Adaptive Resource Allocation for Efficient Patient Scheduling [⋆]

I.B. Vermeulen[1], S.M. Bohte[1], S.G. Elkhuizen[2], J.S. Lameris[2], P.J.M. Bakker[2], and J.A. La Poutré[1]

[1] Centre for Mathematics and Computer Science (CWI),
Kruislaan 413, Amsterdam, the Netherlands.
[2] Academic Medical Center, University of Amsterdam,
Meibergdreef 9, Amsterdam, the Netherlands.
I.B.Vermeulen@cwi.nl

Abstract. Efficient scheduling of patient appointments on expensive resources is a complex and dynamic task. A resource is typically used by several patient groups. To service these groups, resource capacity is often allocated per group, explicitly or implicitly. Importantly, due to fluctuations in demand, for the most efficient use of resources this allocation must be flexible. We present an adaptive approach to automatic optimization of resource calendars. In our approach, the allocation of capacity to different patient groups is flexible and adaptive to the current and expected future situation. We additionally present an approach to determine optimal resource openings hours on a larger time frame. Our model and its parameter values are based on extensive case analysis at the AMC hospital. We have implemented a comprehensive computer simulation of the application case. The results of our simulation experiments show that our approach can effectively schedule patients groups with different attributes and make efficient use of capacity.

Key words: patient scheduling, capacity planning, dynamic optimization, simulation

1 Introduction

High patient service levels are becoming increasingly important in the hospital. At the same time, the demand for health care is increasing and, more and more patients must be treated with the same limited capacity and budget. High efficiency on resources is necessary to provide patients with high quality care including short access times. To this end, improvements on all levels of hospital operations must be made, from strategic innovations and adjustments, to improved day to day scheduling [2].

Efficient scheduling of patient appointments on expensive resources is a complex and dynamic task. Traditional approaches to logistical improvement for increased efficiency are usually not easily applied to the medical domain. In patient scheduling, we have to consider that hospital resources are many: ranging from CT and MRI scanners,

[⋆] A preliminary version of this paper has appeared as [1] to hospital beds, to attending staff, to operating rooms. To achieve a high hospital-wide patient throughput, local resources must maintain short access times. For some resources, this is a more complex problem than for other resources.

A resource is typically used by several patient groups with different properties [3]. Groups can be distinct based on referring departments, inpatients (admitted to the hospital) or outpatients (not admitted), medical constraints, and level of urgency [4]. To allow different norms on what is an acceptable access time per patient group, hospital resource capacity is allocated per group, explicitly or implicitly. However, due to variation in demand, determining the optimal allocating is a complex problem. Importantly, to achieve the best performance, this allocation must be dynamic.

We study the case of scheduling Computer Tomography scans (CT-scans) at the radiology department within the Academic Medical Centre Amsterdam (AMC). Diagnostic resources such as the CT-scanners are literally central in the clinical pathways of many patients. Long access times to such resources are immediately felt as bottlenecks for health care processes in the hospital. In recent years the whole logistical process around the CT-scan in the AMC has already improved substantially [5]. The actual scheduling of appointments is (still) done by human schedulers: they select a timeslot on the resource calendar for each patient, given the scheduling restrictions due to the allocation of capacity. There is often a lack of overview on how these low-level scheduling decisions influence overall performance.

A calendar supervisor determines a long time in advance how to allocate scanner capacity, based on experience, future expectation, and in cooperation with medical experts. Often, the actual realization of patient arrivals does not match the allocation, which results in inefficient use of the capacity and/or long access time for patients. This fact is well known from general queuing theory: a static allocation of capacity will increase variability and can reduce resource efficiency. In current practice, the calendar supervisor can counter such problems by manually adjusting the calendar to adopt the allocation of capacity to variability in demand as best as possible. With constant active – and time consuming – supervision of the calendar, the scheduling and adjustment practice performs satisfactory. However, making good adjustments is critically dependent on the supervisor's expertise: even a short vacation or illness of the calendar supervisor leads to immediate and significant deterioration of the resource efficiency. Additionally it would take a long time to train a new calendar supervisor with similar capabilities. From a planning and sustainability perspective, this is highly unsatisfactory.

As our main contribution, we present an adaptive approach to automatic optimization of resource calendars. In our approach, the allocation of capacity to different patient groups is flexible and adaptive to the current and expected future situation. To maintain high performance levels, our approach shifts capacity between different urgent and non-urgent patients groups. It does not require any rescheduling of patients, or a pool of on-call patients to fill in empty timeslots. Our approach enables the calendar supervisor to quickly implement calendar adjustments, and anticipate – and remedy – the impact of current demand trends to future resource efficiency, as well as assess the impact of possible changes to the calendar. Additionally, we present an approach to determine optimal resource openings hours on a larger time frame. Opening hours can be reduced to increase capacity usage while maintaining high performance levels, or extended to counter increasing access time.

We extensively evaluate our adaptive approaches in a precise simulated environment. Our model and its parameter values are determined from extensive case analysis. Sources include historical data and extensive discussion with experts at various levels in the organization. We evaluate with a comprehensive computer simulation of the application case. This additionally allows us to study various problem scenarios and scheduling approaches. Due to the complexity of our process model, queuing theory [6] cannot provide analytical answers, and modeling the problem as a Markov decision problem [7] results in a state space of unsolvable size.

In the next Section, we will discuss the problem and our case study in more detail. In Section 3 we will present our simulation model built based on our case study. Our adaptive patient scheduling approach is presented in Section 4. We present the results of our experiments in Section 5. We discuss related work in Section 6, and conclude in Section 7.

2 CT-scan Scheduling Model

In this Section, we define our CT-scan scheduling model. From the AMC electronic calendar system[3], we have collected the historical data of the appointments made from October 2005 until March 2006. We have complemented this with data from actual production of CT-scans (November 2005 until January 2006). During this period, some scans were taken without an appointment. We derive the patient arrival process, patient attributes distributions and scheduling practice from this data as well as from site-visits and extensive discussions with the human schedulers, the calendar supervisor, and resource manager.

Our case consists of three main parts, discussed in the following sections. One part is the arriving **patients** that need to be scheduled for a specific scan. Secondly, we describe the available resources, and the way the associated appointment **calendar** is structured. The third part is the **scheduling** process that determines how appointments are made by assigning patients to timeslots on the calendar.

2.1 Patients

An important issue is that there is a great variety in patients and scan attributes. We make the abstraction that a patient always needs to be scheduled for exactly one CT-scan. We therefore model the patient and his/her scan as a unity, which we from now on refer to as ’patient’. The patient attributes most important are listed in Table 1. We make an abstraction from a patient’s physical arrival time and consider the request time of when the actual request for a CT-scan is made. Medical attributes include whether the patient is admitted to the hospital or not (inpatient versus outpatients), which has influence on the duration of the appointment needed. If a patients needs to be injected with intravenous contrast (ivc) before the scan can be taken, a doctor must be present. The

[3] X/CARE, McKesson

urgency, or acceptable access time, is expressed with a planning window (PLANWIN) in which the appointment must be scheduled.

Table 1. Patient attributes

attribute	**description**
request time	date and time when request for CT-scan is made
in-/outpatient	is the patient admitted in the hospital?
contrast needed? (± ivc)	does intravenous-contrast need to be injected?
planning window (PLANWIN)	expresses urgency of patient.
duration	of the needed appointment

Table 2. Patient groups

group	**urgency**	PLANWIN (fraction)	**duration**	**size(%)**
OUT+IVC	normal	$(2, 14)$	15 mins	$52\% \pm 6\%$
OUT–IVC	normal	$(2, 14)$	15 mins	$23\% \pm 4\%,$
URGENT	high	$(0, 1)(33\%), (0, 2)(33\%),$ or $(0, 3)(33\%)$	15 mins	$10\% \pm 3\%$
CLINIC	high	$(0, 1)(40\%),$ or $(0, 2)(60\%)$	30 mins	$6\% \pm 2\%$
SPECIAL$_n$	n.a.	n.a.	$duration_n$	$9\% \pm 2\%$

We structure patients and their attributes in different patient groups. Table 2 lists these groups and their specific properties. The group size is given relative to the total number of patients, with its variation between weeks. These groups are defined based on the groups used in practice and a medical and scheduling perspective. Based on all their attributes, patient can be grouped in many different ways. Patient group definitions can be unclear or incorrect in practice, which can cause a negative effect on efficiency. Defining the correct groups is important for efficiency, and finding the best definition of groups can be a complex problem. In our definition of patient groups, the most important attributes are medical constraints, and urgency. We additionally considered the compatibility with the current schedule procedure in the hospital, aggregating the URGENT and CLINIC groups for instance, would require additionally changes in the instructions for the human schedulers.

The largest group – OUT+IVC – is comprised of non-urgent outpatients who need intravenous contrast (ivc). Non-urgent outpatients who do not need intravenous-contrast are in the group OUT–IVC. All urgent outpatients form the group URGENT. The fourth group – CLINIC – consists of all inpatients.

Besides these four groups, there are a number of smaller, highly specific groups: SPECIAL$_n$. These include patients taking part in special programs, and patients who need a very specific treatment for making a CT-scan. E.g., one special group is the group of patients, usually children, who need to be sedated while making the CT-scan.

Urgency of patients is defined in terms of planning windows (PLANWIN) with different sizes. Besides medical urgency, a planning window also expresses the norm on acceptable access time. These norms are indicators, used to evaluate and compare hospitals on a national level. In the AMC the norm for non-urgent outpatients is two weeks: OUT+IVC and OUT–IVC have a planning window of $(2, 14)$, which means that the appointment must be scheduled between 2 days and 14 days after the request for the scan is made. The planning window starts from day 2 such that outpatients do not have to return to the hospital within a day, but have some time to plan things around the appointment at home or work. Urgent outpatients (URGENT) and inpatients (CLINIC) have high urgency. They can be planned from the request day (0), and have varying due-date of either 1, 2, or 3 days after the request date. Patients from special groups do not have specific planning windows and are always scheduled to the first available timeslot of matching type. In our model, we do not consider patients with a urgency of less than one day, or those that need to be scanned immediately without an appointment. For these patients there is an additional CT-scanner available in the emergency room of the AMC.

2.2 Resource Calendar

Patients must be scheduled to a timeslot on the calendar. The total resource capacity is given by the number of actual CT-scanners m (for the radiology department at the AMC $m = 2$) and the opening hours. It is not allowed by the hospital to make appointments on the CT-scanner in the emergency room, we therefore do not include this resource in our model.

The resource calendar used in practice is a grid of timeslots of varying sizes. Restrictions on the scheduling of patients are enforced by defining blocks of timeslots of a specific type. Timeslots of different types are used differently. In this way, resource capacity is allocated to patient groups. An initial allocation is determined months in advance, based on historical data and hospital policy. Short term adjustments of this allocation are currently done manually by the calendar-supervisor if problems occur.

We model a standard calendar, structured in days and weeks. The time on the calendar is partitioned into timeslots of different sizes. All timeslots have a size of a multitude of the time unit tu. (on the CT-scan calendar tu is 15 minutes, and there are timeslots of sizes $1tu$ up to $4tu$.) The parameters in Table 3 define the resource calendar. The parameters m and tu are fixed for long periods, the other parameters can be changed dynamically. Adjustments to the openings hours must be known at least one week in advance to plan staff. In general we assume that the m actual resources are interchangeable.

Timeslot Type Specification CT-scan capacity is allocated to different patients groups, and these allocations serve medical restrictions (e.g. due to preparation constraints for narcosis), as well as a scheduling goal (e.g. reserve timeslots for urgent patients). The allocations include: three timeslots are reserved on all Thursday mornings for patients from a SPECIAL$_n$ group, who need to be sedated while making the CT-scan; during lunch time, radiologists schedule meetings and other activities, therefore, OUT+IVC patients, who need to be injected with intravenous contrast for which a radiologist must be present, cannot be scheduled during lunch; in the afternoon of every day a number of timeslots is reserved for URGENT patients, such that urgent CT-scans ordered during the day can be performed on the same day as much as possible.

Table 3. Calendar parameters

parameter	**description**
m	number of resources
$o_{j,d}$	opening time of resource j on day d
$c_{j,d}$	closing time of resource j on day d
tu	unit size timeslots
TTS	timeslot type specification

Table 4. Timeslot Types

Timeslot-Type	**allowed patients**	**size**
TTout	OUT+IVC, OUT–IVC, URGENT	$1tu$
TT-ivc (during lunch)	OUT–IVC, URGENT (no ivc)	$1tu$
TTurgent	URGENT	$1tu$
TTclinic	CLINIC	$2tu$
TTspecial$_n$	SPECIAL$_n$	1–$4tu$

We model this allocation by using a timeslot-type specification (TTS). A timeslot-type specifies which patient can be scheduled to a certain timeslot (Table 4). The TTS thus determines how much of the resource capacity is allocated to the patient groups. See Figure 1 for an example TTS in the CT-scan resource calendar as used in practice. The TTS is not necessarily fixed as the capacity allocation can be dynamically altered.

The TTspecial$_n$ type of timeslots can only be used by very specific types of patients. For each TTspecial$_n$ type there is a rule which states that if there are still any free slots remaining r_n days in advance, these slots are changed to TTout type of timeslots, to not waste the capacity otherwise. This rule is currently the only automatic TTS adjustments in operation at the hospital.

2.3 Scheduling

Patient scheduling is the process of assigning patients to timeslots on the calendar thus setup, i.e. making appointments. In the case we describe, two elements influence overall scheduling performance: first, how well the TTS matches the actual patient arrival, and second, the used scheduling method (the selection of a timeslot per patient given the TTS).

As in many hospitals, the actual scheduling of appointments for CT-scans is done manually in the AMC. Requests arrive either by telephone or via request form. Patients are scheduled in turn by human schedulers, who look at the calendar for the availability of a suitable slot, or use the search function of the electronic calendar system. The search function returns a list of the first available suitable timeslots. For urgent patients the search function is usually not used, instead, the schedulers look at the first few days of the calendar and select a free timeslot by hand. Often the human schedulers deviate from the offered list for non-urgent patients as well. They can take a patient's personal attributes into account, including a patient's preference (e.g. for a specific day, or time).

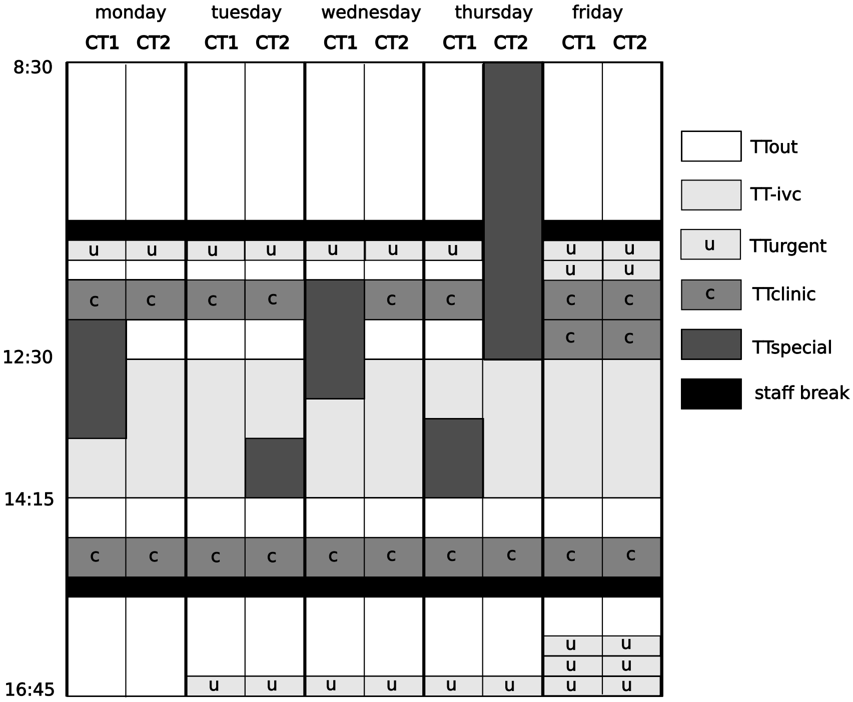

Fig. 1. **Example TTS (capacity allocation) on CT-scan calendar.**

3 Simulation

Based on the case study we have implemented a patient scheduling simulation, see Figure 2. We use the simulation in the evaluation of different scheduling and capacity allocation approaches. The case inputs of our simulation model are based on the case we studied in the previous Section. These elements together with our adaptive model (Section 4) are the inputs of our simulation. We discuss the main parts of our simulation in more details next.

3.1 Patient Arrival Simulation

With our model of patient properties and the distributions over patient attributes (Table 2) derived from analyzing the historical data, we can simulate the stochastic arrival process of patients. To simulate the stochastic arrival process, including trends where some periods of weeks are busier than others, we model the number of patients per week by means of a random walk.

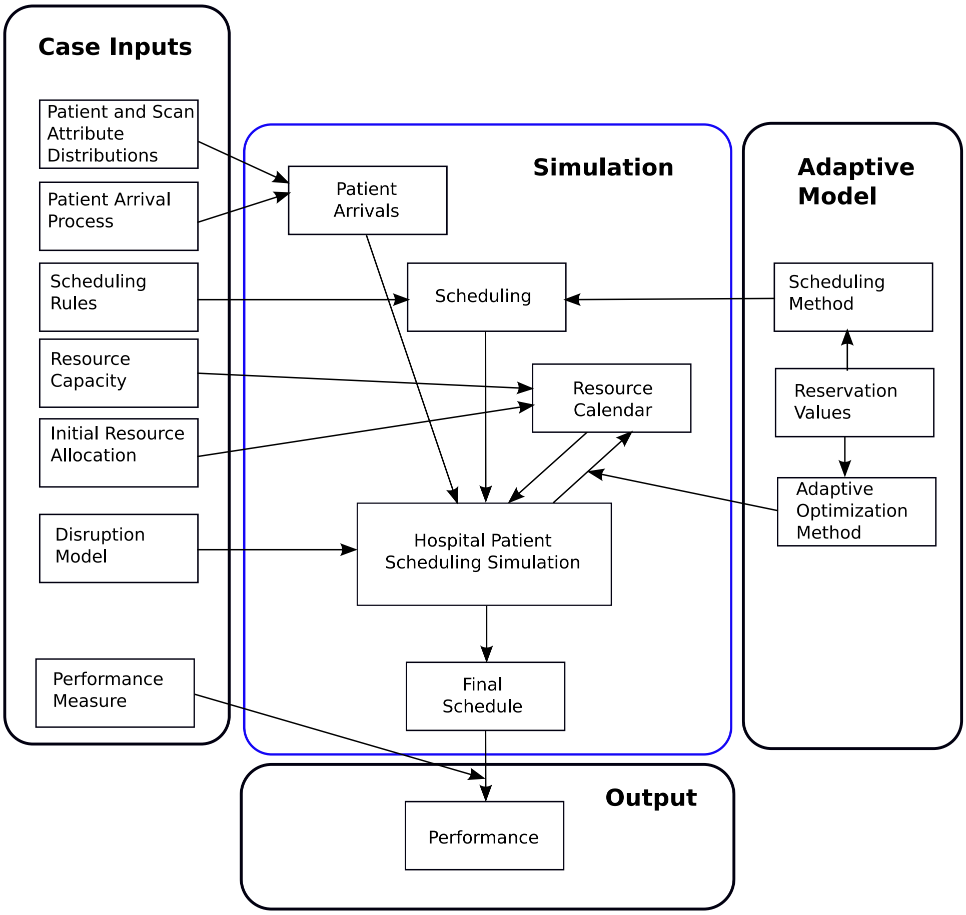

Fig. 2. Overview of Hospital Patient Scheduling Model.

A standard random walk with a drift τ towards the average $\bar{n}$ fits the distribution over the number of patients arrivals per week. The number of patients for next week (n_{w+1}) is determined as a function of the current patient arrivals n_w as:

$$n_{w+1} = n_w + \mathcal{N}(0, \sigma) + \frac{\bar{n} - n_w}{\tau},$$

where $\mathcal{N}(0, \sigma)$ a normally distributed fluctuation of patient arrivals. We set: $\bar{n} = 250$, $n_0 = \bar{n}$, $\sigma = 30$, and $\tau = 3$.

Furthermore, arrival of patients within the week is stochastic, simulating that some days have more arrivals than other days. Because of extra rounds for CLINIC patients on Monday and Friday, patient arrival is slightly structured during the week. On Monday and Friday, twice as many requests for CT-scans of CLINIC patients are ordered compared to the other three weekdays. Request for urgent outpatients arrive following a uniform distribution over the week. Because of the relative small number of URGENT and CLINIC patients there can occur a large difference in number of arrivals between days. Non-urgent outpatient have a planning window of two weeks, their arrival distribution over the week is of little influence on performance, we assume that request arrivals for these non-urgent outpatients follow a uniform distribution over the week. The resource is closed on Saturday and Sunday, appointments for URGENT and CLINIC patients requested on a Friday with a PLANWIN of (0,1) or (0,2) must be scheduled the same day.

3.2 Resource Calendar

In our simulation, we use a resource calendar, which is practically the same as the calendar used in practice. Opening time on the calendar is 8:30, while the resource closes at 16:45. The TTS used (which defines the capacity allocation) is set such that there is no other TTS with a better performance given the stochastic patient arrival and optionally an adjustment method. The best initial TTS was determined experimentally.

3.3 Scheduling

In our experiments, we want to simulate current scheduling practice to evaluate it in scenarios different from current practice. In current scheduling practice there is considerable variation in the rules for selection timeslots. Often patient preferences and other considerations are taken into account while selecting timeslots. We define a schedule method that simulates this in the following way:

*First Come Randomly Served (*FCRS*)* Patients are scheduled in order of arrival. A patient is assigned to a timeslot within his planning window, randomly selected from all the free timeslots of allowed types in his planning window. If there are no free timeslots, the first free timeslot after the planning window is selected.

For URGENT and CLINIC patients, patient preferences are of little importance because they have a high urgency. However due to the mix of urgency within URGENT and CLINIC patients, which often the human schedulers do not take into account, these patients are also scheduled randomly to allowed timeslots in the planning window. In Section 4.1 we discuss a dynamic approach to the scheduling of URGENT and CLINIC patients.

3.4 Performance Measure

Our performance measure expresses that patients must be scheduled within their planning windows. As a performance measure per patient group, we take the percentage of patients scheduled on time (within their planning window). We call this percentage the service level (SL) of the group. This is a typical performance indicator in the hospital. It is important that each group G has a high service level. Hospital policy determines the importance of each group in the overall scheduling objective. In discussion with hospital experts at the AMC, all groups were given equal weights, and the lowest service level between groups, the most indicative of a department's performance. We therefore define our objective to maximize the minimum service level (MSL) of the four main groups of Table 2:

$$MSL = \min_G \left(\frac{|\text{patient } \in G = \text{ontime}|}{|G|} \right),$$

where ontime is defined as scheduled within the planning window of the patient. In our AMC case study, groups with higher urgency are also of smaller size, while all groups have equal weight in MSL. Therefor, to maximize performance, it is often more important to schedule a single urgent patient on time than it is to schedule a single non-urgent patient on time.

4 Adaptive Model

In the hospital, the allocation of capacity, through the TTS on the resource calendar, can be separated in different timeframes:

1. Long term (months): the initial overall allocation is determined, based on long-term expectations of patient arrivals and hospital policy.
2. Medium term (weeks): adjustments can be made for known future events, e.g. holiday periods, additional workload, or planned maintenance of the machines. These adjustments could include adjusting the openings hours to optimize performance. (In some initial experiments, we observed that adjusting the allocation weekly, based on the realization of patient arrival, had a limited effect.)
3. Short term (days): small adjustments to the allocation are made daily, based on the realized and expected patient arrivals.

In our research, our focus is on an adaptive approach for short term adjustments. Additionally we look at medium term adjustments of opening hours. We discuss these in turn next.

4.1 Short-term adjustments

Our short-term adaptive approach consists of two parts: a scheduling method, and a method for adjusting capacity between patient groups. In our approach URGENT and CLINIC patients are scheduled by taking the expected number of patient arrivals per day and their specific planning window into account. Secondly, we use the expectation values to compare available capacity with needed capacity on the first three days, and change capacity between groups.

Adaptive Urgent Scheduling To schedule urgent and clinic patients on time, the allocated capacity must be large enough. Some patients with different PLANWINs use the same type of slots: URGENT with PLANWINs of (0,1), (0,2), or (0,3) use timeslots of type TTurgent; CLINIC patients with PLANWINS of (0,1) or (0,2) use timeslots of type TTclinic. Because of this mix of urgency, there is a trade-off between scheduling patients to the earliest timeslots available to not waste capacity and keeping timeslots open for the possible arrival of more urgent patients. In current hospital practice, human schedulers do not have an overview to solve this efficiently (either too many low-urgency patients are scheduled in place of high-urgency ones or too much capacity is wasted).

In our approach, we solve the problem of scheduling patients with mixed urgency, by virtually dividing urgent capacity while scheduling: a number of timeslots are specifically reserved for patients with a certain PLANWIN (for each day of arriving patients). In Table 5 we show the matrix of reservations for URGENT patients, $R(\text{TTurgent})_{reqd}^{\text{PLANWIN}}$, for each PLANWIN and request day $reqd$ (relative to the current day 0). We similarly define a reservation matrix for CLINIC patients on TTclinic capacity.

Table 5. Reservations within TTurgent type timeslots

	request day		
PLANWIN	**0**	**1**	$reqd$
(0,1)	$R(\text{TTurgent})_0^{(0,1)}$	$R(\text{TTurgent})_1^{(0,1)}$	$R(\text{TTurgent})_{reqd}^{(0,1)}$
(0,2)	$R(\text{TTurgent})_0^{(0,2)}$	$R(\text{TTurgent})_1^{(0,2)}$	$R(\text{TTurgent})_{reqd}^{(0,2)}$
(0,3)	$R(\text{TTurgent})_0^{(0,3)}$	$R(\text{TTurgent})_1^{(0,3)}$	$R(\text{TTurgent})_{reqd}^{(0,3)}$

The number of timeslots reserved per urgency and arrival day (size of reservation) is the expectation of the number of patients and some additional surplus capacity. The expected number of patients is small for a specific reservation and has high variability. To select which timeslots are reserved, we use the following heuristic for placing the reservations over timeslots on the calendar: the reservations are placed on the last day of the PLANWIN ($R(\text{TTurgent})_0^{(0,1)}$ on day 1, $R(\text{TTurgent})_0^{(0,2)}$ on day 2, etc) see Figure 3. By placing the reservations at the end of the planning window, variability of their usage is minimized. Note that because the resource is closed in the weekend, a large number of reservations is positioned on Friday; this corresponds to practice, where on Fridays a large capacity is allocated to URGENT and CLINIC patients.

Given the reservations, URGENT patients are scheduled first come first serve (FCFS), to a timeslot either not reserved or reserved and matching the patient's entry-date and PLANWIN. To make this method more flexible and reduce utilization variability, a reservation violation is allowed if the patient is not scheduled on time otherwise. Specifically, if there are no allowed timeslots available in the patient's planning window, the patient is scheduled to the day within his PLANWIN which has the most available timeslots regardless of reservations. We call this scheduling with flexible reservation (**FlexRes**), see Algorithm 1. We use a similar algorithm for scheduling CLINIC patients to TTclinic type timeslots.

Adjusting Capacity In the previous section we discussed how we use reservations of timeslots in scheduling URGENT and CLINIC patients. Crucially, if timeslots within the reservations are not used, these could be made available for other groups. In our

Algorithm 1 FlexRes: Scheduling with Flexible Reservations for URGENT patients.

```
$p$ is the current to be scheduled patient at day 0
$R(\text{TTurgent})^{\text{PLANWIN}}_{reqd}$ is the number of TTurgent slots reserved for patients with PLANWIN and have a request date of $reqd$.
$FREE(\text{TTurgent})_d$ is the number of free TTurgent timeslots on day $d$
$TS$ = the first available TT urgent timeslot
if PLANWIN $== (0,2)$ OR PLANWIN $== (0,3)$ then
    if ($TS$ is on day 1 AND $FREE(\text{TTurgent})_1 \leq R(\text{TTurgent})^{(0,1)}_{reqd=0}$) then
        $TS$ = the first available TTurgent timeslot after day 1
if PLANWIN $== (0,3)$ AND $TS$ is on day 2 AND ($FREE(\text{TTurgent})_2 \leq R(\text{TTurgent})^{(0,1)}_{reqd=1} + R(\text{TTurgent})^{(0,2)}_{reqd=0}$ then
    $TS$ = the first available TTurgent timeslot after day 2
if $TS$ is outside PLANWIN then
    $D$ is day within PLANWIN with most free TTurgent slots
    if $FREE(\text{TTurgent})_D > 0$ then
        $TS$ = the first available TTurgent timeslot on day D
schedule $p$ to $TS$
```

Algorithm 2 Dynamic: Adjusting capacity between patient groups.

```
change all $FREE(\text{TTout})_1$ (on day 1) timeslots into TTurgent
if $FREE(\text{TTurgent})_1 > R(\text{TTurgent})^{(0,1)}_0$ then
    change $(FREE(\text{TTurgent})_1 - R(\text{TTurgent})^{(0,1)}_0)$ number of TTurgent timeslots into TTclinic
if $FREE(\text{TTclinic})_1 > R(\text{TTclinic})^{(0,1)}_0$ then
    change $(FREE(\text{TTclinic})_1 - R(\text{TTclinic})^{(0,1)}_0)$ number of TTclinic timeslots into TTurgent
$CHSLOTS = \sum_{d \leq 2} FREE(\text{TTurgent})_d - \sum_{d \leq 2} R(\text{TTurgent})^{0,d}_{d-1}$
change $(min(FREE(\text{TTurgent})_2, CHSLOTS)$ number of TTurgent timeslots on day 2 into TTout
```

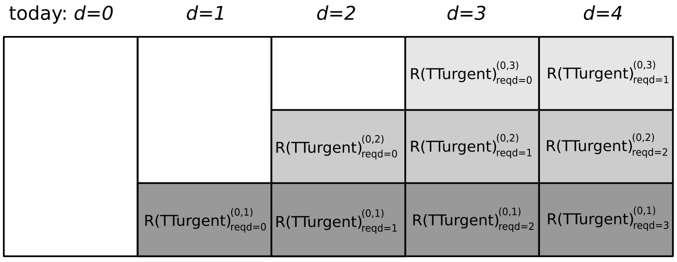
today: *d=0*
d=1
d=2
d=3
d=4
$R(TTurgent)^{(0,3)}_{reqd=0}$
$R(TTurgent)^{(0,3)}_{reqd=1}$
$R(TTurgent)^{(0,2)}_{reqd=0}$
$R(TTurgent)^{(0,2)}_{reqd=1}$
$R(TTurgent)^{(0,2)}_{reqd=2}$
$R(TTurgent)^{(0,1)}_{reqd=0}$
$R(TTurgent)^{(0,1)}_{reqd=1}$
$R(TTurgent)^{(0,1)}_{reqd=2}$
$R(TTurgent)^{(0,1)}_{reqd=3}$

Fig. 3. Positioning of reservations within TTurgent.

approach, we dynamically manage allocated capacity to be adaptive to stochastic patient arrival.

To maintain high MSL, at the beginning of the current day 0, capacity is shifted (Algorithm 2) between timeslot types on the days that their planning windows overlap: on day 1 all remaining TTout and TT-ivc capacity is changed into TTurgent capacity; on day 1 capacity can be shifted between TTurgent and TTclinic; on day 2 some TTurgent timslots can be changed into TTout (see Table 6). Note that because Algorithm 1 does not reserve timeslots on day 0 (see Figure 3) shifting capacity between timeslot types TTurgent and TTclinic on the current day 0 is not neccessary in our approach. Because adjustments are made at least one day in advance, it is sufficient to adjust once a day instead of continuously.

Thresholds for reallocating capacity between timeslot types are based on the reservations discussed above. The goal of the adjustments is to reallocate capacity, such that all reservations (which include surplus) can be made within the capacity allocated per group. In this step, in order to optimize allocation, reservations are no longer necessarily placed on the last day of the planning window. Any timeslot not needed for reservation can be changed into a timeslot of another type.

Table 6. Adjustment of capacity, increased (+) or reduced (-), for different timeslot types and days (with corresponding line in Algorithm 2)

	Day		
Capacity Type	**0**	**1**	**2**
TTout		− (1)	+ (7)
TTurgent		+ (1), − (3), + (5)	− (7)
TTclinic		+ (3), − (5)	

4.2 Adjusting Opening Hours

Adjusting the calendar on a medium term time scale, is suited for reacting to busy periods, holidays, planned resource maintenance, or increased workload due to additional patient programs. Here we focus on adjusting the opening hours for quiet and busy periods, which is important for maintaining short access times and high resource and staff efficiency. This is currently not done efficiently in hospital practice, because it is difficult to oversee the effect of an adjustment long enough in advance. In busy periods, when the total demand reaches or exceeds resource capacity, access time increases rapidly. With little extra capacity this can usually be avoided. Reducing opening hours in slow periods can compensate increased staff working hours in busy periods.

Although fixed working hours are still preferred by the staff, there is some flexibility in the working hours if the changes are known in advance. We assume that for the planning of staff, the actual opening hours of the resource must be known at least one week in advance. In our approach we fix the opening time of the resource, and adjust the closing time. We set a parameter OH_w which defines the total amount of openings hours of week w. Before the beginning of week $w - 1$ we determine the best value for OH_w.

We use a standard bi-directed search method, with a discrete step size of $stepsize$, where performance of different values of OH_w are determined by a number of simulation runs using our patients scheduling simulator. We search for the smallest OH_w that has an MSL performance of at least P_{pref} (preferred performance level). In these simulations we schedule patients arriving over 2 weeks, since the opening hours are adjusted for the second week. The current partially filled-in calendar and the estimation on the expectation of future patient arrivals are used as the starting point of the simulation. We use the number of patients from the previous weeks, as an estimate for future weeks. Conceivably, we can easily use more specific estimates for known holidays etc. When adjusting the openings hours on the calendar the total hours OH_w are divided equally over the days.

This approach takes into account that on each day the closing time can not be reduced further than the latest appointment already scheduled in the partially filled-in calendar. To make it possible to reduce the openings hours in quiet periods for increased capacity usage, a small adjustment to the scheduling method is used: for weeks of which the opening hours can still be adjusted, patients are preferably scheduled to timeslots before 3pm, to allow opportunities for earlier closing times.

5 Experiments

We conduct computer experiments to evaluate our adaptive optimization of the scheduling process. In our simulations, we generate realistic problem runs. We compare the performance of our fully adaptive approach to benchmark approaches. Performances are averaged over 70 runs. Within each run, patients arrive during 20 weeks. To avoid start-up effects, we start with a partially filled-in calendar, and measure average performance (MSL) only over the last 10 weeks of the simulation run.

We use a TTS optimized for a stochastic arrival of patients, including:

- 18 TTclinic timeslots reserved for an average of 14 (± 4) CLINIC patients per week,
- 34 TTurgent timeslots reserved for an average of 25 (± 8) URGENT patients per week.

No-shows and machine downtime are not included in simulation runs presented here, these had only a limited effect in our experiments, other that creating more busy periods which is captured by our arrival model.

We determined the best sizes of the reservations experimentally, see Table 7 for the used values in comparison with the expected number of patients per reservation. Because we use patient expectation per request-date and urgency (see Table 5) the actual reservation size is a small number. Determining the optimal size of a reservation is relatively easy due to the discreteness of timeslots. The smallest planning window (0,1) has the least average number of patients and largest variability and therefore needs the most relative surplus. Larger planning windows (lower urgencies) have less variability, and can also make use of surplus capacity reserved for higher urgency if necessary (Algorithm 1).

Table 7. Reservation sizes (expected n.o. patients).

	TTurgent	**TTclinc**	**TTclinc**
planwin	all days	mon, fri	tue, wed, thu
(0,1)	3 (1.6)	4 (2)	2 (1)
(0,2)	2 (1.6)	2 (2)	1 (1)
(0,3)	2 (1.6)		

First, we show our main results for short-term scheduling methods and adaptive allocation of capacity. Second, we show how opening hours can be adjusted to maintain high MSL or increases resource usage.

5.1 Short-term

We present average performances of three scheduling approaches with a static allocation, and the same three approaches with capacity dynamically adjusting by the method presented in this paper. The first benchmark is a baseline approach using FCRS for all patients (see Section 3.3). This approach is similar to the practical case in a hospital where there is no staff to adjust the calendar dynamically, or where the calendar supervisor is absent due to illness of vacation. The second benchmark is the standard scheduling rule First Come First Serve (FCFS), which optimizes resource efficiency but does not consider any stochastic element in the scheduling process. The third approach is our scheduling method for URGENT and CLINIC patient based on flexible reservations (FlexRes), Algorithm 1. All three approaches are evaluated with either a static calendar or in combination with our approach to dynamic adjustments of capacity (Dynamic), Algorithm 2.

We present the results of three different scenarios for the number of patients arriving per week: n_w is given by a random walk (see Section 3.1), n_w is constant with $n_w = 250$, and n_w is constant with $n_w = 270$. The average performances (MSL), standard deviation (stdv), and average capacity usage (cu) are presented in Table 8. We additionally compare performances to the baseline approach (FCRS with static calendar) with an additional capacity of 2.5 hours per week (6% extra capacity).

Table 8. Performances (MSL) averaged over 70 runs, with standard deviation (stdv.) and average capacity usage, for three different scenarios, given 41h15min openings hours per week.

	Random Walk		
Approach:	**performance MSL**	**stdv.**	**cap.usage**
FCRS static	0.79	0.19	0.91
FCFS static	0.78	0.25	0.93
FlexRes static	0.77	0.24	0.91
FCRS Dynamic	0.88	0.14	0.91
FCFS Dynamic	0.88	0.10	0.93
FlexRes Dynamic	**0.96**	0.07	0.91
FCRS static + 2,5h	0.96	0.03	0.86
	Constant 250		
Approach:	**performance MSL**	**stdv.**	**cap.usage**
FCRS static	0.88	0.06	0.91
FCFS static	0.92	0.05	0.93
FlexRes static	0.85	0.16	0.91
FCRS Dynamic	0.94	0.03	0.91
FCFS Dynamic	0.95	0.02	0.92
FlexRes Dynamic	**0.98**	0.01	0.91
FCRS static + 2,5h	0.97	0.03	0.86
	Constant 270		
Approach:	**performance MSL**	**stdv.**	**cap.usage**
FCRS static	0.56	0.16	0.97
FCFS static	0.28	0.26	0.97
FlexRes static	0.42	0.29	0.97
FCRS Dynamic	0.76	0.07	0.97
FCFS Dynamic	0.68	0.12	0.99
FlexRes Dynamic	**0.93**	0.03	0.97
FCRS static + 2,5h	0.96	0.02	0.92

The results in Table 8 show that our dynamic approach to capacity allocation in combination with flexible reservations, has a very high performance close to a MSL of 1.0, even in the busiest (constant with $n_w = 270$) and most stochastic (random walk) scenarios. Even though standard deviation is generally high, due to the wide range of problem instances created in our simulation, our dynamic approach has the lowest performance-variability. The performance of FlexRes Dynamic is significantly better than FCRS static (p-value = $< 10^{-10}$), FlexRes Dynamic better than FCRS dynamic (p-value = $< 10^{-6}$), and FCRS Dynamic better than FCRS static (p-value $< 10^{-6}$), using the two-sample Kolmogorov-Smirnov test with significance level $\alpha = 0.01$ for the random walk scenario.

In the random walk scenario, FCRS with a static calendar has an average MSL performance level of 0.79: of the worst-off patient group only 79% of patients are scheduled on time. With dynamic adjustments and flexible reservation performance increases to 0.94: even of the worst-off patient group 94% of patients is scheduled on time. The capacity of the static baseline approach has to be increased with 6% to achieve similar performance as our dynamic approach.

With a static allocation, our scheduling approach with flexible reservation (FlexRes) achieves performance similar to the scheduling benchmarks. However, our dynamic adjustments approach performs far better in combination with FlexRes, than any of the benchmark schedulers.

5.2 Medium-term

When more patients arrive than expected, access time increases exponentially [6]. Adding extra capacity temporarily can prevent this from happening. Our approach (Section 4.2) proposes changes in openings hours to resource managers to maintain high performance. We show the experimental results for an example scenario of 16 weeks with a short busy period. The number of patients n_w per week in this scenario is given by:

$$n_w = 200 | w \leq 4, n_w = 300 | 6 \leq w \leq 11, n_w = 250 | w = 5, w \geq 12$$

In Figure 4 we show the performances (averaged over 10 runs) of the baseline approach and our dynamic approach with fixed capacity, against our dynamic approach with adjustable openings hours (see Section 4.2, and the parameters in Table 9). We plot the extra time (in minutes) used by our dynamic approach with adjustable openings hours, per week and averaged over the weeks, in Figure 5.

Table 9. Adaptive approach parameter values.

parameter	value
$OH_{standard}$	41 hours, 15 minutes (from 8:30 till 16:45)
$stepsize\ (OH)$	30 minutes
P_{pref}	0.95 (MSL)

It is clear that a busy period results in a great decline in performance for the baseline approach. Our fully adaptive approach with fixed capacity does decline in performance but reaches good performance quickly after the busy period. The fully adaptive approach with adjustable openings hours can adjust capacity such that high performance is maintained over all weeks. Summed over all 16 weeks, it uses little more than the total capacity used by approaches with fixed capacity.

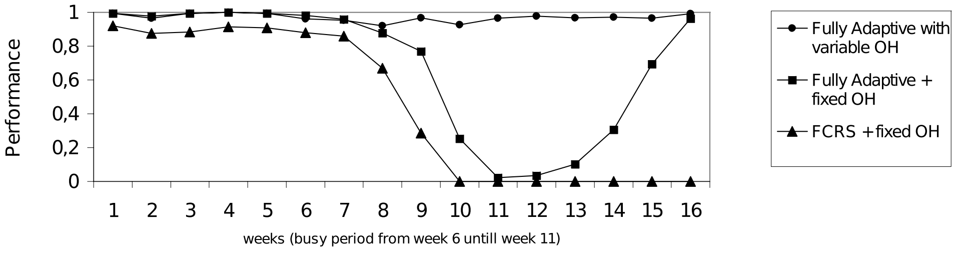

Fig. 4. Performance over weeks with variable and fixed OH.

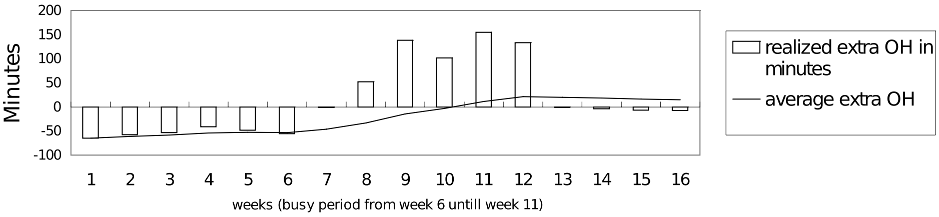

Fig. 5. Performance over weeks with variable and fixed OH.

6 Related Work

There is a number of research fields closely related to our work. Much literature considers the hospital capacity planning problem on a strategic level. On an operational level patient scheduling is researched, either specific scheduling methods or the scheduling process in hospital practice. Additionally there is work on coordination of scheduling multiple patient appointments and optimizing patient flow.

Capacity planning in hospitals at a strategic level is extensively studied in the literature, e.g. [8] [9] [10], for an overview see [2] and [11]. Most approaches consider the capacity allocation problem on a strategic level; the allocation is static on the operational level. Here we focus on short-term dynamic adjustments to the initial allocation.

Exceptions to the strict separation of capacity planning and operational scheduling are [12], [13], [14]. However, all three papers only consider allocation capacity to two priority classes, where we consider multiple priorities with additional medical constraints. In [12], the authors similarly consider a CT-scan scheduling problem. The approach assumes the use of a pool of on-call outpatients that can be scheduled to unused timeslots. The results show the benefit of a flexible approach compared to a static allocation. The work however, does not consider a full scheduling problem as the authors assume all patient arrivals are known at the beginning of each day. Furthermore, the authors use a more abstract case with only two types of reservations and measure performance in growth rate of access time. In [14] the authors consider a model similar to ours, but focus on optimizing the usage of overbooking and overtime, without dynamic rules for scheduling and capacity allocation. In [15], the authors discuss a profit maximization problem of a MRI scheduling problem for three classes of patients. Their more abstract model requires setting specific revenue and penalty functions, for which the authors identify properties of an optimal solutions.

A more abstract approach to capacity planning can also be taken from a queuing theory point of view [6]. Although realistic models are too complex to be analyzed mathematically, the problem and solutions are related: overflow rules between queues can correspond to a dynamic usage of capacity. The definition of queues and servers [16][17], corresponds to the problem of defining patient groups and timeslot types. However, queuing systems do not consider specific timeslots and appointments, and therefore do not capture the full scheduling problem.

The patient scheduling problem is not solved with optimal capacity allocation alone, the actual method of scheduling determines whether the allocated capacity is efficiently used. Scheduling methods are studied for various problem properties and objective measures, including online problems, for an overview see [18]. We have partly based our scheduling approach on insights from scheduling theory, specifically scheduling problems with objectives related to MSL. Furthermore, the scheduling method can be optimized for other considerations such as minimizing doctor and patient idle time during the execution of a schedule [19] [20]. Additionally, optimizing the logistical process in hospital practice can also be largely beneficial for resource efficiency [5] [21].

Short access time to all resources is necessary for high patient throughput in the hospital. Optimally coordinating patient paths between resources is an additional problem [22] [23]. In our approach, the human schedulers are still responsible for coordination. Multi-agent approaches seem promising to solve this distributed and dynamic coordination problem [24] [25] [26], and are part of our current research.

7 Conclusions

We presented a detailed model for scheduling multiple patient groups to a hospital resource. Specifically we presented the details of the CT-scan scheduling case at the academic hospital AMC. Short access time to central diagnostic resources is crucial for high patient throughput in the hospital. Arriving patients have varying attributes, including their urgency, corresponding to the group they belong to. Patients are scheduled to a resource calendar with capacity allocated per group. This capacity allocation must be flexible to achieve high service levels for all groups. We have implemented a realistic simulation of our case study to analyze the problem and evaluate approaches.

Given our practical case, model validation is a complex issue. The current practice and historical data provide only a single instance, and it is difficult to identify appropriate performance indicators for a wide range of settings. Recent organizational changes in the department limit the availability or usability of historical data. Additional to historical data, for which the average capacity usage was the most indicative, we evaluated model elements in numerous discussions with hospital experts with many years of detailed experience.

We developed a dynamic approach to adjusting the allocation on the calendar. We focus on short term adjustments given the current state of the calendar and the expectation of future patient arrival. We create flexible reservations for patients per request-date and urgency. Patients are scheduled based on these reservations, and the reservations determine how much capacity can be shifted between different patient groups. Additionally we use our simulation to determine the best medium-term adjustment of openings hours for maintaining high service levels, which can serve as proposals to the resource manager.

The results of our simulation experiments show that our approach can effectively schedule patients groups with different attributes and make efficient use of capacity. By dynamically adjusting capacity allocation, overall, all patient groups benefit. We have shown that there is a significant improvement over static capacity allocation. In current practice, adjusting the calendar manually requires constant attention and is critically dependent on the expertise of the calendar supervisor.

In our experiments, we focus on measuring the minimum service level of patient groups. This objective expresses the goal of the AMC to have short access times for all groups, where short can be differently defined per group. In general, the objective of our approach is efficiency of scheduling and capacity usage. By using FCRS for scheduling outpatients we simulate the effect of including patient preferences in the objective.

Many resources in the hospital are used by multiple patient groups, with different attributes such as urgency. Implicitly or explicitly, the resource capacity must be allocated to these groups of patients. Our approach can readily be applied to these problems, given the patient group definitions and parameter values. In general, when capacity is allocated, dynamically adjusting the allocation increases efficiency.

Our approach is on a operational level. Furthermore, our approach matches the current schedule procedure in the hospital. An approach that improves, not replaces, the current scheduling process is most beneficial. Human schedulers, as well as doctors, are used to working with an allocation of capacity. This is important for flexibility in usage and acceptance of the system. Furthermore it will not cause any disruptions on existing coordination with external logistics in other departments, and personal schedules. This is important for user acceptance and fast implementation. Notably, based on our results, the AMC hospital has started cooperation with a third-party software-company to fully develop our dynamic approach into implementation.

In future work we want to develop our dynamic approach to capacity allocation further. We will focus on a more general method for flexible usage of capacity, where the parameters of our approach are fine-tuned automatically. This will coincide with more case studies at different departments of the AMC. We will extend the scheduling method to take patient preferences into account. Based on our results for efficient resource usage locally we will also scale the scheduling problem to multiple departments and research mechanisms for coordination between departments.

Acknowledgement We would like to thank Han Noot for his programming contributions, and Anieke Eikelenboom for all of her expert knowledge.

References

1. Vermeulen, I., Bohte, S.M., Elkhuizen, S.G., Lameris, J.S., Bakker, P.J.M., La Poutré, J.A.: Adaptive optimization of hospital resource calendars. In Bellazzi, R., Abu-Hanna, A.,

Hunter, J., eds.: 11th Conference on Artificial Intelligence in Medicine. Volume 4594 of Lecture Notes in Computer Science. Springer (2007) 305–315

2. Vissers, J., Beech, R.: Health Operations Management. Routledge (2005)
3. Maruster, L., Weijters, T., de Vries, G., van den Bosch, A., Daelemans, W.: Logistic-based patient grouping for multi-disciplinary treatment. Artificial Intelligence in Medicine **26**(1-2) (2002) 87–107
4. Bowers, J., Mould, G.: Managing uncertainty in orthopaedic trauma theatres. European Journal of Operational Research **154**(3) (2004) 599–608
5. Elkhuizen, S.G., van Sambeek, J.R.C., Hans, E.W., Krabbendam, J.J., Bakker, P.J.M.: Applying the variety reduction principle to management of ancillary services. Health Care Management Review **32**(1) (2007) 37–45
6. Hopp, W.J., Spearman, M.: Factory Physics: The Foundations of Manufacturing Management. McGraw-Hill (2000)
7. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley and Sons, New York (1994)
8. VanBerkel, P.T., Blake, J.T.: A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. Health Care Management Science **10**(4) (2007) 373–385
9. Harper, P.R., Shahani, A.K.: Modelling for the planning and management of bed capacities in hospitals. Journal of the Operational Research Society **53**(1) (2002) 11–18
10. Bretthauer, K.M.: A model for planning resource requirements in health care organizations. Decision Sciences **29**(1) (1998) 243–270
11. Smith-Daniels, V.L., Schweikhart, S.B., Smith-Daniels, D.E.: Capacity management in health care services: Review and future research directions. Decision Sciences **19** (1988) 889–918
12. Patrick, J., Puterman, M.L.: Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization. Journal of the Operational Research Society **58**(Feb) (2007) 235–245
13. Gerchak, Y., Gupta, D., Henig, M.: Reservation planning for elective surgery under uncertain demand for emergency surgery. Management Science **42** (1996) 321–334
14. Rohleder, T., Klassen, K.: Rolling horizon appointment scheduling: a simulation study. Health Care Management Science **5** (2002) 201–209
15. Green, L., Savin, S., Wang, B.: Managing patient service in a diagnostic medical facility. Operations Research **54** (2006) 11–25
16. Rothkopf, M.H., Rech, P.: Perspectives on queues: Combining queues is not always beneficial. Operations Research **35**(6) (1987) 906–909
17. van Dijk, N.M.: Making simulation relevant in business: to pool or not to pool? ”the benefits of combining queuing and simulation”. In: WSC ’02: Proceedings of the 34th conference on Winter simulation, Winter Simulation Conference (2002) 1469–1472
18. Pruhs, K., Torng, E., Sgall, J.: Online scheduling. In Leung, J.Y.T., ed.: Handbook of Scheduling: Algorithms, Models, and Performance Analysis. CRC Press (2004) 15.1–15.41
19. Ho, C.J., Lau, H.: Minimizing total cost in scheduling outpatient appointments. Management Science **38** (1992) 1750–1765
20. Kaandorp, G.C., Koole, G.: Optimal outpatient appointment scheduling. Health Care Management Science **10**(3) (2007) 217–229
21. Kopach, R., DeLaurentis, P.C., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., Wan, H., Intrevado, P., Qu, X., Willis, D.: Effects of clinical characteristics on successful open access scheduling. Health Care Management Science **10**(2) (2007) 111–124
22. Marinagi, C., Spyropoulos, C.D., Papatheodorou, C., Kokkotos, S.: Continual planning and scheduling for managing patient tests in hospital laboratories. Artificial Intelligence in Medicine **20**(2) (2000) 139–154

23. Policella, N., Oddi, A., Smith, S., Cesta, A.: Generating robust partial order schedules. In Wallace, M., ed.: Principles and Practice of Constraint Programming. Volume 3258 of Lecture Notes in Computer Science. Springer (2004) 496–511
24. Decker, K., Li, J.: Coordinating mutually exclusive resources using gpgp. Autonomous Agents and Multi-Agent Systems **3**(2) (2000) 133–157
25. Paulussen, T.O., Jennings, N.R., Decker, K., Heinzl, A.: Distributed patient scheduling in hospitals. In Gottlob, G., Walsh, T., eds.: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann (2003) 1224–1232
26. Vermeulen, I.B., Bohte, S.M., Somefun, D.J.A., La Poutré, J.A.: Multi-agent pareto appointment exchanging in hospital patient scheduling. Service Oriented Computing and Applications **1**(3) (2007) 185–196