig. 1.** Illustration of all possible arrival moments for virtual patient $k$ if he is unpunctual, is a no-show or cancels. We assume here that $\overline{u} = -\underline{u} = 5$ for all patients and $M = -5$. If patient $k$ does not show, his arrival time is equal to $\tau_k + \overline{u}_k$ as the doctor cannot separate him from a late patient before this slot. Likewise, if patient $k$ cancelled the appointment, his arrival moment lies before the possible arrival of any following patients.

of the patient. We assume that cancellations are known well before the arrival of any following patient and consequently, that the doctor does not unnecessarily delay any subsequent appointment. To limit the computational effort, we choose $U_{\text{cancel}} = \min\limits_{1\le k\le K} \underline{u}_k = M$ which means that cancellations occur $-M$ slots before their appointment time and ensures that no patient arrives in the virtual system before a cancellation of an earlier scheduled patient. In summary, a no-show is modeled as an extremely late patient while a late cancellation is interpreted as an extremely early patient, both with zero consultation time. This is illustrated in Fig. 1 where the three scenarios and the respective possible arrival moments of the virtual patients are shown.

Let $p_k^{\text{n}}$ and $p_k^{\text{c}}$, respectively, denote the probability that the $k$th patient is a no-show or a late cancellation, then the unpunctuality distribution $\tilde{U}$ of a virtual patient has pmf

$$\tilde{u}_k(j) = (1 - p_k^{\text{n}} - p_k^{\text{c}})\ u_k(j) + p_k^{\text{c}}\, \mathbb{1}_{\{j=M\}} + p_k^{\text{n}}\, \mathbb{1}_{\{j=\overline{u}_k\}}\,, \tag{1}$$

with, $\mathbb{1}_{\{.\}}$ denoting the indicator function which evaluates to 1 if all of its arguments are true and to 0 if this is not the case. The set of possible unpunctuality times of the $k$th virtual patient is given by $\Omega_k = \{M, \underline{u}_k, \underline{u}_k + 1, \ldots, \overline{u}_k\}$.

Note that the consultation time and unpunctuality of patients are no longer independent in the virtual system. By applying Bayes' theorem we find the conditional consultation time distribution $\tilde{S}_k|\tilde{U}$:

$$\begin{aligned}\tilde{s}_k(n|j) &= \Pr[S_k = n|\tilde{U}_k = j] = \frac{\Pr[S_k = n, \tilde{U}_k = j]}{\tilde{u}_k(j)} \\ &= \frac{(1 - p_k^{\text{c}} - p_k^{\text{n}})\ u_k(j)}{\tilde{u}_k(j)}\ s_k(n) \\ &\quad + \frac{p_k^{\text{c}}\, \mathbb{1}_{\{j=M\}} + p_k^{\text{n}}\, \mathbb{1}_{\{j=\overline{u}_k\}}}{\tilde{u}_k(j)}\, \mathbb{1}_{\{n=0\}},\end{aligned} \tag{2}$$

This approach incorporates the effects that no-shows and late cancellations have on the waiting times of subsequent patients.

In the remainder of this section we define the performance measures of this virtual system and show that they can be expressed in terms of the moments and probabilities of the random variables discussed above. Note that when we talk about patients in the calculations below, we mean the patients in the virtual system.

**System equations**. Let $A_k$ denote the interarrival time (expressed in terms of the slot length $\Delta$) between the $k$th and $(k+1)$st patient. In view of the modeling assumptions on the unpunctuality, we have,

$$A_k = a_k - \tilde{U}_k + \tilde{U}_{k+1}\,, \tag{3}$$

for $k = 1, \ldots, K$. Let the waiting time of the $k$th patient $W_k$ be defined as the number of slots between the arrival of this patient and the start of his consultation. Consecutive waiting times then satisfy Lindley's recursion,

$$W_{k+1} = (W_k - A_k + \tilde{S}_k)^+ = (Q_k)^+\,, \tag{4}$$

for $k = 1, \ldots, K$, with $(\cdot)^+$ a shorthand notation for $\max(0,\ \cdot)$, $Q_k$ being an auxiliary variable and with,

$$W_1 = (\Theta - \tau_1 - U_1)^+\,. \tag{5}$$

Note that, from the system equations (3) and (4), the consecutive waiting times clearly do not constitute a Markov chain. However, a Markov chain is obtained if we augment the waiting time with the corresponding unpunctuality: $(W_k, U_k)$ denotes the state of the chain for patient $k$. We now express the moments and probabilities of this chain for the $(k+1)$st patient in terms of moments and probabilities of the preceding patient.

**Waiting times**. First, consider the moments of the waiting time, conditioned on the unpunctuality. By conditioning on the unpunctuality of the $(k+1)$st patient and by accounting for the cases where $Q_k$ is negative, we find the following expression for the $q$th conditional moment,

$$\begin{aligned}\mathcal{W}_{k+1}^{(q)}(j) &= \mathrm{E}\left[W_{k+1}^q \mathbb{1}_{\{\tilde{U}_{k+1}=j\}}\right] = \mathrm{E}\left[Q_k^q \mathbb{1}_{\{\tilde{U}_{k+1}=j, Q_k\geqslant 0\}}\right] \\ &= -\,\mathrm{E}\left[Q_k^q \mathbb{1}_{\{\tilde{U}_{k+1}=j, Q_k<0\}}\right] + \mathrm{E}\left[Q_k^q \mathbb{1}_{\{\tilde{U}_{k+1}=j\}}\right] \\ &= -\ell_k^{(q)}(j) + \tilde{u}_{k+1}(j) \\ &\qquad \times \mathrm{E}[(W_k + \tilde{S}_k - a_k + \tilde{U}_k - j)^q | \tilde{U}_{k+1} = j], \\ &= -\ell_k^{(q)}(j) + \tilde{u}_{k+1}(j) \sum_{i\in\Omega_k} \sum_{r=0}^{q} \sum_{m=0}^{q-r} \binom{q}{r} \binom{q-r}{m} \\ &\qquad \times \mathrm{E}[\tilde{S}_k^m | \tilde{U}_k = i]\ (i - a_k - j)^{q-r-m}\ \mathcal{W}_k^{(r)}(i)\,,\end{aligned} \tag{6}$$

with,

$$\begin{aligned}\ell_k^{(q)}(j) &= \tilde{u}_{k+1}(j) \\ &\quad \times \mathrm{E}[(W_k + \tilde{S}_k - a_k + \tilde{U}_k - j)^q \mathbb{1}_{\{W_k+\tilde{S}_k+\tilde{U}_k<j+a_k\}} | \tilde{U}_{k+1} = j] \\ &= \sum_{i\in\Omega_k} \tilde{u}_{k+1}(j)\, \mathrm{E}[(W_k + \tilde{S}_k - a_k + i - j)^q \mathbb{1}_{\{\tilde{U}_k=i, W_k+\tilde{S}_k<a_k+j-i\}}] \\ &= \sum_{i\in\Omega_k} \sum_{r=0}^{a_k+j-i-1} \sum_{m=0}^{r} \tilde{u}_{k+1}(j)\ \tilde{s}_k(r-m|i) \\ &\quad \times (r - a_k + i - j)^q\ w_k(m,i).\end{aligned} \tag{7}$$

Here, $w_k(m, i) = \Pr[W_k = m, \tilde{U}_k = i]$ denotes the joint probability of waiting time and unpunctuality of patient $k$. These joint probabilities of consecutive patients relate as,

$$w_{k+1}(n, j) = \sum_{i\in\Omega_k} \sum_{m=0}^{n+a_k-i+j} \tilde{u}_{k+1}(j)\ \tilde{s}_k(n + a_k - i + j - m|i) w_k(m, i), \tag{8}$$

for $n > 0$ and $j \in \Omega_{k+1}$, whereas for the boundary case $n = 0$, we have,

$$w_{k+1}(0, j) = \sum_{i\in\Omega_k} \sum_{m=0}^{a_k-i+j} \sum_{\ell=0}^{a_k-i+j-m} \tilde{u}_{k+1}(j)\ \tilde{s}_k(\ell|i)\ w_k(m, i)\,, \tag{9}$$

for $j \in \Omega_{k+1}$. The expressions for the moments and probabilities of the first patient are treated separately. If the first patient arrives after the doctor, he is served right away. The $q$th moment of his waiting time is thus given by:

$$\begin{aligned}\mathcal{W}_1^{(q)}(j) &= -\tilde{u}_1(j) \sum_{n=0}^{\tau_1+j-1} \theta(n)\ (n - \tau_1 - j)^q \\ &\quad + \tilde{u}_1(j) \sum_{m=0}^{q} \binom{q}{m} \mathrm{E}[\theta^m]\ (-\tau_1 + j)^{q-m}\,,\end{aligned} \tag{10}$$

$$w_1(n,j) = \begin{cases} \tilde{u}_1(j)\ \theta(n+\tau_1+j)\,, & \text{for } n>0,\\ \tilde{u}_1(j)\sum_{m=0}^{\tau_1+j}\theta(m)\,, & \text{for } n=0\,. \end{cases} \tag{11}$$

Eqs. (6)–(11) now allow to determine the conditional moments of the waiting times of the consecutive patients. Clearly, by (6), we can recursively determine the conditional moments of the waiting times provided the moments of the waiting time of the first patient and a number of probabilities $w_k(m,i)$ are known, see (7). We would like to stress that only a finite number of probabilities needs to be determined, even if the consultation time is unbounded. Calculating the moments of the waiting time of the dummy patient $K+1$ requires to calculate the probabilities $w_K(m,i)$ for $m = 0 \to \tilde{a}_K - 1$ and $i \in \Omega_K$. This in turn requires calculation of $w_{K-1}(m,i)$ for $m = 0 \to \tilde{a}_K + \tilde{a}_{K-1} - 1$ and $i \in \Omega_{K-1}$, calculation of $w_{K-2}(m,i)$ for $m = 0 \to \tilde{a}_K + \tilde{a}_{K-1} + \tilde{a}_{K-2} - 1$ and $i \in \Omega_{K-2}$ and so on. Finally, all probabilities can be expressed in terms of $w_1(m,i)$.

Once the conditional moments of the waiting times are retrieved, we immediately find the (unconditional) moments,

$$\mathrm{E}[W_k^q] = \sum_{j\in\Omega_k} \mathcal{W}_k^{(q)}(j)\,, \tag{12}$$

**Idle times**. The idle time $I_k$ of the $k$th patient is defined as the time (expressed in terms of slots) the doctor has to wait for the arrival of this patient. By similar arguments as those for Lindley's equation (4), we find,

$$I_{k+1} = (A_k - \tilde{S}_k - W_k)^+ = (-Q_k)^+\,,$$

such that $W_{k+1} - I_{k+1} = Q_k$. Clearly, a non-zero waiting time $W_k > 0$ implies a zero idle time $I_k = 0$ and vice versa. Hence, $\mathrm{E}[W_k^q I_k^r] = 0$ for $q$, $r \geq 1$, which further yields for $q \geq 1$,

$$\begin{aligned} \mathrm{E}[W_{k+1}^q] + (-1)^q\,\mathrm{E}[I_{k+1}^q] &= \mathrm{E}[((Q_k)^+)^q] + (-1)^q\,\mathrm{E}[((-Q_k)^+)^q] \\ &= \mathrm{E}[(W_k - A_k + \tilde{S}_k)^q]\,. \end{aligned}$$

Solving for $\mathrm{E}[I_{k+1}^q]$ and using (6) and (7), we find,

$$\begin{aligned} \mathrm{E}[I_{k+1}^q] &= (-1)^q\big(\mathrm{E}[(W_k - A_k + \tilde{S}_k)^q] - \mathrm{E}[W_{k+1}^q]\big) \\ &= (-1)^q \sum_{j\in\Omega_{k+1}} \ell_k^{(q)}(j)\,. \end{aligned}$$

Finally, the conditional moments of the idle time of the first patient can easily be obtained as,

$$\mathcal{I}_1^{(q)}(j) = \tilde{u}_1(j) \sum_{m=0}^{\tau_1+j} \theta(m)\,(\tau_1 + j - m)^q\,.$$

**Modified waiting time**. Some studies exclude any waiting prior to the appointment time (Cayirli et al., 2006, 2008). From the clinic's perspective, waiting due to early arrival may indeed not be perceived as a shortcoming of the appointment system. To this end, we define the modified waiting time of a patient as the time between the appointment time and the start of the consultation if the patient is early, and to the arrival instant of the patient and its consultation time if this is not the case. Clearly, the modified waiting time $\mathring{W}_k$ relates to the waiting time as follows,

$$\mathring{W}_k = \mathbb{1}_{\{-\tilde{U}_k < W_k\}}(W_k - (-\tilde{U}_k)^+)\,,$$

such that the moments of the modified waiting time can be calculated in the same way as (6) and (7),

$$\begin{aligned} \mathring{\mathcal{W}}_{k+1}^{(q)}(j) &= \mathrm{E}[\mathring{W}_k^q\,\mathbb{1}_{\{\tilde{U}_k=j\}}] \\ &= -\mathring{\ell}_k^{(q)}(j) + \sum_{i\in\Omega_k}\sum_{r=0}^{q}\sum_{m=0}^{q-r} \tilde{u}_{k+1}(j)\binom{q}{r}\binom{q-r}{m} \\ &\quad \times \mathrm{E}[\tilde{S}_k^m | \tilde{U}_k = i]\,(i - a_k - (j)^+)^{q-r-m}\,\mathcal{W}_k^{(r)}(i)\,, \end{aligned}$$

with,

$$\begin{aligned} \mathring{\ell}_k^{(q)}(j) &= \tilde{u}_{k+1}(j) \\ &\quad \times \mathrm{E}[(W_k + \tilde{S}_k - a_k + \tilde{U}_k - (j)^+)^q\,\mathbb{1}_{\{W_k+\tilde{S}_k+\tilde{U}_k<(j)^++a_k\}} \\ &\quad \times | \tilde{U}_{k+1} = j] \\ &= \sum_{i\in\Omega_k}\ \sum_{r=0}^{a_k+(j)^+-i-1}\ \sum_{m=0}^{r} \\ &\quad \tilde{u}_{k+1}(j)\tilde{s}_k(r-m|i)(r - a_k + i - (j)^+)^q w_k(m,i). \end{aligned}$$

When $q = 1$, the expected modified waiting time can also be related to the expected waiting time as,

$$\begin{aligned} \mathring{\mathcal{W}}_{k+1}^{(1)}(j) &= \mathbb{1}_{\{j<0\}}\big[\tilde{u}_{k+1}(j)\,j + \mathcal{W}_{k+1}^{(1)}(j) + \ell_k^{(1)}(j) - \mathring{\ell}_k^{(1)}(j)\big] \\ &\quad + \mathbb{1}_{\{j\geqslant 0\}}\mathcal{W}_{k+1}^{(1)}(j)\,. \end{aligned}$$

For the first patient, the moments of the modified waiting time equal,

$$\mathring{\mathcal{W}}_1^{(q)}(j) = \mathbb{1}_{\{j<0\}} \sum_{r=0}^{q}\binom{q}{r} j^{q-r}\,\mathcal{W}_1^{(r)}(j) + \mathbb{1}_{\{j\geqslant 0\}}\mathcal{W}_1^{(r)}(j)\,.$$

### 4.2. Translating performance measures back to actual problem

After calculating the performance measures in the virtual system, we are now interested in their counterparts for the original problem. Fig. 2 depicts both systems for an example path. It can be seen that a waiting time and idle time is associated with a virtual patient regardless of whether or not he shows in the actual system. In what follows we calculate the various measures for the actual system from the ones in the virtual system.

**Waiting times**. Since we are only interested in the waiting time that is actually experienced by patients, we need to disregard any waiting time that no-shows and late cancellations encounter in the virtual system. Let $\mathbb{W}_k$ denote the actual waiting time of the $k$th patient, then from (12) we find:

$$\mathrm{E}[\mathbb{W}_k^q] = \mathrm{E}[W_k^q] - \frac{p_k^{\mathrm{n}}\,\mathcal{W}_k^{(q)}(\overline{u}_k)}{\tilde{u}_k(\overline{u}_k)} - \frac{p_k^{\mathrm{c}}\,\mathcal{W}_k^{(q)}(M)}{\tilde{u}_k(M)}\,,$$

with $q$ being the order of the conditional moment and for $k = 1,\ldots,K$. Likewise, we find the $q$th moments of the actual modified waiting time $\mathring{\mathbb{W}}_k^q$ by subtracting the conditional moments of the modified waiting times of no-shows and late cancellations.

**Idle times**. For the idle times, the situation is a little more complex. In the virtual system, we defined the idle time of the $k$th patient as the period before the arrival of the $k$th patient in which the doctor is waiting because the consultation of patient $k-1$ is already finished. However, as depicted in Fig. 2, the actual idle time $\mathbb{I}_k$ of the $k$th patient may be longer when patient $k-1$ fails to show and the doctor was already idle. Thus, in the presence of no-shows and late cancellations, the moments of the idle times of virtual patients do not correspond with the actual idle times. Fortunately, the sum of the first moments are the same:

$$\sum_{k=1}^{K}\mathrm{E}[\mathbb{I}_k] = \sum_{k=1}^{K}\mathrm{E}[I_k]\,.$$

Note that if patient $K$ does not show, $\mathbb{I}_K$ is measured as the time between the last consultation and the end of the session. The doctor remains available until $\tau_K + \overline{u}_K$, i.e. the latest possible arrival time of the patient.

**Overtime**. The overtime $O$ is equal to the amount of time that the doctor has to remain present at the hospital after the previsioned session length. Recalling the dummy punctual patient added at the end of the session, it is easy to see that the waiting time of this patient exactly corresponds to the doctor's overtime. If

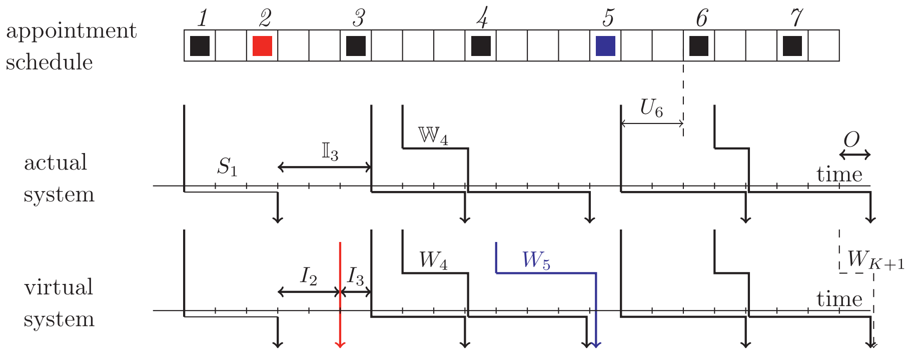

**Fig. 2.** Illustration of the transformation to the virtual system for a schedule with $K = 7$ and where the second patient (red) fails to show without notice and the fifth patient cancelled the appointment (blue). For all patients, we assume $\overline{u} = -\underline{u} = 3$. In the virtual system, an arrival time, waiting and idle time is also associated with no-shows and cancellations. The waiting time of the dummy patient at the end of the schedule corresponds to the doctor's overtime. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a patient were to be scheduled at the end of the session, this patient must wait till the overtime is completed. Hence, we directly find,

$$\mathrm{E}[O^q] = \mathrm{E}[W_{K+1}^q] = \mathcal{W}_{K+1}^{(q)}(0)\,.$$

### 4.3. Numerical complexity

We discuss the numerical complexity under the assumption that (i) the unpunctuality ranges from $-u$ to $u$ for all patients and (ii) all $a_k = a = 2u$ are equal. As the main computational effort comes from the computation of the probabilities $w_k(n, i)$, we focus on the number of floating point multiplications that are needed to calculate all probabilities required to calculate the moments of waiting and idle times of all patients. Carefully counting the number of multiplications in (8) and (9) and only accounting for the probabilities needed in (7), yields a total of $64/3\,K^3u^4 + O(u^3K^3)$ floating-point multiplications. Since the number of patients in a session is practically limited, $K^3$ is not overly large, and poses no computational issues. Given the amount of time between patients, the value of $u$ (or $a$) depends on the discretization of time, that is, on the value of $\Delta$. Small $\Delta$ yield large $a$ and $u$: increasing the accuracy leads to considerably more computational effort. To give an idea, a large schedule of 4 hours with a granularity of $\Delta = 5$ minutes, $K = 12$ equally spaced patients and with $-\underline{u}_k = \overline{u}_k = 30$ minutes can be calculated by a moderate-speed computer in about 1 second.

## 5. Numerical experiments

In this section we illustrate the usefulness of our approach by some numerical results. In particular, we focus on studying the effects of unpunctuality on patient scheduling. Parameters are based on empirical results and assumptions made in prior studies. We consider a 4-hour session ($t_{\max}$=240). The distributions for the consultation times follow a lognormal distribution, $S^{(c)} \sim \mathrm{LogN}(\mu_S,\, \sigma_S)$ with $\mu_S = \mathrm{E}[S^{(c)}]$ and $\sigma_S^2 = \mathrm{Var}[S^{(c)}]$. This is supported by empirical data in the literature (Cayirli et al., 2006; Klassen & Rohleder, 1996). The coefficient of variation CV ($=\sigma_S/\mu_S$) is selected from the set {0.4, 0.6, 0.8} which reflects the range of CVs identified by Cayirli and Veral in their extensive literature review Cayirli and Veral (2003). Patient no-shows and late cancellations are modeled at three levels: 0%, 10% and 20%. Doctor lateness $\Theta$ is modeled by a normal distribution with $\mu_\Theta = 0$ minutes and $\sigma_\Theta = 15$ minutes. This results in the doctor being late about 50% of the time with an expected lateness of $\mathrm{E}[\Theta|\Theta > 0] = \sqrt{\frac{2}{\pi}}\sigma_\Theta \approx 12$ minutes, which is consistent with health care literature (Fetter & Thompson, 1966; Klassen & Yoogalingam, 2014).

Finally, patient unpunctuality is modeled by a normal distribution $U^{(c)} \sim \mathrm{N}(\mu_U,\, \sigma_U)$ with $\mu_U$ equal to 0 and $-15$ minutes and $\sigma_U$ respectively equal to 10 and 20 minutes. These distributions are then truncated to the interval $[-30\,\text{minutes}, 30\,\text{minutes}]$. Moreover, it is assumed that patients' unpunctuality distributions are equal across the session, which is consistent with prior studies that found unpunctuality to be independent of appointment time (Cayirli et al., 2006; Klassen & Yoogalingam, 2014; White & Pike, 1964).

The discretized consultation time and unpunctuality distributions can then be obtained from their continuous counterparts $S^{(c)}$ and $U^{(c)}$, respectively as,

$$s(n) = \Pr[(n - \tfrac{1}{2})\Delta < S^{(c)} < (n + \tfrac{1}{2})\Delta], \qquad n \geqslant 0\,,$$

$$u(n) = \frac{\Pr[(n - \frac{1}{2})\Delta < U^{(c)} < (n + \frac{1}{2})\Delta]}{\Pr[\underline{u} - \frac{1}{2} < U^{(c)} < \overline{u} + \frac{1}{2}]}, \qquad \underline{u} \leqslant n \leqslant \overline{u}\,.$$

As many practical settings do not require appointment times with a greater accuracy than 5-minute intervals, we set the time unit $\Delta$ equal to 5 minutes during the optimization process. Finally, recall that we assume that the doctor handles an AO queueing discipline and that patients can arrive out of order since we allow that the unpunctuality intervals overlap.

### 5.1. Effects of patient unpunctuality

In this section, we examine the effects of several factors on the performance measures for a particular schedule. As a base case scenario, we assume a session with $K$=12 identical patients with a consultation time distribution $S^{(c)} \sim \mathrm{logN}(25{,}15)$ and unpunctuality distribution $U^{(c)} \sim \mathrm{N}(-15, 20)$. Their no-show probability $p^{\mathrm{n}}$ and late cancellation probability $p^{\mathrm{c}}$ are both equal to 10% ($M = -30$). Patients are scheduled every 20 minutes: $a_k = a = 20$.

First, we look at the impact of the unpunctuality distribution. Fig. 3 shows the performance measures of the schedule for different unpunctuality distributions. Fig. 3a depicts the case when all patients are punctual for comparison. Obviously, for punctual patients (including the dummy patient at the end), the mean waiting time equals the mean modified waiting time. It can be observed that both the mean waiting times and mean modified waiting times increase indefinitely in accordance with known results in queuing theory as the inter-appointment times are larger than the mean consultation time. However, here the 'steady-state' waiting time limit will not exist as the number of patients is finite.

Fig. 3b to d shows the different unpunctuality distributions under consideration: N($-15$, 20), Uniform($-30$, 30) and N(0,10), all truncated to the interval $[-30\,\text{minutes}, 30\,\text{minutes}]$. As can be seen, unpunctuality has mostly an effect on the mean waiting

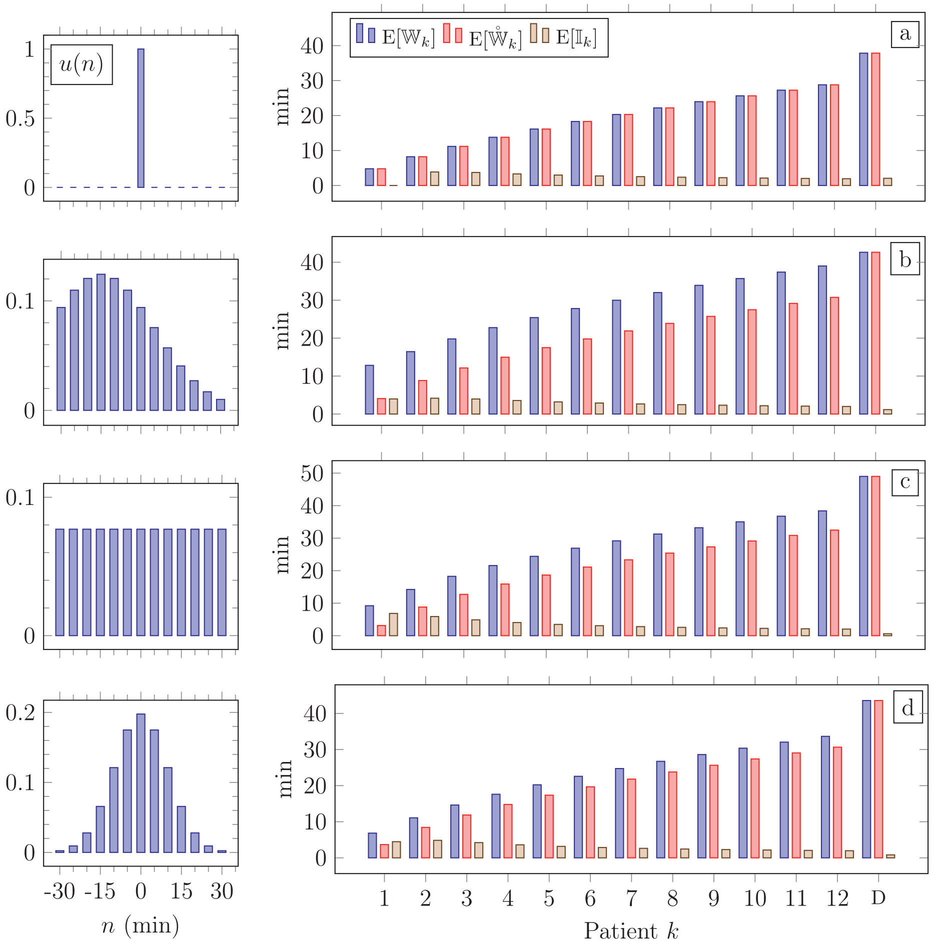

**Fig. 3.** The mean waiting times, mean modified waiting times and mean idle times of a schedule with $a = 20$ for (a) the punctual case, (b) $U^{(c)} \sim \mathrm{N}(-15{,}20)$, (c) $U^{(c)} \sim \mathrm{Uniform}(-30{,}30)$ and (d) $U^{(c)} \sim \mathrm{N}(0{,}10)$. ‘D’ stands for the dummy patient at $\tau_{K+1} = t_{\max}$, the end of the session. Recall that the waiting time of this patient corresponds to the overtime of the session.

**Table 1**
Performance measures for different cancellation and no-show probabilities.

| $(p^c;\, p^n)$ | $\sum_{k=1}^{K} \mathrm{E}[\mathbb{W}_k]$ | $\sum_{k=1}^{K} \mathrm{E}[\mathring{\mathbb{W}}_k]$ | $\sum_{k=1}^{K} \mathrm{E}[\mathbb{I}_k]$ | $\mathrm{E}[O]$ |
|---|---|---|---|---|
| (0.1; 0.1) | 333.1 | 236.2 | 35.4 | 42.6 |
| (0.2; 0.0) | 308.4 | 218.8 | 31.0 | 39.8 |
| (0.0; 0.2) | 355.1 | 252.4 | 38.9 | 45.1 |

times. Patients wait considerably longer than in the case of punctuality. Most of the extra mean waiting time is explained by the fact that patients arrive early. This is clear from $\mathrm{E}[\mathring{\mathbb{W}}]$ which shows the mean modified waiting times and its difference with $\mathrm{E}[\mathbb{W}]$. From the vantage point of the doctor, Fig. 3 also shows that unpunctuality increases the expected overtime (depicted as $\mathrm{E}[\mathbb{W}_D]$) while the idle times only slightly differ.

In addition to the unpunctuality distribution, the probabilities of no-show and late cancellation also have a big impact on the performance of a schedule. Table 1 shows that the performance worsens considerably when patients do not notify the doctor about their non-attendance, as is expected because with AO forces the doctor will wait for a no-show until the upper limit of that patient’s unpunctuality interval. For example, the expected overtime increases by 13% when none of the patients notify the doctor.

### 5.2. Degree of unpunctuality and queueing discipline

In this section, we measure the impact of varying the size of the unpunctuality interval $[\underline{u}_k, \overline{u}_k]$. A large interval allows patients to arrive far from their appointment time, but it also causes the doctor to stay idle for longer before deeming patients to be no-shows; a short interval ensures that the doctor stays idle for a short period of time, but it prevents patients to arrive far from their appointment time. Note that we assume that the doctor follows the queueing discipline AO.

Fig. 4 shows the effect of increasing the range of possible unpunctuality times on the mean (modified) waiting times, idle times and overtime. We assume the unpunctuality interval is symmetric, i.e. $\overline{u}_k = -\underline{u}_k$, and unpunctuality is $U^{(c)} \sim$ N$(-15, 20)$. As expected, the mean waiting time as well as the overtime increase for increasing $\overline{u}_k$. The mean modified waiting time and mean idle time however, are less affected by the unpunctuality limits. The same trend was found for other unpunctuality distributions, even when patients arrive late on average. Thus an increase in unpunctuality worsens the performance of the schedule.

Next, we look at the impact of the queueing discipline. Recall from Section 3 that the formulas no longer apply for the queueing discipline AOWC because AO and AOWC may take different actions when the doctor is idle. Because there is no analytic approach available for AOWC, a simulation is carried out to estimate the performance of this discipline. Fig. 5 demonstrates the differences for both the modified waiting time and the overtime between the two disciplines when $\overline{u}_k = 15, 30$. Clearly, AOWC results in a lower overtime because it is a work-conserving discipline while AO is not. It can be seen that the differences in waiting times increase with increasing $\overline{u}_k$. That is, when the intervals $[\tau_k + \underline{u}_k, \tau_k + \overline{u}_k]$ become

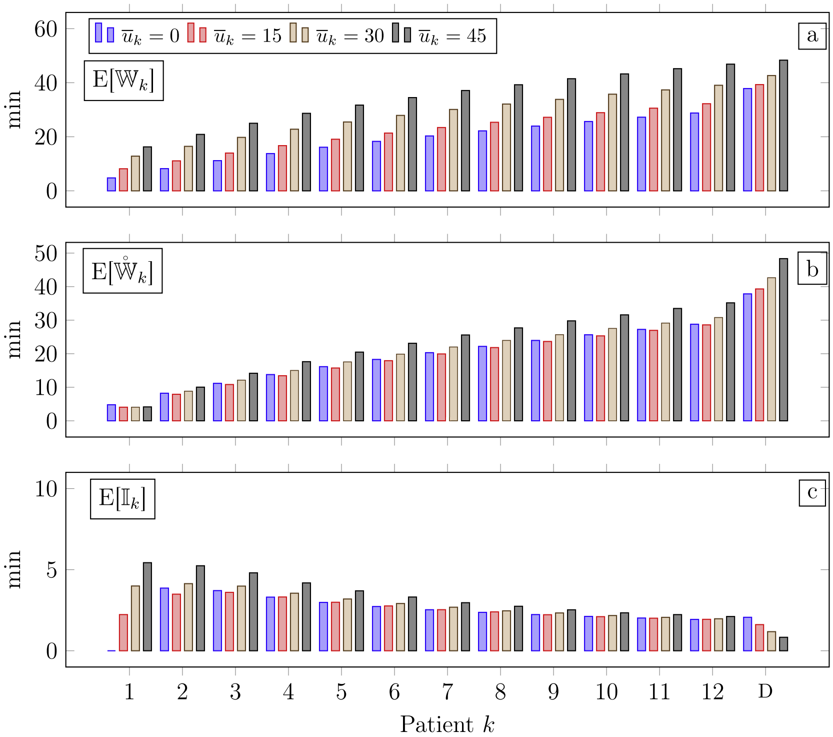

**Fig. 4.** Effect of the unpunctuality intervals on (a) the mean waiting times, (b) the mean modified waiting times and (c) the mean idle times with $U^{(c)} \sim N(-15, 20)$.

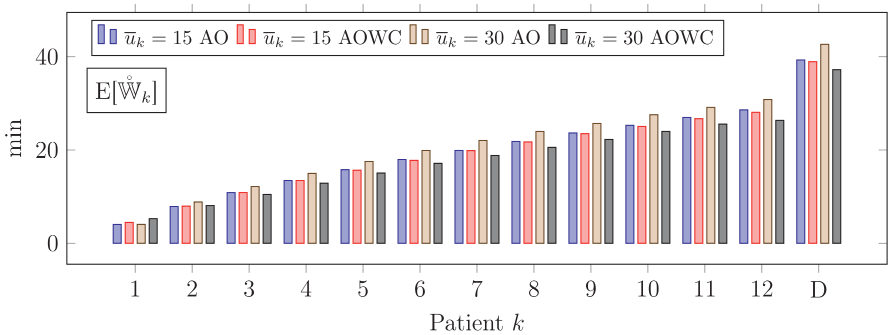

**Fig. 5.** Mean waiting time in case of AO and AOWC scheduling disciplines for a schedule with $a = 20$ and $U^{(c)} \sim N(-15,20)$.

broader and start overlapping more. This should be taken into account when unpunctuality distributions are broad compared to the inter-appointment time.

### 5.3. The local search algorithm

In the following sections, we measure the performance of the evaluation algorithm when it is included in an optimization process. In this section we first discuss how the evaluation algorithm is used in our optimization algorithm. As mentioned earlier, the scheduling process typically involves two decisions: ‘sequencing’ and ‘appointment allocation’. For the first decision, we follow the findings of Klassen and Rohleder (1996) and sequence the patients based on their consultation time variability. That is, patients with low variability are either scheduled at the beginning (LVBEG-rule) or at the end (LVEND-rule). The second decision then consists of determining the optimal vector of appointment times $\boldsymbol{\tau}$ for this given sequence.

Even if we limit ourselves to a fixed sequence, there are a still a huge number of possible schedules $\boldsymbol{\tau}$. This requires the use of a heuristic method which is capable of finding a good solution in a reasonable amount of time. To this end, we choose a local search procedure as they are proven to perform well for this type of problems (Kaandorp & Koole, 2007; Koeleman & Koole, 2012b). The main idea of such algorithms is to perform an iterative search throughout the solution space, by continuously evaluating and applying small adjustments to a solution.

Our method starts by choosing a good initial candidate solution. To this end, we select the best solution $\boldsymbol{\tau}_0$ from a set of some traditional appointment rules such as Bailey’s rule and the individual-block/fixed-interval rule (IBFI) (Cayirli et al., 2006). We then iteratively move to a neighbor solution, with the neigborhood of a schedule $\boldsymbol{\tau}$ consisting of all schedules that follow the neigborhood relation $\mathcal{N}$:

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

A neighbor solution $\boldsymbol{\tau}'$ can thus be interpreted as moving one of the appointment times of $\boldsymbol{\tau}$ from time slot $\tau_k$ to $\tau_k + 1$ or $\tau_k - 1$. As mentioned earlier, 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 includes only the first moments of the performance

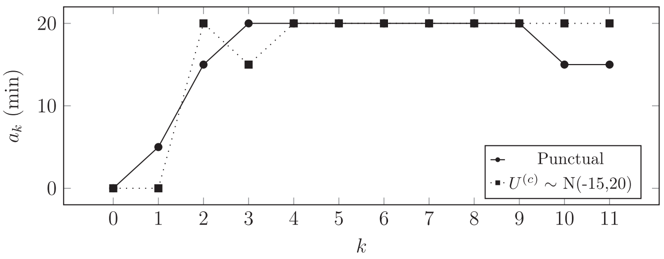

**Fig. 6.** Comparison of the heuristic solutions of the base case scenario when assuming punctual patients and unpunctual patients. Instead of the appointment times $\boldsymbol{\tau}$ themselves, the inter-appointment times for both schedules are shown.

**Table 2**
Cost reductions obtained by incorporating unpunctuality into the model in function of the unpunctuality distribution and number of patients $K$ and averaged over all parameters ($c_I$, $CV$ and $p$).

| | N($-15, 20$) | | N(0,10) | | Uniform($-30, 30$) | |
|---|---|---|---|---|---|---|
| $K$ | Avg. gap (%) | Max. gap (%) | Avg. gap (%) | Max. gap (%) | Avg. gap (%) | Max. gap (%) |
| 6 | 0.27 | 1.55 | 0.17 | 1.66 | 0.50 | 3.27 |
| 8 | 0.19 | 1.78 | 0.13 | 1.05 | 0.85 | 5.87 |
| 10 | 0.24 | 2.63 | 0.27 | 1.45 | 1.44 | 7.10 |
| 12 | 0.43 | 3.13 | 0.48 | 3.55 | 1.96 | 12.04 |

measures:

$$\mathrm{TC}(\boldsymbol{\tau}) = c_W \,\mathrm{E}\left[\sum_k \mathbb{W}_k\right] + c_I \,\mathrm{E}\left[\sum_k \mathbb{I}_k\right] + c_O \,\mathrm{E}[O], \tag{13}$$

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). Obviously, the relative importance of each term will greatly depend on the type of service and organization. In most environments, however, 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. As a final remark, it should be noted that, in general, local search cannot guarantee optimality unless patients are punctual and identical (Kaandorp & Koole, 2007; Koeleman & Koole, 2012b).

### 5.4. Cost reductions incorporating unpunctuality

In this section, we use the local search procedure described in Section 5.3 and measure the benefit of considering unpunctuality. We use the modified waiting time $\mathrm{E}[\mathring{W}_k]$ instead of $\mathrm{E}[W]$ as performance measure in the optimization process. That is, we exclude waiting prior to the appointment time in our objective function. For the base case scenario of Section 5.1, Fig. 6 shows the solution of the local search heuristic as well as the one in case all patients are punctual. On the one hand, the inter-appointment times depict a dome-shaped pattern for the punctual case, which corresponds with earlier works for punctual patients (Cayirli et al., 2006). On the other hand, Bailey's rule can be identified for the unpunctual case. The first patients are scheduled closer together to offset the negative effects of patients who arrive late. This example demonstrates that the heuristic solution may slightly differ when we incorporate unpunctuality into the model.

To quantify the cost reductions by our approach, we developed numerous test instances which capture a diverse set of environments. For each instance, we performed our local search procedure twice, respectively under the assumption of punctual and unpunctual patients which resulted in the schedules $\boldsymbol{\tau}^{(0)}$ and $\boldsymbol{\tau}^{(u)}$. The cost of assuming punctuality $\mathrm{TC}^{(0)}$ is then obtained by re-evaluating the schedule $\boldsymbol{\tau}^{(0)}$ for unpunctual patients, while $\mathrm{TC}^{(u)}$ is the cost corresponding with $\boldsymbol{\tau}^{(u)}$. The relative increase or gap between these two costs is then calculated as follows:

$$\text{gap} = \frac{\mathrm{TC}^{(0)} - \mathrm{TC}^{(u)}}{\mathrm{TC}^{(u)}} \, 100\%$$

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.
- The probability that a patient fails to show up $p_k$ was selected from the set {0, 0.1, 0.2} and we assumed that the fraction of late cancellations $\frac{p_k^c}{p_k^n}$ equalled 0, 0.5 or 1.
- The *consultation times* follow a lognormal distribution with a mean equal to $\frac{240}{K(1-p_k)}$, resulting in an average of 240 minutes of work. The standard deviations $\sigma$ are calculated in order to get coefficients of variation equal to one of the following values: {0.4, 0.6, 0.8}.
- We consider three unpunctuality distributions: N($-15, 20$), N(0,10) and Uniform($-30, 30$).

The waiting time costs $c_W$ are assumed identical for all the patients and without loss of generalization, they are set equal to 1. Following earlier works, we fix the $c_O/c_I$ ratio at 1.5 to reflect the common practice that overtime cost is valued 50% higher than regular cost (Cayirli et al., 2006). Finally, to capture the relative importance of doctor's time and patient's time, $c_I$ is chosen from the set {1, 2, 5, 10}. This represents a total of 1008 different environments.

In Table 2 both the average and maximum gaps are given for the different unpunctuality distributions under consideration. For all environments with a particular $K$ and unpunctuality distribution, the average and maximum gap is shown. Overall, for homogeneous patients, the cost reduction that is obtained by incorporating unpunctuality is rather small. The average gap over all instances is equal to 0.57%.

However, the cost reduction that can be obtained is strongly dependent on the environment. This is shown in Fig. 7, which depicts the impact of several parameters on the average gap. The results indicate that unpunctuality should especially be accounted for environments where consultation times have low variability and there are few no-shows or late cancellations. That is, when there are few other sources of uncertainty. Practitioners should thus decide whether it is worthwhile to account for unpunctuality in their particular case.

### 5.5. Heterogeneous patients

So far, we assumed that patients are identical. However, as argued earlier, our evaluation algorithm supports heterogeneity in patient characteristics. In this section, we will look at the case where patients differ in no-show probability, unpunctuality and consultation time distribution.

To this end, we consider the case where there are two types of patients. For example, it is common to divide patients into newly referred ('new') patients and follow up ('return') patients. New patients often require a longer consultation time compared to returning patients. We set the ratio of the mean consultation time of new patients to the mean consultation time of return patients equal to 1.5, based on the findings of Partridge (1992), who reported that this ratio ranged from about 1.3 to 1.9. Empirical evidence also suggests that new patients are less likely to show up than return patients (Sawyer, Zalan, & Bond, 2002). We therefore choose the following parameters: $S_{\text{new}}^{(c)} \sim$ LogN(30,24) with $p_{\text{new}}^{\text{c}} = p_{\text{new}}^{\text{n}}$=0.1 and $S_{\text{return}} \sim$ LogN(20,16) with $p_{\text{return}}^{\text{n}} = p_{\text{return}}^{\text{c}}$=0.05. The percentage of new patients is equal to 25%. The other parameters remain the same as the base case scenario in Section 5.1.

We look at the performance of seven different policies. Policy 1 involves running the local search algorithm under the assumption of identical and punctual patients. As all patients are con-

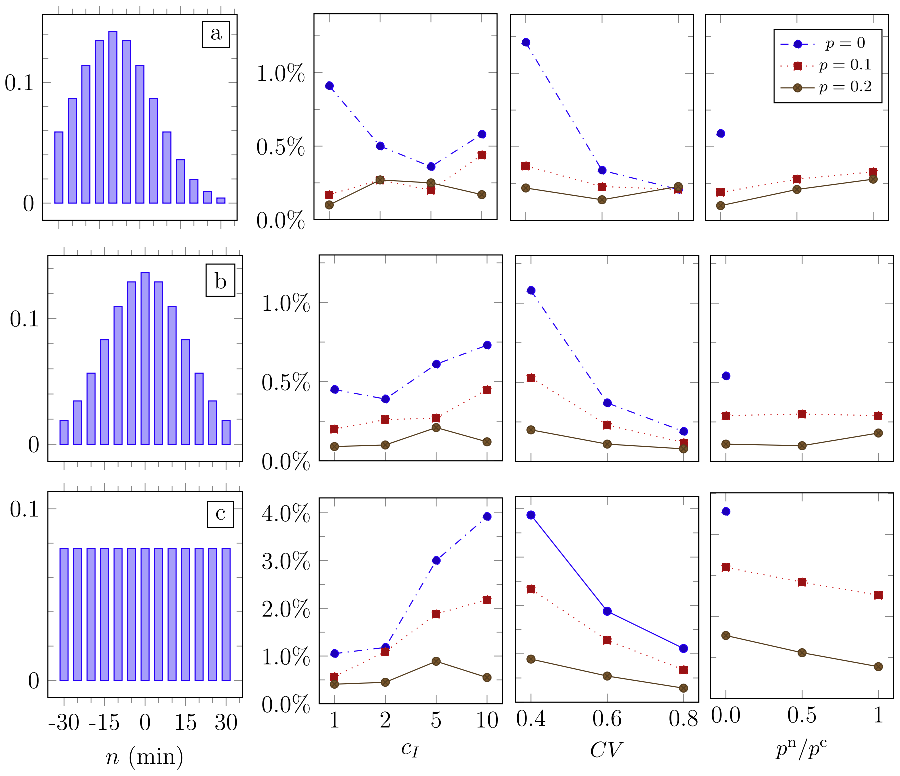

**Fig. 7.** Average gap (%) by incorporating unpunctuality compared to assuming punctual patients with unpunctuality distributions (a) N(−15, 20), (b) N(0,15) and (c) Uniform(−30, 30). Results are obtained for different levels of the parameters $c_I$, $CV$ and the fraction of no-shows $p^n/p^c$.

sidered equal for this policy, the patients' consultation time distributions are the weighted mixture of the distributions $S_{\text{new}}$ and $S_{\text{return}}$ with no-show probability and late cancellation probability equal to their average probabilities: 0.0625. Since one patient cannot be distinguished from another, no sequencing decisions are made and the actual realized sequence can be any of the possible sequences. The total cost TC, however, will obviously depend on the actual sequence. Therefore we re-evaluate the schedule and calculate the cost TC for both the best-case scenario and the worst-case scenario in terms of sequencing decisions under the assumption that patients are heterogeneous and unpunctual. For the cases considered here, the best-case scenarios correspond with using the LVBEG-rule. Policies 2 and 3, respectively, run the local search algorithm for punctual and unpunctual patients and take patient classification into account. Here, the 'heuristic solution' does fix the sequence and thus best-case and worst-case scenarios are the same. Finally, we also compare the performance of our approach with some traditional appointment rules. Policies 4 and 5 refer, respectively, to Bailey's rule and IBFI, while for policies 6 and 7 we apply their heterogeneous counterpart (Cayirli et al., 2008), in which appointment intervals are adjusted based on the expected consultation times of the different patient types.

**Table 3**
Total cost TC for different scheduling policies in case of heterogeneous patients.

| Policy | Class? | Unp? | $(c_I = 2; c_O = 3)$ | | $(c_I=5; c_O = 7.5)$ | |
|---|---|---|---|---|---|---|
| | | | Best-case | Worst-case | Best-case | Worst-case |
| 1 | No | No | 423.2 | 480.5 | 675.4 | 738.0 |
| 2 | Yes | No | 420.1 | 420.1 | 665.9 | 665.9 |
| 3 | Yes | Yes | 419.8 | 419.8 | 662.8 | 662.8 |
| Bailey | No | – | 435.0 | 532.7 | 677.2 | 736.4 |
| IBFI | No | – | 425.9 | 479.3 | 753.9 | 766.5 |
| Bailey-adj | Yes | – | 461.6 | 461.6 | 664.0 | 664.0 |
| IBFI-adj | Yes | – | 424.3 | 424.4 | 702.4 | 702.4 |

**Table 4**
Inter-appointment times for the different policies in case of heterogeneous patients for $c_I$=5 and $c_O$=7.5.

| Policy | $a_0$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ | $a_{10}$ | $a_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 10 | 15 | 15 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 15 |
| 2 | 0 | 10 | 15 | 15 | 15 | 20 | 15 | 20 | 15 | 15 | 20 | 20 |
| 3 | 0 | 5 | 15 | 20 | 15 | 20 | 20 | 15 | 20 | 15 | 20 | 20 |
| Bailey | 0 | 0 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| IBFI | 0 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| Bailey-adj | 0 | 0 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 24 |
| IBFI-adj | 0 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 24 | 24 |

Table 3 shows the objective functions of each policy for both the best-case and worst-case scenario for two cost structures. It can be seen that patient classification can indeed lead to substantial cost reductions and that its impact may be much bigger than incorporating unpunctuality. Given the big differences between best-case and worst-case scenarios, a good sequencing rule is clearly essential for good performance of the appointment system. Thus when consultation times do depend on the type of service or characteristics of the patient, practitioners should especially be focussing on sequencing decisions as well as estimating the different distributions effectively. Furthermore, the performances of the (adjusted) Bailey's rule and IBFI clearly depend on the cost structure. As expected, Bailey's rule performs better when the value of the doctor's time increases. For both cases, our heuristic approach outperforms these traditional rules. Similar results were obtained for other parameter settings. Finally, Table 4 gives the allowances found by the optimization process for the different policies when $c_I$=5 and $c_O$=7.5. Note that a dome-shaped pattern can be identified again for the punctual homogeneous case (policy 1).

Secondly, we investigate the effect of heterogeneous unpunctuality distributions. Preliminary tests indicated that practitioners can also benefit by making sequencing decisions based on the un-

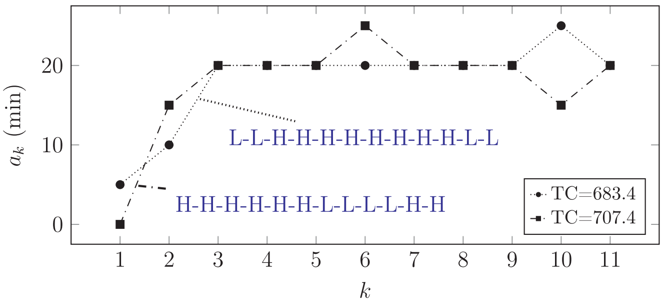

**Fig. 8.** Inter-appointment times and sequence of the heuristic solutions when patients are classified based on their unpunctuality distributions. The unpunctuality distribution of the patients either follows $U_L^{(c)} \sim$ N$(-10, 15)$ or $U_H^{(c)} \sim$ N(0,20).

punctuality of patients. We make the same assumptions as the base case scenario in Section 5.1, but now assume that patients may differ in their unpunctuality behavior. Each patient has one of the following unpunctuality distributions: $U_L^{(c)} \sim$ N(−10,15) or $U_H^{(c)} \sim$ N(0,20). We will refer to the former as 'Low-Variance' (L) and the latter as 'High-Variance' (H). Let us further assume that 4 of the patients are classified as 'L' while 8 patients are classified as 'H'. Next, we perform our local search heuristic for different sequences and compare the objective functions. Fig. 8 depicts the inter-appointment times of the best and worst sequence. A cost difference of 3.5% was found between the two heuristic solutions. It can be seen that it is better to schedule patients who are early on average and have low unpunctuality variability at the beginning to minimize the risk of idle time, but also in the end to reduce the expected overtime.

## 6. The wait-preempt dilemma

In many applications, the unpunctuality distributions of patients are quite broad and patients frequently arrive in a different order than the order in which they are scheduled. Every time this happens and the doctor becomes idle between the two arrivals, the so-called "wait-preempt dilemma" arises, as explained by Samorani and Ganguly (2016). So far, we assumed that the doctor resolves such dilemmas by (1) following AO and *waiting* for the first scheduled patient or (2) following AOWC and *preempting* by seeing the early patient right away. Practitioners typically opt for the latter option and handle an "always-preempt" policy (Cayirli & Veral, 2003). However, Samorani and Ganguly (2016) shows that the optimal solution is more complex and depends on the specific time interval as well as some environmental parameters. It is demonstrated that under some conditions (e.g. relatively low overtime cost, long consultation times) it is actually beneficial to stay idle during certain time intervals. In this section, we show that the mathematical framework of Sections 3 and 4 can also be used to solve this dilemma and we present some numerical examples. The analytical method in Samorani and Ganguly (2016), which considers deterministic and identical consultation times, is extended here to include general, stochastic consultation times and unpunctualities.

### 6.1. Analytical description

Suppose that at some time during the session $t_0$, the doctor becomes idle and the next patient $k$ has not arrived yet, but patient $k+1$ is already in the waiting room. Or equivalently, patient $k+1$ arrives at time $t_0$ while the doctor was idle. Then the problem consists of determining a waiting time $x$ ($x \geq 0$), for which the doctor remains idle and does not preempt by serving patient $k+1$.

Before we can use the formulas in Section 4, we have to make certain modifications. First of all, we have to update the patient characteristics to take into account all the available information at time $t_0$. For example, the unpunctuality of patient $k+1$ is known and fixed at this time and his no-show probability $p_{k+1}^{\mathrm{n}}$ changes to zero. For the $k$th patient, the modified unpunctuality distribution $\tilde{U}_k$, given by Eq. (1), has to be conditioned on the fact that $\tilde{U}_k$ is greater than $t_0 - \tau_k$. Here, Bayes' theorem applies and we can divide $\tilde{u}_k(n)$ as well as $p_k^{\mathrm{n}}$ by $\sum_{i=\underline{u}_k}^{t_0-\tau_k} \tilde{u}_k(i)$ for $n > t_0 - \tau_k$. The probability that patient $k$ will not show up thus increases with $t_0$. Given the knowledge that he did not arrive in the interval $[\tau_k + \underline{u}_k, t_0]$, it is more likely that he will not show up altogether. To avoid burdening the notation, we assume that all variables are updated with the information at time $t_0$.

Now, we make the distinction between the two possible scenarios:

[S1] The $k$th patient arrives at time $t_0 + t$ $(0 < t \leq x)$ and is served right away. The sequence of the patients remains unchanged with $W_k = 0$ and $I_k = t$. The cost of this scenario $\mathrm{TC}^{(w)}(t)$ is determined by calculating the cost TC($\boldsymbol{\tau}^*$) of the remaining schedule $\boldsymbol{\tau}^*= \{\tau_k, \ldots, \tau_K\}$. We set $\theta(t_0) = 1$ to reflect that the doctor is available since $t_0$ and fix $U_k$ at $t_0 - \tau_k + t$.

[S2] The $k$th patient does not arrive before $t_0 + x$ and we find the cost of this scenario $\mathrm{TC}^{(p)}$ by calculating the cost of the remaining schedule $\boldsymbol{\tau}^*= \{t_0 + x, \tau_{k+1}, \tau_k, \tau_{k+2}, \ldots, \tau_K\}$ for which we interchange the position of the $k$th and $(k+1)$th patient. Note that we also introduce a dummy patient at $t_0 + x$ with consultation time zero to account for the idle time during $[t_0, t_0 + x]$. Finally, we again set $\theta(t_0) = 1$.

Here, it should be noted that we ignored situations where the dilemma arises between two non-adjacent appointments.

Given these scenarios, the problem now translates to finding $x$ which minimizes TC:

$$\mathrm{TC}(x) = \left[1 - \sum_{t=1}^{x} \tilde{u}_k(t_0 - \tau_k + t)\right] \mathrm{TC}^{(p)}(x) + \sum_{t=1}^{x} \tilde{u}_k(t_0 - \tau_k + t)\, \mathrm{TC}^{(w)}(t)$$

The optimal value $x^*$ is the time that the doctor should wait for the arrival of the $k$th patient before she starts serving the $k+1$th patient. Clearly, when $x^* = 0$, the doctor should serve patient $k+1$ right away.

Finally, Samorani and Ganguly (2016) has proven that the minima and maxima of this cost function are independent of $t_0$. Therefore, it suffices to solve the dilemma just once by analyzing the cost function TC($x$) and determining the optimal decisions for each time interval. When there is no time in the future that has a lower cost, the doctor should preempt. Otherwise, she should wait until that time is reached to incur the lowest cost possible.

### 6.2. Numerical examples wait-preempt dilemma

In what follows we will illustrate our approach by some numerical examples. As a base case scenario we make the following assumptions: $K = 6$ identical patients with $S^{(c)} \sim$ logN(40,8), $U^{(c)} \sim$ N$(-10, 15)$, $\overline{u}_k = -\underline{u}_k = 60$, $p^{\mathrm{n}}$=20% and no cancellations, $t_{\max} = 240$ and set $\Delta$ equal to 5. This parameter configuration results in a single "wait-preempt" dilemma being resolved within 1 second. Finally, we use the modified waiting time and set $c_I$ and $c_O$, respectively, to 0 and 2. This represents clinics where patients' times are highly valued.

As a first example we look at the case where patients are scheduled with an equal inter-appointment time $a_k = a$ of 40 minutes: $\boldsymbol{\tau}$={0, 40, 80, 120, 160, 200}. Now, suppose that the first 3

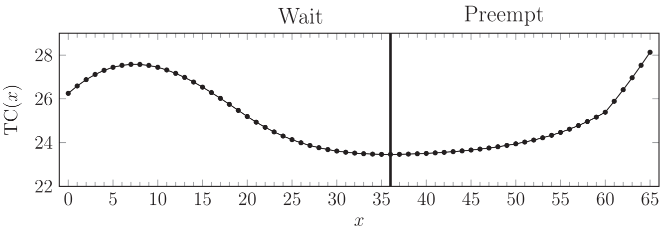

**Fig. 9.** Numerical example of “Wait preempt dilemma” at $t_0 = \tau_4 - 20$ with $a_k = a = 40$ and $\Delta = 1$ minute.The consultation times and unpunctuality times of the patients are independent and are respectively distributed as $S^{(c)} \sim$ logN(40,8) and $U^{(c)} \sim$ N(−10, 15). The overtime cost $c_O$ and idle time cost $c_I$ are respectively equal to 2 and 0. It can be seen that $x^*$= 36 which means that the doctor should wait for 36 minutes.

**Table 5**
Behavior of TC($x$) for multiple parameter settings.

| $c_O$ | $\mu_S$ | $CV$ | $p^n$ | $\sigma_U$ | Behavior of TC($x$) |
|---|---|---|---|---|---|
| 2 | 40 | 0.2 | 0.2 | 15 | Increases in [0,7) and (36,65],<br>Decreases in (7,36) |
| **5** | 40 | 0.2 | 0.2 | 15 | Increases in [0,15) and (24,65]<br>Decreases in (15,24) |
| **10** | 40 | 0.2 | 0.2 | 15 | Increases in [0,65) |
| 2 | 40 | 0.2 | 0.2 | **10** | Increases in [0,6) and (39,65],<br>Decreases in (6,39) |
| 2 | 40 | **0.6** | 0.2 | 15 | Increases in [0,14) and (32,65],<br>Decreases in (14,32) |
| 2 | 40 | 0.2 | **0.0** | 15 | Increases in [0,9) and (34,65],<br>Decreases in (9,34) |
| 2 | **20** | 0.2 | 0.2 | 15 | Increases in [0,65] |
| 2 | **60** | 0.2 | 0.2 | 15 | Decreases in [0,49),<br>Increases in (49,65] |

patients have been served and that the dilemma arises in the time interval [100, 165]. That is, the doctor is idle with patient 5 present in the waiting room. What are the time intervals where it is optimal for the doctor to wait and for which time intervals it is better to preempt? Fig. 9 depicts TC($x$) for $x$ between 0 and 65 minutes and $\Delta$ is set equal to 1 minute here. It can be seen that the lowest cost is obtained for $x = 36$ and thus whenever the dilemma arises before this moment ($t_0 < 136$), the doctor should wait until $t = 136$. By doing this, the system can reduce its cost by 11% compared to the AOWC discipline at $t_0$ and even 18% when the dilemma arises 8 minutes later ($t = t_0 + 8$). On the other hand, if $t_0 \geq 136$, the doctor should preempt right away because TC($x$) strictly increases in the time interval [136, 165]. Note that for $x > \tau_5 - t_0$ =60, the cost function TC($x$) clearly increases at a higher rate. This is explained by the fact that from this time the waiting patient starts to incur a waiting cost as well.

In Table 5 we look at what happens if we change some of these parameters. As expected, the doctor should preempt earlier when $c_O$ increases. For $c_O$=10, we find that the doctor should always preempt immediately. Furthermore, it can be seen that not only the overtime cost $c_O$, but also the parameters $\mu_S$, $CV$ and $\sigma_U$ have a large effect on the optimal decision. The results show that when $\mu_S$ increases, the doctor should wait for a longer period. In contrast, adding variability to either the consultation time or unpunctuality reduces the waiting intervals. It can be concluded that the analytical method should only be considered in environments with long consultation times, low overtime cost, high probability of no-shows and few sources of variability. This corresponds with the findings of Samorani and Ganguly (2016). Finally, we investigated the impact of $K$ on the behavior of TC($x$). As expected, waiting intervals become smaller when the number of remaining patients increases, since more patients may see their expected waiting time increase. This means that at the beginning of a session, the doctor should preempt earlier compared to the same situation occurring at a later moment during the session.

Given the results in Section 5.5, we may also be interested in the effect of heterogeneity for this problem. Assuming that the same dilemma arises as discussed above (i.e. waiting for patient 4 or preempting with patient 5), we varied different parameters for patients 4 and 5 separately. Preliminary tests indicated that the consultation time variability of patient 4 has a big impact on the decision. To confirm these results, we set up a second experiment in which we vary the consultation time variability of patients 4 and 5.

We consider the following scenarios: (S1) patient 4 has a higher consultation time variability ($\sigma_S$=27) than the other patients ($\sigma_S$=7); (S2) patient 4 has a lower consultation time variability ($\sigma_S$=5) than the other patients ($\sigma_S$=20); (S3) patient 5 has a higher consultation time variability ($\sigma_S$=27) than the other patients ($\sigma_S$=5); (S4) patient 5 has a lower consultation time variability ($\sigma_S$=5) than the other patients ($\sigma_S$=20). These values are chosen so that the total consultation time variability of the remaining schedule is about the same for every scenario to allow for fair comparison. For each scenario, we then calculate the cost function TC($x$) for $0 \leq x \leq 65$ and divide each value by TC($x^*$). The resulting graph for every scenario is shown in Fig. 10. It can be seen that at time $t_0$, the doctor should only preempt in scenarios (S1) and (S4), i.e. when the first-scheduled patient has a high consultation time variability compared to the other patients. In this case, it only becomes optimal to wait if the dilemma arises at $t \geq t_0 + 5$. Note that the cost function is quite flat in the region [25, 55] and both decisions have similar cost. A similar observation can be made for scenario (S4), where the waiting patient has a low consultation time variability compared to the other patients. In all of these cases, the

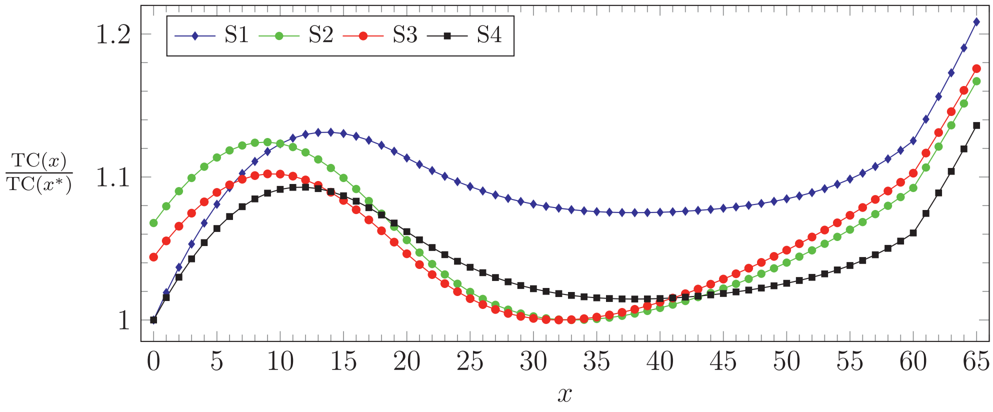

**Fig. 10.** Overview of ratios TC($x$)/TC($x^*$) for scenarios (S1)–(S4).

doctor should serve the patient with the lowest consultation time variability first. Note that in case we scheduled according to the LVBEG-rule, the doctor should wait and follow the AO discipline, since the patients are scheduled with increasing consultation time variability.

## 7. Conclusions

This paper introduces an effective numerical method to assess the moments of the waiting times of patients as well as the idle times and overtime of the doctor under fairly general conditions. Patients may have general, distinct consultation time distributions. There are no restrictions on these consultation times other than that they are independent and their distributions are known. The doctor may have a stochastic starting time and the appointment times can be freely chosen. Most importantly however, patients do not have to be punctual. The approach requires no restrictions on the unpunctuality distribution, but it should be noted that patient unpunctuality can lead to the reordering of the patients.

The algorithmic approach advocated here is fast in comparison with simulations and was included in a local search algorithm. Some numerical examples are presented in which we particularly focus on the effects of unpunctuality. Although the bulk of the literature focusses on punctual patients, our analytical results suggest that practitioners can benefit by incorporating unpunctuality into the model. This is especially true for environments with low no-show rates, low consultation time variability and when the provider's time is highly valued. It is shown, however, that judicious patient sequencing is often more important than taking unpunctuality into account. The cost of a schedule can be dramatically reduced by choosing a good sequence and by correctly exploiting the knowledge that patients belong to a certain class.

Finally, we show that our analysis can also be used to address the problem of dealing with patients which arrive out of order. It is shown that under certain situations the optimal decision for the doctor is to stay idle and wait for the next scheduled patient rather than serving early patients right away. This supports our assumption that patients are served following the appointment order.

A possible extension of the study could be to relax the independence assumption of the consultation times. Modeling the dependence between consultation time and unpunctuality is an unexplored domain and the model at hand could easily be extended to include this. Doctors may, for example, speed up the consultation to some extent if the patient was late to reduce the waiting time for other patients. Future work could also focus on methods to estimate individual consultation time distributions and no-show probabilities based on patient related characteristics. Using data about different treatments and types of patients will reduce the variance of the stochastic variables. Other possible directions of future research include investigating different objective functions and queueing disciplines.

## References

Ahmadi-Javid, A., Jalali, Z., & Klassen, K. J. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research, 258*(1), 3–34.

Alexopoulos, C., Goldsman, D., Fontanesi, J., Kopald, D., & Wilson, J. R. (2008). Modeling patient arrivals in community clinics. *Omega, 36*(1), 33–43.

Bar-dayan, Y. (2002). Waiting time is a major predictor of patient satisfaction in a primary military clinic. *Military Medicine, 167*(10), 842.

Brailsford, S., & Vissers, J. (2011). Or in healthcare: A European perspective. *European Journal of Operational Research, 212*(2), 223–234.

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

Cayirli, T., Veral, E., & Rosen, H. (2006). Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science, 9*(1), 47–58.

Cayirli, T., Veral, E., & Rosen, H. (2008). Assessment of patient classification in appointment system design. *Production and Operations Management, 17*(3), 338–353.

Chen, R. R., & Robinson, L. W. (2014). Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management, 23*(9), 1522–1538.

Cheong, S., Bitmead, R. R., & Fontanesi, J. (2013). Modeling scheduled patient punctuality in an infusion center. *Lecture Notes in Management Science, 5*, 46–56.

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

Creemers, S., Colen, P., & Lambrecht, M. (2012). Evaluation of appointment scheduling rules: a multi-performance measures approach. Available at SSRN 2086264.

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

Erdogan, S. A., Gose, A., & Denton, B. T. (2015). Online appointment sequencing and scheduling. *IIE Transactions, 47*(11), 1267–1286.

Fetter, R. B., & Thompson, J. D. (1966). Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research, 1*(1), 66.

Fiems, D., & De Vuyst, S. (2013). Efficient evaluation of out-patient scheduling with unpunctuality. In *Analytical and stochastic modeling techniques and applications* (pp. 171–182). Springer.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives, 25*(4), 191–209.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics, 115*(3), 791–810.

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

Jouini, O., & Benjaafar, S. (2009). Appointment scheduling with non-punctual arrivals. *IFAC Proceedings Volumes, 42*(4), 235–239.

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

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

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

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

Kocas, C. (2015). An extension of osuna's model to observable queues. *Journal of Mathematical Psychology, 66*, 53–58.

Koeleman, P., & Koole, G. (2012a). Appointment scheduling using optimisation via simulation. In *Proceedings of the winter simulation conference* (p. 22).

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

Kuiper, A., Kemper, B., & Mandjes, M. (2015). A computational approach to optimized appointment scheduling. *Queueing Systems, 79*(1), 5–36.

Meza, J. P. (1998). Patient waiting times in a physician's office. *The American Journal of Managed Care, 4*(5), 703–712.

Osuna, E. E. (1985). The psychological cost of waiting. *Journal of Mathematical Psychology, 29*(1), 82–105.

Partridge, J. W. (1992). Consultation time, workload, and problems for audit in outpatient clinics. *Archives of Disease in Childhood, 67*(2), 206–210.

Perros, P., & Frier, B. M. (1996). An audit of waiting times in the diabetic outpatient clinic: Role of patients' punctuality and level of medical staffing. *Diabetic Medicine, 13*(7), 669–673.

Rohleder, T. R., & Klassen, K. J. (2000). Using client-variance information to improve dynamic appointment scheduling performance. *Omega, 28*(3), 293–302.

Salzarulo, P. A., Mahar, S., & Modi, S. (2016). Beyond patient classification: Using individual patient characteristics in appointment scheduling. *Production and Operations Management, 25*(6), 1056–1072.

Samorani, M., & Ganguly, S. (2016). Optimal sequencing of unpunctual patients in high-service-level clinics. *Production and Operations Management, 25*(2), 330–346.

Samorani, M., & LaGanga, L. R. (2015). Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research, 240*(1), 245–257.

Sawyer, S., Zalan, A., & Bond, L. (2002). Telephone reminders improve adolescent clinic attendance: A randomized controlled trial. *Journal of Paediatrics and Child Health, 38*(1), 79–83.

Tai, G., & Williams, P. (2012). Optimization of scheduling patient appointments in clinics using a novel modelling technique of patient arrival. *Computer Methods and Programs in Biomedicine, 108*(2), 467–476.

White, M. J. B., & Pike, M. C. (1964). Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Medical Care, 2*, 133–145.

Williams, K. A., Chambers, C. G., Dada, M., McLeod, J. C., & Ulatowski, J. A. (2014). Patient punctuality and clinic performance: Observations from an academic-based private practice pain centre: a prospective quality improvement study. *BMJ Open, 4*(5), e004679.