Received: 8 November 2019 | Revised: 23 December 2020 | Accepted: 29 December 2020

DOI: 10.1111/deci.12517

DECISION SCIENCES
A JOURNAL OF THE DECISION SCIENCES INSTITUTE
WILEY

ORIGINAL ARTICLE

A flexible and optimal approach for appointment scheduling in healthcare

Alex Kuiper[1] | **Michel Mandjes**[1,2] | **Jeroen de Mast**[3] | **Ruben Brokkelkamp**[4]

[1] Department of Operations Management, Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands

[2] Korteweg de-Vries institute, University of Amsterdam, Amsterdam, The Netherlands

[3] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

[4] Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands

Correspondence
Alex Kuiper, Department of Operations Management, Amsterdam Business School, University of Amsterdam, Plantage Muidergracht 12, 1018TV, Amsterdam, The Netherlands.
Email: a.kuiper@uva.nl

ABSTRACT

Appointment scheduling is generally applied in outpatient clinics and other healthcare services. The challenge in scheduling is to find a strategy for dealing with variability and unpredictability in service duration and patient arrivals. The consequences of an ineffective strategy include long waiting times for patients and idle time for the healthcare provider. In turn, these have implications for the perceived quality, cost-efficiency, and capacity of healthcare services. The generation of optimal schedules is a notoriously intractable problem, and earlier attempts at designing effective strategies for appointment scheduling were based on approximation, simulation, or simplification. We propose a novel strategy for scheduling that exploits three tactical ideas to make the problem manageable. We compare the proposed strategy to other approaches, and show that it matches or outperforms competing methods in terms of flexibility, ease of use, and speed. More importantly, it outperforms competing approaches nearly uniformly in approaching the desired balance between waiting and idle times as specified in a chosen objective function. Therefore, the strategy is a good basis for further enrichments.

KEYWORDS
Healthcare optimization, convexity, appointment scheduling, phase-type distributions, queueing

1 | INTRODUCTION

Most healthcare services are nonprofit in nature and exist to serve their communities. The rising expenditures for healthcare, however, have created general awareness that their performance should be evaluated in terms of the delivered care relative to the expenses incurred (Porter, 2010). This in turn has drawn attention to the performance of the operating and management practices involved. Two highly influential reports of the Institute of Medicine (IOM, 2001, 2006) have urged the use of operational management methods and information technologies to improve the quality and efficiency in healthcare, and healthcare applications of operations management theory and techniques have become a thriving field.

One of the central challenges in healthcare-operations management is to match capacity and demand under variability and unpredictability. With some exceptions such as emergent demand, healthcare services generally apply appointment scheduling to synchronize patient visits with the availability of specialists, facilities, and resources. Primary and specialty care in outpatient clinics are often scheduled, as well as inpatient care such as elective surgeries; see, for example, May et al. (2000) and Denton and Gupta (2003). On the one hand, appointments should be set up such that excessive waiting times for patients are avoided, as these are an important determinant of the perceived service quality and satisfaction (Anderson et al., 2007; Huang, 1994). On the other hand, the scheduling approach should maximize the utilization of specialists, staff, and facilities by avoiding idle time, that is, time lost waiting for patients. Utilization is an important factor in the unit-costs of delivered care, and in addition, it essentially determines the total capacity of the service in question (and thus affects admission times and availability of care). Appointment scheduling, therefore, directly impacts the perceived quality, cost-efficiency, and capacity of a substantial part of healthcare services.

A key difficulty in scheduling is to deal effectively with uncertainty and variability. Variability in the service times is typically substantial in specialty care and surgery, and scheduling in healthcare is often confronted with random no-shows, cancelations, walk-ins, and emergencies (Çayirli &

Veral, 2003; Çayirli et al., 2006; Deceuninck et al., 2018; Gupta & Denton, 2008; Zacharias & Pinedo, 2014).

This article presents a systematic framework for setting up appointment schedules in healthcare applications. Our approach has been designed to be *flexible*, in that it is able to incorporate a wide range of sources of variability (including no-shows and walk-ins) and can deal with operational restrictions (e.g., that time slots should be multiples of 5 min). Second, the framework is *robust*. In many studies it is assumed that the service–time distribution is fully known, whereas in practice often limited information is available. Our approach requires only the first two moments, and turns out to perform highly accurately. We refer to, for example, Begen et al. (2012) and Mak et al. (2015) for other approaches relying on limited distributional information. Third, we offer our approach in a form that is *computationally fast* and *easy to use* for the general public: it has been implemented in a webtool that generates schedules instantaneously.

In its full form, the scheduling problem is to determine optimal appointed arrival times for all n patients in a session, which is an optimization in $n-1$ dimensions. In this general form, the problem was found to be analytically intractable. Many approaches have been proposed to make the problem tractable; see Ahmadi-Javid et al. (2017) for an extensive overview of these attempts. Early work has simplified the problem to the form where appointed arrivals are equidistant with slot lengths equal to the mean service time. The first slot is then overbooked by one or more patients (Bailey, 1952; Welch & Bailey, 1952). In these simple heuristics, variability can only be accounted for by scheduling more patients per slot. Charnetski (1984), Hassin and Mendel (2008), and others have studied a simplification where the appointed arrival times are still equidistant, but the number of patients per slot is limited to one, and it is the (constant) length of the slots that is the parameter that is optimized.

Recent literature has considered the problem in its full-fledged form, where appointed arrivals are not necessarily equidistant, and have attempted to make the problem tractable by applying approximations or simulations, such as Wang (1997), Robinson and Chen (2003), Kaandorp and Koole (2007), De Vuyst et al. (2014), Çayirli et al. (2012), and Kuiper et al. (2015). Across a wide variety of settings, they found that optimal schedules turn out not to have equidistant appointment times. But the optimal interarrival times typically have the shape of a dome: shorter slots in the beginning and toward the end of a session, and longer in the middle.

We propose a novel strategy for scheduling that exploits three tactical ideas to make the problem manageable. First, we develop a method to determine the expected waiting and idle times analytically by approximating the service time's distributions by phase-type distributions. This approximation is both accurate and analytically convenient. Second, we extend the approach to incorporate no-shows and walk-ins. Our approach for incorporating no-shows and walk-ins is general: it can be used to augment alternative scheduling strategies as well. And third, we offer an algorithm that generates near-optimal schedules instantaneously from a grid of precalculated schedules. We compare our strategy to other approaches, and show that it matches or outperforms competing methods in terms of flexibility and ease of use (it can be applied in a wide range of scenarios), robustness (its performance is not very sensitive to critical model assumptions), and speed (it does not require lengthy simulation procedures). More importantly, it outperforms competing approaches nearly uniformly in approaching the desired balance between waiting and idle times as specified in a chosen objective function.

An important idea driving our approach is to approximate the distribution of service times by phase-type distributions. This type of approximation offers important advantages for our purpose. First, phase-type distributions are computationally convenient, because they are mixtures and/or convolutions of exponential distributions, and as a result, the computation of expected idle and waiting times becomes analytically tractable. Second, any positive distribution can be approximated with arbitrary accuracy by a phase-type approximation. And third, it turns out that for finding an appropriate phase-type distribution, we need to estimate only the first two moments of the service–time distribution. Therefore, phase-type approximations are an attractive way for making the scheduling problem tractable, as they are accurate and convenient. In the next section, we start by defining the problem carefully, and placing it in accepted theory in operations management.

2 | THEORY AND PROBLEM DEFINITION

In this section, we define the problem in mathematical terms and state the problem to be solved. Furthermore, we contribute a rigorous proof of the convexity of the objective function under consideration.

2.1 | Preliminaries

Figure 1 illustrates the setting of the problem, with a session consisting of n slots with lengths $x_1, \ldots, x_{n-1}$, from which the patients' scheduled arrival times $t_i := \sum_{j=1}^{i-1} x_j$ follow, with the convention of an empty sum being zero. The realized service times are indicated by the widths of the solid rectangles, and the waiting times by the widths of the dark-gray rectangles. The total expected waiting and idle times in the session are determined by the schedule $(x_1, \ldots, x_{n-1})$ and by the components of process variability and their stochastic properties. In the first place, we have the distribution F_B (assuming i.i.d. for now) of the service times B_i, with mean $\mu = \mathbb{E}B_i$ and with squared coefficient of variation ϱ (the ratio of the variance of B_i and its squared mean). In addition, we have a per-patient probability $q \in [0, 1)$ of a no-show, and a per-slot probability v that an unscheduled patient arrives (a walk-in).

In correspondence with Figure 1, we define by W_i the net *waiting time* of the i-th patient, that is, the time in between her scheduled arrival and the moment she receives service,

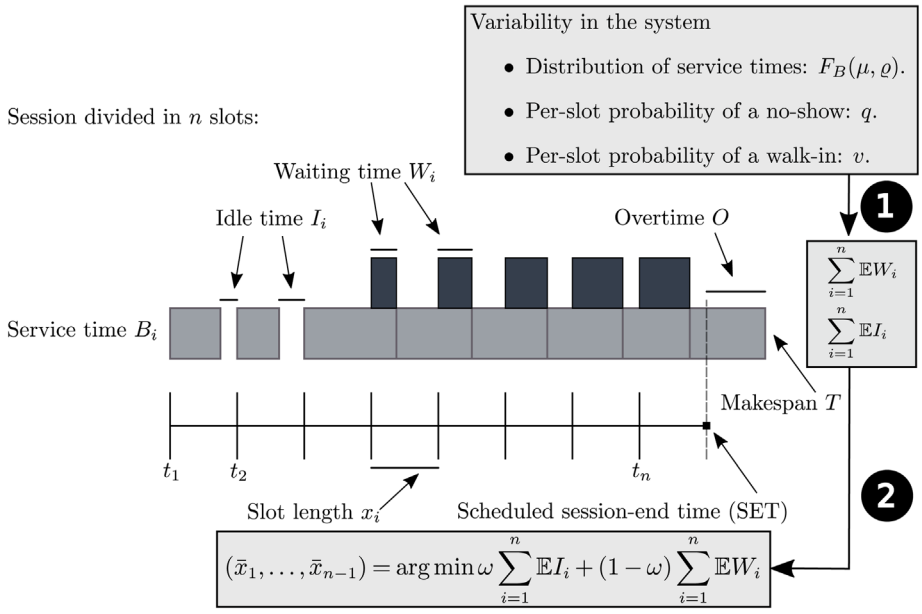

FIGURE 1 The setting of the problem, in which n slots with lengths $x_1, \ldots, x_{n-1}$ determine the appointment schedule. This in turn should be optimized such that it minimizes the objective function

where we set $W_1 = 0$. As a consequence earliness is on the patient's account. Define I_i as the server *idle time* prior to the i-th patient's arrival, with $I_1 = 0$. It is a standard result that, by virtue of the *Lindley* recursion, the I_i can be determined recursively:

$$I_i = \max\{x_{i-1} - W_{i-1} - B_{i-1}, 0\}.$$

Likewise,

$$W_i = \max\{W_{i-1} + B_{i-1} - x_{i-1}, 0\}. \tag{1}$$

Evidently, we cannot have that both W_i and I_i are strictly positive. This observation leads to the following identities, where $S_i = W_i + B_i$ denotes the *sojourn time* of the i-th patient:

$$I_i + W_i = |\, S_{i-1} - x_{i-1} \,| \quad \text{and} \quad W_i^2 + I_i^2 = (S_{i-1} - x_{i-1})^2.$$

The *makespan* or *session end time* (SET), defined as the epoch T that patient n has been fully served, can be written in two alternative ways, noting that $\sum_{i=1}^{n-1} x_i = t_n$,

$$T := \sum_{i=1}^{n} B_i + \sum_{i=1}^{n} I_i = \sum_{i=1}^{n-1} x_i + S_n. \tag{2}$$

2.2 | Objective function and convexity

In our approach the schedules are generated so as to optimize a specific objective function. We now argue that suitable objective functions for appointment schedules in healthcare are a weighted average of idle time for the servers and waiting time for patients. Designing an appointment schedule is an instance of the more general problem of dealing with process variability; see Hopp and Spearman (2008). A sensible first step in dealing with variability and uncertainty is to try to reduce them. Some outpatient clinics, for instance, bring down no-shows and last-minute cancelations by employing reminders and/or sanctions (Barron, 1980). A second step is to counterbalance variability by flexibility. Healthcare providers sometimes handle peak loads by stretching their working day or shrinking lunch time. Oppositely, they may put unanticipated idle time to effective use by catching up on administrative work or other pending tasks.

After variability has been reduced or counterbalanced as far as possible, the *variability buffering law* (Hopp & Spearman, 2008) predicts that the remaining variability will be absorbed by a combination of three buffers. In the first place, the provider may build up in advance an inventory of finished "products" as a buffer to absorb peaks in demand. This is rarely an option, however, for the type of services that we consider. Products in our setting are treatments and diagnoses, and production cannot start until patient and doctor come together. This leaves us with two other types of buffering:

- a queue of patients waiting to get served and
- an excess of unutilized capacity of the healthcare provider (which implies idle time).

As a consequence, a schedule's performance degradation due to variability is a combination of waiting time for patients and idle time for servers, which act as communicating vessels. Given the stochastic characteristics of service times and patient arrivals (including no-shows and walk-ins), and weighting the relative importance of idle and waiting times by $\omega \in (0, 1)$, this performance degradation is expressed by

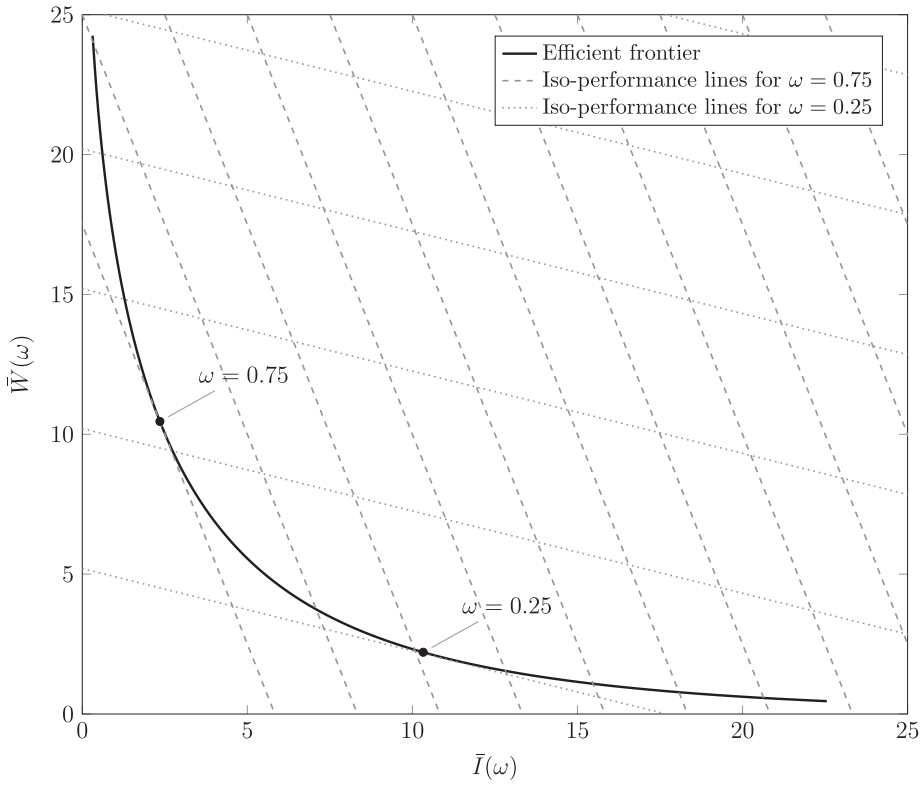

FIGURE 2 The efficient frontier showing the trade-off in terms of idle and waiting times when 10 patients are scheduled optimally, given ω, with ω ranging from 0.05 to 0.95. Here exponential service times are assumed. In gray the iso-performance lines are drawn for $\omega = 0.75$ (dashed) and for $\omega = 0.25$ (dotted)

the objective function

$$\mathcal{F}[x_1, \ldots, x_{n-1}] = \omega \sum_{i=1}^{n} \mathbb{E}I_i + (1-\omega) \sum_{i=1}^{n} \mathbb{E}W_i. \qquad (3)$$

For a given weight ω, the optimal schedule is the sequence $\bar{x}_1, \ldots, \bar{x}_{n-1}$ that minimizes $\mathcal{F}[x_1, \ldots, x_{n-1}]$. Define $\bar{I}(\omega) = \sum_{i=1}^{n} \mathbb{E}I_i$ and $\bar{W}(\omega) = \sum_{i=1}^{n} \mathbb{E}W_i$ as the mean total idle and waiting time of the optimal schedule $\bar{x}_1, \ldots, \bar{x}_{n-1}$ for the weight ω. Generally, when ω approaches 1 (i.e., the situation in which the value of the objective function is essentially determined by the idle times only), $\bar{W}(\omega)$ explodes; cf. the *utilization law* of Hopp and Spearman (2008). On the other hand, when ω approaches 0, the contribution of the mean total idle time experienced by the server, that is, $\bar{I}(\omega)$, increases sharply.

The curves in Figure 2, consisting of combinations $(\bar{I}(\omega), \bar{W}(\omega))$ for $\omega \in (0, 1)$, are named the *efficient frontier* in the management literature. All feasible schedules correspond to combinations on or above this curve. The efficient frontier conceptualizes that some differences between the performance of schedules are due to the trade-off between $\bar{I}(\omega)$ and $\bar{W}(\omega)$ and other differences are due to suboptimality. The first is expressed by ω and corresponds to a position on the curve. The latter are represented by the iso-performance lines (the dashed or dotted lines in Figure 2). Schedules that lie on such a line result in equivalent trade-offs that have the same value in terms of the objective function and therefore the line's angle is determined by the ω of choice. The given objective function thus breaks down the performance of schedules into a trade-off component (which ultimately is a strategic decision) and an optimality component (which is a matter of superiority of one schedule compared to another). The efficient frontier itself can be moved in the direction of the lower left corner by reducing the variability in the process.

Interestingly, the optimization problem in Equation (3) can be rewritten in terms of only (expected) waiting times:

$$\begin{aligned}(\bar{x}_1, \ldots, \bar{x}_{n-1}) &= \underset{x_1,\ldots,x_{n-1}}{\arg\min}\, \mathcal{F}[x_1, \ldots, x_{n-1}], \\ &= \underset{x_1,\ldots,x_{n-1}}{\arg\min} \left(\mathbb{E}W_n + \sum_{i=1}^{n-1} (1-\omega)\mathbb{E}W_i + \omega \sum_{i=1}^{n-1} x_i \right), \end{aligned} \qquad (4)$$

as can be seen as follows. From Equation (2), by comparing the makespan up to patient i with that up to patient $i-1$, we obtain

$$B_i + I_i = \left(\sum_{j=1}^{i-1} x_j + S_i \right) - \left(\sum_{j=1}^{i-2} x_j + S_{i-1} \right) = x_{i-1} + S_i - S_{i-1}.$$

This directly leads to

$$I_i = x_{i-1} + W_i - W_{i-1} - B_{i-1}. \tag{5}$$

Relation (4) now follows by taking expectations in (5) and by noting that a number of terms (in a telescopic series) vanish. The $\mathbb{E}B_i$ terms can be dropped from the resulting expression for the objective function, as these constants do not affect the optimal interarrival times $\bar{x}_1, \ldots, \bar{x}_{n-1}$.

Because we aim to minimize the objective function in (4), it is important to know whether the function is convex in its arguments $x_1, \ldots, x_{n-1}$. Convexity of the optimization problem has attracted substantial attention in the past (Hassin & Mendel, 2008). For a specific objective function (different from ours), convexity has been proven in Wang (1993). We present the following result.

Theorem 1. *$\mathcal{F}[x_1, \ldots, x_{n-1}]$ is convex in $\boldsymbol{x} \equiv (x_1, \ldots, x_{n-1})$, and consequently, there is a unique minimum on $\mathbb{R}_+^{n-1}$.*

Proof. Define by $W_i(\boldsymbol{x})$ the waiting time of the i-th patient if the vector of interarrival times is given by $\boldsymbol{x}$. For a given i define

$$Z_j = \sum_{k=i-j+1}^{i} B_k, \quad y_j(\boldsymbol{x}) = \sum_{k=i-j+1}^{i} x_k.$$

The following distributional equality can be obtained after repeated iteration of Equation (1):

$$W_i(\boldsymbol{x}) \overset{\mathrm{d}}{=} \max_{j\in\{0,1,\ldots,i-1\}} \{Z_j - y_j(\boldsymbol{x})\}. \tag{6}$$

Here we have followed the convention that empty sums are defined equal to 0; this correctly yields that $W_1 = 0$, that is, the first patient does not have to wait.

Now using this equality and the fact that y_j is linear in each of the interarrival times x_k, we have with $\nu \in [0, 1]$ given and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}_+^{n-1}$ that

$$\begin{aligned}
&\mathbb{E}\left[W_i(\nu\boldsymbol{x}_1 + (1-\nu)\boldsymbol{x}_2)\right] \\
&= \mathbb{E}\left[\max_{j\in\{0,\ldots,i-1\}} \{Z_j - y_j(\nu\boldsymbol{x}_1 + (1-\nu)\boldsymbol{x}_2)\}\right] \\
&= \mathbb{E}\left[\max_{j\in\{0,\ldots,i-1\}} \{\nu(Z_j - y_j(\boldsymbol{x}_1)) + (1-\nu)(Z_j - y_j(\boldsymbol{x}_2))\}\right] \\
&\leqslant \mathbb{E}\left[\max_{j\in\{0,\ldots,i-1\}} \{\nu(Z_j - y_j(\boldsymbol{x}_1))\}\right] \\
&\quad + \mathbb{E}\left[\max_{j\in\{0,\ldots,i-1\}} \{(1-\nu)(Z_j - y_j(\boldsymbol{x}_2))\}\right],
\end{aligned}$$

which equals $\nu\,\mathbb{E}[W_i(\boldsymbol{x}_1)] + (1-\nu)\,\mathbb{E}[W_i(\boldsymbol{x}_2)]$. This directly implies that $\mathbb{E}[W_i(\boldsymbol{x})]$ is convex. Because of Equation (4), the objective function is a linear combination of expected waiting times (with positive weights), minus a function that is linear in $\boldsymbol{x}$, and therefore convex as well. As a consequence, there is a unique minimum on $\mathbb{R}_+^{n-1}$. □

A weighted linear objective function is the standard and unchallenged option in the scheduling literature. For possible future consideration in research, we raise the possibility of weighted-quadratic objective functions:

$$\mathcal{F}^{(\mathrm{q})}[x_1, \ldots, x_{n-1}] = \omega \sum_{i=1}^{n} \mathbb{E}I_i^2 + (1-\omega) \sum_{i=1}^{n} \mathbb{E}W_i^2.$$

The potential of such nonlinear objective functions is that they could capture nonlinear utility loss of waiting or idle time, as anticipated, for example, in the finding of Ahmadi-Javid et al. (2017) that a patient's perception of waiting time is likely to be nonlinear. In this article, we focus on the weighted-linear objective function, but the weighted-quadratic, and also mixed versions (linear-quadratic or quadratic-linear) are readily incorporated in our approach. These alternatives are available in the online webtool implementation of our approach.

In the next section, we elaborate on our approach to compute these schedules. The first idea is to approximate F_B by mixtures of phase-type distributions, with correct μ and ϱ, which allow an analytical computation of $\sum_{i=1}^{n} \mathbb{E}I_i$ and $\sum_{i=1}^{n} \mathbb{E}W_i$. Next, we show how the approach can be extended to incorporate no-shows and walk-ins. The approach is relatively fast, but to arrive at a procedure that computes schedules instantaneously, our last tactical idea is to create a set of precalculated schedules for a grid of values of ω, ϱ, and n, and to determine schedules from there by means of interpolation. In Section 5 we will systematically compare the effectiveness of our approach with competing methods.

3 | PROPOSED APPROACH

This section presents our approach, based on an approximation of F_B by phase-type distributions. Thereupon, we demonstrate how no-shows and walk-ins can be incorporated, and how session overtime can be integrated in the objective function. Finally, we discuss how the approach can be implemented by using our webtool.

3.1 | Phase-type approximation

Unfortunately, no analytical procedures are available to determine, for generally distributed service times, the distributions of the idle times I_i, and the waiting times W_i. We remedy this by approximating the actual service times by their so-called phase-type counterparts. This approach uses the well-known fact that phase-type distributions are capable of approximating any positive distribution with arbitrary accuracy; see, for

example, Asmussen et al. (1996). The resulting queueing system allows (semi)explicit computation of the objective function, as pointed out in, for example, Kuiper et al. (2015).

We characterize the service–time distributions by fitting a phase-type distribution with the correct first two moments; the values of these moments can be chosen in line with, for instance, the findings of Çayirli and Veral (2003). This choice is motivated by the fact that it is cumbersome to estimate higher moments, and moreover, we anticipate that those higher moments have only modest impact on the performance of an appointment schedule (Kuiper et al., 2017).

In line with the literature on scheduling, we represent the first two moments by (i) the mean μ and (ii) the *squared coefficient of variation* (ς), a unitless quantity that is defined as the ratio of the variance and the square of the mean. We follow the standard procedure, advocated in Tijms (1986) and also used in Kuiper et al. (2015), to approximate the service time by a mixture of two Erlang random variables if it has a ς smaller than 1, and by a hyperexponential random variable if it has a ς larger than 1. The case of $\varsigma = 1$ corresponds to exponential service times (as was used in Figure 2). In more detail, the approximation is constructed as follows.

- In case ς is smaller than 1, the service–time distribution is approximated by a mixture of Erlang distributions: it is an Erlang distribution with $K-1$ phases and mean $(K-1)/\lambda$ with probability $p \in [0,1)$, and an Erlang distribution with K phases and mean K/λ with probability $1-p$. This mixture of Erlang random variables has mean $(K-p)/\lambda$ and variance $(K-p)/\lambda^2$. As a consequence, its squared coefficient of variation is $1/(K-p)$, which due to $p \in [0,1)$ lies in the interval $[1/K, 1/(K-1))$, for $K \in \{2,3,...\}$. As a result, we can identify unique K, λ, and p such that our mixture of Erlangs has the desired mean μ and squared coefficient of variation ς. For instance, if $\mu = 1$ and $\varsigma = 0.4$, we have to pick $K = 3$ (as $0.4 \in (s\frac{1}{3}, 1/2]$), $p = 0.5$ (so that $\varsigma = 0.4 = 1/(3-p)$), and $\lambda = 2.5$ (so that the mean service time is $(3-0.5)/\lambda = 1$).
- In the other situation, in which ς is larger than 1, the service time is approximated by a hyperexponential random variable, which is constructed as an exponential random variable with mean λ_1^{-1} with probability $p \in [0,1]$, and an exponential random variable with mean λ_2^{-1} with probability $1-p$. By imposing *balanced means*, $\lambda_1 = 2p\lambda$ and $\lambda_2 = 2(1-p)\lambda$ for some $\lambda > 0$, one reduces the number of free parameters from three to two, so that for each μ and ς, a unique hyperexponential distribution can be determined. In detail, the mean and variance of the hyperexponential random variable are given by $p/\lambda_1 + (1-p)/\lambda_2$ and $p/\lambda_1^2 + (1-p)/\lambda_2^2$, respectively, so that the corresponding squared coefficient of variation equals

$$\frac{p\lambda_2^2 + (1-p)\lambda_1^2}{p^2\lambda^2 + 2p(1-p)\lambda_1\lambda_2 + (1-p)^2\lambda_1^2},$$

which under balanced means reduces to

$$\frac{1}{4p(1-p)}.$$

Notice that $p(1-p) \in [0, 1/4]$ for $p \in [0,1]$, so that the squared coefficient of variation can attain any value larger than 1. We again illustrate the procedure to select the parameters by an example. Suppose that $\mu = 1$ and $\varsigma = 1.125$. Then we have to pick $p = 1/3$ to make sure that the squared coefficient of variation has the right value, and $\lambda = 1$ to make sure the mean has the right value.

As mentioned above, the appeal of using phase-type distributions lies in the fact that they produce accurate fits even with a relatively small number of parameters. In various specific contexts, extensive experiments have been performed to assess the performance of the phase-type approximation. In Asmussen et al. (1996) and also in more recent papers, in-depth numerical studies confirm the good fit for practically relevant distributions such as lognormal and Weibull. This is in line with our own experiments, reported later in this article, in which we work with lognormally distributed service times.

For the sake of completeness, we include in the appendix (supplementary file) the procedure for evaluating the objective functions introduced earlier for the case that the service–time distributions are of phase type; for more detail on how to use this type of procedures, we refer to Wang (1997) and Kuiper et al. (2015).

3.2 | Employing a precalculated grid

The minimization of the objective function, applying the algorithm thus outlined, can be performed using standard software (such as `MATLAB`). Normally the algorithm arrives at the optimal solution within a minute, but there are instances which require longer computation times (up to 15 min for a single problem instance). The numerical minimization is typically slower when

- n is larger than, say, 30, in which case the vector over which the optimization is performed is of (relatively) high dimensionality;
- ς is relatively small. Consider, for instance, the case that $\varsigma \in (0.2, 0.25]$ and $n = 25$. Then potentially five exponential phases enter the system with each arrival, such that the state space of the sojourn time of the n-th patient is 125-dimensional.

To avoid these long computation times for practitioners, our third tactical idea is to exploit a set of precalculated schedules for a grid of values of $\varsigma \in [0.1, 1.5]$ (step size 0.1), $\omega \in [0.05, 0.99]$ (step size 0.05), and $n = 2, \ldots, 35$. Then cubic Hermite spline interpolation is used to find a suitable schedule for values of (ς, ω) that are not on the grid. This

interpolation relies on the values and derivatives of two neighboring points in the following way.

First consider the situation of a real-valued grid $(s_i)_{i=0}^I$ such that $s_0 < s_1 < ... < s_I$ at which we have the precomputed schedules $(\boldsymbol{x}_i)_{i=0}^I$. Define the Hermite basis functions $h_{00}(\cdot)$, $h_{10}(\cdot)$, $h_{01}(\cdot)$, and $h_{11}(\cdot)$ by $h_{00}(y) = 2y^3 - 3y^2 + 1$, $h_{10}(y) = y^3 - 2y^2 + y$, $h_{01}(y) = -2y^3 + 3y^2$, and $h_{11}(y) = y^3 - y^2$. Then the interpolation for an $s \in [s_i, s_{i+1}]$ is

$$\begin{aligned}\boldsymbol{x}(s) = h_{00}(y)\boldsymbol{x}_i + h_{10}(y)(s_{i+1} - s_i)\boldsymbol{m}_i \\ + h_{01}(y)\boldsymbol{x}_{i+1} + h_{11}(y)(s_{i+1} - s_i)\boldsymbol{m}_{i+1},\end{aligned}$$

with $y := (s - s_i)/(s_{i+1} - s_i)$ and $\boldsymbol{m}_i$ the tangent in $\boldsymbol{x}_i$. This tangent can be approximated by using the grid of precomputed schedules again:

$$\boldsymbol{m}_i = \frac{\boldsymbol{x}_i - \boldsymbol{x}_{i-1}}{s_i - s_{i-1}} \quad \text{and} \quad \boldsymbol{m}_{i+1} = \frac{\boldsymbol{x}_{i+2} - \boldsymbol{x}_{i+1}}{s_{i+2} - s_{i+1}}. \tag{7}$$

In our case the grid is two-dimensional; recall that we use the interpolation to obtain schedules for pairs (ϱ, ω). Observe however that the above procedure can be extended to multiple dimensions by applying the approach in a nested manner.

The precalculated schedules are stored in and exploited by a tool that can be accessed via http://www.appointmentscheduling.info. An extensive validation, by checking middle points in between the grid, revealed that the discrepancy with the optimal schedule is negligible (well below 1%, but typically 0.05% or less).

3.3 | No-shows and walk-ins

We now point out how no-shows and walk-ins can be dealt with in a generic way, beginning with the former. We assume a per-patient probability $q \in [0, 1)$ of a no-show. We apply the developed machinery, but with adjusted service times $\bar{B}_i = B_i$ with probability $1 - q$, and $\bar{B}_i = 0$ with probability q. We have $\mathbb{E}\bar{B}_i = \mu(1 - q)$, and the squared coefficient of variation becomes

$$\begin{aligned}\bar{\varrho}(q) &= \frac{(1-q)\,\mathbb{E}B^2 - (1-q)^2\mu^2}{(1-q)^2\mu^2} \\ &= \frac{(1-q)\,\varrho + (1-q)q}{(1-q)^2} = \frac{\varrho + q}{1-q}.\end{aligned} \tag{8}$$

Although ϱ is typically smaller than 1 in healthcare, the adjusted $\bar{\varrho}$ *can* be larger than 1. For this reason, when using the phase-type approximations , the situation that $\varrho > 1$, where service times are approximated by a hyperexponential variable, *is* relevant.

Next, we incorporate walk-ins, where in addition to the patients scheduled, also unscheduled patients arrive during the session. Let $v \in [0, 1]$ be the probability that an unscheduled patient is added to an appointment slot. The revised service time $\tilde{B}_i$ is therefore equal to:

(i) two i.i.d. copies of the service time B_i with probability $(1 - q)v$ (no no-show and a walk-in),
(ii) equal to one service time B_i with probability $(1 - q)(1 - v) + qv$ (either no no-show and no walk-in, or a no-show and a walk-in),
(iii) equal to 0 with probability $q(1 - v)$ (a no-show and no walk-in).

As a consequence, the expected service time becomes

$$\mathbb{E}\tilde{B}_i = 2(1-q)v\mu + ((1-q)(1-v) + qv)\mu = (1 - q + v)\mu.$$

Along the same lines, the second moment becomes, with B' an independent copy of B,

$$\begin{aligned}\mathbb{E}\tilde{B}_i^2 &= (1-q)v\,\mathbb{E}\left[(B + B')^2\right] + ((1-q)(1-v) + qv)\,\mathbb{E}B^2 \\ &= (1-q)v\left(2\,\mathbb{E}B^2 + 2\mu^2\right) + ((1-q)(1-v) + qv)\,\mathbb{E}B^2,\end{aligned}$$

so that

$$\begin{aligned}\mathbb{V}\text{ar}\tilde{B}_i &= (1-q)v\left(2\,\mathbb{E}B^2 + 2\mu^2\right) \\ &\quad + ((1-q)(1-v) + qv)\,\mathbb{E}B^2 - (1 - q + v)^2\mu^2 \\ &= (1 - q + v)\,\mathbb{E}B^2 - (v^2 + (1-q)^2)\mu^2.\end{aligned}$$

The squared coefficient of variation thus becomes

$$\begin{aligned}\tilde{\varrho}(q, v) &= \frac{(1 - q + v)\,\mathbb{E}B^2 - (v^2 + (1-q)^2)\mu^2}{(1 - q + v)^2\mu^2} \\ &= \frac{(1 - q + v)\,\varrho + q(1-q) + v(1-v)}{(1 - q + v)^2}.\end{aligned}$$

Indeed, the case $v = 0$ is equivalent to the situation with no-shows only. This method can generally be applied to accommodate no-shows and walk-ins into methods that compute schedules, and in fact, we use this approach in Section 5 to augment methods already presented in the literature (Bailey, 1952; Hassin & Mendel, 2008; Mak et al., 2015; Robinson & Chen, 2003).

Figure 3 demonstrates the effects of no-shows and walk-ins on the optimal schedule. The graph shows the slot lengths $\bar{x}_i$ for patient $i = 1, ..., n - 1$ (with $n = 20$) in the optimal schedule in four scenarios. For all four scenarios, the optimal slot lengths follow the dome shape, with shorter slot lengths in the beginning and end of the schedule, and longer in the middle. The scenarios are defined by combinations of v and q, where $q = 0.0$ or $q = 0.4$ and $v = 0.0$ or $v = 0.4$. These ranges demarcate the values found in practice (Çayirli et al., 2012). In all scenarios, $\varrho = 0.5$ (variation in service times) and $\omega = 5/6$ (weight for idle versus waiting time).

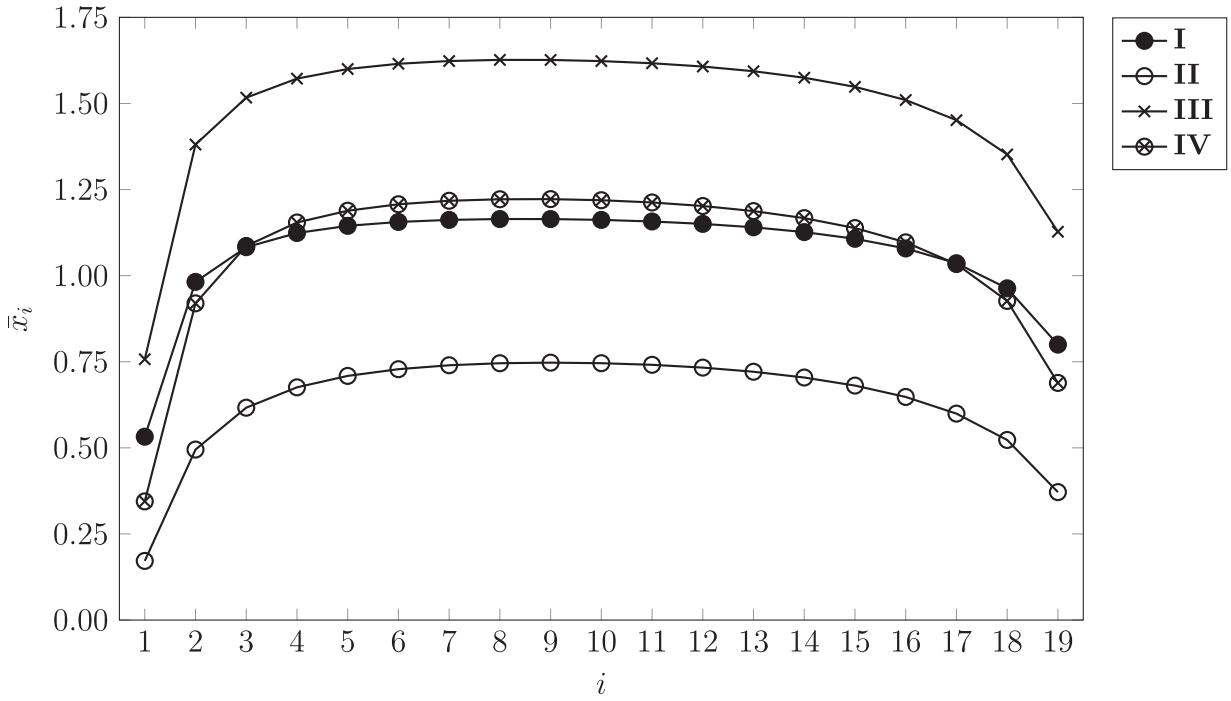

FIGURE 3 Various environments by varying the no-show rate q and walk-in rate v are considered for $\mathcal{F}$ with $\varrho = 0.5$, $\omega = 5/6$, and $n = 20$. The mean of the service times is normalized to 1

TABLE 1 A consideration of various scenarios to study the impact of environmental parameters on our approach and key metrics of the resulting schedule

	Environment			Rev. parameters		Session metrics		Performance	
Scenario	ϱ	q	v	$\mathbb{E}\tilde{B}$	$\tilde{\varrho}$	$\mathbb{E}N$	$\mathbb{E}T$	$\sum \mathbb{E}I_i$	$\sum \mathbb{E}W_i$
I	0.5	0.0	0.0	1.0	0.50	20	22.84	2.84	18.38
II	0.5	0.4	0.0	0.6	1.50	12	14.50	2.50	20.04
III	0.5	0.0	0.4	1.4	0.48	28	31.92	3.92	25.13
IV	0.5	0.4	0.4	1.0	0.98	20	23.78	3.78	26.46

The obvious effects of $q = 0.4$ (40% no-shows as in Scenario II) are that the expected number $\mathbb{E}N$ of patients seen in the session is smaller than in Scenario I, and that the allotted slot times per patient are shorter, thus accommodating the expected percentage of no-shows. Table 1 shows that the expectation $\mathbb{E}\tilde{B}$ of the adjusted service times $\tilde{B}_i$ is 0.6 instead of 1.0, so scheduling in the face of 40% no-shows is similar to scheduling with zero percent no-shows and 40% shorter expected service times. This is shown in Figure 3 as graph II is much lower than graph I. Analogously, graph III is much higher than graph I, illustrating that walk-ins are accommodated in the optimal schedule by making slot lengths longer.

Graph IV is the scenario of $q = 0.4$ no-shows and $v = 0.4$ walk-ins. The expected session length $\mathbb{E}T$ and the expected number of patients served are similar to those of Scenario I, as the expected no-shows and walk-ins cancel each other out. The variability in Scenario IV, however, is much larger than in Scenario I, which is reflected in $\tilde{\varrho} = 0.98$ as opposed to $\varrho = 0.50$. This, in turn, results in a more pronounced dome shape for the optimal schedule. This insight holds in general: larger variability in patient arrivals or service times (larger $\tilde{\varrho}$) results in the dome shape to be more pronounced. The rightmost two columns in Table 1 demonstrate the effects of larger variability due to walk-ins and no-shows on the performance of a schedule. Scenario IV serves the same expected number of patients as Scenario I, but with larger expected idle time for the service provider (3.78 instead of 2.84) and much larger expected waiting times for patients (26.46 instead of 18.38). The reason that the expected waiting time is much more affected by variability than the expected idle time is the chosen value $\omega = 5/6$, which makes idle time the overriding consideration in optimizing the schedule.

Figure 3 illustrates a session of size $n = 20$ scheduled patients, and the observed behavior is the same for other session sizes. For increasing n, the middle part of the dome shape rises up to an upper bound $\bar{x}_\infty$. This upper bound is the optimal slot-length in a steady-state model, where a session has an infinite number of patients ($n \to \infty$). In that situation the

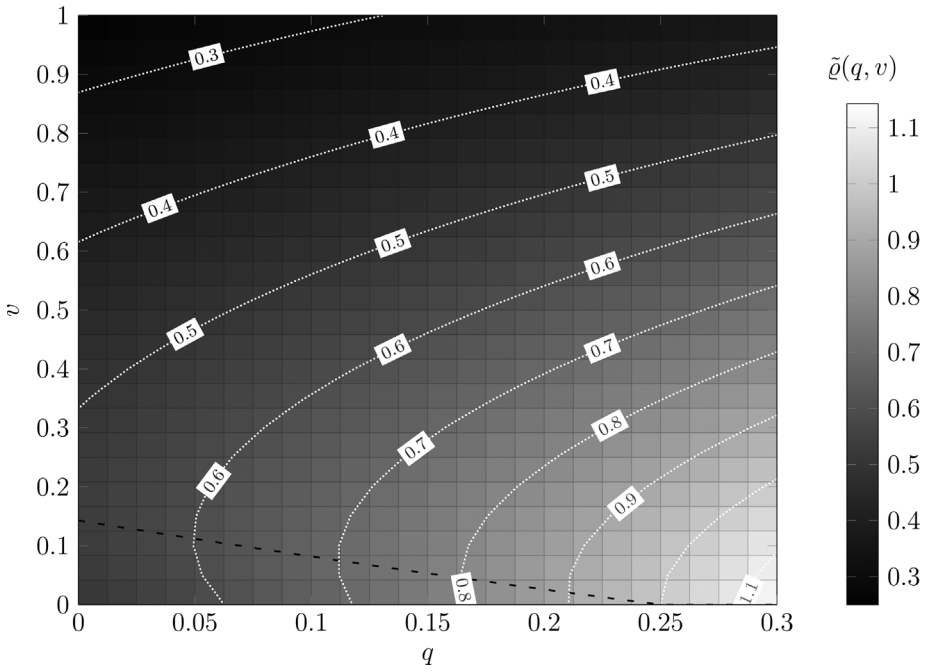

FIGURE 4 Density plot with contours of the $\tilde{\varrho}$ as a function of q and v, ϱ is set to 0.5. The dashed line demarcates two regions, below walk-ins aggravate the variation, conversely above an increase in v reduces the coefficient of variation

optimal slot lengths $\bar{x}_\infty$ are the same for every slot, and therefore, the profile of interarrival times over a session is flat instead of a dome shape (Kuiper et al., 2017). For sessions of finite size, the dome shapes are below this upper bound, and for increasing session size n, the middle part of the dome flattens out as optimal slot lengths in the middle of the session approach the upper bound.

As an aside, intuitively one could have expected that the occurrence of no-shows and walk-ins should always increase the relative variability in the process, but this does not turn out to be the case. For example, comparing Scenario IV to II in Table 1, one sees that a 0.4 walk-in rate reduces $\tilde{\varrho}$. Figure 4 shows $\tilde{\varrho}$ for a range of values for q and v, taking $\varrho = \frac{\mathbb{V}\text{ar}B}{\mathbb{E}B^2} = 0.5$. Increasing the probability q of a no-show always increases $\tilde{\varrho}$, implying that the variability increases, and as a consequence, the interarrival times in the optimal schedule have a more pronounced dome shape (Kuiper et al., 2015). The graph shows that the effect of increasing the probability v is not necessarily monotone, though. The dashed line in Figure 4 is the solution to $\partial\tilde{\varrho}(q, v)/\partial v = 0$, which gives

$$v_0(q) = \max\left\{q - 1 + \frac{4(q-1)^2}{3 - 2q + \varrho}, 0\right\}.$$

For $\varrho = 0.5$, in the area above the dashed line, $v \geq v_0(q)$, an increase in the probability v of a walk-in *decreases* the variation $\tilde{\varrho}$.

3.4 | Overtime

The makespan (SET) T as defined in Equation (2) determines how long the service provider should be available. Overtime occurs if $T >$ SET, with SET the scheduled or targeted SET. Overtime is an undesirable effect, but it is not taken into account in objective functions of the form Equation (3). Here, we show how overtime can be incorporated by a minor modification.

An objective function that penalizes overtime is

$$\begin{aligned}\mathcal{F}[x_1, \ldots, x_{n-1}] + \bar{\omega}\,\mathbb{E}O &= \omega \sum_{i=1}^{n} \mathbb{E}I_i + (1-\omega)\sum_{i=1}^{n}\mathbb{E}W_i + \bar{\omega}\,\mathbb{E}O \\ &= \omega \sum_{i=1}^{n} \mathbb{E}I_i + (1-\omega)\sum_{i=1}^{n}\mathbb{E}W_i + \bar{\omega}\max\left\{\sum_{i=1}^{n-1} x_i + S_n - \text{SET}, 0\right\} \\ &\approx \omega \sum_{i=1}^{n} \mathbb{E}I_i + (1-\omega)\sum_{i=1}^{n}\mathbb{E}W_i + \bar{\omega}\left(\sum_{i=1}^{n}\mathbb{E}B_i + \sum_{i=1}^{n}\mathbb{E}I_i - \text{SET}\right)\end{aligned} \tag{9}$$

for some scalar $\bar{\omega} > 0$ and scheduled SET. Given the fact that the $\mathbb{E}B_i$ s and SET are not affected by the choice of the schedule, it is seen that minimizing the above objective function is equivalent to minimizing

$$\frac{\omega + \bar{\omega}}{1 + \bar{\omega}} \sum_{i=1}^{n} \mathbb{E}I_i + \frac{1-\omega}{1+\bar{\omega}} \sum_{i=1}^{n} \mathbb{E}W_i.$$

This objective function is of the same form as the original problem, but with adapted weights. As a consequence, the techniques we developed can be used in the setting that incorporates overtime in the objective function. Indeed, Klassen and Yoogalingam (2014) conclude by simulation that the incorporation of overtime in the objective function has a similar effect as increasing the weight put on idle time.

3.5 | Possible extensions

To explicitly show the flexibility of our model, we outline some extensions in several directions. In the first place, while we concentrate in this article on the situation that the patients' service times are i.i.d., our set-up naturally extends to heterogeneous service times; more specifically, the result of Theorem 1 does not require the service times to be i.i.d., and the evaluation techniques also generalize to heterogeneous service times. This observation in particular implies that the parameters q and v could be chosen slot-dependent. This could be used if it is empirically observed that there are some systematic variations in these parameters. For instance, if the no-show probability at the first slot, early in the morning, is higher than later in the day, then this can be taken care of by giving the parameter q corresponding to the first patient a suitably chosen higher value.

Above we assumed that at most one walk-in patient could be added to each appointment slot. One could, however, consider the (more realistic) setup in which walk-ins follow a Poisson process with parameter κ. This means that the number of unscheduled patients entering in $[t_i, t_{i+1})$ follows a Poisson distribution with parameter κx_i, with $x_i = t_{i+1} - t_i$ denoting the slot length. By Equation (8) we know for the situation with only no-shows (i.e., no walk-ins), the mean $\mathbb{E}\bar{B}$ equals $(1-q)\mu$, whereas the variance $\mathbb{V}\mathrm{ar}\bar{B}$ equals $(\varrho + q)(1-q)\mu^2$. Incorporating Poisson walk-ins, the mean (of the work brought along by all patients in this slot) becomes

$$\mathbb{E}\tilde{B}_i = \mu(1 - q + \kappa x_i),$$

whereas (relying on the variance version of Wald's equality for a random sum of random variables) the variance becomes

$$\begin{aligned}\mathbb{V}\mathrm{ar}\tilde{B}_i &= \mathbb{V}\mathrm{ar}\bar{B} + \kappa x_i\, \varrho\mu^2 + \kappa x_i\, \mu^2 \\ &= (\varrho + q)(1-q)\mu^2 + \kappa x_i(\varrho + 1)\mu^2.\end{aligned}$$

When dividing $\mathbb{V}\mathrm{ar}\tilde{B}_i$ by $(\mathbb{E}\tilde{B}_i)^2$, we thus obtain

$$\tilde{\varrho}_i(q, \kappa) = \frac{(\varrho + q)(1-q) + \kappa x_i(\varrho + 1)}{(1 - q + \kappa x_i)^2}.$$

The procedure can even be further generalized by allowing the arrival rate κ to be time dependent (which is in particular useful if there is, say, a day profile in the arrival process of the walk-ins). If the walk-in arrival rate at time s is $\kappa(s)$, then κx_i in the above expression should be replaced by $\int_{[t_i, t_{i+1}]} \kappa(s)\, \mathrm{d}s$.◇

4 | OPERATION OF THE WEBTOOL

We explain the use of the webtool in practice. First, the user enters appropriate values for μ $(= \mathbb{E}B_i)$ and ϱ, preferably based on the estimated first two moments of measured service times. Typical values for ϱ in healthcare are between 0.1225 and 0.7225 (Çayirli & Veral, 2003), where larger ϱ implies less consistent and less predictable service times. Both μ and ϱ need to be estimated from recorded service times, from which the average $\bar{X}$ and standard deviation S can be used to estimate μ and ϱ (the latter as $\frac{S^2}{\bar{X}^2}$). The user also enters estimates for the per-slot probabilities q and v of a no-show or a walk-in.

Second, the user enters two out of the triple n (the number of patients to be scheduled), ω (the desired weight), and $\mathbb{E}T$ (the expected SET). As explained before, ω is ultimately a strategic choice reflecting the clinic's value proposition. Robinson and Chen (2011) studied mean queue lengths and utilizations in practice to determine how the trade-off is implicitly made in reality. They report implied values up to $\omega = 0.98$, corresponding to situations where idle times are minimized by allowing long waiting times for patients. Depending on which two parameters are entered, the webtool produces the following results:

1. If n and ω are provided, the webtool generates the resulting optimal schedule as well as its expected makespan $\mathbb{E}T$.
2. Entering n and $\mathbb{E}T$ (where, evidently, $\mathbb{E}T > n\,\mathbb{E}B$), the tool determines the implied value of ω and returns the optimal schedule that matches the expected makespan $\mathbb{E}T$.
3. The final option is to select $\mathbb{E}T$ and ω to find out how many patients can optimally be scheduled such that the expected makespan remains below $\mathbb{E}T$, given the weight ω.

Note that in the latter two use cases the requirement set on the makespan can be thought of as complying to a target on the SET. The tool produces schedules in continuous time. In practice, however, slot lengths are typically discrete multiples of a resolution Δ (for instance 5 min), which the user can specify. An idea is to round the arrival epochs $t_1, \ldots, t_n$ to multiples of Δ. Such a procedure is computationally considerably more efficient than solving the corresponding integer-programming problem. As illustrated in Table 2, rounding typically leads to near-optimal solutions: the scheduled arrival epochs of the rounded schedule and those of the optimal discrete solution differ only for one (for $\omega = 0.5$) or two patients (for $\omega = 0.8$). Moreover, the difference in terms of the objective function is marginal.

TABLE 2 Both tables give values for the interarrival and arrival times, for the optimal continuous schedule, the optimal discrete schedule, and the rounded schedule. Left table corresponds to $\omega = 0.5$ and right table to $\omega = 0.8$. In the tables, ϱ is assumed 0.5 and $\Delta = 5$

$\varrho = 0.5$ and $\omega = 0.5$

Patient	Continuous		Discrete		Rounded	
i	$\bar{x}_i$	$\bar{t}_{i+1}$	x_i	t_{i+1}	x_i	t_{i+1}
1	15.93	15.93	15	15	15	15
2	20.76	36.69	20	35	20	35
3	21.48	58.17	20	55	25	60
4	21.73	79.90	25	80	20	80
5	21.81	101.71	20	100	20	100
6	21.82	123.54	25	125	25	125
7	21.77	145.31	20	145	20	145
8	21.65	166.96	20	165	20	165
9	21.42	188.38	25	190	25	190
10	20.97	209.35	20	210	20	210
11	19.99	229.34	20	230	20	230
12	17.03	246.37	15	245	15	245
$\mathbb{E}T$	268.92		268.51		268.55	
Objective function	66.57		67.04		67.04	

$\varrho = 0.5$ & $\omega = 0.8$

Patient	Continuous		Discrete		Rounded	
i	$\bar{x}_i$	$\bar{t}_{i+1}$	x_i	t_{i+1}	x_i	t_{i+1}
1	8.82	8.82	10	10	10	10
2	15.32	24.14	15	25	15	25
3	16.64	40.79	15	40	15	40
4	17.13	57.91	20	60	20	60
5	17.31	75.22	15	75	15	75
6	17.33	92.55	20	95	20	95
7	17.24	109.78	15	110	15	110
8	17.02	126.81	20	130	15	125
9	16.66	143.46	15	145	20	145
10	16.05	159.51	15	160	15	160
11	14.96	174.47	15	175	15	175
12	12.42	186.89	15	190	10	185
$\mathbb{E}T$	222.30		223.74		222.42	
Objective function	52.46		52.77		52.79	

5 | PERFORMANCE EVALUATION

We discuss now how our approach performs. The second part of this section presents a quantitative comparison of our approach against a selection of competing methods. Not all approaches published in the literature solve exactly the same problem, however, and to avoid a comparison of apples and oranges, the first part of this section presents a collation of a wider range of approaches against qualitative criteria.

5.1 | Qualitative comparison

We first compare approaches in terms of the generality of the scheduling problem that they can solve, and their ease of use. Table 3 collates approaches presented in the recent literature (Çayirli et al., 2012; Denton & Gupta, 2003; Hassin & Mendel, 2008; Kaandorp & Koole, 2007; Klassen & Yoogalingam, 2009; Mak et al., 2015; Robinson & Chen, 2003). We also included Bailey's rule (Bailey, 1952), where two patients are scheduled in the first slot, and subsequent arrival times are set at intervals equal to the mean service time. This rule marks the beginning of scientific interest in the topic and has had a big impact in the field. We include the rule as a benchmark against which the newer approaches can be compared.

The first part of the qualitative comparison addresses the generality of the scheduling problem that the approaches can solve. The first criterion summarizes the components of the objective function that is optimized, where $\mathbb{E}W$ is a shorthand for the sum of expected waiting times, $\mathbb{E}I$ for the sum of expected idle times, and $\mathbb{E}O$ for overtime. Note that all approaches incorporate expected waiting time in their objective function, and either the expected idle time, or overtime, or both. All approaches offer at least one weight parameter for balancing these subgoals.

The criteria *No-shows* and *Walk-ins* indicate whether these phenomena are incorporated in the approach. The criterion of *Service times* summarizes the distributional assumptions that the approach makes about F_B. Most approaches claim that they are applicable for a wide range of distributions (described as "general" in the table). Other approaches are based on specific distributional assumptions, such as the lognormal (which seems realistic in many situations) or the exponential distribution (which will rarely be realistic). *Approximations and simplifications* summarizes the sort of tactics used to make the problem tractable.

The next three criteria reflect how easily the approaches can be used. *Required input* summarizes the parameters that the user needs to estimate and offer as input, such as μ, ϱ, q, and v. Approaches that need fewer estimates are easier to use, whereas approaches such as Klassen and Yoogalingam (2009) and Denton and Gupta (2003) require the user to determine the whole distribution F_B before the approach can be used. *Computation time* qualifies how fast the approach is, ranging from methods that return a schedule almost instantaneously, to approaches that are so slow that for larger session sizes n, the computation time becomes prohibitive, especially if users want to experiment with the options and try out various alternative schedules. *Implementation* describes the form in which the approach is offered to users, which may be in the form of a webtool, a closed-form expression, or an implementation requiring specialist software. The last criterion in Table 3, *Performance*, summarizes the results of the quantitative comparisons explained in the second part of this section.

From the table we conclude that all approaches have a more narrow application domain than our method, although Çayirli et al. (2012) (restricted to lognormally distributed service

TABLE 3 Qualitative comparison between recent approaches

	Proposed approach	Mak et al. (2015)	Cayirli et al. (2012)	Klassen & Yoog. (2009)	Hassin & Mendel (2008)	Kaandorp & Koole (2007)	Denton & Gupta (2003)	Robinson & Chen (2003)	Bailey (1952)
Problem generality									
Objective function	$\mathbb{E}W, \mathbb{E}I, \mathbb{E}O$	$\mathbb{E}W, \mathbb{E}O$	$\mathbb{E}W, \mathbb{E}I, \mathbb{E}O$	$\mathbb{E}W, \mathbb{E}I, \mathbb{E}O$	$\mathbb{E}W, \mathbb{E}I$	$\mathbb{E}W, \mathbb{E}I, \mathbb{E}O$	$\mathbb{E}W, \mathbb{E}I, \mathbb{E}O$	$\mathbb{E}W, \mathbb{E}I$	$\mathbb{E}W, \mathbb{E}I$
No-shows	Yes	No	Yes	Yes	Yes	Yes	No	No	No
Walk-ins	Yes	No	Yes	No	No	No	No	No	No
Service times	General	General	Lognormal	General	Exponential	Exponential	General	Specific Lambda	General
Approximations and simplifications	Phase-type approx. for F_B	Worst-case distribution	Least-Sq. approximation	Simulation, scatter search	Sequential quadratic progr.	Discretized time, local search	Partitioning, Bender's decomp. (LP)	Two atheoretic functions	Simple heuristic
Ease of use									
Required input	μ, ϱ, q, v	μ, σ	μ, ϱ, q, v	F_B, q	μ, q	μ, q	F_B	None	None
Computation time	Negligible	Negligible	Negligible	Prohibitive	Considerable	Considerable	Prohibitive	Negligible	Negligible
Implementation	Webtool	Spreadsheet	Closed-form expression	Specialist software	Specialist software	Specialist software	Specialist software	Closed-form expression	Simple rule
Performance									
In 162 test cases	Benchmark	24.6%[1)]	5.8%		9.5%[1)]			17.7%[1)]	9.5%[1)]

[1)]Performance when the method is improved so as to incorporate no-shows and walk-ins, following the ideas outlined earlier.

times) and Klassen and Yoogalingam (2009) (do not incorporate walk-ins) are rather similar. In terms of ease of use, the differences are large, with on one extreme, some approaches offering an easy-to-use closed-form expression or a webtool, but on the other extreme, approaches that require specialist software or that quickly become very slow for all but the smallest session sizes n. Again Çayirli et al. (2012) is comparable to our approach, and also Mak et al. (2015) and Robinson and Chen (2003) have comparable ease of use. Combining the two perspectives, we believe that only Çayirli et al. (2012) competes with our approach, all the others either solving a more restricted problem, or offering inferior ease of use.

5.2 | Quantitative comparison of performance

Based on the qualitative comparison in Table 3, we selected the *universal appointment rule* CAY (Çayirli et al., 2012) as the main competitor. CAY produces dome-shaped schedules, where the arrival times are set at

$$t_i := \max\left\{0, k(i-1)\mathbb{E}\tilde{B} - \sigma'\sqrt{i}\cdot\frac{n+i}{n-1}\right\}, \tag{10}$$

where time is normalized such that $\mathbb{E}B_i = 1$ and σ' is a revised standard deviation. The schedule is additionally adjusted for situational characteristics by the scalar k. It is a value that depends on ϱ, ω, q, v, and n, and which optimizes a linear-weighted objective function. The relationship between k and the parameters ϱ, ω, q, v, and n is approximated by a nonlinear regression equation, based on extensive simulations, and achieving an R^2 of 0.9529. CAY optimizes a weighted-linear objective function of expected idle time, waiting time, and overtime.

The CAY rule approximates the optimal schedule for a situation by the best fitting schedule adhering to Equation (10), which in turn is approximated by the nonlinear regression equation for k. This fitted approximation has an R^2 of 0.9529 when service times are lognormal, but its precision is unknown otherwise. The proposed approach approximates service times by their phase-type counterparts, and returns the corresponding optimal schedule (exactly on the grid points, or by interpolation otherwise).

The numerical performance comparison is based on $3 \times 3 \times 3 \times 2 \times 3 = 162$ test cases that Çayirli et al. (2012) claim to be representative for outpatient clinics. These 162 situations correspond to ϱ equal to 0.16, 0.36, and 0.64; no-show probabilities q of 0.05, 0.20, and 0.40, and walk-in probabilities v of 0.0, 0.2, and 0.4. The number n of patients to be scheduled equals 10 or 20. The weight parameter $\omega/(1-\omega)$ has three levels: 2, 5, and 10 (and hence ω equals $2/3$, $5/6$, and $10/11$). To enable the comparison with CAY we include and fix the relative cost of overtime (w.r.t. the cost of idle time) by setting $\bar{\omega} = 1.5\omega$. As a consequence, the optimization problem uses a revised weight parameter $\omega^\star :=$

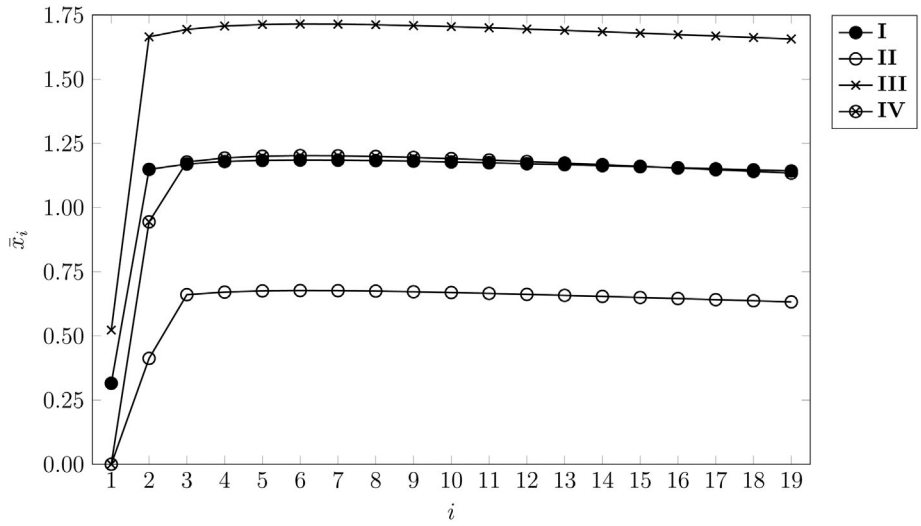

FIGURE 5 The appointment schedules generated by the CAY rule, with same settings as used in Figure 3

$2.5\omega/(1+1.5\omega)$, which is the result of Equation (9). In our cases this leads to the following $\omega^\star$ values $5/6$, $25/27$, and $26/27$. Furthermore, in line with Çayirli et al. (2012), the test cases involve a lognormal distribution for the service times. Note that this set-up for the comparison favors the CAY rule, as this rule was optimized for lognormal services times. The lognormal distribution is not a phase-type distribution, so the test cases are unfavorable to our approach.

Figure 5 depicts the interarrival times by using the CAY rule with the same settings as used earlier. A comparison of Figures 3 and 5 shows that both approaches lead to dome-shaped patterns. The interarrival times by the CAY rule (Figure 5) start low and then jump to a plateau after which the slot lengths gradually decrease, which is a direct consequence of Equation (10) that imposes this structure. In our approach, however, the dome shape is endogenous: we do not a priori impose this shape, but it instead follows by optimizing our objective function.

For each of the 162 test situations, we computed a schedule following our approach, and following the CAY rule. We express the performance of both schedules in terms of the objective function given in Equation (9). To evaluate the objective function, given one of both schedules, we simulate 10,000 sessions, where the service times are drawn from a lognormal distribution (with squared coefficient of variation ϱ), and no-shows and walk-ins are incorporated as realizations of Bernoulli variables with probabilities q and v. From these 10,000 realizations, $\sum_{i=1}^{n} \mathbb{E}I_i$ and $\sum_{i=1}^{n} \mathbb{E}W_i$ can be found. Thus, for each of the 162 test cases, we determined the achieved values $\mathcal{F}_{\texttt{CAY}}$ and $\mathcal{F}_{\texttt{PRO}}$ of the objective function in Equation (9) for the schedules generated by the CAY rule and by the proposed approach, respectively.

Table 4 presents the relative difference in the performance achieved by our schedules compared to those generated from the CAY rule in the 162 test cases. The table's entries are $100\% \times (\mathcal{F}_{\texttt{CAY}} - \mathcal{F}_{\texttt{PRO}})/\mathcal{F}_{\texttt{PRO}}$. The proposed approach outperforms CAY in 154 of the 162 test cases; in the other eight cases, the relative difference is below 1%. Averaged over all 162 cases, CAY performs about 5.8% worse than the proposed approach, even though the test cases are favorable to the CAY rule.

Zooming in on the realized schedules by both approaches, we observe that differences between the scheduled appointment times are relatively small. However, the effects of these small differences on the corresponding $\mathbb{E}I_i$ and $\mathbb{E}W_i$ typically accumulate over a session, which in many cases lead to a substantially different value of the objective function. The resulting differences in terms of expected idle and waiting times are most pronounced in the last slots, where the idle times fall and waiting times rise more sharply for the PRO approach than for the CAY rule.

A similar numerical study shows that the proposed approach outperforms Bailey's rule by over 21.2%. We improved Bailey's rule by modifying the slot lengths so that they incorporate the probabilities of no-shows and walk-ins (following the ideas outlined in the section that deals with no-shows and walk-ins). The improved rule is still outperformed by the proposed approach by 9.5% (this result is reported in the bottom row of Table 3). In similar vein, we improved the methods of Mak et al. (2015) and Robinson and Chen (2003) to take no-shows and walk-ins into account, and Hassin and Mendel (2008) to take walk-ins into account. In the improved versions, these methods are outperformed by 24.6%, 17.7%, and 9.5%.

From the qualitative and quantitative comparisons in this section, we conclude that the proposed approach is general and easy to use, and that the produced schedules outperform competing approaches almost uniformly.

TABLE 4 Performance of our approach relative to the CAY rule. The percentages are the gain of our approach compared to CAY

Environment				$n = 10$			$n = 20$		
				$\omega^\star = 5/6$	$\omega^\star = 25/27$	$\omega^\star = 25/26$	$\omega^\star = 5/6$	$\omega^\star = 25/27$	$\omega^\star = 25/26$
#	ϱ	q	v	%					
1	0.16	0.05	0.00	17.15	25.79	34.65	7.11	12.56	20.20
2	0.16	0.05	0.20	8.97	13.04	20.04	2.65	3.54	8.89
3	0.16	0.05	0.40	11.06	11.85	17.57	4.06	3.34	6.74
4	0.16	0.20	0.00	1.75	6.13	15.51	0.87	0.57	4.78
5	0.16	0.20	0.20	2.39	2.40	6.12	2.13	0.13	0.79
6	0.16	0.20	0.40	3.62	2.13	4.53	1.21	0.36	0.62
7	0.16	0.40	0.00	−0.92	1.97	10.11	4.19	−0.26	1.44
8	0.16	0.40	0.20	0.43	−0.62	1.57	6.08	1.78	−0.10
9	0.16	0.40	0.40	2.11	−0.08	0.93	3.08	1.67	−0.09
10	0.36	0.05	0.00	7.18	16.30	25.57	2.35	7.46	15.33
11	0.36	0.05	0.20	5.95	10.99	18.86	1.85	4.38	10.16
12	0.36	0.05	0.40	9.11	10.98	17.42	3.49	4.23	9.51
13	0.36	0.20	0.00	2.00	6.31	14.30	1.28	1.17	6.88
14	0.36	0.20	0.20	2.83	3.45	8.79	1.45	0.58	3.34
15	0.36	0.20	0.40	4.73	4.53	8.41	1.31	1.10	3.26
16	0.36	0.40	0.00	−0.33	2.83	10.28	3.28	−0.41	2.99
17	0.36	0.40	0.20	0.90	0.94	4.91	3.03	0.24	0.43
18	0.36	0.40	0.40	2.96	1.93	4.47	1.21	0.25	0.61
19	0.64	0.05	0.00	5.44	11.13	19.56	1.40	6.11	14.58
20	0.64	0.05	0.20	5.48	9.61	17.28	1.85	4.48	11.95
21	0.64	0.05	0.40	9.26	11.56	18.08	3.69	5.98	12.02
22	0.64	0.20	0.00	2.10	5.40	13.88	1.21	1.80	8.14
23	0.64	0.20	0.20	2.82	4.47	10.03	1.10	1.47	6.03
24	0.64	0.20	0.40	5.96	7.50	12.48	1.82	2.25	6.75
25	0.64	0.40	0.00	0.22	3.83	10.72	2.17	0.07	4.46
26	0.64	0.40	0.20	1.51	3.13	7.17	1.52	0.21	2.53
27	0.64	0.40	0.40	5.57	5.29	9.46	0.81	0.70	3.39
Average				4.45	6.77	12.69	2.45	2.44	6.13

6 | DISCUSSION AND CONCLUSIONS

Appointment scheduling directly impacts the perceived quality, cost-efficiency, and capacity of a substantial part of healthcare services. Our account establishes a framework for reasoning about the performance of appointment schedules, and structures the problem of designing a schedule. We frame the problem as one of achieving appropriate buffers to absorb variability and uncertainty in the arrivals and service times of patients.

We further show that the appropriateness of schedules pertains to two issues. First, the achieved balance between idle and waiting time, reflected in the implied ω, should be consistent with the clinic's value proposition. And second, for given ω, the achieved expected idle and waiting times should be as near as possible to the efficient frontier. Our framework identifies a number of situational characteristics besides ω that should be taken into account in the design of a schedule: the number n of patients to be scheduled per session, the mean service time μ, the squared coefficient of variation ϱ of service times, and the probabilities q of a no-show and v of a walk-in.

For the actual generation of schedules we propose an approach that is fast, robust, and flexible. We offer a webtool that produces optimal schedules instantaneously. The scheduled interarrival times follow the familiar dome-shaped pattern and are based on an approximation of the service–time distribution by its phase-type counterpart. The tool offers a number of customizations such as the choice of four objective functions and a preferred resolution on the produced schedules. The tool implements three functionalities: it produces an efficient schedule given n and ω, or the implied ω given n

and the expected SET $\mathbb{E}T$, or the number of patients that can be scheduled for a given value of ω in a session with expected end time $\mathbb{E}T$.

Compared to other approaches, our method is more useful in terms of the generality of the problem that it can solve, and in terms of its ease of use, as shown in Table 3. Numerical simulations demonstrate that the method outperforms its main competitor, the *universal appointment rule* of Çayirli et al. (2012), almost uniformly and by 5.8% on average. The traditional rule of Bailey (Bailey, 1952) is outperformed by 22.1% on average.

This article contributes to the literature presenting and assessing mathematical approaches to the problem of appointment scheduling. A rich literature has built up on this subject, with the approaches included in the performance comparison in the previous section as the most relevant contributions for our setting. Recently, De Snoo et al. (2011), Ahmadi-Javid et al. (2017), and others have drawn attention to clinical practice, raising the question in how far mathematical approaches are, can, and should be adopted in practice. These questions are the focus of a multiple-case-study undertaken by Kuiper et al. (2021), involving 10 outpatient clinics. The authors find a large gap between mathematical theory, which offers a rich edifice of formal mathematical optimization approaches, and clinical practice, which deals with the problem based on experience and improvisation. Across the board, care professionals are unaware of results in operations research theory and there is limited awareness of the concepts on which it is built, such as the pursuit of striking a balance between waiting and idle times.

None of the clinics studied in Kuiper et al. (2021) used any form of theory, data, or formal method to design its scheduling policies, and instead, scheduling is based on experience and practices that have evolved over years. Partly, these findings mirror a conclusion of White et al. (2011), that clinicians' intuition about managing capacity in clinics may differ substantially from best policies.

Kuiper et al. (2021) propose that mathematical optimization approaches, such as the one proposed in this article, are useful especially in clinics whose operations have the characteristics of a high-volume and low-variety process, where patients and resources are interchangeable for the purpose of scheduling. Mathematical optimization may be less useful in settings where the scheduling problem is dominated by complex constraints brought about by idiosyncratic differences between patients, resources, and demands.

We believe that the framework set forth in this article establishes a solid basis for further refinements and additions. We demonstrated how the core procedure for computing suitable schedules lends itself easily to the incorporation of situational specifics such as no-shows, walk-ins, overtime, and discrete time slots. Further research could enrich the approach by incorporating additional relevant phenomena, such as unpunctuality (e.g., Çayirli et al., 2006; Klassen & Yoogalingam, 2014; Deceuninck et al., 2018). Obvious candidates include the situation of multiple interchangeable healthcare providers, heterogeneous patient populations, and processes consisting of more than one stage. The integration of such additions into a single framework and webtool is a strong trump for finding adoption for the approach in practice.

REFERENCES

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

Anderson, R., Camacho, F. & Balkrishnan, R. (2007) Willing to wait?: The influence of patient wait time on satisfaction with primary care. *BMC Health Services Research*, 7(1), 7–31.

Asmussen, S., Nerman, O. & Olssen, M. (1996) Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4), 419–441.

Bailey, N. (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(2), 185–199.

Barron, W. (1980) Failed appointments. Who misses them, why they are missed, and what can be done. *Primary Care*, 7(4), 563–574.

Begen, M., Levi, R. & Queyranne, M. (2012) A sampling-based approach to appointment scheduling. *Operations Research*, 60(3), 675–681.

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

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

Çayirli, T., Yang, K. & Quek, S. (2012) A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management*, 21(4), 682–697.

Charnetski, J. (1984) Scheduling operating room surgical procedures with early and late completion penalty costs. *Journal of Operations Management*, 5(1), 91–102.

De Snoo, C., Van Wezel, W. & Jorna, R.J. (2011) An empirical investigation of scheduling performance criteria. *Journal of Operations Management*, 29(3), 181–193.

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.

Deceuninck, M., Fiems, D. & Vuyst, S.D. (2018) Outpatient scheduling with unpunctual patients and no-shows. *European Journal of Operational Research*, 265(1), 195–207.

Denton, B. & Gupta, D. (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35(11), 1003–1016.

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

Hassin, R. & Mendel, S. (2008) Scheduling arrivals to queues: a single-server model with no-shows. *Management Science*, 54(3), 565–572.

Hopp, W. & Spearman, M. (2008) *Factory Physics*, 3rd edition, Boston, MA: McGraw-Hill.

Huang, X. (1994) Patient attitude towards waiting in an outpatient clinic and its applications. *Health Services Management Research*, 7(1), 2–8.

IOM (2001) *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press.

IOM (2006) *Hospital-Based Emergency Care: At the Breaking Point*. Washington, DC: National Academy Press.

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

Klassen, K. & Yoogalingam, R. (2009) Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management*, 18(4), 447–458.

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

Kuiper, A., de Mast, J. & Mandjes, M. (2021) The problem of appointment scheduling in outpatient clinics: a multiple case study of clinical practice. *Omega*, 98, 102122.

Kuiper, A., Kemper, B. & Mandjes, M. (2015) A computational approach to optimized appointment scheduling. *Queueing Systems*, 79(1), 5–36.
Kuiper, A., Mandjes, M. & De Mast, J. (2017) Optimal stationary appointment schedules. *Operations Research Letters*, 45(6), 549–555.
Mak, H., Rong, Y. & Zhang, J. (2015) Appointment scheduling with limited distributional information. *Management Science*, 61(2), 316–334.
May, J.H., Strum, D.P. & Vargas, L.G. (2000) Fitting the lognormal distribution to surgical procedure times. *Decision Sciences*, 31(1), 129–148.
Porter, M. (2010) What is value in health care? *New England Journal of Medicine*, 363(26), 2477–2481.
Robinson, L. & Chen, R. (2003) Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions*, 35(3), 295–307.
Robinson, L. & Chen, R. (2011) Estimating the implied value of the customer's waiting time. *Manufacturing & Service Operations Management*, 13(1), 53–57.
Tijms, H. (1986) Stochastic Modelling and Analysis—A Computational Approach. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Chichester, UK: John Wiley & Sons.
Wang, P. (1993) Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics*, 40(3), 345–360.
Wang, P. (1997) Optimally scheduling *n* customer arrival times for a single-server system. *Computers & Operations Research*, 24(8), 703–716.
Welch, J. & Bailey, N. (1952) Appointment systems in hospital outpatient departments. *The Lancet*, 259(6718), 1105–1108.
White, D.L., Froehle, C.M. & Klassen, K.J. (2011) The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management*, 20(3), 442–455.
Zacharias, C. & Pinedo, M. (2014) Appointment scheduling with no-shows and overbooking. *Production and Operations Management*, 23(5), 788–801.

AUTHOR BIOGRAPHIES

Alex Kuiper is an assistant professor at the Amsterdam Business School of the University of Amsterdam since 2016. He received his MSc degrees in Stochastics and Financial Mathematics and Econometrics from the University of Amsterdam (2013) and a PhD in Operations Research from the same university (2016). He also works as a Lean Six Sigma consultant for the Institute of Business and Industrial Statistics within the University of Amsterdam. His research includes operations improvement and healthcare optimization.

Michel Mandjes (1970) is a full professor at the Mathematics Department of the University of Amsterdam, and has a part-time appointment at the business school of the same university. He received his MSc degrees in Mathematics and Econometrics from the VU University, Amsterdam (1993), and a PhD in Operations Research from the same university (1996). He has worked as a member of technical staff subsequently at KPN Research, the Netherlands, and at Lucent Technologies/Bell Laboratories, Murray Hill NJ, United States. From 2000 he has been appointed as a full professor, between 2000 and 2004 at the University of Twente, and as of 2004 at the University of Amsterdam (since 2006 full-time). Between 2000 and 2006 he worked at CWI, Amsterdam, where he was department head. In 2008, he was visiting professor at Stanford University, and more recently at Columbia University and New York University. His research focuses on various aspects of stochastic processes and their applications in operations research.

Jeroen de Mast is professor in the Statistics and Actuarial Science Department of the University of Waterloo, and also Academic Director of Professional Education at the Jheronimus Academy of Data Science. His research interests include analytical problem solving, statistical methods for operations improvement, and the statistical evaluation of measurement systems. Besides academic work, he works as a consultant, professional trainer, and scientific director at Holland Innovative.

Ruben Brokkelkamp is a PhD student in the Networks & Optimization group at CWI, Amsterdam. He earned a BSc degree in Computer Science and BSc and MSc degrees in Mathematics from the University of Amsterdam. His research interests include algorithmic game theory and combinatorial optimization.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kuiper A, Mandjes M, de Mast J, Brokkelkamp R. (2023) A flexible and optimal approach for appointment scheduling in healthcare. *Decision Sciences*, 54, 85–100. https://doi.org/10.1111/deci.12517