Special Issue Article

Quality and
Reliability
Engineering
International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1863

Published online 7 September 2015 in Wiley Online Library

Practical Principles in Appointment Scheduling

Alex Kuiper[a]*† and Michel Mandjes[b]

Appointment schedules aim at achieving a proper balance between the conflicting interests of the service provider and her clients: a primary objective of the service provider is to fully utilize her available time, whereas clients want to avoid excessive waiting times. Setting up schedules that strike a good balance is severely complicated by the fact that the clients' service times are random. Because of the lack of explicit expressions, one has often set up schedules relying on simulation techniques. In this paper, we take a radically different approach: we use newly developed analytical techniques to numerically determine optimal schedules (i.e., schedules that optimize a given objective function that incorporates the interests of the service provider as well as the clients) and compare them with a number of easily evaluated heuristics; in our setup, it is throughout assumed that a given fraction of the clients does not show up. Our results are particularly useful in situations in which there is a significant variation in the service times, which is typically the case in various healthcare-related settings. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: appointment scheduling; heuristics; phase-type distribution; healthcare engineering

1. Introduction

One of the challenges in healthcare engineering concerns setting up well-balanced appointment schedules. The goal of such schedules is to regulate supply and demand: when no appointment schedule is used, this likely leads to all sorts of unwanted effects, such as excessive waiting times for patients during peak periods as well as long idle times for the medical staff during quiet hours.

An appointment schedule is essentially the sequence of epochs at which the individual patients are supposed to arrive. Those schedules are meant to soundly balance the interests of the medical staff and its clients. Clearly, the doctor's valuable time should not to be wasted, and therefore, idle time should be avoided. At the same time, it is increasingly realized that one should set up schedules such that they offer the patients an acceptable service level (for instance expressed in terms of the waiting times that they experience).

To the best of our knowledge, the problem of setting up appointment schedules was first studied by Bailey[1] and Welch and Bailey[2] in 1952 and has gained increasing interest ever since; we refer to, for example, Cayirli and Veral[3] for a good overview of the work on appointment scheduling. A wide range of approaches has been developed; we here review the ones that are particularly relevant in the context of the setting discussed in the present paper.

Most papers focus on the situation that patients' service times are random, where it is typically assumed that the individual service times are independent and identically distributed. The variability of the service-time distribution is often expressed in terms of the *squared coefficient of variation*, in the sequel denoted by SCV, defined for a random variable $B \geq 0$ by

$$\text{SCV} := \frac{\mathbb{V}\text{ar}[B]}{\mathbb{E}[B]^2};$$

where CV equals $\sqrt{\text{SCV}}$. In healthcare settings, the CV typically lies in the interval [0.35, 0.85], as reported by Cayirli and Veral.[3]

In many studies, one relies on extensive and often case-specific simulations; see, for example, Bailey,[1] Welch and Bailey,[2] and Ho and Lau.[4,5] A more generic approach is to assume a specific service-time distribution that allows explicit expressions for the waiting-time and idle-time distributions, so as to analytically generate schedules. The easiest distribution to work with is the exponential distribution, as studied by Hassin and Mendel,[6] but this choice, corresponding with CV $= 1$, typically overestimates the variability. One has therefore looked into methods in which the service-time distribution is fitted by a distribution that provides more freedom but that still allows a (semi-)analytic solution. More specifically, a fit with the beta distribution was advocated by Lau and Lau,[7] whereas a phase-type distribution was proposed by, for example, Wang[8] and Kuiper *et al.*[9] In the latter reference, the validity of the phase-type approach, in which the first two moments of the patients' service-time distribution are fit (or, equivalently, the mean and the CV), has been thoroughly examined for typical service-time distributions observed in healthcare. We finally mention that one can rely on discrete-time

[a] *Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Amsterdam, The Netherlands*
[b] *Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands*
**Correspondence to: Alex Kuiper, Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands.*
† *E-mail: a.kuiper@uva.nl*

versions of the continuous schedules, thus facilitating a very fast evaluation of any specific schedule; see, for example, Brahimi and Worthington[10] and De Vuyst *et al.*[11]

A fundamentally different approach was followed by Zacharias and Pinedo.[12] In that setup, service times are deterministic and equal to the slot size, such that the fact the patients possibly not show up is the only stochastic element in the model. No-shows are indeed a prevalent problem in many healthcare scheduling practices as they correspond to typically 5% up to 30% of all patients, as reported by Cayirli and Veral.[3] Moreover, in an assessment by Ho and Lau[4] of environmental factors that affect appointment schedules, it was found that no-shows and the service-time variability have a profound impact on the appointment schedule, which motivates why a proper design should take both factors into account.

As the main contribution of the present paper, we present a technique to generate schedules that incorporate random service times and no-shows. In addition, we provide a comparison of such schedules with those resulting from straightforward heuristics (with scenarios in which the parameters match those observed in practice). Our framework relies on the phase-type approach as proposed by Kuiper *et al.*,[9] which we have adapted to incorporate no-shows. Another approach that covers both stochastic effects is by Vissers and Wijngaard.[13] Their main idea is to modify the mean and the variance of the patients' service times in their simulation studies to account for no-shows.

At the time appointment scheduling research took off, computational power was limited, and one therefore primarily focused on various heuristics. Perhaps the most classical example is the equidistant schedule in which the slot sizes equal the patients' average service time. However, as is known already from the pioneering work of Welch and Bailey,[2] such a scheme performs badly in many cases; to remedy this, they propose to overbook the first slot with an additional patient. It was shown by Ho and Lau[4] that this rule, often referred to as the Bailey–Welch rule, is fairly robust over a broad range of situations. It has been proven by Kemper *et al.*,[14] however, that equidistant schedules ultimately lead to long waiting times when the number of patients grows large (with the mean waiting of the nth patient roughly behaving as $\sqrt{n}$); hence, in that regime, the Bailey-Welch rule may lead to schedules that are highly unattractive to patients.

The remainder of the paper is organized as follows. In Section 2, we introduce the concept of a *risk function* that balances the interests of the service provider (doctor) and the patients, and then we extend the phase-type approach given in Kuiper *et al.*[9] The framework thus obtained enables us to evaluate an optimal schedule, that is, the schedule that minimizes the risk function. Then, in Section 3, we compute commonly used scheduling heuristics and numerically compare them with the optimal schedule. We conclude our paper with a discussion of the results in Section 4.

2. Modeling approach

In this section, we outline the stochastic model and method used for evaluation and optimization of appointment schedules. The main focus is on extending the framework given in Kuiper *et al.*[9] to a setup that incorporates patients' no-shows. This extension, as relevant as it is, requires to deal with some subtleties. First, we describe the risk function that represents the expected loss per patient in terms of mean idle times and mean waiting times. Then we describe the phase-type approach and point out how the recursive method should be adapted to deal with no-shows. We assume patients (also referred to as *clients*) and the specialist or doctor (also referred to as *practitioner*) to be punctual.

2.1. Framework, risk function

In mathematical terms, the appointment scheduling problem aims at determining epochs t_1 up to t_n at which the n clients are supposed to enter. We denote by $\mathcal{V} := (t_1, \ldots, t_n)$ the resulting appointment schedule. In the context of the present paper, the service times B_1 up to B_n are assumed independent and identically distributed (but this assumption can be alleviated). We write I_i for the practitioner's (random) idle time prior to the ith arrival, and W_i for the (random) waiting time of the ith client; it is clearly ruled out that both are positive.

The *risk* associated with client i, defined as weighted sum of the expected idle and the expected waiting time, is given by

$$R_i^{(\alpha)}(t_1, \ldots, t_i) = \alpha\, \mathbb{E}[I_i] + (1-\alpha)\, \mathbb{E}[W_i],$$

where the $\alpha \in (0, 1)$ is a weight factor that embodies the importance of the practitioner's (idle) time versus the clients' (waiting) time. Realize that (obviously) the random variables I_i and W_i are affected by the arrival epochs $t_1, \ldots, t_i$. The *aggregate risk* is given by

$$R^{(\alpha)}(t_1, \ldots, t_n) = \sum_{i=1}^{n} R_i^{(\alpha)} = \sum_{i=1}^{n} (\alpha\, \mathbb{E}[I_i] + (1-\alpha)\, \mathbb{E}[W_i]). \tag{1}$$

Because we consider *expected* idle and waiting times, we do not have to compute explicit idle and waiting-time distributions to evaluate Eqn (1). Instead, we rely on the definition of the *sojourn time* as the sum of waiting and service time, that is,

$$S_i = W_i + B_i \tag{2}$$

in combination with the fact that the *makespan* equals the sum of idle and service times:

$$t_i + S_i = \sum_{j=1}^{i} \left(I_j + B_j \right). \tag{3}$$

If we take the expected value in Eqns (2) and (3), we end up with a formula for the expected waiting time and a recursion for the expected idle time of the ith client in terms of her expected sojourn time:

$$\mathbb{E}[W_i] = \mathbb{E}[S_i] - \mathbb{E}[B]; \tag{4}$$

$$\mathbb{E}[I_i] = t_i + \mathbb{E}[S_i] - i\,\mathbb{E}[B] - \sum_{j=1}^{i-1} \mathbb{E}[I_j]. \tag{5}$$

We thus conclude that having knowledge of the sojourn-time distribution (and in particular its mean) enables a recursive algorithm to find the mean waiting times and the mean idle times, with which we can evaluate our objective function.

The next step is to introduce phase-type distributions, which are intended to approximate the service-time distribution B. As we will see, they are relatively easy to work with; more specifically, we can compute the corresponding sojourn-time distribution (and hence its mean). It is well known that phase-type distributions, which are mixtures and convolutions of exponential distributions, can be used to approximate any distribution with positive support arbitrarily accurately; see, for example, Asmussen *et al.*[15] and Tijms.[16]

2.2. *Phase-time distribution*

We approximate the service-time distribution B by a phase-type counterpart based on the mean and the SCV, in the way proposed by Tijms.[16] The candidate distributions that we rely on in this paper are the mixture of Erlang distributions $E_{K-1,K}(\mu;p)$ and the hyperexponential distribution $H_2(\boldsymbol{\mu};p)$. These phase-type distributions are characterized by an m-dimensional (row) vector $\boldsymbol{\alpha}$, where $m \in \mathbb{N}$, with nonnegative entries adding up to 1, and an $(m \times m)$-dimensional matrix $\boldsymbol{S} = (s_{ij})_{i,j=1}^{m}$ such that $s_{ii} < 0$, $s_{ij} \geq 0$ and $\sum_{j=1}^{m} s_{ij} \leq 0$ for any $i \in \{1, \ldots, m\}$. For the two specific phase-type distributions mentioned in the previous text, the representations in terms of m, $\boldsymbol{\alpha}$, and $\boldsymbol{S}$ are given as follows:

- In case SCV < 1, we use an $E_{K-1,K}(\mu;p)$ distribution. In this case, $m = K$, and the vector $\boldsymbol{\alpha}$ is such that $\alpha_1 = 1$ and $\alpha_i = 0$ for $i = 2, \ldots, K$. In addition, $s_{ii} = -\mu$ for $i = 1, \ldots, K$ and $s_{i,i+1} = -s_{ii} = \mu$ for $i = 1, \ldots, K-2$, while $s_{K-1,K} = (1-p)\mu$; all other entries of $\boldsymbol{S}$ are 0.
- In case SCV ≥ 1, we use a $H_2(\boldsymbol{\mu};p)$ distribution. Then $m = 2$, and $\alpha_1 = p = 1 - \alpha_2$. Also, $s_{ii} = -\mu_i$, for $i = 1, 2$, while the other two entries of $\boldsymbol{S}$ equal 0.
- If SCV $= 1$, then the exponential distribution, Exp(μ), is used.

Observe that the first case (SCV < 1) is particularly relevant in healthcare as it contains the typical CV values in the range of 0.35 to 0.85.

We write for a phase-type distributed random variable B that $B =_{\mathrm{d}} \mathbb{P}\mathrm{h}(\boldsymbol{\alpha}, \boldsymbol{S})$. An attractive feature of phase-type distributions is that the moments have explicit forms (see, e.g., Asmussen[17]); for the mean, we have ($\boldsymbol{e}_m$ being an m-dimensional column vector consisting of ones)

$$\mathbb{E}[B] = -\boldsymbol{\alpha}\boldsymbol{S}^{-1}\boldsymbol{e}_m, \tag{6}$$

which can be evaluated fast for the phase-type representations that are being considered, because $\boldsymbol{S}$ is an upper diagonal matrix in case of a mixture of Erlangs or a diagonal matrix for the hyperexponential distribution.

Now suppose there are *no-shows*, in the sense that each scheduled arrival corresponds to a no-show with probability $q \in (0, 1)$. Then the phase-type distribution should be adapted reflecting the fact that each client requires no service with probability q and a service time B with probability $(1-q)$. As a consequence, the vector $\boldsymbol{\alpha}$ is multiplied by $(1-q)$, that is, $B =_{\mathrm{d}} \mathbb{P}\mathrm{h}((1-q)\boldsymbol{\alpha}, \boldsymbol{S})$.

2.3. *Recursive approach*

The key idea is to use the recursive procedure proposed by Wang[8] to compute each client's sojourn-time distribution. These are of phase-type, and hence, the objective is to identify the $\boldsymbol{\alpha}$ and $\boldsymbol{S}$ featuring in its representation $\mathbb{P}\mathrm{h}(\boldsymbol{\alpha}, \boldsymbol{S})$. The basic idea is that at each moment in time, we keep track of the number of clients in the system together with the phase of the client in service; the current state of the system is given by these two variables. Notice that the ith client's sojourn time is only affected by her $i-1$ predecessors. Because typically the number of clients to be scheduled is relatively small, the dimensionality issue is not prevalent; our numerical techniques provide us with accurate output for problems of a realistic size. When considering very large numbers of clients, one can opt for neglecting some of the dependence between the clients by introducing a *lag order*, as proposed in Vink *et al.*[18]; if the lag order is k, then this means that only clients $i-k$ up to $i-1$ can affect the sojourn time of the ith client.

To outline the procedure under no-shows (with probability q), define the bivariate process $\{N_i(t), K_i(t), t \geq 0\}$ for client $i = 1, \ldots, n$, where $N_i(t) \in \{0, \ldots, i-1\}$ represents the number of clients in front of the ith arriving client, t time units after his or her arrival. The second component, $K_i(t) \in \{1, \ldots, m\}$, represents the actual phase of the client in service at t time units after the arrival. We introduce the corresponding probabilities, for $t \geq 0$, $i = 1, \ldots, n$, $j = 0, \ldots, i-1$, and $k = 1, \ldots, m$:

$$p_{j,k}^{(i)}(t) = \mathbb{P}\left(N_i(t) = j, K_i(t) = k\right).$$

In addition, the vector $\boldsymbol{P}_i(t)$ (of dimension mi) plays a crucial role; it is given by

$$\left(p^{(i)}_{i-1,1}(t),\ldots,p^{(i)}_{i-1,m}(t),p^{(i)}_{i-2,1}(t),\ldots,p^{(i)}_{i-2,m}(t),\ldots,p^{(i)}_{0,1}(t),\ldots,p^{(i)}_{0,m}(t)\right).$$

The sojourn-time distribution of the ith client can be computed from $\boldsymbol{P}_i(t)$ through the following identity, with $\boldsymbol{e}_{mi}$ as an all-ones vector of dimension mi:

$$F_i(t) := \mathbb{P}(S_i \le t) = 1 - \sum_{j=0}^{i-1}\sum_{k=1}^{m} p^{(i)}_{j,k}(t) = 1 - \boldsymbol{P}_i(t)\boldsymbol{e}_{mi}.$$

Considering the first client, which is evidently to arrive at $t_1 = 0$, it is only his or her service-time distribution that determines his or her sojourn time: for $t \ge 0$,

$$\boldsymbol{P}_1(t) = (1-q)\boldsymbol{\alpha}\exp(\boldsymbol{S}t)$$

(which is an m-dimensional object). Considering the second client, arriving $x_1 := t_2 - t_1$ time units after the first client, he or she either shows up with probability $(1-q)$ (thus increasing the number of clients by one), or he or she does not show up with probability q. For any $t \ge 0$, with $\boldsymbol{0}_m$ denoting an all-zeros vector of dimension m, this leads to

$$\boldsymbol{P}_2(t) = ((1-q)(\boldsymbol{P}_1(x_1), \boldsymbol{\alpha}F_1(x_1)) + q(\boldsymbol{0}_m, \boldsymbol{P}_1(x_1)))\exp(\boldsymbol{S}_2 t),$$

which is an object of dimension $2m$; here, with $\boldsymbol{s} := -\boldsymbol{S}\boldsymbol{e}_m$ and $\boldsymbol{0}_{m,m}$ denoting an $(m \times m)$-dimensional all-zeros matrix,

$$\boldsymbol{S}_2 := \begin{pmatrix} \boldsymbol{S} & \boldsymbol{s}\boldsymbol{\alpha} \\ \boldsymbol{0}_{m,m} & \boldsymbol{S} \end{pmatrix}.$$

For the other clients, the vector $\boldsymbol{P}_i(t)$ (dimension mi) can be found from $\boldsymbol{P}_{i-1}(t)$ (dimension $m(i-1)$) by the recursion, for $t \ge 0$,

$$\boldsymbol{P}_i(t) = ((1-q)(\boldsymbol{P}_{i-1}(x_{i-1}), \boldsymbol{\alpha}F_{i-1}(x_{i-1})) + q(\boldsymbol{0}_m, \boldsymbol{P}_{i-1}(x_{i-1})))\exp(\boldsymbol{S}_i t),$$

where $x_{i-1} := t_i - t_{i-1}$ (which is commonly known as the *inter-arrival time*) and the matrix $\boldsymbol{S}_i$ is defined recursively by

$$\boldsymbol{S}_i := \begin{pmatrix} \boldsymbol{S}_{i-1} & \boldsymbol{T}_i \\ \boldsymbol{0}_{m,(i-1)m} & \boldsymbol{S} \end{pmatrix},$$

with $\boldsymbol{T}_i$ a matrix of dimension $(i-1)m \times m$ defined by

$$\boldsymbol{T}_i := (\boldsymbol{0}_{m,m}, \boldsymbol{0}_{m,m}, \ldots, \boldsymbol{0}_{m,m}, \boldsymbol{s}\boldsymbol{\alpha})^{\mathrm{T}}.$$

In the previous text, we have outlined the procedure for evaluating the aggregate risk of *any* schedule $\mathcal{V}$. Using this recursive procedure, we can use standard numerical tools to *optimize* over all possible schedules, so as to find *optimal* schedule. In other words, we identify the $x_i^\star$s that minimize the risk function (for a given weight α), thus finding the optimal schedule $\mathcal{V}^\star = (t_1^\star, \ldots, t_n^\star)$ by $t_i^\star = \sum_{j=1}^{i-1} x_j^\star$ for $i = 2, \ldots, n$. We will use this procedure in Section 3 to evaluate commonly used scheduling heuristics and compare those with the optimal schedule $\mathcal{V}^\star$.

3. Experiments and results

The primary objective of this section is to examine how frequently used scheduling heuristics perform relative to each other, and relative to optimal schedules (i.e., schedules that minimize $R^{(\alpha)}(t_1, \ldots, t_n)$ for some $\alpha \in [0,1]$). We do so by evaluating the so-called *efficient frontier*, consisting of all combinations of the averaged (over all clients) mean waiting times and aggregated mean idle times when varying the weight α. In our computations, we rely on the phase-type approach, augmented to incorporate no-shows, as has been described in Section 2; it was validated in Kuiper *et al.*[9] that the underlying phase-type service-time distributions very well match the Weibull and log-normal distributions that have been observed in practice (as was pointed out by Cayirli and Veral).[3]

We consider five heuristics, each of them based on the average service time $\mathbb{E}[B]$; we refer to, for example, Ho and Lau[4] for a simulation study (using the uniform and exponential distribution) on the performance of various scheduling heuristics of this and related types. It is remarked that in these heuristics, the no-show probability q is not taken into account, despite the fact that the no-shows (obviously) reduce the expected service time needed per client. Later, we also consider the counterparts of these schedules in which it is corrected for no-shows; they are obtained by replacing $\mathbb{E}[B]$ by $(1-q)\,\mathbb{E}[B]$.

$\mathcal{A}$ An equidistant schedule:
$t_i = (i-1)\,\mathbb{E}[B]$ for all i.

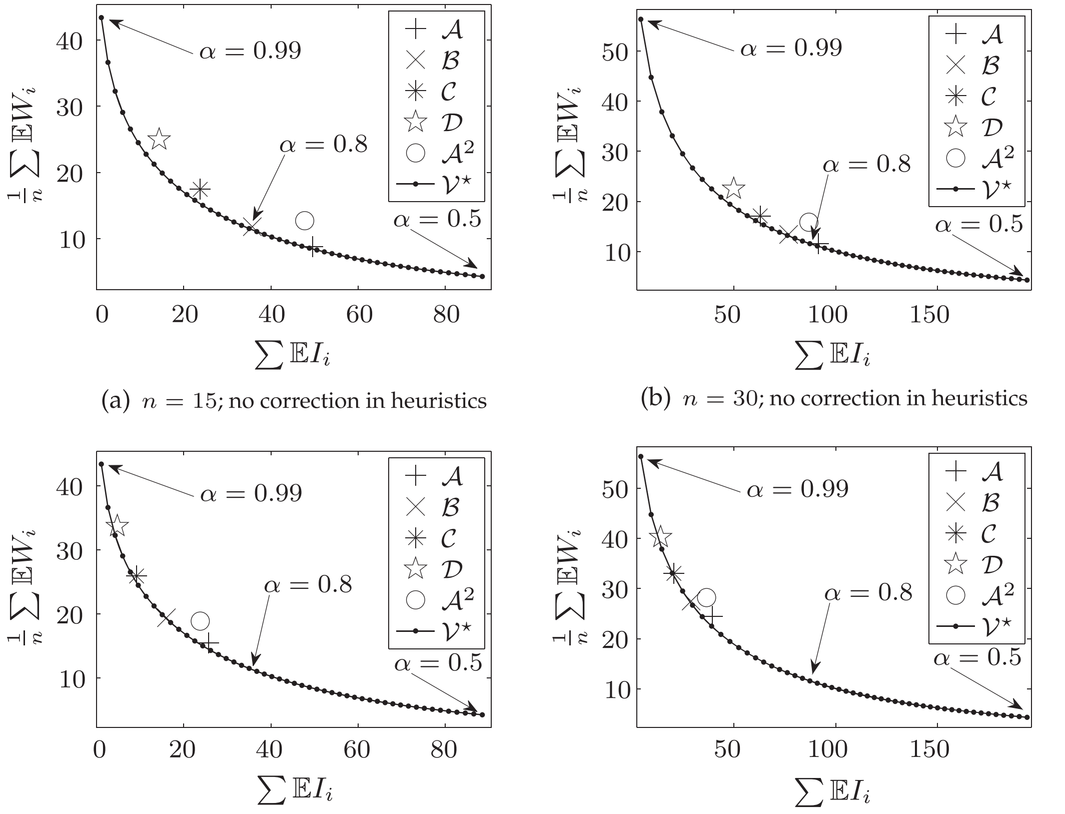
- $\mathcal{B}$ The Bailey–Welch rule with two clients at the first time slot:
  $t_1 = t_2 = 0; t_i = (i-2)\,\mathbb{E}[B]$ for $i > 2$.
- $\mathcal{C}$ Adaptation of the Bailey–Welch rule with three clients at the first time slot:
  $t_1 = t_2 = t_3 = 0; t_i = (i-3)\,\mathbb{E}[B]$ for $i > 3$.
- $\mathcal{D}$ Adaptation of the Bailey–Welch rule with four clients at the first time slot:
  $t_1 = t_2 = t_3 = t_4 = 0; t_i = (i-4)\,\mathbb{E}[B]$ for $i > 4$.
- $\mathcal{A}^2$ Block appointment rule of two clients arriving for a double slot:
  $t_i = t_{i+1} = 2(i-1)\,\mathbb{E}[B]$ for $i = 1, 3, 5, \ldots$.

It was shown by Ho and Lau[4] that the original Bailey–Welch rule (rule $\mathcal{B}$, that is) performs reasonably well in that it generates a value of the risk function that is close to optimal; this is, however, just for a specific choice of α. Rule $\mathcal{A}$ was not studied in their paper, but is, owing to its simplicity, of particular interest to a practitioner. Rules $\mathcal{C}$ and $\mathcal{D}$ are adaptations of the original Bailey–Welch rule, intended to further reduce the idle time (obviously at the expense of additional waiting time). Rule $\mathcal{A}^2$, also known as the *two-at-a-time scheduling rule*, is a naïve solution to reduce the practitioner's idle time; the advantages of rule $\mathcal{A}^2$ are studied in the paper by Soriano[19] and compared with the equidistant schedule as described in $\mathcal{A}$.

In our numerical experiments, we consider the following three scenarios. (i) First we assess the scheduling heuristics based on the average service time $\mathbb{E}[B] = 15$ (min) and 15 clients to be scheduled. (ii) Then, so as to assess the effect of having more clients in the system, we double the number of clients to 30. (iii) Finally, we study schedules in which the *no-show correction* is applied; that is, $\mathbb{E}[B]$ is replaced by $(1-q)\mathbb{E}[B]$.

The experiments are carried out in `MATLAB R2014b`. The optimization procedure consists of four stages, whereas for evaluation of practical scheduling heuristics, only the first three steps are applicable.

1. The n clients' service-time distributions are approximated by phase-type counterparts based on its first two moments, as described in Section 2.2.
2. For each schedule $\mathcal{V}$, formed by $n-1$ inter-arrival times, each client's sojourn-time distribution is computed by the recursive approach outlined in Section 2.3.
3. Using the clients' sojourn-time distributions in Eqn (6), the expected idle and waiting times are computed by the formulas in Eqns (5) and (4).



(a) $n = 15$; no correction in heuristics

(b) $n = 30$; no correction in heuristics

(c) $n = 15$; no-show correction in heuristics

(d) $n = 30$; no-show correction in heuristics

Figure 1. Various appointment schedules in the setting of SCV $= 0.4225$ and no-show probability $q = 17.5\%$, where the number of clients n is varied horizontally, and where there is a correction in the heuristics for the no-show probability q in the bottom graphs

4. By using `MATLAB R2014b`'s built-in optimization routines, for a specific choice of α, the aggregate risk of Eqn (1) is minimized over all possible schedules $\mathcal{V}$ using a start vector in which each entry equals the expected service time with a no-show correction. The resulting optimized appointment schedule is $\mathcal{V}^\star$.

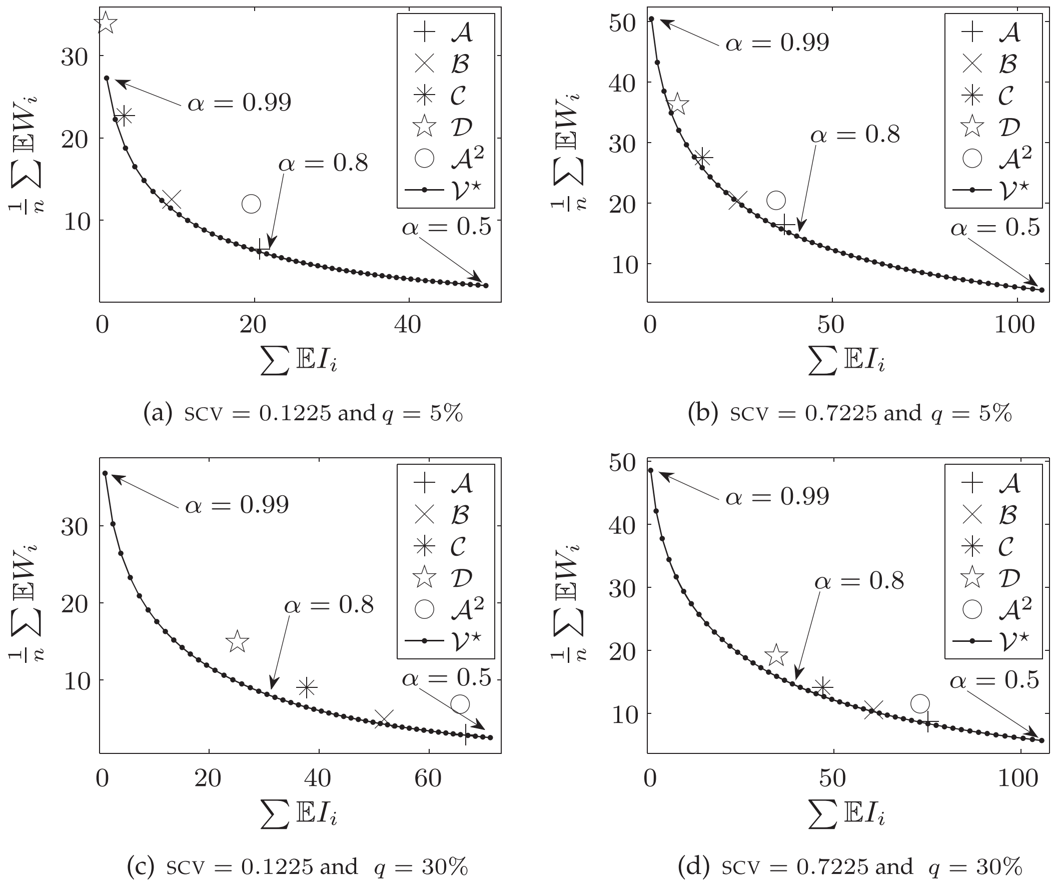
The five scheduling rules ($\mathcal{A}, \dots, \mathcal{A}^2$) are evaluated for specific scenarios (i.e., with particular values of SCV, q, and n), in terms of the expected idle and waiting times they correspond with. In addition, we compute the optimal schedules $\mathcal{V}^\star$ for different αs in $(0, 1)$, so as to minimize Eqn (1), letting α run from 0.5 to 0.99 in steps of 0.01. Informally, for $\alpha = 0.5$, the value assigned to the time of an individual client equals that assigned to the practitioner's time (which is not very common in the medical context); $\alpha = 0.99$ entails that the time of the practitioner is valued $99 = 0.99/0.01$ times more than the time of an individual client. Choosing $\alpha = 1$ corresponds to the trivial schedule in which all clients arrive at time zero, such that the risk function has the value (in the case the schedules are not corrected for no-shows)

$$\frac{1}{n}\sum_{i=1}^{n-1} i\mathbb{E}[B] = \frac{(n-1)}{2}\mathbb{E}[B],$$

as the (expected) idle times are zero; an explicit analysis is also possible in case the no-show probability is incorporated. Computing the optimal schedules $\mathcal{V}^\star$ for each α results in what we call *efficient frontier*; owing to the fact that these are obtained by optimizing the risk function, no schedule can perform better than these schedules.

First, we study the impact of the number of clients on the schedules. To this end, we choose SCV $= 0.4225$ and $q = 0.175$ and set the number of clients first to 15 and then to 30. In Figure 1, we plot the heuristics and the efficient frontier based on the optimized schedules $\mathcal{V}^\star$ for a range a values of α. It is first observed that implementing the no-shows correction in the scheduling rules has a substantial effect. Furthermore, comparing Figure 1(a) with Figure 1(b) shows for the situation that no no-shows correction has been applied that the heuristics converge to each other as the number of customers grows; for large n, the scheduling rules are very similar as they correspond to equal slot sizes in a stationary queue with load smaller than 1 (apart from the beginning of the session). In Figure 1(c) and (d), the heuristics will always differ as the system considered behaves essentially as a queue with load 1, so that there is no convergence to steady state.

In a second series of experiments, we let the number of clients be $n = 15$ and consider in both scenarios (i.e., with or without no-shows correction) *four settings* that match the boundaries settings of typical healthcare situations that have been reported by Cayirli



(a) SCV $= 0.1225$ and $q = 5\%$

(b) SCV $= 0.7225$ and $q = 5\%$

(c) SCV $= 0.1225$ and $q = 30\%$

(d) SCV $= 0.7225$ and $q = 30\%$

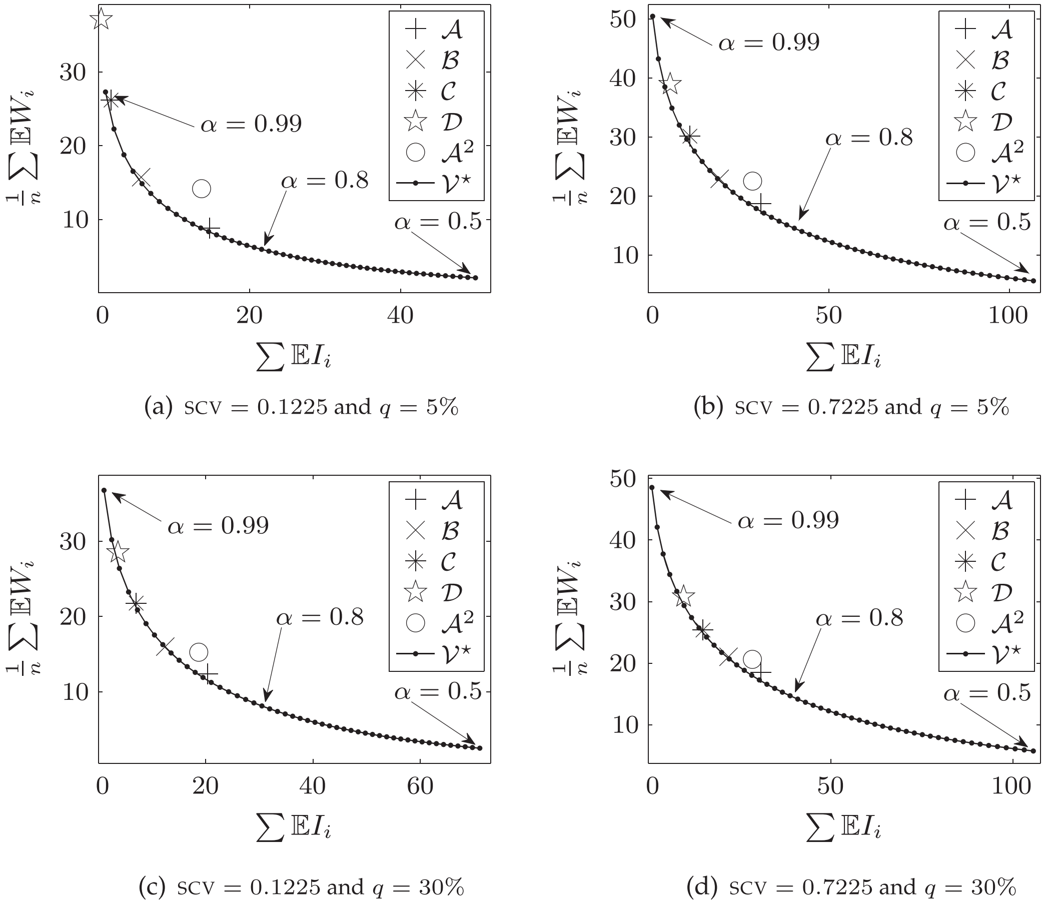
Figure 2. Various appointment scheduling heuristics based on the average service time $\mathbb{E}[B]$, where the number of clients n equals 15, in four healthcare settings: SCV increases from left to right and no-show probability q increases from top to bottom

and Veral[3]: we take (i) the CV (SCV) equal to 0.35 (0.1225) and 0.85 (0.7225), and (ii) we let the no-show probability q have the values 0.05 and 0.30. We do not apply the no-show correction in the heuristics.

Comparing the scheduling rules without a correction, Figure 2, with the scheduling rules that have been corrected for no-shows, Figure 3, it is observed that the corrected scheduling rules are more defensive in that they lead to lower expected idle times (and hence higher expected waiting times). Furthermore, we remark that the corrected scheduling rules cover only a small range of various trade-offs in terms of α. Zooming in on Figure 2(b) and (c) (equivalently, for the corrected versions, Figure 3(b) and (c)), one finds that from the two environmental factors an increase in the service-time variability has a greater impact on the schedule than an increase in the no-show probability.

It is remarkable that many heuristics lie close to the efficient frontier, indicating that the risk function Eqn (1) is relatively flat. This does not apply to $\mathcal{A}^2$, which schedules clients evenly over the session but now with two clients per double slot. At first sight, it seems to improve upon rule $\mathcal{A}$ in the sense that it reduces the practitioners idle time, but it has the drawback that if both clients show up one of the clients has an expected waiting time that is at least equal to the average service time.

Finally, all figures show that the *ordering* of the scheduling rules (except for rule $\mathcal{A}^2$) is preserved in all settings. Comparing the heuristics in terms of idle time, the expected idle time is reduced by planning additional clients at the beginning of the session, and therefore, the ordering from low to high in expected idle times is $\mathcal{D}, \mathcal{C}, \mathcal{B}, \mathcal{A}$. Analogously, planning additional clients at the beginning results in an increase in the waiting times. So the ordering from low to high in expected waiting times is $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$. For rule $\mathcal{A}^2$, we know that it has less idle time than $\mathcal{A}$, because even numbered clients arrive one time slot earlier than in rule $\mathcal{A}$. On the other hand, rule $\mathcal{B}$ has lower expected idle times than $\mathcal{A}^2$, because all arrival epochs are set equal to or earlier than the epochs set by $\mathcal{A}^2$. For the waiting times, such a comparison cannot be made, because under the no-shows rule, $\mathcal{A}^2$ does not necessarily lead to lower waiting times than rule $\mathcal{B}$ (or $\mathcal{C}$ and $\mathcal{D}$).



(a) SCV $= 0.1225$ and $q = 5\%$

(b) SCV $= 0.7225$ and $q = 5\%$

(c) SCV $= 0.1225$ and $q = 30\%$

(d) SCV $= 0.7225$ and $q = 30\%$

Figure 3. Various appointment scheduling heuristics based on the *no-show corrected* average service time $(1-q)\mathbb{E}[B]$, where the number of clients n equals 15, in four healthcare settings: SCV increases from left to right and no-show probability q increases from top to bottom

4. Conclusion and discussion

In this paper, we have considered appointment schedules in a setting with service-time variability. We have pointed out how to adapt the approach proposed by Kuiper *et al.*[9] so as to incorporate the possibility of clients not showing up. Despite the fact that service-time variability and no-shows are highly relevant to healthcare we are among the first to systematically assess both effects in a computational study.

Phase-type distributions are used to approximate the clients' service-time distributions based on the first two moments. It is known from literature that the first two moments mainly dictate the performance of an appointment schedule as reported by Cayirli and Veral.[3] Furthermore, they point out that typical distribution functions in healthcare are either Weibull or log-normal. In Kuiper *et al.*,[9] the choice of using phase-type distributions is justified by a series of numerical experiments. These experiments assess the performance of the phase-type approach with the simulated optimal appointment schedules.

In our numerical study, we have compared a number of heuristic schedules by evaluating for each of them the sum of the expected idle times as well as the average of the expected waiting times. From the two stochastic components, we find that the service-time variability (expressed in terms of the SCV) has a more significant impact than the no-show probability. Secondly, the optimization procedure enables us to find the optimal schedule for a specific choice of the weight factor α that sets the trade-off between idle and waiting times. Therefore, we are able to relate practical principles in appointment scheduling to their intrinsic trade-offs in terms of idle and waiting times. Finally, the optimization procedure itself can be used to generate optimal appointment schedules under service-time variability and in the presence of no-shows.

There are several directions for further research. In the first place, one could consider alternative risk functions. In the present paper, we have considered the sum of the expected idle and waiting times, but in principle, there is no clear reason for this choice (besides perhaps this form having become the standard risk function, or the fact that it may be easier to evaluate than other functional forms). Given the fact that, for both the practitioner and clients, a modest amount of slack time is hardly negatively perceived, one could argue that a quadratic loss function may be more appropriate than a linear one. The choice of the risk function, however, may have a very substantial impact on the optimal schedule, as can be found in Kuiper *et al.*[9]

Another relevant feature concerns the choice of an appropriate weight factor, embodied by the parameter α in the context of our paper. This parameter essentially represents the value of the practitioner's time relative to clients' time. There is obviously no clear recipe to choose the 'right' α. It is increasingly felt that clients should not suffer from inefficiencies, but there is also a strong societal pressure to use the practitioner's time well so as to prevent excessive healthcare costs. For example, Robinson and Chen[20] present attempts to find the 'implied value' of the practitioner's time with respect to the clients' time.

References

1. Bailey NTJ. 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)* 1952; **14**(2):185–199.
2. Welch JD, Bailey NTJ. Appointment systems in hospital outpatient departments. *The Lancet* 1952; **259**(6718):1105–1108.
3. Cayirli T, Veral E. Outpatient scheduling in healthcare: a review of literature. *Production and Operations Management* 2003; **12**(4):519–549.
4. Ho CJ, Lau HS. Minimizing total cost in scheduling outpatient appointments. *Management Science* 1992; **38**(12):1750–1764.
5. Ho CJ, Lau HS. Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *European Journal of Operational Research* 1999; **112**(3):542–553.
6. Hassin R, Mendel S. Scheduling arrivals to queues: a single-server model with no-shows. *Management Science* 2008; **54**(3):565–572.
7. Lau H, Lau A. A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. *IIE Transactions* 2007; **32**(9):833–839.
8. Wang PP. Optimally scheduling *n* customer arrival times for a single-server system. *Computers & Operations Research* 1997; **24**(8):703–716.
9. Kuiper A, Kemper B, Mandjes M. A computational approach to optimized appointment scheduling. *Queueing Systems* 2015; **79**(1):5–36.
10. Brahimi M, Worthington DJ. Queueing models for out-patient appointment systems – a case study. *The Journal of the Operational Research Society* 1991; **42**(9):733–746.
11. De Vuyst S, Bruneel H, Fiems D. Fast evaluation of appointment schedules for outpatients in healthcare. *Proc. ASMTA*, Venice, Italy, 2011,113–131.
12. Zacharias C, Pinedo M. Appointment scheduling with no-shows and overbooking. *Production and Operations Management* 2014; **23**(5):788–801.
13. Vissers J, Wijngaard J. The outpatient appointment system: design of a simulation study. *European Journal of Operational Research* 1979; **3**(6): 459–463.
14. Kemper B, Klaassen CAJ, Mandjes M. Optimized appointment scheduling. *European Journal of Operational Research* 2014; **239**(1):243–255.
15. Asmussen S, Nerman O, Olssen M. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics* 1996; **23**(4):419–441.
16. Tijms H. *Stochastic Modelling and Analysis - A Computational Approach*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons: Chichester, UK, 1986.
17. Asmussen S. *Applied Probability and Queues, Stochastic Modelling and Applied Probability* (2nd edn), Vol. 51. Springer-Verlag: New York, NY, USA, 2003.
18. Vink W, Kuiper A, Kemper B, Bhulai S. Optimal appointment scheduling in continuous time: the lag order approximation method. *European Journal of Operational Research* 2015; **240**(1):213–219.
19. Soriano A. Comparison of two scheduling systems. *Operations Research* 1966; **14**(3):388–397.
20. Robinson LW, Chen RR. Estimating the implied value of the customer's waiting time. *Manufacturing & Service Operations Management* 2011; **13**(1): 53–57.

Authors' biographies

Alex Kuiper obtained his Master's degrees in mathematics and econometrics at the University of Amsterdam in 2013. Currently, he works for the Institute for Business and Industrial Statistics as a Lean Six Sigma consultant and is a PhD student at the University of Amsterdam. His current research focuses on appointment scheduling in healthcare.

Michel Mandjes obtained a PhD in Operations Research and Applied Probability from Vrije Universiteit, Amsterdam. After having worked as a Member of Technical Staff at KPN Research (Leidschendam, the Netherlands), and Lucent Technologies/Bell Laboratories (Murray Hill, NJ, United States), and several academic positions, he is now a full professor in Applied Probability at the University of Amsterdam. His research focuses on queueing theory and stochastic operations research, predominantly applied in the design of communication networks, but also in finance/risk, as well as in the production and service systems. He is the author of the published book ŚLarge Deviations for Gaussian Queues' (Wiley Online Library, 2007).