Health Care Manag Sci (2017) 20:403–418
DOI 10.1007/s10729-016-9361-7

Strategies for interday appointment scheduling in primary care

Lara Wiesche[1] · **Matthias Schacht**[1] · **Brigitte Werners**[1]

Received: 15 July 2015 / Accepted: 3 February 2016 / Published online: 21 March 2016

Abstract When faced with a medical problem, patients contact their primary care physician (PCP) first. Here mainly two types of patient requests occur: non-scheduled patients who are walk-ins without an appointment and scheduled patients with an appointment. Number and position of the scheduled appointments influence waiting times for patients, capacity for treatment and the utilization of PCPs. As the number of patient requests differs significantly between weekdays, the challenge is to match capacity with patient requests and provide as few appointment slots as necessary. In this way, capacity for walk-ins is maximized while overall capacity restrictions are met. Decisions as to the optimal appointment capacity per day on a tactical decision level has gained little attention in the literature. A mixed integer linear model is developed, where the minimum number of appointments scheduled for a weekly profile is determined. We are thus able to give the answer as to how many appointments to offer on each day in a week in order to create a schedule that takes patient preferences as well as PCP preferences into account. Appointment schedules are often influenced by uncertain demands due to the number of urgent patients, interarrivals and service times. Based on an exemplary case study, the advantages of the optimal appointment schedule on different performance criteria are shown by detailed stochastic simulations.

Keywords Interday appointment scheduling · Decision support · Capacity allocation · Mathematical optimization · Stochastic simulation

✉ Lara Wiesche
lara.wiesche@rub.de

Matthias Schacht
matthias.schacht@rub.de

Brigitte Werners
or@rub.de

[1] Ruhr University Bochum, Chair of Operations Research, Universitätsstraße 150, 44801 Bochum, Germany

1 Introduction

Primary care faces multiple challenges meeting patient and practitioner needs with regard to service-time planning. During the week, there is an imbalance between available capacity and treatment requests with extensive treatment demand on Mondays and Tuesdays. As a result, urgent requests from patients cannot be met immediately and long waiting times occur. Additionally, the workload for primary care physicians (PCPs) varies substantially. During a busy day the PCPs have to work overtime, whereas at less frequented times there is underutilization. Appropriate optimization models which consider both sides' preferences support decision makers in determining efficient appointment schedules. The objective is to deduce a suitable solution for patients as well as for staff members. This paper presents a model to determine the optimal capacity for different types of patients, to improve the availability for urgent requests and to balance PCPs workload. Different types of arrivals exist in primary care [24]:

1. uncontrollable arrivals: patients seeking treatment on the same day and therefore arriving without further notice, and
2. controllable arrivals: patients booking appointments in advance.

The problem of varying demand is present in healthcare management applications [7, 16, 23]. Degel et al. present

a new methodology to ensure a maximum of coverage by location and relocation of emergency medical services when demand varies during the day [7]. Requests for primary care arise at different times during the day and are likely to follow a specific arrival pattern. Requests are more frequent at certain hours of the day and on certain days in the week (e.g. [12, 13]). Almost 50 % of patients request arrive on Monday or Tuesday [15], and more than 70 % of patients have a treatment request during morning hours [4]. According to a given capacity limit and to prevent extensive overtime, some of the requests from Mondays and Tuesdays have to be re-arranged for other days of the week. Predetermined appointment slots allow the physician to shift patients' requests from days with a high number of requests to less frequented days. As a result, capacity and demand can be matched in an optimal way, while an adequate amount of available service time for uncontrolled arrivals is guaranteed. The number and position of appointment slots affect the performance of a PCP's entire schedule. On the one hand, due to appropriate positioning, the arrival of patients with appointments can be controlled and scheduled to certain days. In addition, appointments improve the workload predictability and scheduling becomes less dependent on walk-ins with uncontrollable arrivals. On the other hand, previous research (e.g. [11, 19, 22]) clearly shows that the further the appointed date is put back, the higher the chance that the patient will cancel the appointment or will not show up (no-shows). Beyond that, if the amount of reserved capacity for appointments is increased, less time for the walk-in blocks can be provided. Taking into account these advantages and disadvantages, one goal of a PCP might be to offer only as many appointment slots as necessary and thus be able to serve all patients requesting urgent treatment on the same day.

Appointment scheduling for primary care practices is a complex undertaking, due to the interdependencies of the appointment slots offered and the capacity for walk-ins. If patient demand could be anticipated, the reservation of appointment slots would allow the treatment of patients to be shifted in such a way that capacity restrictions are met. The ratio of time for appointments and time for walk-ins is, therefore, crucial for the performance of the system. In the literature, variations of this ratio are not discussed systematically and its effects have received little attention. Thus, highly disparate patient occurrence cannot be avoided. As a result, long waiting times occur in peak hours while at other times the physician is underutilized. By partly controlling patient occurrence, waiting time as well as PCP utilization can be improved significantly. This contribution focuses on the tactical decision of how many appointment slots have to be offered for each workday in a week, given that the physician has to see all urgent patients, while other patients can be shifted. We analyze how the number and position of appointment slots on different weekdays affect the ability of a PCP to satisfy same-day demand. Therefore, we develop an optimization model for appointment scheduling that meets both patient and physician preferences and we present the trade-offs between both sides and deduce managerial insights. Solutions of the developed optimization models are evaluated by a comprehensive stochastic simulation model.

The paper is structured as follows: Section 2 gives a review of relevant literature on appointment scheduling in the primary healthcare sector. In Section 3, a new optimization model *ORCA* (Optimal Reservation of Capacity for Appointments) for interday appointment scheduling with consideration of patient preferences is developed. It is extended by $ORCA^{+}$ taking the PCP objectives into account. Section 4 provides results and implications from an exemplary case study. A stochastic simulation model was implemented for the evaluation. The conclusion in Section 5 summarizes the new findings and managerial insights and gives a short overview of potential future work in terms of enhancing interday appointment scheduling.

2 Appointment scheduling in the literature

2.1 Patient structure

Different types of requests exist in primary care. In contrast to, for example, the U.S. [8], in Germany it is quite common to reserve time for walk-in patients showing up without an appointment and advanced notification, which will be considered in this contribution. Patients with *urgent* requests must receive care on the same day or as soon as possible and be admitted directly during a walk-in block. Urgent or same-day requests are often categorized in terms of medical reasons or preferences by the patients. Their main objectives are both a short waiting time and treatment on the day of request [8]. We assume that all urgent patients decide to walk-in. If such urgent patients cannot be treated due to a lack of capacity, they have to be shifted to the next day. This should be avoided whenever possible. We refer to urgent patients being shifted on the day after they requested a treatment as *urgent overflow patients*. The difference between patient arrival time and the time the patient is served by the PCP on the same day is the *direct waiting time*. *Routine* patients require follow-up treatment at a later day and need appointments after current treatment has been finished. Their main objective is to get an appointment on a preferred day in the future. The time between a patient requesting an appointment and the time of that appointment is referred to as *indirect waiting time*. In the literature, a predetermined and fixed differentiation of urgent patients and patients requesting an appointment is assumed [1, 8, 9]. However,

in real-life - especially with primary care physicians - there are *regular* patients who want to be consulted by the PCP, but prefer to get a scheduled appointment on short notice compared to coming in on the same day in order to avoid a long wait during the walk-in block. Their main objective is to get a scheduled appointment over the following days. Willingness on the part of regular patients to wait for an appointment not only depends on medical reasons but also on personal preferences. Depending on urgency and status in the system, regular patients will decide to come directly to the walk-in block or to take an appointment at a later date. Note that the difference between regular and urgent patient is that none of the urgent requests will be assigned to an appointment since their need for a treatment is so pressing that a delayed appointment is not suitable. This is the case, for example, if a patient needs a sick certificate for his or her employer. Figure 1 visualizes the relationship between the different patient types considered.

There is, accordingly, a number of patients who join the walk-in block without an appointment and a share of scheduled patients with an appointment for a later date - all sharing the same capacity. Emergency patients, for example, are not considered because they are treated by emergency doctors. For further patient differentiation, which are not relevant to primary care under consideration, see also [13].

2.2 Intraday scheduling

The literature that deals with scheduling different types of patients reflects a topic of much debate over the past decades and has been growing over the last few years. A good number of reviews and surveys provide an overview of various

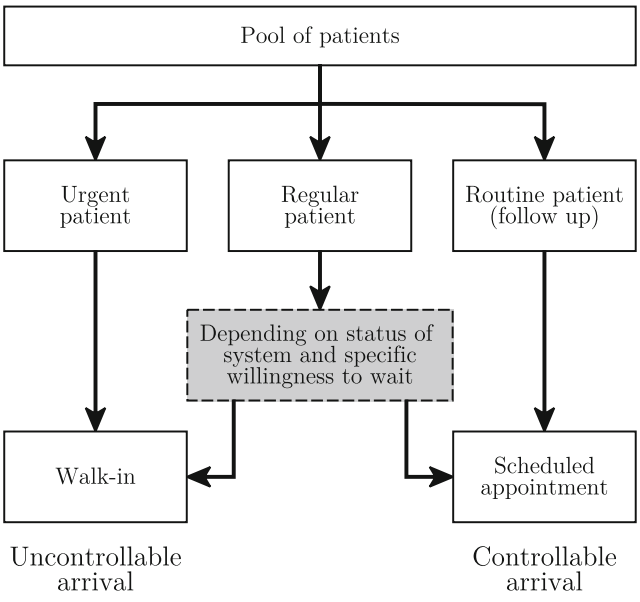

Fig. 1 Structure of patient types in primary care

models featuring distinctive key aspects, and we refer here to [5, 12, 29]. There are several appointment scheduling systems which can be broadly characterized as the traditional model, the carve-out model and the advanced-access model [20]. In the traditional model, the entire working time of the PCP is completely booked up at the beginning of the workday. These appointments have been assigned in advance and urgent patients are often scheduled via appointment double booking to avoid idle time in case of no-shows. The carve-out model operates in a similar way to the traditional model, but reserves a fixed capacity for urgent care. In contrast, the idea behind the advanced-access or same-day scheduling model is to schedule patients within 12–72 hours of request. The key principle is to see patients as soon as possible. For a deeper discussion about the advantages and disadvantages of each appointment scheduling system, see [8, 20, 25]. Feldman et al. [9] describe this stream of literature as *intraday* scheduling with the decision to arrange patient appointments within a day and with the objective of finding a good trade-off between direct waiting time and physician utilization. The main aim of intraday scheduling is to assign arrival requests over the day counteracting no-shows, delays and uncertain service times with predefined assignment rules (e.g. [6]).

2.3 Interday scheduling

Like Feldman et al. [9], our work concentrates on a higher planning level called *interday* scheduling with a focus on allocation and management of daily service capacity. They maximize the net revenue per day, calculated as the revenue obtained from each patient treated and the costs related to the service of scheduled patients. Patient preferences are considered for assigned appointments, no-show rates and patient cancellations. A static mathematical programming model is developed and extended to a dynamic one. In contrast to our approach, they assume a fixed number of urgent and routine patients and therefore a fixed number of appointments, whereas we consider the fact that regular patients either walk in or choose an appointment. Otherwise, there are only few other papers focusing on interday scheduling. Qu et al. identify the optimal percentage of appointment slots that a practice should keep open over a day to maximize the number of patients treated with respect to no-shows [22]. Based on a fixed number of appointment slots available each day, the aim is to find the optimal constant number of slots for routine patients. By contrast, we allow a varying number of appointment slots to be scheduled each day depending on the respective average day and time-dependent patient requests. Dobson et al. assume a fixed number of scheduled appointments for each day and study different measures as a function of the number of slots reserved for same-day patients [8]. They formulate a

stochastic model to quantify **(i)** the *effect of reserving slots* for routine patients on the average number of urgent patients who are not treated during normal office time and **(ii)** the *average queue length* for routine patients. Without providing a concrete schedule, they determine upper and lower bounds for the number of appointment slots. Balasubramanian et al. present a mathematical framework and show that the location of appointment slots has a significant impact on urgent same-day treatments and the waiting time [1]. They evaluated different block policies for appointment slots and found out that a two-block policy with blocks in two clusters of early morning and early afternoon works well. The key difference to our work is that Balasubramanian et al. assume an equally fixed number of scheduled appointments on each day focusing on the location of these appointments. In this contribution, we additionally determine the optimal appointment capacity on each day, due to varying patient requests throughout the weekday.

2.4 Planning level

The complex decision-making process behind capacity planning in primary care takes place on three planning levels [2]. On the strategic level, decisions as to the capacity dimension are determined by the size of the clinic and overall treatment. On a tactical level, capacity allocation has to be taken into account. Decisions as to time for walk-ins as well as the number of scheduled appointment slots for the entire group of patients sharing the same capacity have to be taken. The assignment of each individual patient to specific time slots is carried out on the operational level. On an operational level it is impossible to vary capacity if PCPs and the staff do not wish to change working hours significantly. It is on this decision level that various scheduling policies for assigning specific patients can be tested [1, 15, 24, 27]. Based on a given capacity dimension, this contribution focuses on the optimal appointment capacity on the tactical decision level. The advantages of the suggested optimal tactical schedule are shown by a simulation study on the operational level given a well-known scheduling policy.

2.5 Performance criteria

In terms of evaluating a specific schedule, various performance criteria can be found in the literature [5, 17]. It is essential for patients who need to be seen by a PCP urgently to get treatment on the day of request [19, 21, 22]. Which is why the number of overflow patients is a crucial indicator for the service level. Furthermore, the waiting time in the practice is an important performance criterion [13]. For other patient performance criteria, we refer to [5]. Next to patient preferences, there are PCP preferences such as high utilization and no overtime - aspects which are not necessarily associated with patient preference [26]. A PCP preference for high and balanced utilization will lead to a large number of patients per day and might result in long waiting times and low capacity for urgent patients who need same-day treatment. At times of high utilization, therefore, it is likely that patients will be shifted to the next day. Even though a PCP might prefer full utilization, overtime for the practice as a whole should be avoided. Clearly, there are some trade-offs which have to be taken into account when planning an appointment schedule for PCPs.

A weekly profile for appointments is generated that is thought to be in use every week. In order to test our optimal weekly profile as a framework for the operational level, we developed a simulation model to analyze the day-to-day situation of the system, where the relevant dynamics of the system and the influences of uncertainties on the operational level are modeled.

3 Optimal interday appointment scheduling

An optimal appointment schedule takes patients as well as PCP preferences into account. We present an optimization model for Optimal Reservation of Capacity for Appointments (*ORCA*) to support time management on a tactical level. The innovative concept behind *ORCA* is an efficient linear optimization approach that tackles the most critical task facing appointment scheduling in primary care: matching capacity and demand in such a way that walk-ins do not have to be moved to the next day. The option of offering appointments is exploited to even out patient occurrence during the week. Appointments are therefore offered on days with lower patient demand in order to save as much capacity as possible for treatment of walk-ins on days with high demand. The number of appointments which have to be offered is determined, while capacity limitations are met at all times, so that no overflow patients occur and the maximum capacity for possible walk-ins is provided in order to take non-scheduled patients into account as effectively as possible.

Since patients do not often change their PCP - even if a practice includes several PCPs [3] - we focus on a single server in a primary care practice. Five workdays with a given demand and capacity are modeled, where the number of appointment slots on each day has to be predetermined. The schedule determined gives back the optimal number of appointments for each day in a week resulting in a weekly profile. The *ORCA* thus supports the decision maker deploying a tactical plan for the appointment schedule as a weekly profile which is valid for several weeks or months as long as the demand remains comparable. We assume that the

demand for treatment can be anticipated using mean values and develop a deterministic model. Note that our model will not schedule individual patients; it rather determines the capacity for appointments that should be offered on any given day in a week.

In Section 3.1, a first model for an idealized situation is developed presenting the overall idea. A detailed discussion of the patient structure and the patient booking process in the optimization models is provided in Section 3.2 before the *ORCA* model is described in Section 3.3. In order to consider PCP preferences as well, the extended model $ORCA^+$ is presented in Section 3.4.

3.1 Configuration of a weekly appointment profile

This section develops a modeling approach to determine a weekly profile for appointments which takes into account the effects of shifting patients from any given day to another day or even another week of treatment via the use of appointments. The problem of matching capacity and demand in primary care is structured and the idea of interday appointment scheduling is presented.

Problem description and basic structure of appointment profile strategies

Figure 2 depicts a given week with a fixed working-time capacity $(c_1, \ldots, c_5)$ illustrated by the filled boxes. The boxes show a varying number of patient requests on different days during the week - on Mondays, for instance, when demand exceeds capacity for treatment. In Fig. 2a, consequences are presented for the situation where no appointments are offered and all patients are treated as walk-ins: If overtime is disallowed, patients have to be treated on Tuesday even though they requested a treatment on Monday because demand exceeds capacity. Due to this delay in treatment, those patients become overflow patients. As a consequence, not all patients from Tuesday can be treated on the same day since there are too many overflow patients from Monday to be seen by the PCP. In other words, some Tuesday requests will be handled on Wednesday, even though capacity and demand on Tuesday would match without overflow patients from Monday. The occurrence of overflow patients shows that the appointment schedule of single days has an impact on the other days and so these appointments have to be considered simultaneously. If a PCP follows a strategy without offering any appointments, a high number of overflow patients, long waiting times and unbalanced utilization can occur systematically. In contrast, Fig. 2b illustrates how overflowing demand of Monday could be scheduled to Wednesday via the appointment slots offered. This means that we can satisfy capacity constraints on Monday, increase utilization on Wednesday and avoid overflow patient occurrence, while all Tuesday requests are met on the same day. The basic idea of shifting demand in order to match capacity will now be presented in a mathematical model which minimizes the number of patients that have to be shifted by use of appointments. Note that in a first step we will assume to be able to explicitly appoint selected patients to specific appointments.

Capacity reservation for appointments in a weekly profile

In order to match capacity with demand on each workday, the idea of appointment scheduling in primary care practices is to offer as many appointments as necessary. The idea of offering more appointment slots on days with fewer requests was first proposed by Rising/Baron/Averill [24]. They tested different scheduling policies, whereas in our approach an optimal appointment capacity for scheduled patients is determined. If a decision maker can schedule each patient to an appointment, robust schedules with a constant utilization can be determined [23].

We start with the assumption that the decision maker is able to decide which patient gets an appointment for a specific day. We first consider only the regular patients who

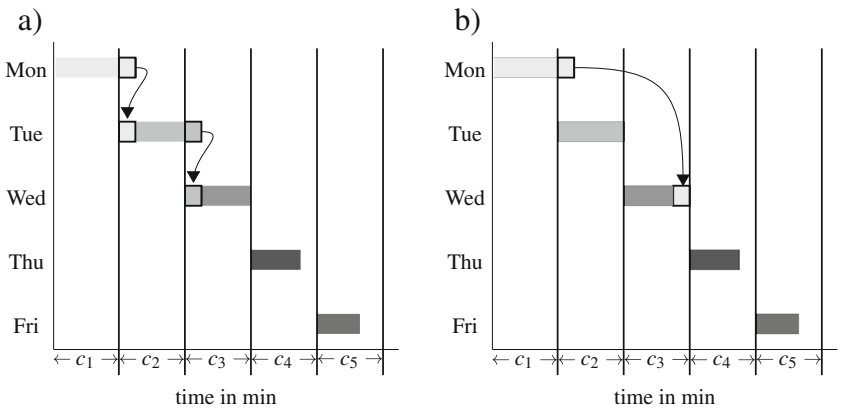

Fig. 2 Strategies for interday appointment scheduling in order to allocate capacity for appointment slots. **a**) A weekly profile without any appointment slots: overflow patients are shifted from Monday to Tuesday and from Tuesday to Wednesday. By this, some patients have to be shifted as overflow patients to the next day. **b**) Parts of the patients' requests from Monday were scheduled to an appointment slot on Wednesday

would take an appointment. In the following, we look at a representative repeating week for which the capacity for appointments is reserved and dynamics between the weeks are considered. Since we seek to model a weekly profile, we introduce the index set $\mathcal{I} := \{1, \ldots, 5\}$ in order to model the five workdays from Monday to Friday. We assume to know the number of patient requests per day i denoted by d_i. Let c be the capacity a PCP can treat on one day (in minutes) and let b be the service time for one patient. Then $d_i\, b$ gives the service time for all patients requesting a treatment on one day i. To match capacity with demand, $d_i\, b \leq c$ has to hold true for all $i \in \mathcal{I}$. Note that the number of patient requests can vary during the week while capacity is constant in this case. For cases in which the upper restriction does not hold true, we introduce the decision variable $y_{i,k} \in \mathbb{N}_0$ describing the number of requests per day $i \in \mathcal{I}$ being scheduled to day $k \in \mathcal{I}$. If patient requests can be scheduled by the decision maker, $y_{i,k}$ describes the number of patients being shifted from day i to day k. In other words, it gives back the number of appointment slots on day k assigned with patients from day i. As a result, the demand in day i can be reduced so that $d_i - \sum_{\ell \in \mathcal{I}} y_{i,\ell}$ describes the number of patients requesting on day i, who are treated on the same day. We define the number of patients who have a request on a certain day i and who are not scheduled to any appointment as walk-ins $w_i \in N_0$: $w_i = d_i - \sum_{\ell \in \mathcal{I}} y_{i,\ell}$. The overall number of patients being treated on day i is the sum of all walk-ins plus the scheduled appointments on that day: $w_i\, b + \sum_{k \in \mathcal{I}} y_{k,i}\, b$ which has to be less or equal to the given capacity.

Note that this way of modeling a weekly profile features the consideration of carry-overs: patients who request an appointment in one week are scheduled to an appointment in one of the following weeks. If the number of requests on a Thursday is higher than the capacity ($d_4 b > c$), shifting requests from Thursday to any other day of the week allows the capacity restriction to be matched. And so some requests from Thursday can be scheduled to appointments on Wednesday by $y_{4,3}$ being positive. Note that in this case, if for $y_{i,k}$, i is greater than k, requests are scheduled to an appointment in the next week. The corresponding capacity for appointment slots is reserved in the weekly profile on day k. Otherwise, if $i < k$, requests are shifted from the day of request to a later day of the week. This basic idea of shifting demand from one day to another day in a weekly profile allows us to create an idealized model for every week which minimizes the number of appointment slots:

$$\min \quad \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{I}} y_{i,k} \tag{1}$$

$$\text{s.t.} \quad w_i = d_i - \sum_{\ell \in \mathcal{I}} y_{i,\ell} \qquad \forall i \in \mathcal{I} \tag{2}$$

$$w_i\, b + \sum_{k \in \mathcal{I}} y_{k,i}\, b \leq c \qquad \forall i \in \mathcal{I} \tag{3}$$

$$y_{i,k} \in \mathbb{N}_0 \qquad \forall i, k \in \mathcal{I} \tag{4}$$

$$w_i \in \mathbb{N}_0 \qquad \forall i \in \mathcal{I} \tag{5}$$

The basic model minimizes the number of appointment slots (1) in order to offer as much capacity as possible for walk-ins, as long as capacity constraints for each day are satisfied. The number of appointment slots is determined by the number of requests and the walk-in patients (2). The sum of walk-in patients and patients with appointments has to be equal to or less than the number of patients a PCP can treat on day i (3). The domains of the decision variables are described in (4) and (5).

The results support the decision maker to schedule the optimal number of appointments $\sum_{i \in \mathcal{I}} y_{i,k}$ on day k and leaving the rest open for walk-ins, which means offering a walk-in block for the rest of the day. If $y_{i,k} = 0 \ \forall i \in \mathcal{I}$ there will only be a walk-in block on day k and no scheduled appointments.

When applying these determined appointment slots, our operational dynamic simulation showed rather poor results. This is due to the fact that the first model neglects three crucial elements for a real-world PCP: Firstly, we assume the ability to schedule as many patients as we wish to certain time slots, neglecting that some urgent cases will not be willing to take an appointment, since they seek same-day treatment. Secondly, we neglect the occurrence of routine patients. Thirdly, we assume the shifting of patients to a certain day in the week. But in reality not all patients are willing to take an appointment on a given day. The next section tackles the last assumption and presents an approach where a PCP cannot schedule each request into an appointment slot on a specific day.

3.2 Booking process of appointments for regular patients

In this section, we describe the booking process of appointments in a primary care practice. The focus of this general idea is on regular patients. Such patients can either be scheduled to an appointment or will join the walk-in block. And so we distinguish between urgent and regular requests on day i in our modeling approach so that d_i^u and d_i^r denote the number of urgent and regular patient requests on day i.

The former decision variable $y_{i,k}$ is not suitable when it comes to reflecting patient influence on the booking process since the decision maker cannot guarantee that an appointment will be booked by a patient with a request for a specific day. Meaning that, if an appointment slot is

offered on a Wednesday, the decision maker has no influence on the patient's choice so that this appointment can also be booked via a patient request on, say, Monday or Thursday. Our optimization model anticipates this by offering as many appointments as necessary in order to assure a reasonable appointment capacity. While in Section 3.1 it was, for example, sufficient to schedule one appointment on Wednesday to shift a patient from Monday to Wednesday. Now, to anticipate patient choice, a Wednesday appointment could be chosen by any patient in the week so that more slots on Wednesday are necessary to get the intended effect by way of capacity smoothing. We assume that regular patients do not have a strong preference for certain days of the week, but will prefer a short indirect waiting time and accept the first appointment date offered by the PCP if the waiting time is shorter than their willingness-to-wait. Therefore, a regular patient calls with a request and decides depending on the actual status of the system, whether he or she takes an appointment or directly joins the walk-in block. Travel times are negligible so that patients who do not take an appointment immediately join the walk-in block. This assumption is reasonable for German PCPs. Note that long travel times to the clinic might affect the performance of the appointment schedule, which could be taken into account in future research. For an integration of patient preferences for a specific day and time for appointments in the model, see for example [9]. The following effect is observable: if an appointment slot is offered on a Wednesday, it can be booked by a patient from any day of the week. Since there is a varying number of patient requests during the week, chances vary that a slot is booked by patients with requests for different days. We define the *appointment request proportion* $p_i := \frac{d_i^r}{\sum_{i \in \mathcal{I}} d_i^r}$ as the ratio of the number of all regular requests on day i to the total regular requests during the week. If there is a slot free on a certain day, the chance that a patient from day i will take it is p_i. For example, if 40 regular patient requests out of a total of 160 regular patient requests per week arise on Monday, the probability that one appointment slot on a Wednesday or any other day of the week, including Monday is chosen by a patients' request from Monday equals $\frac{40}{160} = 25\ \%$. Thus, four appointment slots have to be offered in order to schedule on average one Monday request on a Wednesday. Figure 3 illustrates the probability of appointment booking by workdays: each appointment slot will be booked out of the pool of regular requests with varying weights (corresponding to the number of requests). The position of the appointment slot does not affect the probability of which day of request will be assigned to it. It is rather the case that the share of requests with respect to the pool of requests per week determines the ratio. As mentioned earlier, this approach considers carry-overs, meaning that Friday requests lead to an appointment on Wednesday (by a 15 % chance) in the following week.

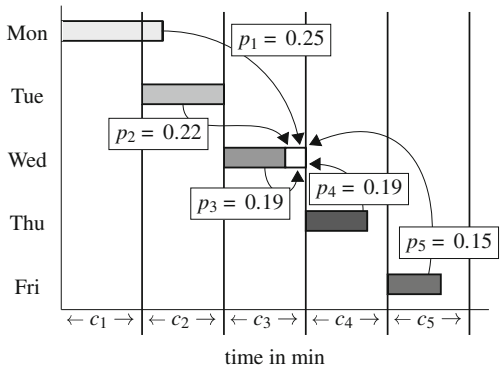

Fig. 3 Applying the appointment request ratio (p_i) to determine the probability that any appointment slot (here on a Wednesday) is taken by a regular patient request on day i

3.3 Reserving optimal appointment capacity to match capacity limits

Section 3.2 has shown that the basic model is not sufficient because patient preferences and thus the appointment request ratios have to be taken into account. In this subsection, we present *ORCA*, a linear mixed integer deterministic optimization model which minimizes the number of appointments in order to match capacity and patient requests. By determining the minimum number of appointments necessary in order to match capacity constraints, we find a solution that allows us to treat as many walk-in patients with same-day requests as possible. Interarrival times of such walk-ins cannot be influenced and PCPs can only substantially influence direct waiting times by reserving a reasonable share of capacity for walk-in patients. In this way, the approach particularly takes patients objectives into account. Instead of determining the day an individual patient is scheduled, *ORCA* determines the optimal appointment capacity on each day i to be offered in order to match capacity constraints.

To formulate the model, we introduce further notations for the parameters and variables involved. As described in Section 1, patient demand not only varies for different workdays but also throughout any day. Since there is a significantly higher demand in the morning than in the afternoon session, it is crucial to distinguish these two sessions. For this reason in line with Klassen and Rohleder [14], we consider the placement of slots either in the morning session or in the afternoon session for urgent patients. Each workday i is divided into two time periods, denoted by $j \in \mathcal{J} := \{1, 2\}$, where $j = 1$ indicates the morning session and $j = 2$ the afternoon session. This set can easily be extended to differentiate between more sessions per day.

***ORCA** – Optimal Reservation of Capacity for Appointments*

$$\min \quad \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{i,j} + \sum_{i\in\mathcal{I}} v_i \tag{6}$$

$$\text{s.t.} \quad (d^u_{i,1} + d^r_{i,1})\, b + x^r_{i,1}\, b - v_i\, b - p_{i,1} \sum_{\ell\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{\ell,j}\, b + x^{rout}_{i,1} b \le c_{i,1} \qquad \forall i \in \mathcal{I} \tag{7}$$

$$(d^u_{i,2} + d^r_{i,2})\, b + x^r_{i,2}\, b + v_i\, b - p_{i,2} \sum_{\ell\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{\ell,j}\, b + x^{rout}_{i,2} b \le c_{i,2} \qquad \forall i \in \mathcal{I} \tag{8}$$

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^{rout}_{i,j} = D^{rout} \tag{9}$$

$$v_i \le d_{i,1} \qquad \forall i \in \mathcal{I} \tag{10}$$

$$x^r_{i,j}, x^{rout}_{i,j} \in \mathbb{N}_0 \qquad \forall i \in \mathcal{I}, \forall j \in \mathcal{J} \tag{11}$$

$$v_i \in \mathbb{N}_0 \qquad \forall i \in \mathcal{I} \tag{12}$$

We assume that the capacity $c_{i,j}$ of the PCP per day i and session j has to be matched with the overall demand of patient requests per week which is subdivided according to the following aspects:

- *number of urgent patients* ($d^u_{i,j}$) who cannot be scheduled to a later appointment since they need to be seen by the PCP on the day of request,
- *number of regular patients* ($d^r_{i,j}$) who either join the walk-in block or who can be scheduled to a later appointment and
- *number of routine patients* (D^{rout}) who do not join the walk-in block and have to be scheduled to specific appointment slots.

Hence, the total *average* number of patients (D) being treated over a week is the sum of all urgent patients, of all regular patients over all weekdays and sessions, and of all routine patients:

$$D = \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} (d^u_{i,j} + d^r_{i,j}) + D^{rout}$$

Note that the day of request for regular and urgent patients will be explicitly modeled, whereas demand for routine patients will be modeled for the whole week. As mentioned before, routine patients will be scheduled to appointments which take place so far in the future that we assume patients are free on that date. Consequently, our optimization model has to ensure that the number of appointments for routine patients matches demand per week.

The decision variable $x^r_{i,j} \in \mathbb{N}_0$ denotes the number of appointment slots *for regular patients* planned for day i and session j. All other regular patients, who will not be scheduled to an appointment, will be defined as walk-in patients. The decision variable $x^{rout}_{i,j} \in \mathbb{N}_0$ denotes the number of appointment slots for routine patients. Note that the division of the workday into morning and afternoon sessions leads to a break in working time. We assume that walk-ins who arrive in the morning are willing to wait in order to be treated in the afternoon session. The decision variable $v_i \in \mathbb{N}_0$ denotes the number of walk-ins in the morning who are treated in the afternoon session on day i. In this way we incorporate the shift of patients from the morning to the afternoon session and define these patients as *intraday overflow patients*. For a given i and j, $(d^u_{i,j} + d^r_{i,j})b$ describes the time (in min) needed to treat all requests made by urgent and regular patients (when no appointments are offered to regular patients). $c_{i,j}$ is used for the capacity (in min) of the PCP on day i and session j.

The appointment request proportion can be adjusted by incorporating the morning and afternoon sessions as follows: $p_{i,j} = \frac{d^r_{i,j}}{\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} d^r_{i,j}}$. To ensure there is sufficient appointment time for patient requests on a certain day i and session j, the variables for appointments are multiplied by the respective proportion: $p_{i,j}\sum_{\ell\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{\ell,j}$. We can formulate the capacity constraints for a morning session (7) on friday ($i = 5$) as follows:

$$(d^u_{5,1} + d^r_{5,1})\, b + (x^r_{5,1} + x^{rout}_{5,j})\, b - v_5\, b - p_{5,1} \sum_{\ell\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{\ell,j}\, b \le c_{5,1}.$$

The left-hand side of the constraint denotes the duration of patients being treated in the morning on friday i. The total service time for patients requesting in the morning is expressed by $(d^u_{5,1} + d^r_{5,1})\, b$. Additional service

time for appointments (regular and routine) in the session is considered by adding $(x^r_{5,1} + x^{rout}_{5,j})\, b$. The service time for patients who booked an appointment is denoted by $p_{5,1} \sum_{\ell \in \mathcal{I}} \sum_{j \in \mathcal{J}} x^r_{\ell,j}\, b$ and subtracted on the left-hand side, as well as time for walk-ins who are shifted to the afternoon on the same day, described by $v_5\, b$. Service time for all patients in the morning has to be less or equal to the available time $c_{5,1}$. Capacity constraints for the afternoon are slightly different to the morning, since no patients can be shifted to a later session. Moreover, intraday overflow patients have in fact to be treated. Note that in this consideration with an average weekly appointment profile, we do not model any occurrence of overflow patients from one day to another. The linear integer optimization model *ORCA* can be formulated as follows:

Objective (6) minimizes the number of intraday overflow patients and the number of appointment slots for regular patients in order to have as much capacity as possible for walk-ins. The maximum amount of capacity is thus reserved for walk-ins per session. In equations (7) and (8) describe the capacity constraints. Constraint (9) reserves appointment slots for all routine patients per week. Constraints (10) prevent shifting more requests from the morning to the afternoon session than there are requests in the morning. The number of appointments (11) as well as the number of intraday overflow patients (12) have to be integer and non-negative.

The deployment of *ORCA* supports the decision maker on a tactical level. It determines the optimal appointment capacity that should be offered. In contrast to other appointment scheduling literature, there is no focus on the assignment of individual patients on an operational level to an appointment slot. The presented approach can easily be applied on an operational level by adding scheduling policies (e.g. [15, 24, 27]) to assign individual patients to specific time slots during a session.

In foregoing subsection, we present an innovative linear integer model that minimizes the number of necessary appointments and the number of intraday overflow patients. In the next subsection, we extend our model in order to incorporate PCP balanced utilization during the week and during daily sessions. In order to keep the following extension simple, we will not consider routine patients in the upcoming model. The integration is straightforward and easy to adopt.

3.4 Consideration of PCP preferences

After having introduced *ORCA*, the model is extended to a multi-criteria model called *ORCA*$^+$. By additionally integrating a balanced utilization throughout the day and week patient as well as PCP objectives are considered. PCP prefers the same number of treatments per day in order to have a balanced utilization during the week [17]. This is obtained when the relative difference between maximum available service time and actual service time is identical for all days and sessions. Therefore, the average working time has to be determined. Empirical investigations have shown a difference between the service time for patients with scheduled appointments and that for walk-in patients – with a significant lower service time for the latter [24]. Such information can be integrated into the optimization model in order to have a more realistic consideration of service times and we distinguish between duration times for walk-ins b^w and appointments b^a. The total capacity per week (in min) is denoted by $C := \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} c_{i,j}$. The overall average utilization $\bar{u}$ can then be calculated by dividing the service time for all walk-ins and appointments by the total capacity per week dependent on the respective schedules:

$$\bar{u} = \frac{1}{C} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \left((d^u_{i,j} + d^r_{i,j} - x^r_{i,j})\, b^w + x^r_{i,j}\, b^a \right).$$

We refer to the deviation of the average utilization $\bar{u}$ by $s^+_{i,j} \geq 0$ or $s^-_{i,j} \geq 0$. Where $s^+_{i,j}$ becomes positive if the actual utilization of session j on day i is above average utilization, $s^-_{i,j}$ is positive if the utilization is below average utilization. For morning sessions, Fig. 4 exemplarily visualizes the relation between the average working time per day and $s^+_{i,1}$ or $s^-_{i,1}$ with a high utilization on Tuesday morning ($s^+_{2,1} > 0$) and a relatively low utilization on Friday ($s^-_{5,1} > 0$).

By multiplying the average utilization $\bar{u}$ of a week with the capacity $c_{i,j}$ for day i and session j, we get the service time of a session if the PCP's utilization this day is in line with the average for the week. The utilization constraints for morning sessions can be set out as follows:

$$(d^u_{i,1} + d^r_{i,1})\, b^w + x^r_{i,1}\, b^a - v_i\, b^w - p_{i,1} \sum_{\ell \in \mathcal{I}} \sum_{j \in \mathcal{J}} x^r_{\ell,j}\, b^w - s^+_{i,1} + s^-_{i,1} = c_{i,1}\, \bar{u} \qquad \forall i \in \mathcal{I}.$$

The extended model *ORCA*$^+$ is formulated as follows:

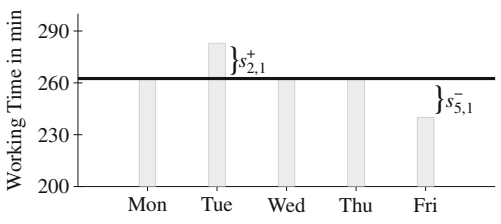

Fig. 4 Utilization ( ) in morning sessions with deviations ($s^+_{i,1}$ and $s^-_{i,1}$) from the average working time $\bar{u}\, c_{i,1}$ (—)

ORCA^{+}*– Optimal Reservation of Capacity for Appointments + balancing utilization for PCPs*

$$\min \quad \lambda\Big(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{i,j} + \sum_{i\in\mathcal{I}} v_i\Big) + (1-\lambda)\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}(s^+_{i,j} + s^-_{i,j}) \tag{13}$$

$$\text{s.t.} \quad (d^u_{i,1} + d^r_{i,1})\, b^w + x^r_{i,1}\, b^a - v_i\, b^w - p_{i,1}\sum_{\ell\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{\ell,j}\, b^w - s^+_{i,1} + s^-_{i,1} = c_{i,1}\bar{u} \qquad \forall i \in \mathcal{I} \tag{14}$$

$$(d^u_{i,2} + d^r_{i,2})\, b^w + x^r_{i,2}\, b^a + v_i\, b^w - p_{i,2}\sum_{\ell\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{\ell,j}\, b^w - s^+_{i,2} + s^-_{i,2} = c_{i,2}\bar{u} \qquad \forall i \in \mathcal{I} \tag{15}$$

$$(d^u_{i,1} + d^r_{i,1})\, b^w + x^r_{i,1}\, b^a - v_i\, b^w - p_{i,1}\sum_{\ell\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{\ell,j}\, b^w \le c_{i,1} \qquad \forall i \in \mathcal{I} \tag{16}$$

$$(d^u_{i,2} + d^r_{i,2})\, b^w + x^r_{i,2}\, b^a + v_i\, b^w - p_{i,2}\sum_{\ell\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^r_{\ell,j}\, b^w \le c_{i,2} \qquad \forall i \in \mathcal{I} \tag{17}$$

$$\bar{u} = \frac{1}{C}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}((d^u_{i,j} + d^r_{i,j} - x^r_{i,j})\, b^w + x^r_{i,j}\, b^a) \tag{18}$$

$$v_i \le d^u_{i,1} + d^r_{i,1} \qquad \forall i \in \mathcal{I} \tag{19}$$

$$x^r_{i,j} \in \mathbb{N}_0,\ s^+_{i,j}, s^-_{i,j} \ge 0 \qquad \forall i \in \mathcal{I}, \forall j \in \mathcal{J} \tag{20}$$

$$v_i \in \mathbb{N}_0 \qquad \forall i \in \mathcal{I} \tag{21}$$

Constraints (14) and (15) are adjusted with parameters for service time b^w and b^a and extended by multiplying the capacity of each day by the overall utilization $\bar{u}$, determined in constraints (18). (16) and (17) ensure that the PCP has no systematic overtime in any slot.

The objective function (13) integrates two criteria by a convex combination: minimizing the number of required appointments in a week and a balanced utilization during the week. This is incorporated by minimizing the sum of all deviation of utilization from the average $\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}(s^+_{i,j} + s^-_{i,j})$. By weighting the two-criteria (minimum number of appointments *and* controlling PCP utilization) with the parameter $\lambda \in [0, 1]$, the decision maker can adjust the model with respect to his or her preferences. We provide results with varying values for λ in the next section.

Weekly appointment profiles can be generated on the basis of few data which are also easy to acquire from reality – which means that this model can be used effectively with real-world applications. This tactical plan is suitable for use as the framework for an intraday appointment schedule, which has been the subject to several studies. With the results of *ORCA*$^{+}$, appointment scheduling on an operational level can assign individual patients to one of the given appointment slots. The focus on an average weekly scheduling profile allows the use of general average data, which can be anticipated to medium-term changes in patient demand. In contrast to the tactical level, there are significant stochastic influences which have to be considered on the operational level, such as varying service and interarrival times. As stated above the proposed approach on the tactical level is a deterministic formulation with the use of few data. In the next section, the results from this approach are analyzed in an operational stochastic framework in a simulation study. The focus of this contribution lies in the general innovative idea and an easily to implement deterministic model. Comparisons with stochastic extensions can be future research.

4 Analyzing the performance of weekly appointment profiles

In order to test our findings of the weekly profile as a framework for the operational level, we developed a simulation model to analyze the day to day situation of the system, where relevant dynamics of the system, influences of uncertainties and individual patient scheduling is modeled in detail. Especially the uncertainty with respect to patient arrivals and varying service times helps to analyze whether solutions generated with *ORCA* and *ORCA*$^{+}$ are well suited for real-world PCPs. In Section 4.1, the framework for the analysis is presented and the simulation model is described. Compared to the existing scheduling rules found in the literature – with a fixed number of appointments during the week – the advantages of the optimization models *ORCA* and *ORCA*$^{+}$ are shown in Sections 4.2 and 4.3.

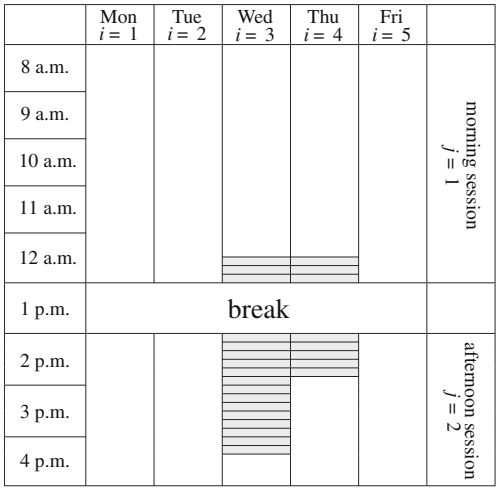

Fig. 5 Weekly appointment profile with the optimal number and allocation of offered appointments (appointment slots ▭ and walk-in blocks ▭)

4.1 Experimental environment

In this subsection, we present the framework for an exemplary case study and explain how results from the optimization models were evaluated. We consider a PCP where patients either call to make appointments for a visit at a later date or become direct walk-ins. Given the current appointment schedule, the administrative staff schedules each incoming request for an appropriate day and updates the schedule accordingly. The total number of appointments that can be scheduled in a session is constant over the whole year for one schedule tested.

Framework and parameters for the optimization

The PCP has a certain amount of patient requests on average per week and decides on the number of appointment slots per day and session. Each appointment slot can be booked. Appointment scheduling studies have shown that ten-minute appointment slots are the norm. As in Klassen/Rohleder, we model an eighthour day, resulting in a maximum of 48 possible appointment slots spread over a five-hour morning and three-hour afternoon session [15]. In accordance to Rising/Baron/Averill we assume a lower average service time for walk-ins of 8.5 minutes than the average service time for scheduled appointments of 10 minutes [24]. On average, we expect the number of patient requests to be 255 per week with a relative allocation on Monday / Tuesday (each 25%) and Wednesday / Friday (each 16.7%) [27]. The overall service time requested per week is slightly lower than the capacity per week (90% load) so that a steady-state status for the system can be attained. The amount of requests in the morning session is significantly higher than that in the afternoon session ($\frac{5}{6}$ to $\frac{1}{6}$). A multi-period environment with a five day period is modeled and patients can be scheduled on any of these five days, including the workday they call, in a following week. After they have been scheduled, patients directly arrive at the appointed time and day. Note that in this simulation we will focus on urgent and regular patients and not on routine patients – and so keep results simple and comprehensible. The incorporation of routine patients would, however, pose no difficulties.

Criteria

We test optimal weekly profiles obtained from the optimization models in an extensive simulation study using different criteria to quantify the quality of the solutions. On the patient side, we look at the average waiting time in a walk-in block, at the average waiting time across all patients in minutes and at the respective 95 percent quantiles. We also look at the number of overflow patients. From the PCP perspective, we focus on the sum of deviations from average utilization which is determined by $\sum_{i\in\mathcal{I},\, j\in\mathcal{J}}(s_{ij}^{+}+s_{ij}^{-})$

Table 1 Performance criteria for $\lambda = 1$, with the number of patients with a request on a day (rows) and the actual day of treatment (columns) for **a**) walk-ins and **b**) appointments

Day of	a) Walk-ins: Day of treatment						b) Appointments: Day of treatment					
request	Mon	Tue	Wed	Thu	Fri	OFLOW	Mon	Tue	Wed	Thu	Fri	App.
Mon	2249	257	0	0	0	257	0	0	238	21	0	259
Tue	0	2119	348	2	0	350	0	0	168	109	0	277
Wed	0	0	1135	550	2	552	0	0	64	103	0	167
Thu	0	0	0	1412	307	307	0	0	93	61	0	154
Fri	69	0	0	0	1634	69	0	0	142	13	0	155
$\sum$	2318	2376	1483	1964	1943	1535	0	0	705	307	0	1012

a) If the day of request deviates from the day of treatment, these patients are denoted as overflow patients. **b**) Treatments of patients with appointments are only on days with appointment slots, but day of request for such appointments can occur from Monday to Friday. All values below the main diagonal refer to appointments in a following week

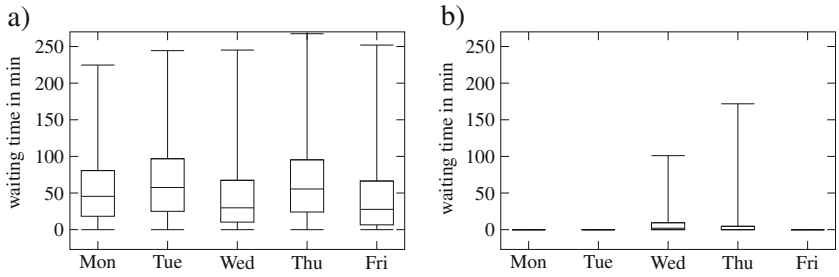

Fig. 6 Boxplots for the waiting time for **a**) walk-in patients and **b**) scheduled patients

and the average overtime per day in minutes. Note that the objective is to minimize each of these criteria.

Simulation

In order to evaluate and analyze the results from the models *ORCA* and *ORCA*$^+$, we deployed a discrete event simulation capable of modeling stochastic parameters and the dynamics of a queuing system in order to evaluate patient waiting time for an appointment or direct waiting time. As empirical studies suggest, we assume that the service time for walk-ins follows a lognormal distribution [5] with an expected value of 8.5 minutes and a standard deviation of five minutes, for scheduled appointments accordingly an expected value of ten minutes and a standard deviation of five minutes. Interarrival time for patient requests described as above follows an exponential distribution resulting in 255 patients per week. In the literature, proportions between 10 and 25% are assumed for urgent patients (e.g. [8, 15]). We assume 20% of all patients to be urgent patients seeking same-day treatment. Regular patients will either take an appointment or walk-in. This depends on their willingness-to-wait for an appointment and the actual status of the system. If indirect waiting time for an appointment is larger than the maximum willingness-to-wait, the request becomes a walk-in. Otherwise, an appointment is booked. By this, we model patient choice depending on the current status of the queuing system, namely the time one has to wait for an appointment. The maximum amount of days a patient is willing to wait is modeled by a Weibull distribution with $\alpha = 2$ and $\beta = 15$.

To avoid high overtime, walk-in patients are accepted if expected duration of treatment for patients in the waiting room is smaller than the remaining time in the walk-in block. Otherwise, the patient will be shifted to the next walk-in block on the same (or the next) day. Each simulation run represents one year with 52 weeks and $52 \times 255 = 13260$ patients (including a four week warm-up and a four week cool-down phase). For all results, simulation runs were repeated twenty times and expected values were calculated. For all of our results, the 95% confidence intervals were within at most 5% of their expected values.

System analysis showed for all results a steady-state status. Since parameters change for single workdays, the steady-state distribution through the week differs but for each individual workday it is the same over all weeks. Therefore, the output of the simulation cycled through the same five series of steady-states with respect to the five workdays so that a steady-state cycle is reached.

4.2 Results for ORCA tested in the simulation

This subsection presents results from our proposed new modeling approach, which explicitly takes interday scheduling into account. In order to emphasize the performance of the model, different evaluation criteria are used as pointed out in 4.1. The optimal solution of *ORCA* was determined using Fico Xpress-Optimizer 7.3 [10] and evaluated with the discrete event simulation tool WITNESS [18]. Note that the optimization models determine the optimal appointment capacity on each day i and session j whereas the appropriate location is chosen due to practical reasons. Klassen/Rohleder studied, on the one hand, a single period and found out that placing the appointment slots at the beginning of a day leads to shorter waiting times for all clients and also resulted in many untreated urgent patients [14]. On the other hand, the effect of planning urgent

Table 2 Number of treated patients and PCP utilization for $\lambda = 1$

Treated patients	Monday	Tuesday	Wednesday	Thursday	Friday
Walk-ins	2317	2376	1483	1964	1943
Scheduled	0	0	704	308	0
PCP utilization	89 %	91 %	87 %	90 %	75 %

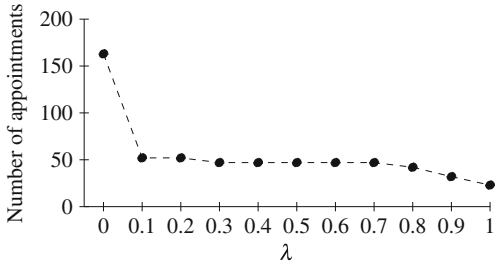

Fig. 7 Number of appointments for different λ values

patients at the end of a day works well for treating many urgent cases, but results in a significantly higher waiting time for all patients. Their study showed that placing urgent slots in the middle of the period results in the best overall performance. Furthermore, Rohleder/Klassen analyzed a dynamic situation and found almost the same results [27]. They explicitly analyzed the optimal placement of a predetermined number of urgent slots finding an optimal strategy with urgent appointments over the whole day. Nevertheless, their analysis shows that the final allocation of urgent slots during the day has no major impact on performance. Following Balasubramanian et al. and Klassen/Rohleder, the appointments from the optimization models are scheduled directly before and after the break [1, 14]. The optimal appointment capacity is visualized in Fig. 5. The minimum number is equal to 23 appointments allocated on Wednesday and Thursday.

The results of the key performance measures were recorded and the expected values are used to evaluate the different schedules. Table 1 presents the expected values for different performance criteria where a) shows that minimizing the number of appointments leads to a sum of 1535 overflow patients. With respect to one workday, this means we have less than six overflow patients per day. Thus, a large share of patients can be treated on the day even if they call in the afternoon and throughout the year only 15% of the walk-in patients are shifted by appointments. Due to the appointments offered, patient requests from Monday (238) and Tuesday (168) will be shifted to Wednesday and 130 (Mon + Tue: 21 + 109) patients to Thursday (Table 1 b)), to gain more flexibility for walk-in patients. Note that for example 13 patients requesting on Friday accept an appointment on Thursday in a following week.

Patient waiting time is an important criterion besides the number of overflow patients. Fig. 6 shows boxplots for the waiting time for walk-in patients (left) and patients with appointment (right). As expected, the average waiting time for scheduled patients is substantially lower than the waiting time during walk-in blocks. Furthermore, our results show that the average waiting time on Mondays and Tuesdays is longer than on the other days corresponding to a higher overtime for the physician on Mondays and Tuesdays with an average of 13.2 minutes. This is in line with Rohleder/Klassen, depending on the scheduling policy [27].

The strength of this solution is a low number of appointments, no substantial overtime for the PCP and very few overflow patients. The utilization of the PCP is not modeled in *ORCA* whereby *ORCA* is equivalent to *ORCA*$^+$ with $\lambda = 1$. The results for the utilization are presented in Table 2 and show that utilization can be shifted from Monday and Tuesday to Wednesday and Thursday by offering appointment slots. Even so, utilization varies between 75% on Friday and 91% on Tuesday, a rather unbalanced workload for the PCP during the week. We present the results of our extended model *ORCA*$^+$ to test whether a solution can be determined that serves patient and PCP goals.

4.3 Comparison of results and analysis for *ORCA*$^+$

ORCA$^+$ considers a PCPs' balanced workload throughout the week and day. Corresponding results are presented and

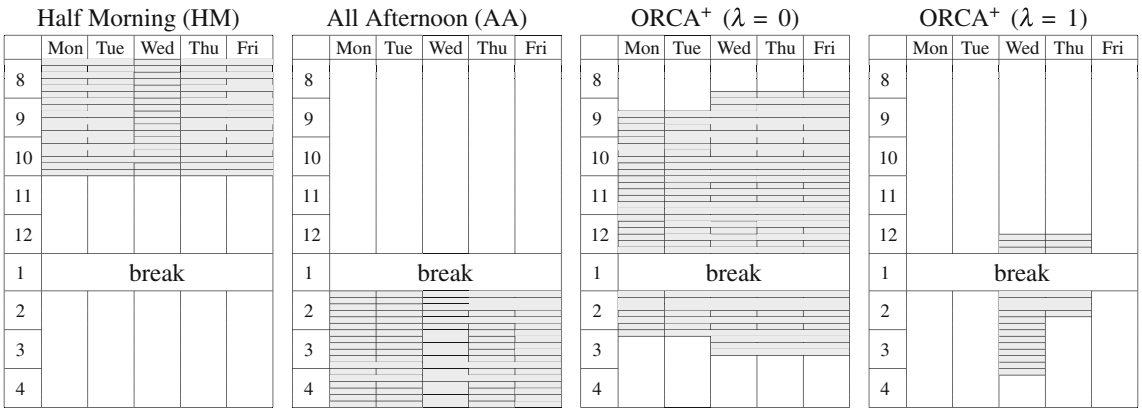

Fig. 8 Overview of different weekly appointment profiles tested (appointment slots ▭ and walk-in blocks ▭)

compared to other solutions in this subsection. Depending on the value of λ, the objectives *overflow patients* (λ = 1) and *a balanced PCPs' workload* (λ = 0) can be considered simultaneously (see again the objective function (13)).

$$\min \quad \lambda\Big(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} x^{r}_{i,j} + \sum_{i\in\mathcal{I}} v_i\Big) + (1-\lambda)\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}(s^{+}_{i,j} + s^{-}_{i,j})$$

Fig. 7 presents solutions of $ORCA^{+}$ for varying λ. In comparison to the solutions of *ORCA* (λ = 1) many more appointments are offered when choosing λ = 0. As expected, appointments can be used to balance the utilization and to shift patients.

Compared to the existing scheduling rules in the literature with a fixed number of appointments during the week, advantages of $ORCA^{+}$ are shown. Appointment rules of Balusabramanian et al. are used, namely *half morning* (HM) and *all afternoon* (AA) policies and compared with the results of the optimization models for λ = 1 and λ = 0 [1]. The different scheduling policies are visualized in Fig. 8.

Crucial performance results for each scheduling alternative are summarized in Table 3. Differentiated by day, number of appointments, average number of overflow patients, the average utilization deviation, waiting time for walk-in patients with the corresponding 95% quantile and average overtime are presented. Results show that the objective for λ = 1 is achieved with a small number of appointments corresponding with few overflow patients during the week compared to the other scheduling alternatives. Next to that, the second lowest waiting time on average and for its 95% quantile is observable for λ = 1. If we focus solely on a PCP balanced workload, the advantages of $ORCA^{+}$ for λ = 0 become clear: deviation from average utilization is on average only 2.82 minutes and substantially lower compared to the other scheduling rules. Of course, this scheduling strategy results in a high number of overflow patients but still leads to the lowest average waiting time for walk-ins in comparison to the HM and AA strategy. By varying the weighting parameter λ in the multi-criteria model, decision maker preferences can easily be adopted – which renders our model suitable for different decision maker objectives and generally valid.

Further simulation studies have shown that decision support is of special importance in case of high utilization of the PCP as is 90 % in this study. With growing demand, the number of overflow patients and length of waiting time increase considerably. That being so, allowing even a little overtime reduces these values to an acceptable level again so that a PCP can react to short-term changes in demand [28]. The weekly profile generated for the appointment schedule has been optimized by the use of mean data

Table 3 Comparison of evaluation criteria for different schedules

	HM	AA	$ORCA^{+}$ λ = 0	$ORCA^{+}$ λ = 1
# appointments				
Mon	18	18	29	0
Tue	18	18	29	0
Wed	18	18	35	16
Thu	18	18	35	7
Fri	18	18	35	0
total	**90**	**90**	**163**	**23**
mean # of overflow patients				
Mon	10	16	14	6
Tue	16	22	15	8
Wed	10	15	11	13
Thu	6	11	11	7
Fri	3	9	11	2
total	**45**	**73**	**62**	**36**
mean deviation from $\bar{u}$ (minutes)				
Mon	12.38	7.14	3.85	14.32
Tue	14.63	15.6	1.76	23.1
Wed	7.03	8.1	3.43	5.31
Thu	10.04	9.92	2.4	11.05
Fri	23.67	20.92	2.68	53.04
average	**13.55**	**12.34**	**2.82**	**21.36**
mean waiting time during walk-in block (minutes)				
Mon	70.7	55.4	42.9	53.66
Tue	73.16	86.3	43.52	65.1
Wed	69.91	98.08	31.42	45.69
Thu	63.03	72.88	30.72	64.68
Fri	56.02	54.34	31.45	44.08
average	**66.56**	**73.40**	**36.00**	**54.64**
waiting time during walk-in block 95% quantile (minutes)				
Mon	155.71	153.5	96.08	136.3
Tue	160.53	200.67	96.19	155.92
Wed	156.26	228.12	72.87	143.72
Thu	146.65	207.4	72.32	161.27
Fri	133.46	177.73	73.97	147.97
average	**150.52**	**193.48**	**82.29**	**149.04**
mean overtime per week (minutes)				
Mon	7.02	14.62	7.67	4.93
Tue	21.27	20.72	25.97	10.58
Wed	19.6	20.65	29	14.83
Thu	18.43	19.76	28.68	13.99
Fri	15.73	17.44	28.63	4.82
total	**82.05**	**93.19**	**119.95**	**49.15**

and the tactical plan is tested in a stochastic environment using Monte-Carlo simulation for operational realization. The results of the case study strongly suggest that a tactical strategy using *ORCA* or *ORCA*$^+$ outperforms the use of other strategies which do not anticipate varying demand throughout the week by adjusting the number of appointment slots offered in a session. The proposed new model not only has a positive impact on the avoidance of overflow patients, but also decreases the waiting time significantly and leads to more balanced utilization. It is evident that minimizing the number of appointments leads to a high capacity for urgent patients and therefore to a high rate of patients treated on the day of request. Remember that the simulation study as presented refrained from modeling routine patients. Considering additional appointment slots for routine patients, significantly more potential to match capacity and demand is offered, since routine patients can be planned well in advance.

5 Conclusion and outlook

This contribution set out to present an innovative method for interday appointment scheduling. In the primary care sector, *ORCA* and *ORCA*$^+$ offer decision support on a tactical level where demand and capacity have to be matched. We determined the optimal appointment capacity that should be offered during a week. This aspect has received little attention in the literature until now. The analyses provide managerial insights into real-world applications and physicians can use our model to determine the optimal appointment capacity based on a small amount of data that exists in almost every practice. The innovative contribution of this approach is that the ratio of time offered for appointments and time for walk-ins is not kept fixed, but will be determined in an optimization model that performs well even on the basis of few data. The integration of our model into appointment scheduling increases patient satisfaction as well as practice objectives such as short waiting time, balanced utilization and little overtime.

ORCA considers patient preferences by minimizing the number of overflow patients in the system, while practice capacity constraints are satisfied. It has been shown that offering of appointment slots partly enables patients to be shifted from the more frequented to the less frequented days. In this way, the practice can handle as many urgent patients as necessary on the same day, resulting in a small number of overflow patients. *ORCA* is extended to *ORCA*$^+$ which considers additionally practice objectives. A simulation study has shown that objectives can be attained satisfactorily, even with a small amount of data. This is especially valuable for real-world applications since data such as number of requests and average service times are easily available, while further information about patient preferences are difficult to record. If information on varying relative lengths for treatment are available the model can be adopted by e.g. integrating varying treatment duration for first time and follow-up patients. By this, the generated appointment schedule can be very well suited for a specific PCP with an individual patient pool. An incorporation of different travel times from request to arrival could generalize the model additionally. Furthermore, our model can be extended to take varying urgency and different patient preferences into account. Modeling patient preferences with respect to a special day or time can make the model even more realistic. We assumed that each request has an equal chance of getting an appointment, no matter on which day requested. In reality, a large number of urgent patients is likely on Mondays because people need sick certificate for his or her employer or important primary care treatment. Fewer requests from Monday are scheduled to an appointment. As a result, the consideration of the varying urgency of patients with requests on different days can provide valuable insights in order to promote the quality of strategies for interday appointment scheduling even further. The deterministic modeling approach of *ORCA* and *ORCA*$^+$ is straightforward and therefore it is easy to incorporate to real world application. Furthermore, results from the exemplary case study are very promising and it has been shown that this deterministic approach on a tactical level is suitable. In future studies a stochastic extension of our approach to optimally allocate capacity can be developed and it should be tested whether results can be additionally improved.

While our models focus on the optimal appointment capacity offered on a tactical level, where demand and capacity have to be matched, intraday scheduling optimizes the scheduling of patients on a single day to a specific time slot. In essence, interday planning considers capacity and workload constraints; intraday planning takes short-term uncertainties such as no-shows into account. The combination of the presented *ORCA*$^+$ models with intraday planning would seem to be a highly promising strategy. As we expect substantial improvements when combining these approaches, such incorporation could also be the subject of future research.

Acknowledgments The authors gratefully acknowledge the helpful comments of the associate editor and referees. Thanks are also due to Priv.-Doz. Dr. med. Birgitta Weltermann, Institut für Allgemeinmedizin, Universiätsklinikum, Essen, Germany, for providing valuable insights into medical practice.

References

1. Balasubramanian H, Biehl S, Dai L, Muriel A (2014) Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments. Health Care Manag Sci 17(1):31–48
2. Balasubramanian H, Muriel A, Ozen A, Wang L, Gao X, Hippchen J (2013) Capacity allocation and flexibility in primary care. In: Denton BT (ed) Handbook of healthcare operations management, volume 184 of International Series in Operations Research and Management Science, pp 205–228
3. Brailsford S, Kozan E, Rauner MS (2012) Health care management. Flex Serv Manuf J SI 24(4):375–378
4. Cao W, Yi W, Haibo T, Shang F, Liu D, Tan Z, Sun C, Ye Q, Xu Y (2011) A web-based appointment system to reduce waiting for outpatients: A retrospective study. BMC Health Serv Res:11
5. Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. Prod Oper Manag 12(4):519–549
6. Cayirli T, Veral E, Rosen H (2006) Designing appointment scheduling systems for ambulatory care services. Health Care Manag Sci 9(1):47–58
7. Degel D, Wiesche L, Rachuba S, Werners B (2014) Time-dependent ambulance allocation considering data-driven empirically required coverage. Health Care Manag Sci:1–15
8. Dobson G, Hasija S, Pinker EJ (2011) Reserving capacity for urgent patients in primary care. Prod Oper Manag 20(3):456–473
9. Feldman J, Topaloglu NLH, Ziya S (2014) Appointment scheduling under patient preference and no-show behavior. Oper Res 62(4):794–811
10. Fico Xpress-optimizer referenz manual. Fico Xpress Optimization Suite (www.fico.com), Release 20.00, 3 June 2009
11. Gallucci G, Swartz W, Hackerman F (2005) Brief reports: impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. Psychiatr Serv 56(3):344–346
12. Gupta D, Denton B (2008) Appointment scheduling in health care: challenges and opportunities. IIE Trans 40(9):800–819
13. Harper P, Gamlin HM (2003) Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. OR Spectrum 25(2):207–222
14. Klassen KJ, Rohleder TR (1996) Scheduling outpatient appointments in a dynamic environment. J Oper Manag 14(2):83–101
15. Klassen KJ, Rohleder TR (2004) Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. Int J Serv Ind Manag 15(2):167–186
16. Krueger U, Schimmelpfeng K (2013) Characteristics of service requests and service processes of fire and rescue service dispatch centres. Health Care Manag Sci 16:1–13
17. LaGanga L, Lawrence S (2007) Clinic overbooking to improve patient access and increase provider productivity. Decis Sci 38(2):251–276
18. Lanner Witness simulation tool. Power with Ease Release 1.0, Witness (www.lannersimtech.de), June 2009
19. Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. M&SOM-Manufacturing & Service Operations Management 12(2):347–364
20. Murray Mark, Berwick DM (2003) Advanced access: reducing waiting and delays in primary care. JAMA 289(8):1035–1040
21. Ozen A, Balasubramanian H (2013) The impact of case mix on timely access to appointments in a primary care group practice. Health Care Manag Sci 16(2):101–118
22. Xiuli Q, Rardin RL, Williams JAS, Willis DR (2007) Matching daily healthcare provider capacity to demand in advanced access scheduling systems. Eur J Oper Res 183(2):812–826
23. Rachuba S, Werners B (2015) A fuzzy multi-criteria approach for robust operating room schedules. Ann Oper Res:1–26
24. Rising EJ, Baron R, Averill B (1973) A systems analysis of a university-health-service outpatient clinic. Oper Res 21(5):1030–1047
25. Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. M&SOM-Manufacturing & Service Operations Management 12(2):330–346
26. Rohleder TR, Klassen KJ (2000) Using client-variance information to improve dynamic appointment scheduling performance. Omega 28(3):293–302
27. Rohleder TR, Klassen KJ (2002) Rolling horizon appointment scheduling: a simulation study. Health Care Manag Sci 5(3):201–209
28. Schacht M, Wiesche L, Werners B, Weltermann B (2015) Influence of appointment times on interday scheduling. In: Operations Research Proceedings 2015, accepted for publication
29. De Vuyst S, Bruneel H, Fiems D (2014) Computationally efficient evaluation of appointment schedules in health care. Eur J Oper Res 237(3):1142–1154