Optimal outpatient appointment scheduling with emergency arrivals and general service times

Paulien M. Koeleman, Ger M. Koole

Abstract

In this paper we study the problem of deciding at what times to schedule patients when there are emergency arrivals following a non-stationary Poisson process. The service times can be any given distribution. The objective function consists of a weighted sum of the waiting times, idle time and tardiness. We prove that this objective function is multimodular, and then use a local search algorithm which in that case is guaranteed to find the optimal solution. Numerical examples show that this method gives considerable improvements over the standard even-spaced schedule, and that the schedules for different service time distributions can look quite different.

keywords Appointment scheduling, local search, health care, multimodularity, emergencies.

1 Introduction

Outpatient appointment scheduling is a subject of great interest to hospitals and other medical institutions. Most doctors and diagnostic facilities use appointments, but also dentists and general practitioners do. Outside of the medical world the problem occurs as well, for example in the scheduling of loading and unloading ships. No wonder that the problem has been an object of study for a long time, starting with the work of Bailey and Welch in the early fifties [19]. From that time on many papers have been written studying this appointment scheduling in many settings, and with many different assumptions and methods.

In the problem of appointment scheduling, the goal is to balance the interests of the patients with those of the doctors. The patients want long intervals between appointments, as this minimises their waiting time. The doctors on the other hand wish to have as little idle time and overtime as possible, and therefore they like shorter intervals better. If there are emergency arrivals, which have to be seen as soon as possible, this complicates the situation further. How much idle time should be scheduled, and when?

The research on appointment scheduling has a long history, starting with the work of Welch and Bailey. Their most famous result is the so-called Bailey-Welch appointment schedule, which states that two patients should be planned at the start of the day, and the other patients evenly spaced throughout the day, to offset the bad effects of no-shows and patient lateness.

A large part of the literature concerns simulation models for evaluating the performance of appointment schedules. Examples are Fetter and Thompson [5] and Vissers and Wijngaard [16]. They study the influence of several factors on system performance. Ho and Lau [8] compare different rules for making appointments in different settings using simulation, and conclude that no single rule works best in all situations, though the Bailey-Welch rule works fine in many cases.

Other works consider finding the optimal schedule. Stein and Côté [15] use an analytical method for finding the optimal schedule in a case with exponentially distributed service times and the restriction that the resulting schedule be even-spaced. Wang [18] finds an optimal schedule without this restriction and with a Coxian distribution for the service times, and shows that this is an improvement over an even-spaced schedule. He gives the optimal schedule in terms of interarrival times in continuous form. For larger numbers of customers he gives an approximation for the optimal interarrival times.

One of the few practical implementations is the one by Rising et al. [14]. They try to smooth the number of appointments over the week and over the day to complement the number of walk-in patients and emergencies, and find a good schedule by trial and error in a simulation model. The implementation of their schedule gave good results in terms of waiting times and tardiness of clinics.

An overview of the important issues to consider when designing an appointment system can be found in Gupta and Denton [7]. For a thorough review of the literature on appointment scheduling we refer to Cayirli and Veral [3]. They present the research done in the second half of the last century and mention some directions for future work.

In this paper we present a method to find the optimal appointment schedule in a situation with emergency arrivals and general service times. It has been found by O'Keefe [13] that the coefficient of variation of the service times in practice is considerably smaller than 1, as in the exponential distribution used in many studies. He has found it to be more in the order of 0.5. The exact form of the distribution can differ, but Ho and Lau [8] find that only the coefficient of variation has a significant influence on the performance of the appointment schedule. According to Denton and Gupta [4] higher moments of the service time distribution are only important in the case were the costs of patient waiting are high relative to cost of server idle time. Emergency arrivals or other disturbances such as phone calls are also known to have a considerable influence

on the waiting times of patients as also discussed by O'Keefe [13], which can also be seen from our experiments presented below.

The method used by us is a generalisation of the local search method used by Kaandorp and Koole [10], who study the case with only scheduled arrivals and exponential service times. Because we use the amount of work present in the system as a state description instead of the number of patients, the service times can have any positive distribution, and can differ for the scheduled and the emergency patients. Related to [10] is the work of Vanden Bosch et al. [2], who use a different method that is much quicker, for Erlang distributed service times. Unfortunately there is a small problem in one of the proofs, which means their method is less fast than they claim (see section 3.2). This means that their method is fast but does not find the optimal solution. Neither of these papers includes emergencies. Begen and Queyranne [1] do include optimisation with emergencies, but only if they arrive during the service of scheduled patients. This can be a real restriction if the service times of emergency patients is longer than that of scheduled patients. We do not use this assumption.

In the somewhat related area of OR scheduling more work has been done on accounting for emergency arrivals. Examples of this are Gerchak et al. [6] and Lamiri et al. [12]. However, this is different because here the assumption is that the emergencies should be done on the same day instead of as soon as possible, as in the case of appointment scheduling.

The remainder of this paper is organised as follows: in section 2 we describe the model and give a method for evaluating the performance of a given schedule. In the next section we give a method for finding the optimal schedule. Some numerical results will be presented in section 4, and we end with some conclusions and suggestions for further work in section 5.

2 Model description

To model the problem we divide the day (or part of a day) that the doctor is seeing patients into T intervals of length d. In this time window we want to schedule N patients, and assume a certain number of emergency patients will arrive. The emergency patients are assumed to arrive according to a Poisson process with rate λ per interval. This can easily be generalized to interval-dependent arrival rates, but to avoid further complication of the notation we decided to present the results for homogeneous emergency arrivals.

Emergency patients are served as soon as possible, meaning that they wait only for the current patient in service to be finished. If more than one emergency patient is present, they are served in order of arrival. All scheduled patients wait for emergency patients arriving during their waiting

time, and they again are served in order of arrival.

Scheduled patients have a service time that has a known distribution with mean β_s, and emergency patients have a service time that is distributed according to a second known distribution with mean β_e. Each scheduled patient is assumed to have a probability q of not showing up for his appointment.

The number of patients scheduled at the beginning of interval t is denoted by $x_t \in \{0, \ldots, N\}$, $t = 1, \ldots, T$. A complete schedule then is described by a vector $x = (x_1, \ldots, x_T)$ with $\sum_{t=1}^{T} x_t = N$.

In this model we assume all scheduled patients who actually show up for their appointment to arrive exactly on time.

Based on the schedule and the parameter values we calculate the expected waiting time $W(x)$, the expected idle time $I(x)$ and the expected lateness $L(x)$. Then the cost function for the schedule becomes $C(x) = \alpha W(x) + \beta I(x) + \gamma L(x)$, for any $\alpha, \beta, \gamma \geq 0$. The weights can be used to give relative importance to the three objectives. We are looking to minimise this cost function, so the problem then becomes

$$\min\{C(x) | \sum_t x_t = N, x_t \in \mathbf{N}_0\}.$$

To calculate the results for a given schedule, we use the probabilities that there is a certain number of minutes of work in the system at the moment just before or just after an arrival time. These are:

$$p_{t^-}(i) = \mathbf{P}(i \text{ minutes of work in the system just before any arrivals on time } t),$$

$$p_{t^+}(i) = \mathbf{P}(i \text{ minutes of work in the system just after any arrivals on time } t).$$

The probabilities can be calculated as follows: let

$$v_k(i) = \mathbf{P}(\text{number of arriving minutes of work including emergency work is } i \mid$$
$$k \text{ patients scheduled to arrive}).$$

Then

$$p_{1^-}(0) = 1,$$

because we assume the system starts empty.

$$p_{1^+}(i) = v_{x(0)}(i),$$

$$p_{t^-}(0) = \sum_{k=0}^{d} p_{t-1^+}(k), \quad t = 2, \ldots, T+1,$$

$$p_{t^-}(i) = p_{t-1^+}(i+d), \quad t = 2, \ldots, T+1,$$
$$p_{t^+}(i) = \sum_{k=0}^{\infty} \sum_{j=0}^{k} p_{t^-}(j) v_{x(t)}(k-j), \quad t = 2, \ldots, T.$$

To compute $v_k(i)$ we need to compute the expected number of minutes of arriving work coming from emergency patients at one interval, and the same for the k scheduled patients arriving at one interval, and then take the convolution of these two to get the distribution of the total amount of arriving work.

Let us consider first the amount of work related to emergency patients. Because emergency patients are assumed to arrive according to a Poisson process, but are modelled to only arrive at the start of intervals, the number arriving at the start of an interval has a Poisson distribution with expectation λ. The assumption is that if we divide a day in enough small intervals, the difference between this method and arrivals at any moment will be negligible. Let the number of arriving emergency patients be Y. Then the amount of work arriving is the Y-fold convolution of the vector representing the service time for emergency patients, $s_e^{(Y)}$. In this vector the jth element s_{ei} denotes the probability that the service time of an emergency patient is j minutes. Then the distribution of the amount of emergency work arriving at the start of any interval is given by:

$$v_0(i) = \sum_{y=1}^{\infty} s_e i^{(y)} \mathbf{P}(Y = y) = \left(\sum_{y=1}^{\infty} \mathbf{P}(Y = y) s_e^{(y)} \right)_i.$$

The amount of work of a scheduled patient is 0 with probability q, the probability of a no-show, and otherwise his service time distribution is represented by the vector s_s. This goes for every patient arriving at any interval independently. So the total amount of work arriving at a given interval with k patients scheduled to arrive becomes:

$$v_k(i) = \left(\left(\sum_{y=1}^{\infty} \mathbf{P}(Y = y) s_e^{(y)} \right) * ((1-q)s_s + q e_0)^{(k)} \right)_i.$$

2.1 Tardiness and idle time

The expected tardiness is the expected amount of time the doctor has to work later than time $T+1$, or the scheduled end of the day. This is the same as the expectation of the number of minutes of work in the system at time $T+1$:

$$L(x) = \sum_{k=1}^{\infty} k p_{T+1^-}(k).$$

To calculate the expected idle time, we use the expected tardiness. The total time the doctor is working is this tardiness plus the scheduled duration of the day, which is $T * d$. From this we

substract the expected time the doctor has to work, and we get:

$$I(x) = Td + \sum_{k=1}^{\infty} k p_{T^-}(k) - N\beta_s - \lambda T \beta_e.$$

2.2 Waiting time

The waiting time of a patient depends on the number of minutes of work in the system at the time of his arrival, any patients arriving simultaneously with him, and any emergency patients arriving before the start of his service. This makes it harder to compute the waiting time of a patient, as it depends not only on the amount of work present on his arrival but also on the interval in which he arrives.

The first patient to arrive at any given interval waits for at least all the work already present and the emergency work arriving simultaneously with him. This is given by:

$$\mathbf{P}(w_1 = k) = \sum_{j=0}^{k} p_t^-(j) v_0(k-j).$$

For the ith patient to arrive at any interval the waiting time for the service the $i-1$ patients before him has to be added:

$$\mathbf{P}(w_i = k) = \sum_{j=0}^{k} \mathbf{P}(w_{i-1} = j)\mathbf{P}(s_s = k-j).$$

Now that we know $\mathbf{P}(w_i = k)$ we can compute the distribution of the complete waiting time of patient i arriving at interval t with the following procedure. Let $wt_i(k)$ denote the probability that the actual waiting time of patient i arriving at interval t is k minutes. Then we can derive the values of $wt_k(k)$ as follows:

1. $time = t$.
2. for $k = 0 \ldots d-1$ $wt_k(k + (time - t)d) = w_i(k)$.
3. $w_i^*(k) = \sum_{j=0}^{k} w_i(j+d)v_0(k-j)$ for $k = d, d+1, \ldots$.
4. $wt_i(k) = w_i^*(k)\ \forall k \geq (time - t + 1)d - 1$.
5. $time = time + 1$; if $time = T + 1$ then stop, else go back to step 2.

Now that we can evaluate a given schedule, we can use this to search for the optimal schedule, defined as the schedule with the lowest objective value, in an efficient way. To do this, we use certain properties of the objective function, as explained in the next section.

It should be noted that looking at idle time and tardiness at the same time does not make sense in this case, because they are strongly related. In most practical situations one of the two should be chosen as a performance measure, according to the objective in the situation in question.

3 Solution method

Even for relatively small numbers of patients to schedule and intervals to schedule them in, the number of possible schedules becomes too large to make enumeration possible. The number of possible schedules is $\binom{N+T-1}{N}$. This means that another method of finding the optimal schedule has to be found.

The method we use here is local search, which has been used before by Kaandorp en Koole [10] in a setting with exponential service times and without emergencies. The local search method starts with some feasible solution, and improves this step-by-step by finding the best solution in its neighbourhood. This is repeated until a local optimum is reached. This local optimum is of course not necessarily the overall best solution, but for a certain suitable neighbourhood it can be shown that the local search algorithm finds the global optimum starting from any initial solution.

The neighbourhood for the local search algorithm is chosen as follows. Define

$$V^* = \left\{ \begin{array}{c} u_1, \\ u_2, \\ \vdots \\ u_{T-1}, \\ u_T \end{array} \right\} = \left\{ \begin{array}{c} (-1, 0, \ldots, 0, 1), \\ (1, -1, 0, \ldots, 0), \\ (0, 1, -1, 0, \ldots, 0), \\ \vdots \\ (0, \ldots, 0, 1, -1, 0), \\ (0, \ldots, 0, 1, -1) \end{array} \right\},$$

and take as the neighbourhood of a solution x all vectors of the form $x + v_1 + \cdots + v_k$ with $v_1, \ldots, v_k \in V^*$ such that $x + v_1 + \cdots + v_k \geq 0$. The algorithm consists of the following steps:

1. Start with some schedule x.
2. For all $U \subsetneq V^*$:

   for $y = x + \sum_{v \in U} v$ such that $y \geq 0$ compute $C(y)$;

   if $C(y) < C(x)$ then $x := y$ and start again with step 2.
3. x is the optimal schedule.

Adding one vector u_t is equivalent to moving the arrival of one patient from interval t to interval $t-1$. The neighbourhood of x consists of all possible combinations of these one-interval shifts of patient arrivals with respect to x.

3.1 Multimodularity and local search

A property needed to prove that local search does indeed find the global optimum is multimodularity. For completeness we repeat the definition of multimodularity and its relations to local search, which were already given in [10].

$$V = \left\{ \begin{array}{c} v_0, \\ v_1, \\ v_2, \\ \vdots \\ v_{m-1}, \\ v_m \end{array} \right\} = \left\{ \begin{array}{c} (-1, 0, \ldots, 0), \\ (1, -1, 0, \ldots, 0), \\ (0, 1, -1, 0, \ldots, 0), \\ \vdots \\ (0, \ldots, 0, 1, -1), \\ (0, \ldots, 0, 1) \end{array} \right\}.$$

Then multimodularity is defined as follows:

Definition 1. *A function $f : \mathbb{Z}^m \to \mathbb{R}$ is called multimodular if for all $x \in \mathbb{Z}^m, v, w \in V, v \neq w$,*

$$f(x + v) + f(x + w) \geq f(x) + f(x + v + w). \tag{1}$$

We also need the concept of an atom, as it forms the basis of our neighbourhood choice.

Definition 2. *For some $x \in \mathbb{Z}^m$ and σ a permutation of $\{0, \ldots, m\}$, the atom $S(x, \sigma)$ is defined as the convex set with extreme points $x + v_{\sigma(0)}, x + v_{\sigma(0)} + v_{\sigma(1)}, \ldots, x + v_{\sigma(0)} + \cdots + v_{\sigma(m)}$.*

In Koole and Van der Sluis [11] it is shown that for a multimodular function f a certain point x is a global minimum if and only if $f(x) \leq f(y)$ for all $y \neq x$ such that y is an extreme point of $S(x, \sigma)$ for some permutation σ.

This means that if we choose as neighbourhood for our local search algorithm all extreme points of all atoms $X(x, \sigma)$ for all possible permutations σ, we are guaranteed to find the globally optimal solution if our cost function is multimodular. The multimodularity of our cost function is what we will prove next.

Because in our problem $x - T$ is determined by $x_T = \sum_{t=1}^{T-1} x_t$ for given $x_1, \ldots, x_{T-1}$, our problem is $T - 1$-dimensional. The set of possible solutions is $\{x \in \mathbb{Z}^{T-1} | x \geq 0, \sum_{t=1}^{T-1} x_t \leq N\}$. This is of course not equal to $\mathbb{Z}^{T-1}$, but it is shown in Koole and Van der Sluis [11] that the above theorem still holds for this subset of $\mathbb{Z}^{T-1}$.

This means that we have to prove that our cost function is multimodular for the $T - 1$-dimensional problem, which is the same as showing that the T-dimensional cost function satisfies equation (1) with $v, w \in V^*$.

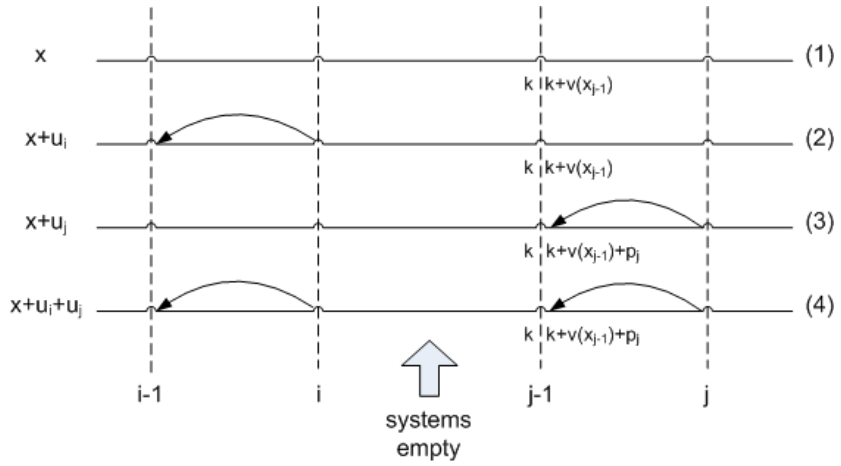


Figure 1: Schedule for case IA.

Theorem 3. *The cost function* $C(x) = \alpha L(x) + \beta I(x) + \gamma W(x)$ *is multimodular for all* $u_i, u_j \in V^*$ *for which* $i \neq j$.

Proof. We prove multimodularity separately for $L(x)$, $I(x)$ and $W(x)$. If two functions are multimodular then so is their sum, and then we have a multimodular cost function. The idle time is related to the makespan, the timespan from the start of the schedule untill the end of service of the last patient, for which multimodularity is easier to prove than for idle time. If the makespan is multimodular then so is the idle time, which means that we have to prove the multimodularity of the waiting time, tardiness and the makespan for every possible i and j for which $1 \leq i < j \leq T$. We use the coupling method to prove this, comparing for given realisations of the service times and emergency arrivals different paths the system follows. We then compare the numbers of minutes of work present to see differences in the waiting time, tardiness and makespan.

We distinguish a number of different cases. These cases differ in the characteristics of their paths, and therefore need to be considered separately, as can be seen in Table 1. The different variables in the table will be introduced below.

I: $x \leq i < j \leq T$

First we look at the case where $2 \leq i < j \leq T$. Here schedule x and schedule $x + u_j$ follow the same path up to time $j - 1$. This means that they have equal total waiting time, say α_1. Also the schedules $x + u_i$ and $x + u_i + u_j$ follow the same path and have equal waiting time, say α_2.

IA: systems empty between i *and* $j - 1$

If the system goes empty between times i and $j - 1$, all four schedules follow the same path from that moment on, see Figure 1. Suppose just before time $j - 1$ k minutes of work are present in all four cases. Then just after $j - 1$ there is present in

I: $2 \leq i < j \leq T$	A: system empties between i and $j-1$		
	B: system does not empty between i and $j-1$	1: $k' - c \geq d$	
		2: $k' - c < d \leq k'$	
		3: $k' < d \leq k' + p_j$	
		4: $d \geq k' + p_j$	
II: $1 = i < j \leq T$	A: system empties before $j-1$	1: system epties between j and T	
		2: system does not empty between j and T	
	B: system does not empty before $j-1$	1: $d \leq k' - c$	
		2: $k' - c < d \leq k'$	a: system empties between j and T
			b: system does not empty between j and T
		3: $k' < d < k' + p_j$	a: system empties between j and T
			b: system does not empty between j and T
		4: $d \geq k' + p_j$	a: system empties between j and T
			b: system does not empty between j and T

Table 1: Cases distinguished

(1): $k + v(x_{j-1})$

(2): $k + v(x_{j-1})$

(3): $k + v(x_{j-1}) + p_j$

(4): $k + v(x_{j-1}) + p_j$

where p_j is the number of minutes of work related to the patient moved from time j to $j-1$. From time $j-1$ (1) and (2) follow the same path and have total waiting time β_1, and (3) and (4) follow the same path and have total waiting time β_2. Then for the total waiting time we have

$$\alpha_2 + \beta_1 + \alpha_1 + \beta_2 \geq \alpha_1 + \beta_1 + \alpha_2 + \beta_1.$$

For the makespan and tardiness we have that $(2)+(3) \geq (1)+(4)$, because the end of the schedule is the same in (1) and (2), and also in (3) and (4).

IB: systems do not empty between i and $j-1$

If the system does not empty between i and $j-1$, there can be at most d minutes of work less present just before time $j-1$ in (1) and (3) compared to (2) and (4), because of the move of one patient from time i to $i-1$. See Figure 2. Then there are present just before time $j-1$ in

(1): k

(2): $k - c$

(3): k

(4): $k - c$

with $c \in \{1, 2, \ldots, d\}$. Let $k' = k + v(x_{j-1})$. Just after $j-1$ then there are present in

(1): k'

(2): $k' - c$

(3): $k' + p_j$

(4): $k' - c + p_j$

IB1: $k' - c \geq d$

In this case d minutes of work is done in all four schedules, and there is no idle time, see Figure 2(a). Then just after time j we have present

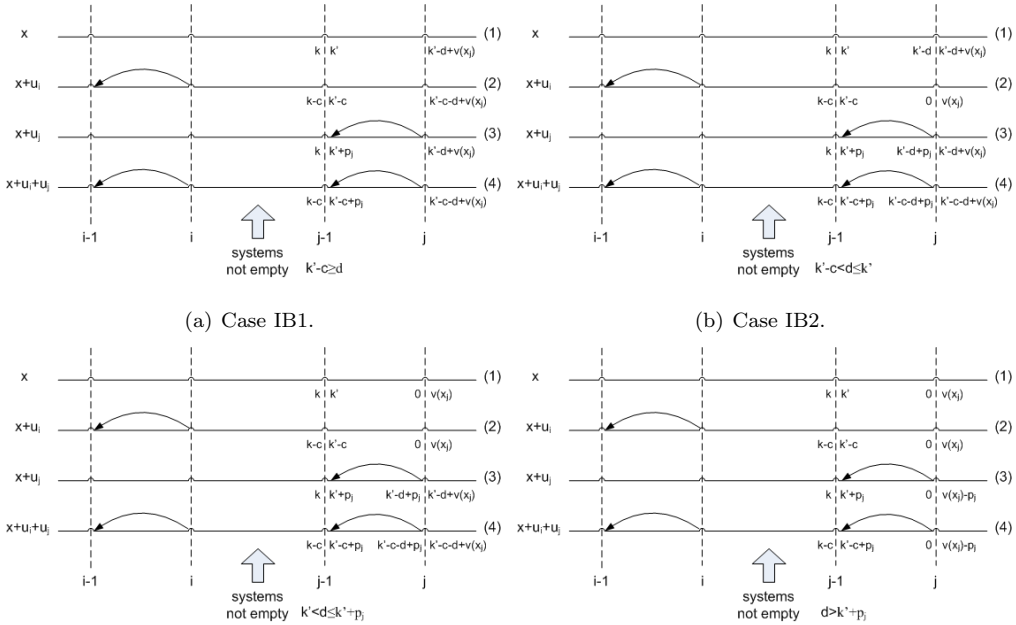
(1): $k' - d + v(x_j)$

(2): $k' - c - d + v(x_j)$

(3): $k' - d + v(x_j)$

(4): $k' - c - d + v(x_j)$

The waiting time between $j-1$ and j changes between schedules (1) and (3) only for the patient moved, and the same goes for schedules (2) and (4). So if the total waiting time between $j-1$



(a) Case IB1. (b) Case IB2.

(c) Case IB3. (d) Case IB4.

Figure 2: Schedules for cases IB1-4.

and j in (1) is β_1, then that in (3) is $\beta_1 + d$, for the moved patient's service cannot start before time j because $k' - c \geq d$. In the same way, if the total waiting time between $j-1$ and j in (2) is β_2, then that in (4) is $\beta_2 + d$.

From time j on schedules (1) and (3) follow the same path and have equal total waiting time, say γ_1. Also schedules (2) and (4) follow the same path and have equal total waiting time, say γ_2.

So we have for the total waiting time

$$(\alpha_2 + \gamma_2 + \beta_2) + (\alpha_1 + \gamma_1 + \beta_1 + d) \geq (\alpha_1 + \gamma_1 + \beta_1) + (\alpha_2 + \gamma_2 + \beta_2 + d).$$

For the makespan and tardiness we have that (2)+(3)≥(1)+(4), because the end of the schedule is the same in (1) and (3) and also the same in (2) and (4).

IB2: $k' - c < d \leq k'$

For this case, see Figure 2(b). Just before time j we have present in the system in

(1): $k' - d \geq 0$

(2): 0

(3): $k' - d + p_j$

(4): $k' - c - d + p_j$

If the total waiting time between $j-1$ and j in (1) is say β_1, then that in (3) is $\beta_1 + d$, because only the moved patient has d minutes more waiting time. In the same way, if the total waiting

time between time $j-1$ and j in (2) is β_2, then that in (4)is $\beta_2 + k' - c$. Just after time j there is present in

(1): $k' - d + v(x_j)$

(2): $v(x_j)$

(3): $k' - d + v(x_j)$

(4): $k' - c - d + v(x_j)$

After time j (1) and (3) have the same amount of work present and follow the same path, and so have the same waiting time, say γ_1. In schedule (4) $k' - c - d$ more work is performed between $j-1$ and j compared to (2), and so in (4) less work is present just after time j. The total waiting time then in (4) is less than in (2). Let the total waiting time after j in (2) be γ_2, then that in (4) is $\gamma_3 < \gamma_2$. For the waiting time we then have

$$(\alpha_2 + \beta_2 + \gamma_2) + (\alpha_1 + \beta_1 + d + \gamma_1) \geq (\alpha_1 + \beta_1 + \gamma_1) + (\alpha_2 + \beta_2 + k' - c + \gamma_3).$$

because $k' - c < d$ and $\gamma_3 < \gamma_2$.

For the makespan and tardiness only the end of the day counts. In (1) and (3) this is equal, so they have equal makespan and tardiness. Just after time j there is less work present in (4) than there is in (2), and the arrival path is equal in both schedules. So the makepsan and tardiness are smaller in (4) than in (2). So for the makespan and tardiness we indeed have that (2)+(3)≥(1)+(4).

IB3: $k' < d \leq k' + p_j$

This case is represented in Figure 2(c). Just before time j there is present in

(1): 0

(2): 0

(3): $k' + p_j - d \geq 0$

(4): $k' - c + p_j - d \geq d - c \geq 0$

If the total waiting time between times $j-1$ and j in (1) is β_1, then that in (3) is $\beta_1 + k'$, and if the total waiting time between $j-1$ and j in (2) is β_2, than that in (4) is $\beta_2 + k' - c$.

Just after time j there is present in

(1): $v(x_j)$

(2): $v(x_j)$

(3): $k' - d + v(x_j) < v(x_j)$

(4): $k' - c - d + v(x_j)$

The total waiting time in (1) and (2) after time j is equal, say γ_1. Just after time j there are in (4) c minutes of work less than there are in (3), so every patient starting his service after this time (and before any empty period, if one occurs) experiences a shorter waiting time in (4) than

in (3). So if the total waiting time after j in (3) is γ_2, then that in (4) is $\gamma_3 < \gamma_2$. Then for the total waiting time we have

$$(\alpha_2 + \beta_2 + \gamma_2) + (\alpha_1 + \gamma_2 + \beta_1 + k') \geq (\alpha_1 + \beta_1 + \gamma_1) + (\alpha_2 + \beta_2 + k' - c + \gamma_3).$$

The makespan and tardiness are equal in (1) and (3), and in (4) they are both smaller than they are in (3). So for the makespan and tardiness we have (2)+(3)≥(1)+(4).

IB4: $d > k' + p_j$

Just before time j there are 0 minutes of work present in all four schedules, see Figure 2(d). For the total waiting time between times $j-1$ and j we know that if this in (1) is β_1, then that in (3) is $\beta_1 + k'$. And also if the waiting time in (2) is β_2, then in (4) it is $\beta_2 + k' - c$.

Just after time j there is present in

(1): $v(x_j)$

(2): $v(x_j)$

(3): $v(x_j) - p_j$

(4): $v(x_j) - p_j$

From time j on (1) and (2) follow the same path, and so have equal total waiting time, say γ_1. Also (3) and (4) have equal waiting time after j, say γ_2. Then for the waiting time we have

$$(\alpha_2 + \beta_2 + \gamma_2) + (\alpha_1 + \beta_1 + k' + \gamma_2) \geq (\alpha_1 + \beta_1 + \gamma_1) + (\gamma_2 + \beta_2 + k' - c + \gamma_2)$$

because $c > 0$.

For the makespan and tardiness we look at the end of the day, and this is equal in (1) and (2) and also in (3) and (4). So we have (2)+(3)=(1)+(4) for the makespan and tardiness.

II: $1 = i < j \leq T$

Now we look at the case where $i = 1$, which means that the one of the patients scheduled at time 1 will be moved to time T. Until time $j-1$ (1) and (3) follow the same path, and have equal total waiting time, say α_1. Also (2) and (4) follow the same path, and have total waiting time α_2.

IIA: systems empty before $j-1$

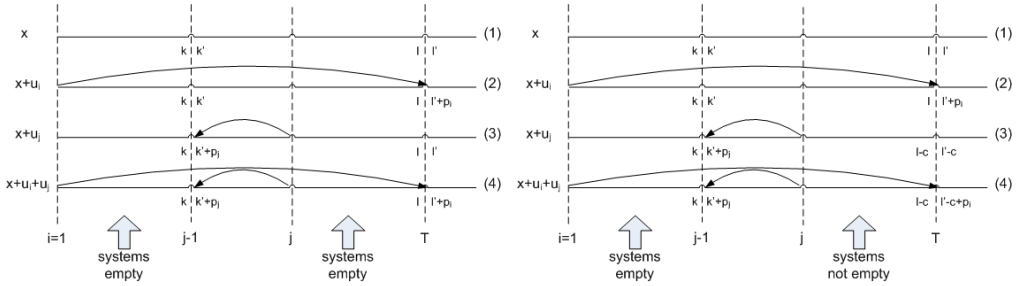
From the moment the system becomes empty until time $j-1$ all four schedules follow the same path, see Figure 3. Let the number of minutes of work present just before time $j-1$ be k. Then just after time $j-1$ there is present in

(1): $k + v(x_{j-1}) = k'$

(2): $k + v(x_{j-1}) = k'$

(3): $k' + p_j$

(4): $k' + p_j$



(a) Case IIA1. (b) Case IIA2.

Figure 3: Schedules for cases IIA1-2.

From $j-1$ to T (1) and (2) follow the same path and have equal total waiting time, say β_1, and (3) and (4) have equal total waiting time, say β_2.

IIA1: systems empty between j and T

From the moment the system becomes empty until time T all four schedules follow the same path, see Figure 3(a). Let l be the number of minutes of work present just before time T. Then just after T there is present

(1): $l + v(x_T) = l'$

(2): $l' + p_i$

(3): l'

(4): $l' + p_i$

So from time T on (1) and (3) follow the same path and have equal total waiting time, say γ_1, and (2) and (4) have equal total waiting time, say γ_2. Then for the waiting time we have

$$(\alpha_2 + \beta_1 + \gamma_2) + (\alpha_1 + \beta_2 + \gamma_1) \geq (\alpha_1 + \beta_1 + \gamma_1) + (\alpha_2 + \beta_2 + \gamma_2).$$

The end is the same in (1) and (3) and also in (2) and (4), so for the makespan and tardiness it holds that (2)+(3)=(1)+(4).

IIA2: systems do not empty between j and T

Between times $j-1$ and j more minutes of work can be done in (3) and (4) than in (1) and (2) because of the patient moved, see Figure 3(b). This amount is smaller than or equal to $min(d, p_j)$. Let the amount of extra work done in (3) and (4) be c. Then just before time T there is present in

(1): l

(2): l

(3): $l - c$

(4): $l - c$

Just after time T there is present in

(1): $l + v(x_T) = l'$

(2): $l' + p_i$

(3): $l' - c$

(4): $l' - c + p_i$

Let the total waiting time after time T in (1) be γ_1, then that in (2) is $\gamma_1 + l'$ because the moved patient waits l' minutes more. In the same way if the total waiting time after time T in (3) is γ_2,, then that in (4) is $\gamma_2 + l' - c$. Then for the waiting time we have

$$(\alpha_2 + \beta_1 + \gamma_1 + l') + (\alpha_1 + \beta_2 + \gamma_2) \geq (\alpha_1 + \beta_1 + \gamma_1) + (\alpha_2 + \beta_2 + \gamma_2 + l' - c),$$

because $c \geq 0$.

For the makspan and tardiness we look at the end of the day, and we see that in (3) the amount of work present is c minutes less after time T than in (1) and also c minutes less in (4) than in (2). So for the makespan and tardiness we have (2)+(3)=(1)+(4).

IIB: systems do not empty before $j - 1$

Because the system does not become empty before time $j - 1$, more work can be prsent in (1) and (3) than in (2) and (4), at most p_i minutes. Let $c \leq p_i$ be the amount of work more present in (1) and (3). Then just before time $j - 1$ there is present in

(1): k

(2): $k - c$

(3): k

(4): $k - c$

Just after time $j - 1$ there is present in

(1): $k + v(x_{j-1}) = k'$

(2): $k' - c$

(3): $k' + p_j$

(4): $k' - c + p_j$

Now we have to distinguish a few different cases, based on the amount of work present after time $j - 1$.

IIB1: $d \leq k' - c$

For this case, see Figure 4. In this case, just before time j there is present in

(1): $k' - d$

(2): $k' - c - d$

(3): $k' + p_j - d$

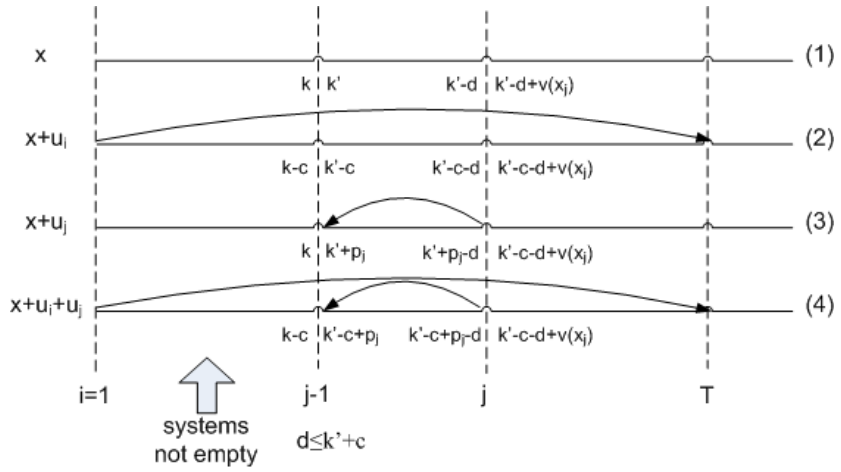


Figure 4: Schedule for case IIB1.

(4): $k' - c + p_j - d$

Just after time j there is present in

(1): $k' - d + v(x_j)$

(2): $k' - c - d + v(x_j)$

(3): $k' - c + p_j - d + v(x_j) - p_j = k' - c - d + v(x_j)$

(4): $k' - c - d + v(x_j)$

If the toal waiting time between $j-1$ and j is β_1, than that in (3) is $\beta_1 + d$. And also if the total waiting time between $j-1$ and j in (2) is β_2, than that in (4) is $\beta_2 + d$. This is true because in (3) and (4) the moved patient waites for d minutes extra, and the waiting time for the other patients does not change at all.

From time j to the end of the day schedules (1) and (3) follow the same path, and have equal total waiting time, say γ_1. Also (2) and (4) have equal total waiting time, say γ_2. For the waiting time we then get

$$(\alpha_2 + \beta_2 + \gamma_2) + (\alpha_1 + \beta_1 + d + \gamma_1) \geq (\alpha_1 + \beta_1 + \gamma_1) + (\alpha_2 + \beta_2 + d + \gamma_2).$$

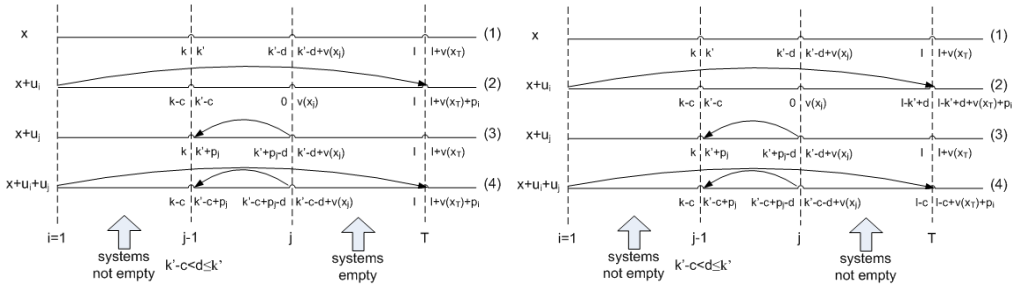
Because the end of the day is exactly the same in (1) and (3) they have equal makespan and tardiness, and the same holds for (2) and (4). Then for the makespan and tardiness it holds that (2)+(3)=(1)+(4).

IIB2: $k' - c < d \leq k'$

For this case, see Figure 5. Just before time j there is present in

(1): $k' - d$

(2): 0



(a) Case IIB2a. (b) Case IIB2b.

Figure 5: Schedules for cases IIB2a-b.

(3): $k' + p_j - d$

(4): $k' - c + p_j - d$

Just after time j there is present in

(1): $k' - d + v(x_j)$

(2): $v(x_j)$

(3): $k' - d + v(x_j)$

(4): $k' - c - d + v(x_j)$

If the total waiting time between times $j-1$ and j in (1) is β_1, then that in (3) is $\beta_1 + d$. And if the total waiting time between times $j-1$ and j in (2) is β_2, then that in (4) is $\beta_2 + k' - c$.

From j until T (1) and (3) follow the same path, and have equal total waiting time, say γ_1. The total waiting time in (4) between j and T is smaller than or equal to that in (2). Let the waiting time between j and T in (2) be γ_2, and that in (4) be $\gamma_3 \leq \gamma_2$. Now again we distuinguish two cases.

IIB2a: system empties between j and T

If the system becomes empty, the same amount of work is present just before time T in all four cases, say l. See Figure 5(a). Then just after T there is present in

(1): $l + v(x_T)$

(2): $l + v(x_T) + p_i$

(3): $l + v(x_T)$

(4): $l + v(x_T) + p_i$

From time T on both (1) and (3) have equal total waiting time, say δ_1, and (2) and (4) have equal waiting time, say δ_2. Then for the total waiting time

$$(\alpha_2 + \beta_2 + \gamma_2 + \delta_2) + (\alpha_2 + \beta_1 + d + \gamma_1 + \delta_1) \geq$$
$$(\alpha_1 + \beta_1 + \gamma_1 + \delta_1) + (\alpha_2 + \beta_2 + k' - c + \gamma_3 + \delta_2),$$

because $\gamma_3 \leq \gamma_2$ and $k' - c < d$.

For the makespan and tardiness we need only look at the end of the day, and this is equal in (1) and (3) and also in (2) and (4). So for the makespan and tardiness (2)+(3)=(1)+(4).

IIB2b: system does not empty between j and T

For this case, see Figure 5(b). Just before time T there is present in

(1): l

(2): $l - k' + d$

(3): l

(4): $l - c$

Just after time T there is present in

(1): $l + v(x_T)$

(2): $l - k' + d + v(x_T) + p_i$

(3): $l + v(x_T)$

(4): $l - c + v(x_T) + p_i$

From time T on (1) and (3) have equal total waiting time, say δ_1. All patients that arrive at time T in (4), which are $x_T + 1$, wait $c - (k' - c)$ minutes less than they wait in schedule (2). So if the total waiting time after time T in (2) is say δ_2, then that in (4) is $\delta_2 - (x_T + 1)(c - k' + d)$. Then for the total waiting time it holds that

$$(\alpha_2 + \beta_2 + \gamma_2 + \delta_2) + (\alpha_1 + \beta_1 + d + \gamma_1 + \delta_1) \geq (\alpha_1 + \beta_1 + \gamma_1 + \delta_1) +$$
$$(\alpha_1 + \beta_1 + k' - c + \gamma_3 + \delta_2 - (x + T + 1)(c - k' + d)).$$

The end of the day follows the same path in (1) and (3), so these two schedules have equal makespan and tardiness. In (4) less work is present after time T than in (2), so (4) has lower makespan and tardiness than (2). Then for the makespan and tardiness we have that (2)+(3)$\geq$(1)+(4).

IIB3: $k' < d \leq k' + p_j$

For this case, see the schedules in Figure 6. Just before time j there is present in

(1): 0

(2): 0

(3): $k' + p_j - d$

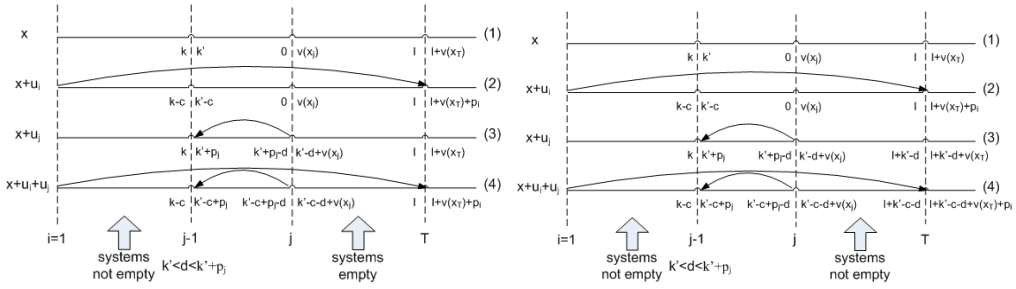
(4): $k' - c + p_j - d$

Just after time j there is present in

(1): $v(x_j)$

(2): $v(x_j)$

(3): $k' - d + v(x_j)$



(a) Case IIB3a. (b) Case IIB3b.

Figure 6: Schedules for cases IIB3a-b.

(4): $k' - c - c + v(x_j)$

If the total waiting time between times $j-1$ and j in (1) is β_1, then that in (3) is $\beta_1 + k'$. And if the total waiting time between times $j-1$ and j in (2) is β_2, then that in (4) is $\beta_2 + k' - c$. From time j to time T schedules (1) and (2) follow the same path, and so have equal total waiting time, say γ_1. If the total waiting time between j and T in (3) is γ_2, then that in (4) is $\gamma_2 - x_j c$. Again we distinguish two cases.

IIB3a: system empties between j and T

Just before time T there are l minutes of work present in all four cases, see Figure 6(a). Just after time T there is present in

(1): $l + v(x_T)$

(2): $l + v(x_T) + p_i$

(3): $l + v(x_T)$

(4): $l + v(x_T) + p_i$

From time T on (1) and (3) have equal total waiting time, say δ_1, and also (2) and (4) have equal waiting time, say δ_2. Then for the total waiting time we have

$$(\alpha_2 + \beta_2 + \gamma_1 + \delta_2) + (\alpha_1 + \beta_1 + k' + \gamma_2 + \delta_1) \geq$$
$$(\alpha_1 + \beta_1 + \gamma_1 + \delta_1) + (\alpha_2 + \beta_2 + k' - c + \gamma_2 - x_j c + \delta_2).$$

From time T on (1) and (3) follow the same path, and also (2) and (4) follow the same path, so for the makespan and tardiness it holds that (2)+(3)=(1)+(4).

IIB3b: system does not empty between j and T

See Figure 6(b) for this case. Just before time T there is present in

(1): l

(2): l

(3): $l + k' - d$

(4): $l + k' - c - d$

Just after T there is present in

(1): $l + v(x_T)$

(2): $l + v(x_T) + p_i$

(3): $l + k' - d + v(x_T)$

(4): $l + k' - c - d + v(x_T) + p_i$

If the total waiting time in (1) after time T is δ_1, then that in (3) is $\delta_1 + x_T(k' - d)$. And if the total waiting time after T in (2) is δ_2, then that in (4) is $\delta_2 + (x_T + 1)(k' - c - d)$. Then for the total waiting time it holds that

$$(\alpha_2 + \beta_2 + \gamma_1 + \delta_2) + (\alpha_1 + \beta_1 + k' + \gamma_2 + \delta_1 + x_T(k' - d)) \geq$$
$$(\alpha_1 + \beta_1 + \gamma_1 + \delta_1) +$$
$$(\alpha_2 + \beta_2 + k' - c + \gamma_2 - x_j c + \delta_2 + (x_T + 1)(k' - c - d)).$$

In (3) there are $d - k' > 0$ minutes less work present after T than there are in (1), and the end of the day is then $d - k'$ earlier. In (4) there are $d + c - k' > d - k'$ minutes less work present than there are in (2), so the end of the day is $d + c - k'$ earlier. Then for the makespan and tardiness it holds that (2)+(3)≥(1)+(4).

IIB4: $d \geq k' + p_j$

Just before time j there are 0 minutes of work present in all four cases, see Figure 7. Then just after j there is present in

(1): $v(x_j)$

(2): $v(x_j)$

(3): $v(x_j) - p_j$

(4): $v(x_j) - p_j$

If the total waiting time between $j - 1$ and j in (1) is β_1, then that in (3) is $\beta_1 + k'$. If the waiting time between $j - 1$ and j in (2) is β_2, then that in (4) is $\beta_2 + k' - c$. From time j to time T (1) and (2) have equal waiting time, say γ_1. And also (3) and (4) have equal waiting time, say γ_2. Here we again distinguish two cases.

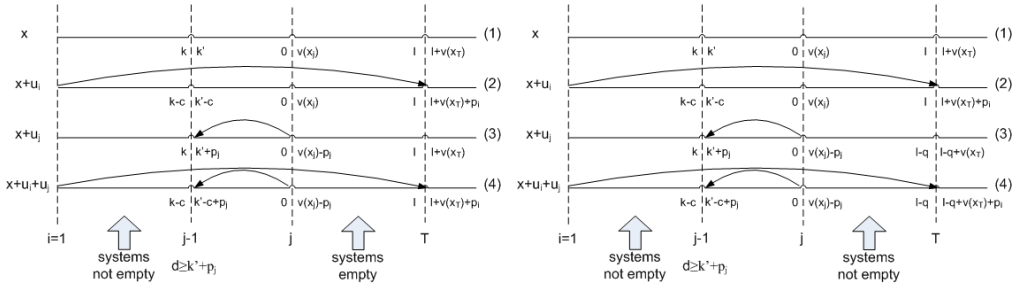
IIB4a: system empties between j and T

Just before time T there are the same number of minutes work present in all four cases, say l. See Figure 7(a). Then just after time T there is present in

(1): $l + v(x_T)$

(2): $l + v(x_T) + p_i$

(3): $l + v(x_T)$



(a) Case IIB4a. (b) Case IIB4b.

Figure 7: Schedules for cases IIB4a-b.

(4): $l + v(x_T) + p_i$

From time T on (1) and (3) follow the same path and so have equal total waiting time, say δ_1. Also (2) and (4) have equal total waiting time, say δ_2. Then for the waiting time it holds that

$$(\alpha_2 + \beta_2 + \gamma_1 + \delta_2) + (\alpha_1 + \beta_1 + k' + \gamma_2 + \delta_1) \geq$$
$$(\alpha_1 + \beta_1 + \gamma_1 + \delta_1) + (\alpha_2 + \beta_2 + k' - c + \gamma_2 + \delta_2).$$

Because after time T (1) and (3) follow the same path, and also (2) and (4) follow the same path, for the makespan and tardiness we have that (2)+(3)=(1)+(4).

IIB4b: system does not empty between j and T

For this case see Figure 7(b). Just before time T there is present in

(1): l

(2): l

(3): $l - q$

(4): $l - q$

with q the extra amount of work done in (3) and (4) compared to (1) and (2) respectively. Just after time T there is present in

(1): $l + v(x_T)$

(2): $l + v(x_T) + p_i$

(3): $l - q + v(x_T)$

(4): $l - q + v(x_T) + p_i$

If the total waiting time after time T in (1) is δ_1, then that in (2) is $\delta_1 + l + v(x_T)$. And if the total waiting time after T in (3) is δ_2, then that in (4) is $\delta_2 + l - q + v(x_T)$. Then for the waiting time we have

$$(\alpha_2 + \beta(2) + \gamma_1 + \delta_1 + l + v(x_T)) + (\alpha_1 + \beta_1 + k' + \gamma_2 + \delta_2) \geq$$
$$(\alpha_1 + \beta_1 + \gamma_1 + \delta_1) + (\alpha_2 + \beta_2 + k' - c + \gamma_2 + \delta_2 + l - q + v(x_T)).$$

The end of the day is in (2) p_i minutes later than in (1), and in (4) p_i than in (3). So for the makespan and tardiness it holds that (2)+(3)=(1)+(4).

We have looked at all different cases, and proved that for every case the waiting time, the makespan and the tardiness are all three multimodular. That means that the sum of these three is multimodular, and this proves the theorem.

The proof for the problem with no-shows is done in exactly the same way by conditioning on the no-show of a patient and following the same method of reasoning with less patients scheduled. □

3.2 Complexity

The complexity of local search algorithms depends on the number of evaluations necessary to check if a given solution is a local optimum. This number is equal to the size of the neigbourhood. For an m-dimensional problem with a multimodular cost function the number of neigbours is $2^{m+1}-2$, so for our problem this is $2^T - 2$.

This is not a number linear in T, which means that it is not possible to check the local optimality of a solution in polynomial time. So our local search algorithm does not belong to the complexity class PLS as defined by Johnson et al. [9]. Only of problems that do fit in PLS something can be said with certainty about the complexity and running times of the algorithm, so we have to assume that (in the worst case) our algorithm has an exponential running time. However, from the numerical experiments we have performed it appears that the running times are still acceptable for problems of realistic size.

The neighbourhood we use here is exact, which means that any local optimum is also a global optimum. It is also possible to use a non-exact neighbourhood. In this case take as neighbourhood of a solution x $y = x + v$ for all $v \in U$ in step two in the above algorithm. This gives much faster results, but there is no guarantee that the solution found is also the optimal solution.

Another method has been proposed by Vanden Bosch et al. [2] that is faster than the method we used. However as we already mentioned above there is a small problem in one of the proofs of this work. In one of the theorems they state that the arrival times of each patient in the so-called early and late schedules differ by at most one interval. Because the early and late schedules are bounds for the arrival times for each patient, this makes the number of possible candidates for the optimal schedule very small and enumeration very fast.

For the proof of this theorem they refer to Wang [17], who proves that the patient waiting time is a convex function in the vector of interarrival times. The problem lies in the fact that [2] does not use the vector of interarrival times like Wang does but rather the vector of the number

of patients scheduled at each interval like we do as well. Consider an example with 2 patients and 4 intervals where the early schedule is given by $S_E = (1, 1, 0, 0)$ and the late schedule by $S_L = (0, 0, 1, 1)$. Then the interarrival times vectors would be $X_E = (0, 1)$ and $X_L = (2, 1)$. Now if we move one patient one interval starting from the early schedule we get $S' = (1, 0, 1, 0)$ and $X' = (0, 2)$. It is clear that X' is not a convex combination of X_E and X_L so the convexity result from Wang cannot be used in this case. This makes that the proof of Theorem 4 in [2] is not complete.

Of course the method in itself still works, only the distance between the early and late schedules is unknown, and so also the computation time.

4 Numerical results

In this section we present the results of some numerical experiments we performed. The starting point for our experiments is a half-day time window, or four hours. This time is split up in 24 intervals of 10 minutes each, so $T = 24$ and $d = 10$. We want to schedule 240 minutes of work in this time period, where each regular appointment takes on average 20 minutes, and each emergency service takes on average 30 minutes. As already mentioned above, it is not really logical in the case with emergencies to look at both tardiness and server idle time simultaneously, so we choose one of the two here, namely the tardiness. However, the results hold true as well if the idle time is taken into account instead.

In this section we first look at how standard even-spaced schedules behave when service times have different distributions, and how performance can be improved by changing the schedule. Then we consider the influence of the amount of emergency work on the performance of the optimal schedule, and also at what times during the four-hour period space should be reserved in the schedule. Then we study how the schedule changes with the relative importance of patient waiting times and server tardiness, and how emergency arrivals influence the optimal schedule for different weights of the two performance measures. Finally we look at how the waiting times for different patients within the schedules compare to each other.

4.1 Standard practice and variability in service times

The most-used schedule is the even-spaced schedule. However, the performance of this schedule can be influenced negatively if there are emergency arrivals, and also by variability in service time. In Table 2 we compare the waiting times and tardiness using the standard schedule in the case with nine scheduled patients and on average two emergency arrivals. The service times are

deterministic in the first case, and exponentially and normally distributed with equal standard deviations in the second and third cases.

Distribution	Waiting time	Tardiness
Deterministic	29.7	18.7
Exponential	44.0	39.8
Normal	30.6	20.2

Table 2: Performance of standard schedule with different service time distributions

In the deterministic case the only influence is that of emergency arrivals, and we can see that this has already a large impact on performance. With the normal distributions for the service times the performance is slightly worse, but in the case with exponential service times the difference is much larger. This might be because the probability of high service times is larger in this case than with normally distributed service times. However, we can see from the results that the largest part of the waiting times and tardiness is caused by the uncertainty in emergency arrivals.

In Table 3 we show the optimal schedule and the performances in the same three cases. We can see that the three schedules are not equal in the three cases, so the schedule adjusts to the service time distribution. There is improvement in performace for all three scenarios. Also, for these scenarios we kept the weights for the tardiness equal, but of course one could adjust the schedule by changing these weights to reflect actual priorities. This is not possible with a fixed schedule. Another thing to note is that these results seem to contradict the statement of Denton and Gupta [4] that higher moments only influence the optimal schedule when waiting costs are high relative to the idle time costs, or in our case overtime costs.

Distribution	Waiting time	Tardiness	Optimal schedule
Deterministic	20.6	21.5	10101010100001010100010
Exponential	39.0	43.3	11010100100100100100100
Normal	22.6	23.8	10101010010010010010100

Table 3: Optimal schedules with different service time distributions

4.2 Influence of emergencies

We noticed above that emergency arrivals can have a large impact on the performance of the schedule. To investigate further how the performance changes when adding emergency work, we look at three different scenarios. The first one is the base scenario described above, with

deterministic service durations and 12 scheduled patients. In the second scenario we schedule three patients less, and we expect on average 2 emergency arrivals. In the third scenario we schedule 6 patients, and expect on average 4 emergency arrivals. In all three scenarios the waiting time and tardiness have equal weight. We choose deterministic service times, so the emergency arrivals are the only source of uncertainty and their influence can be seen more clearly. The results and the optimal schedule in these scenarios are given in Table 4.

Nr. emergencies	Waiting time	Tardiness	Optimal schedule
0	0.0	0.0	101010101010101010101010
2	20.59	21.48	101010101000010101000010
4	39.08	31.13	101010000100001000010000

Table 4: Influence of emergencies on schedule performance

The first scenario is the ideal scenario where the whole schedule is executed according to plan, because there is no variation in either arrival moments or service durations. In this case the optimal scenario is of course to schedule time equal to the service duration for all patients.

In the second and third scenario we can see that performance decreases with emergency arrivals, as is to be expected. The open space in the schedule is concentrated more towards the end of the day than in the beginning, because the probability that an emergency has arrived is very small early in time and open space there would often lead to unnecessary server idle time.

4.3 Relative importance of performance measures

Again we look at three different scenarios, all without emergencies and now with exponential service times. We change the weights for the waiting time and the tardiness to see how the schedule and performance change. The results can be seen in Table 5.

α_W	α_L	Waiting time	Tardiness	Optimal schedule
1	1	39.00	43.26	110101001001001001001000
10	1	19.21	58.07	110100100101001001010102
1	10	46.92	28.12	211101101010100101000000

Table 5: Influence of weights on the optimal schedule

From the results we can see that for higher the relative weight for the waiting time, the arrival times for the patients move more towards the end of the day. This gives longer interarrival times, and so less waiting time for each patient. For the tardiness we see exactly the opposite effect,

the more important the tardiness is the earlier the arrival times. This gives of course the lowest probability and amount of tardiness.

These three scenarios were all without emergencies. To see if the same principles hold in the case with emergencies, we also give the results for three scenarios with 2 and 4 expected emergency arrivals, also all with exponential service times. The results can be seen in Table 6.

Emergencies	α_W	α_L	Waiting time	Tardiness	Optimal schedule
0	1	1	28.18	35.23	20101010101010010101010
2	1	1	39.00	43.26	11010100100100100100100
4	1	1	49.72	52.15	11001000100000100001000
0	10	1	19.21	58.07	11010010010100100101010
2	10	1	30.86	60.60	11001000100010001001011
4	10	1	43.01	66.32	11000010000000010000010
0	1	10	46.92	28.12	21110110101010010100000
2	1	10	58.83	36.14	21101010100100100000000
4	1	10	68.06	45.12	21010010010000000000000

Table 6: Influence of weights on the optimal schedule

We can see from the results that the performance generally gets worse on both waiting time and tardiness as the portion of emergency work increases. It turns out that the scheduling principles that hold in the case without emergencies have an even stronger influence if there are emergencies. So with a strong emphasis on waiting time many patients are scheduled toward the end of the day, and open space for emergencies is concentrated somewhere in the middle. In the case with high weight for tardiness the patients are almost all scheduled in the first half of the available time.

4.4 Waiting times per patient

The objective function contains the expected waiting time averaged over all scheduled patients. It does not take into account how the total waiting time is distributed over the patietns. We consider a case with nine scheduled patients and on average 2 emergency patients per day, and deterministic and exponential and normal distributions for the service times with equal variances. In Table 7 we show the expected waiting times for the nine scheduled patients with the patients numbered in order of arrival when using the optimal schedule shown in Table 3.

We see in all three cases that the waiting times differ considerably per patient. The first patients have low expected waiting times, since they only wait if there is an emergency arrival

Distribution	1	2	3	4	5	6	7	8	9
Deterministic	3.33	9.99	16.65	23.28	29.85	21.04	27.23	32.81	21.13
Exponential	3.20	21.57	32.58	42.70	45.76	48.72	51.07	52.51	52.90
Normal	3.21	11.25	18.84	26.16	27.23	28.51	29.35	29.66	28.94

Table 7: Expected waiting times per patient

immediately at the start of the day. In all three cases the waiting times per patient then increases rapidly and stabilises or decreases slightly for the last patients. As we had already seen for the average waiting times, those with exponentally distributed service times are much higher than in the case with normally distributed served times.

5 Conclusions

In this paper we presented a method for finding the optimal appointment schedule in a setting with emergency arrivals or interruptions. The method uses general service time distributions and can handle no-shows, which makes it suitable for use in practice. It finds the optimal arrival times for a weighted combination of patient waiting time, doctor idle time and tardiness as the objective. The method makes use of a local search algorithm, which for our multimodular objective function is guaranteed to find the global optimum.

From the numerical examples we presented it can be seen that in general more free space for emergencies is reserved towards the end of the day, or in other words, the interarrival times increase over the day. The same goes for space in the schedule reserved for dealing with variation in the appointment durations.

For further research there are a few interesting questions left. One is how to incorporate different patient types with the same urgency but different service time distributions. Another is how to deal with patients that arrive earlier or later than their appointment times. We are planning to address these questions in future work.

References

[1] M.A. Begen and M. Queyranne. Appointment scheduling with discrete random durations. Working paper, Sauder School of Business, University of Britisch Columbia, 2009.

[2] P.M. Vanden Bosch, D.C. Dietz, and J.R. Simeoni. Scheduling customer arrivals to a stochastic service system. *Naval Research Logistics*, 46:549–559, 1999.

[3] T. Cayirli and E. Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12:519–549, 2003.

[4] B. Denton and D. Gupta. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35:1003–1016, 2003.

[5] R.B. Fetter and J.D. Thompson. Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research*, 1:66–90, 1966.

[6] Y. Gerchak, D. Gupta, and M. Henig. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42:321–334, 1996.

[7] D. Gupta and B. Denton. Appointment scheduling in health care: challenges and opportunities. *IIE Transactions*, 40:800–819, 2008.

[8] C.-J. Ho and H.-S. Lau. Minimizing total cost in scheduling outpatient appointments. *Management Science*, 38:1750–1763, 1992.

[9] D.S. Johnson, C.H. Papadimitriou, and M. Yannakakis. How easy is local search? *Journal of computer and system sciences*, 37:79–100, 1988.

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

[11] G. Koole and E. van der Sluis. Optimal shift scheduling with a global serice level constraint. *IIE Transactions*, 35:1049–1055, 2003.

[12] M. Lamiri, X. Xie, A. Dolgui, and F. Grimaud. A stochastic model for operating room planning with elective and emergency demand for surgery. *European Journal of Operational Research*, 185:1026–1037, 2008.

[13] R.M. O'Keefe. Investigating outpatient departments: implementable policies and qualitative approaches. *The Journal of the Operational Research Society*, 36:705–712, 1985.

[14] E.J. Rising, R. Baron, and B. Averill. A systems analysis of a universty-health-service outpatient clinic. *Operations Research*, 21:1030–1047, 1973.

[15] W.E. Stein and M.J. Côté. Scheduling arrivals to a queue. *Computers and Operations Research*, 6:607–614, 1994.

[16] J. Vissers and J. Wijngaard. The outpatient appointment system: design of a simulation study. *European Journal of Operational Research*, 3:459–463, 1979.

[17] P.P. Wang. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics*, 40:345–360, 1993.

[18] P.P. Wang. Optimally scheduling n customer arrival times for a single-server system. *Computers and Operations Research*, 8:703–716, 1997.

[19] J.D. Welch and N.T.J. Bailey. Appointment systems in hospital outpatient departments. *The Lancet*, 259:1105–1108, 1952.