Optimization of Online Patient Scheduling with Urgencies and Preferences

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

[1] Centrum Wiskunde & Informatica (CWI), Amsterdam, the Netherlands, I.B.Vermeulen@cwi.nl,
[2] Academic Medical Centre, University of Amsterdam, the Netherlands

Abstract. We consider the online problem of scheduling patients with urgencies and preferences on hospital resources with limited capacity. To solve this complex scheduling problem effectively we have to address the following sub problems: determining the allocation of capacity to patient groups, setting dynamic rules for exceptions to the allocation, ordering timeslots based on scheduling efficiency, and incorporating patient preferences over appointment times in the scheduling process. We present a scheduling approach with optimized parameter values that solves these issues simultaneously. In our experiments, we show how our approach outperforms standard scheduling benchmarks for a wide range of scenarios, and how we can efficiently trade-off scheduling performance and fulfilling patient preferences.

1 Introduction

Due to increase of demand, improving efficiency in hospitals is becoming increasingly important. Besides the number of patients, the service that patients expect from a hospital is also increasing. Patients want more personalized care, which includes involvement in selecting appointment-times. In addition to high medical quality and resource efficiency, a hospital can compete with other hospitals by providing more patient-oriented services.

Improving efficiency in a hospital can be complex. Due to the distributed nature of a hospital, departments have local objectives and scheduling policies. Local scheduling has to solve a mix of patients with varying properties, hospital-wide performance depends on how local schedulers interact with each other.

We focus on scheduling patients to central diagnostic resources, which is often a bottleneck in patient pathways. Access time to these resources has a large influence on overall hospital performance as it will influence many other departments. The capacity of diagnostic resources is limited, and expensive to extend. To make efficient use of the resource, appointment-based systems are used, although in current practice the actual scheduling is often done by hand.

The basis of our scheduling problem is that different patient groups require different access times. Some patients need an appointment within a few days, others within a few weeks. We focus not only on efficient scheduling, we also want

to have opportunities for patients to select their preferred timeslot. This can be achieved by dynamically controlling the freedom in selecting timeslots. We find a trade-off between scheduling most efficiently and fulfilling patient preferences.

In this paper, we present an approach where the combination of scheduling performance and fulfilling patient preferences is optimized. Our scheduling solution consists of four main parts that we optimize simultaneously. First, resource capacity is allocated to patient groups, which allows us to have different access times per group. Second, we set dynamic rules for when an exception to the allocation can be made, this improves the overall scheduling efficiency. Third, we determine a scheduling heuristic for ordering timeslots based on efficiency. The fourth part is the use this ordering of timeslots in trading-off scheduling efficiency and fulfilling patient preferences. A number of parameters controls each part. We optimize the parameter values to automatically find a complete scheduling solution for each specific problem case we consider.

We show in our simulation experiments how our approach outperforms typical scheduling benchmarks, for a wide range of scenarios. Furthermore, we show how we can efficiently trade-off scheduling performance and fulfilling patient preferences for different patient preference models. Setting this trade-off allows hospital departments to remain in control of the scheduling, which is important for acceptance of our system in practice.

Most approaches to efficiency improvement in the hospital are from the operations research and operations management field [1]. Typical problems are strategic planning, operating room planning, capacity planning, staff scheduling, see e.g. [2]. They mostly focus on static problems, and typically do not consider online decision making, with exceptions such as [3]. They do not consider how to optimize scheduling performance in combination with patient preferences. Solutions for dynamic optimization problems usually come from the field of computational intelligence such as evolutionary algorithms [4, 5].

The theoretical background of resource problems can be found in the field of queuing theory [6]. Related is the question of pooling or separating capacity and dynamic rules such as overflow rules [7]. The difference is that in our problem a timeslot must be determined upon arrival, which in a queuing system is only achieved with observable workload and FCFS scheduling. Our scheduling solution is not bound to FCFS but can select future timeslots per arriving job.

In Section 2 we will discuss the problem and our approach for scheduling patients with urgencies. In Section 3 we discuss how we extend the scheduling problem and our approach to include patient preferences.

2 Scheduling with Urgencies

2.1 Problem definition

The problem we research is how to schedule each arriving patient, such that patients are scheduled on time. For most patients, a diagnostic test must be performed before the next consult with the physician. Most consulting hours

are scheduled on a weekly basis, and the next consult is often in a week or two weeks' time. For patients with more urgent conditions, it is important that the test results are known within a few days. For the most urgent category in the hospital no appointment is scheduled. Either separate capacity is available in the emergency department or reserved on the resource calendar, see [8]. We focus only on patients for which an appointment must be scheduled.

In our model, patients have different urgencies, with urgency defined as the time between a patient's arrival time and required due-date. A patient is scheduled 'on time' if his appointment is before the due-date. Patients with different urgencies are scheduled to the same limited resource capacity. We assume an appointment must be made as soon as the need for the appointment is known, which we also call 'patient arrival'. This allows the hospital to provide the service of immediately communicating the appointment-time to the patient.

For non-urgent patients we set a minimum access time (MAT): the number of days between arrival and the first allowed appointment date. This allows patients to arrange their return visit to the hospital. This means that we can only schedule urgent patients to any timeslots left over on days before MAT.

Part of the problem's stochastic nature is caused by the closure of the resource in the weekend. Urgent patients will often have to be scheduled before the weekend, as the following Monday will be too late given their due-date. This causes an unequal demand over the week: at the end of the week the demand from urgent patients is larger. In our model, without losing the complexity of online scheduling with urgencies, we assume that all resource capacity can be used interchangeably and use unit-time duration for all appointments.

We formulate our model with the following. Patients arrive according to a Poisson process with arrival rate λ. Each patient p belongs to a patient group $g_p \in G$ according to a patient-group distribution D_G. The urgency of a patient is given by its group $u_p = U(g_p)$, with u_p the number of days between the arrival day and due-date. Minimum access time for non-urgent patients is given by MAT in days. Resource capacity is C number of timeslots on each working day.

The performance measure is based on the service levels of patient groups. Service level SL_g is the fraction of patients in group g scheduled on time (before or on their due-date). Scheduling performance is given by $MSL = min(SL_0, ..., SL_{|G|})$, where we take the minimum service level (MSL) over all patient groups.

2.2 Approach

We present a parameterized approach to the scheduling problem outlined above. To enable a different access time per patient group, resource capacity is allocated to groups, and patients are scheduled only to timeslots allocated to their group. In this way, the service level per group SL_g is controlled by allocating capacity $a_{g,d}$ (for group g on weekday d). The relation between service level and capacity depends on group size, urgency level, and stochastic arrival. Finding an optimal allocation is the first step in this scheduling problem (and in many hospitals it is the only step).

In our approach we use a more flexible variation of this static capacity allocation: nested capacity allocation, patients can be scheduled to timeslots allocated to equal and lower urgency levels. Nested capacity is more flexible than strictly separated capacity as timeslots allocated to lower urgencies can be used by more patients. This reduces variability in demand, and improves resource efficiency. The optimal allocation of nested capacity can be different from the optimal allocation of static capacity.

In our approach, capacity usage is made even more efficient with conditional exceptions to the nested capacity allocation: capacity allocated to higher urgencies is also available if its utilization is below a certain threshold $t_{g,d}$. Such dynamic rules have been shown to improve performance [9]. It reduces the chance of timeslots allocated to higher urgencies being wasted.

Besides the number of timeslots allocated, the positioning of those timeslots within a day can also influences scheduling performance. Timeslots positioned at the end of the day have a higher chance of being used by a patient arriving during that day. Therefore the position at the end of the day is most beneficial for more urgent groups. We use this fact in our approach by positioning the timeslots for urgent patient at the end of the day. In Section 3 we will discuss a different positioning to take patient preferences into account.

Capacity allocation (nested with overflow) determines for each patient which timeslots are available for scheduling. To actually schedule we have to select a timeslot from those available timeslots. From a scheduling performance point of view, we want a scheduling method that selects a timeslot such that performance is maximized over time. In Section 3, we discuss how patient preferences are involved in this selection process.

Standard scheduling method First Come First Served (FCFS): patients are scheduled to the earliest available timeslot, maximizes resource utilization. However with FCFS, all timeslots up to a certain point in time are fully utilized, resulting in fewer chances for coping with a peak in demand for more urgent patients. To improve over FCFS, we combine it with a heuristic that counters the negative effects of FCFS, balanced utilization (BU): patients are scheduled to the day with the lowest utilization (before the due-date). With BU any available timeslots are spread out evenly over days, which increases the chances of them being beneficial for overflow from other groups. To combine the two heuristics, available timeslots ts are ordered based on a weighted sum of two normalized values ($w_{g,d} = 0$ equals FCFS, $w_{g,d} = 1$ equals a BU):

$$\text{FCFS+BU}(ts) = (1 - w_{g,d})\text{FCFS}(ts) + (w_{g,d})\text{BU}(ts)$$

$$\text{FCFS}(ts) = \frac{\text{position of } ts \text{ in FCFS ordering}}{\text{total number of timeslots}}$$

$$\text{BU}(ts) = \frac{(\text{utilization of day of } ts) - (\text{lowest utilization})}{(\text{highest utilization}) - (\text{lowest utilization})},$$

where we consider only timeslots and days before the due-date. We find the optimal value of $w_{g,d}$ per group and weekday, and schedule patients to the first

timeslot in the calculated ordering of increasing FCFS+BU values. If there are no available timeslots before the due-date, the patient is scheduled to the earliest available timeslot after his due-date.

Recall that in our approach, we have the following three parameters per patient group per weekday: $a_{g,d}$ (number of timeslots allocated to patient group g each weekday d), $t_{g,d}$ (the utilization threshold for overflow on capacity allocated to group g on weekday d), $w_{g,d}$ (the weight used in FCFS+BU for scheduling patients of group g arriving on weekday d).

3 Patient Preferences

We extend the above scheduling problem with urgencies to additionally include patient preferences. Each non-urgent patient has a preference model P_p over timeslots, P_p states whether a timeslot is preferred for patient p. We focus on boolean-type preference model: either a patient is scheduled to a preferred timeslot or to a non-preferred timeslot. The alternative of quantifying preferences, for instance with utilities, is hard because it is difficult for patients to put values on preferences. Moreover it is hard to compare preferences-values between patients.

With taking patient preferences into account, the overall objective O is now a weighted combination of scheduling performance (MSL), see Section 2, and patient preferences fulfillment (PP), the fraction of non-urgent patients that are scheduled to a preferred timeslot: $O = (\beta) * SP + (1 - \beta) * PP$. By setting β a hospital department can set a preferred combination of objectives. In our experiments we show the resulting trade-off by varying the value of β.

To maximize O, that is to include patient preferences, we extend our approach the following way. Instead of scheduling the patient to the first timeslot given the FCFS+BU ordering, we let the patient select from a number of timeslots: all timeslots with value of at most the lowest value plus m_g are available for selection. With parameter m_g we can control how much selection-freedom a patient has. Lower values will results in more efficient scheduling, while higher values will result in more fulfilled patient preferences.

A patient will select a preferred timeslot if it is in the set of offered timeslots, it will choose uniformly random from multiple preferred timeslots, and uniformly random if there are no preferred timeslots available.

Some patients could prefer a timeslot at the end of the day, which is incompatible with the way we position timeslots within the day (urgent timeslots at the end of the day, see Section 2.2). We therefore alter the method for positioning timeslots within the day to the following: the k_d number of latest timeslots on weekday d are reserved for non-urgent patients the rest of the timeslots is positioned as in Section 2.2. Setting the value of k_d in our approach makes a trade-off between scheduling performance ($k_d = 0$) and patient preferences ($k_d > 0$).

In our experiments, we use three patient preference models P_p based on discussions with human schedulers in the hospital, described in the following paragraphs.

Work/non-work A fraction of patients (NONWORK) is available during the day and prefers an appointment on the middle of the day, avoiding morning and afternoon traffic rush-hour while traveling to the hospital. The remaining fraction (1 − NONWORK) prefers an early or late appointment to minimize the effect on their working days. Given the resource openings hours between 8am and 5pm, early is defined as before 9am, midday as between 10am and 3pm, and late is defined as after 4pm. Note that in this model there are timeslots which are not preferred by any patient. We show experimental results for different values of NONWORK.

Preferred-day In the preferred-day model, patients have one or more preferred weekday(s). All timeslots on a preferred weekday are preferred timeslots. The days are uniformly random drawn. We show experimental results for model instances where patients each have one or two preferred weekdays.

Patient-calendar In the patient-calendar model, which can be viewed as a combination of the two previous models, we model black-spots in a patients calendar. We divide a week in ten parts, 5 weekdays × 2 day-half's (morning/afternoon). On a number of those ten day parts the patient will be unavailable (uniform randomly drawn). We show experimental results with varying number of black-spot day parts per patient.

4 Optimization

Our approach is parameterized and we use a search method to find the best parameter values given a scenario. We found that the problem surface was relatively smooth, and that there was an area of solutions which performed not significantly worse than the best found solution. Although we had to optimize over 50 parameters, it was still possible to find a good set of parameters values in reasonable time (< 24 hours) for a specific scenario. (Note that in practice the parameter values should be updated only as often as a few times per year.)

In the presented results below, we used an Estimation of Distribution Algorithm (EDA), see [4], with a population size of 150 and 15000 evaluations. This is a population based search method, where the distribution of each parameter value in a selection of the population is updated each generation, and used to generate individuals in the next generation. We used pair-wise comparison during selection, with a different random seed in each generation.

In our experiments, we show results of how our optimized approach (FCFS-BUdynamic) as described above, compares to the performance of three typical benchmarks each having their parameter values optimized using the EDA:

- FCFSstatic: scheduling patients First Come First Serve (FCFS) strictly to capacity allocated to their group. Capacity allocation is optimized.
- FCFSnested: scheduling patients FCFS to capacity of equal or lower urgency. Capacity allocation is optimized.
- FCFSdynamic: scheduling patient FCFS to capacity of equal or lower urgency with dynamic overflow. Capacity allocation and overflow thresholds are optimized.

5 Experiments

We have conducted many experiments to test different properties of our approach, due to space limitations we only report our main findings. Although with our EDA we automatically obtain an optimized schedule approach, we can study the found solutions and their properties. We can make the following practical conclusions based on observations in our found solutions:

- More urgent timeslots are reserved at the end of the week.
- More urgent timeslots are reserved on Thursday than on Friday.
- Overflow thresholds for urgent groups are relatively constant over the week.
- Overflow thresholds for non-urgent groups are lower on Wednesday.
- For urgent groups scheduling FCFS is more efficient than scheduling BU.
- For non-urgent groups scheduling FCFS and scheduling BU is relatively balanced, except on Fridays where it is more important to schedule FCFS.
- At the end of the week it is more important that urgent timeslots are placed at the end of the day (variable $k_d = 0$).

In our experiments we use four patient groups, $|G| = 4$, two urgent (urgencies $U_1 = 2$ days, $U_2 = 3$ days) an two non urgent groups (urgencies $U_3 = 5$ days, $U_4 = 10$ days), with relative groups sizes: $D_G : \{D_1 = 0.14, D_2 = 0.14, D_3 = 0.28, D_4 = 0.43\}$. Having more than four different urgencies within a two-week period has little practical meaning: if groups are too similar in due-date they can be considered the same group. Minimal access time (MAT) for non-urgent patients is two days. Resource capacity C is 60 timeslots per day on weekdays, closed on the weekend. Each patient needs an appointment of one timeslot. We experiment over a number of scenario's in which we change the arrival rate, the relative group sizes (D_G), and group urgencies (U_G).

First we present results on schedule performance without patient preferences. In Figure 1 we show the average performance (simulation length is 50,000 patients, averaged over 250 simulation runs) of the four approaches for different ρ's, ρ is ratio between the average number of arriving patients and the number of available timeslots (service rate). For all ρ's we see our approach clearly outperforms the benchmarks. The difference between using static capacity and our dynamic solution can be very large. Importantly, due to stochastic patient arrival, above a certain ρ performance will not be stable but decrease over time (a queuing effect where access time builds up). Our experiments indicate (not shown here) that performance is no longer stable with a ρ of 0.99 or larger.

To show our results are robust for different settings, we compare our approach with the three benchmarks in nine different scenarios. The scenarios differ in relative group sizes and group urgencies: we increase or decrease the due-dates for all groups; we vary the group sizes to have more or less urgent patients relative to non-urgent patients. In Table 1, we show the average performance of the four approaches with $\rho = 0.98$, for nine different scenarios. Our approach FCFSBUdynamic has the best performance in all scenarios, although the difference is not significant in one scenario. The relative ordering of the approaches is almost the same in all scenarios, FCFSdynamic is not always significantly better than

Fig. 1. Main Results

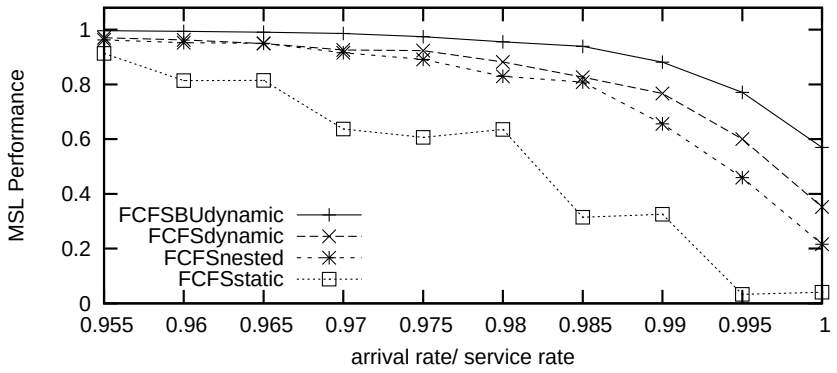

Table 1. MSL performance for nine scenarios

urgency	**urg. group size**	FCFSBUdynamic	FCFSdynamic	FCFSnested	FCFSstatic
normal	normal	**0.96**	0.87	0.86	0.57
normal	smaller	**0.95**	0.86	0.63	0.02
normal	larger	**0.96**	0.88	0.88	0.24
higher	normal	**0.74**	0.70	0.68	0.38
higher	smaller	**0.62**	0.56	0.00	0.00
higher	larger	**0.70**	**0.70**	**0.70**	0.15
lower	normal	**0.99**	0.96	0.88	0.48
lower	smaller	**0.98**	0.94	0.73	0.65
lower	larger	**0.98**	0.95	0.96	0.65

FCFSnested. These results show the adaptiveness of our approach in general, for various settings and using our optimizer, our approach can be implemented to achieve the best scheduling results.

We next discuss results of optimizing the trade-off between schedule performance (MSL) and satisfying non-urgent patient preferences. Two additional variables k_d, and m_g have to be optimized, see Section 3. Given our three patient preference models we optimize solutions for different values of β (the weight in the overall objective) to get a trade-off between the two objectives.

In Figure 2a we show the trade-off between schedule performance and patient preferences, given that a non-urgent patient has a preference for either a daytime appointment or an early or late appointment (work/non-work model). We show results for three values of the non-work fractions: 0.5, 0.75, 0.9. Note that the fraction of daytime-timeslots on a day is 0.55 and the fraction of early/late-timeslots on a day is 0.22. For all three values we see a clear trade-off between patient preferences and schedule performance. The maximum satisfaction level corresponds with a decrease in MSL of around 0.1. However, because the trade-off is optimized, we can get close to the maximum amount of fulfilled preferences with only a very small decrease in schedule performance.

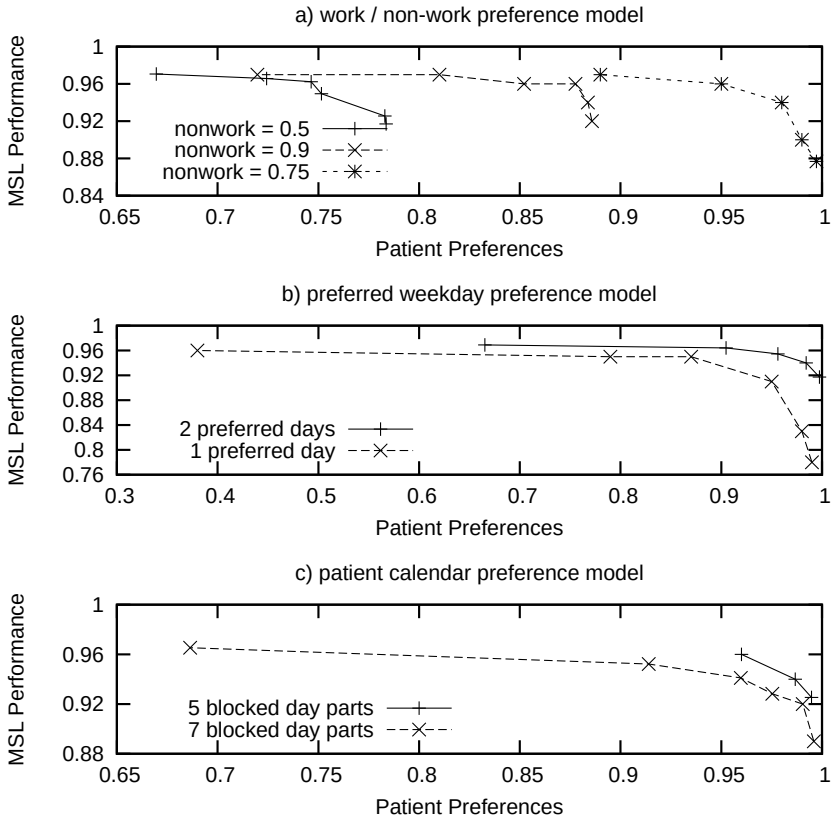

Fig. 2. Schedule Performance vs. Patient Preferences

In Figure 2b we show the trade-off for our preferred weekday model, where we consider patients having one or two preferred weekdays. If patients have only a single preferred weekday, the trade-off is on a large scale. Satisfying 80% of all patients' preferences is relatively easy, but an significant decrease in schedule performance has to be expected if all patients want get a preferred timeslot. However this effect disappears if patients have two preferred weekdays. In Figure 2c we show the trade-off for our patient calendar model, where we vary the number of day parts a patient is unavailable. Even if patients are unavailable for 7 out of 10 day parts, 90% of the patients can get an preferred appointment, with a limited decrease in schedule performance.

6 Discussion and Conclusions

We provide an automatic optimized solution for the problem of scheduling patients with different urgencies and preferences. We show how we outperform

benchmarks, independent of scenario specifics. We are able to find any preferred trade-off between schedule performance and providing patients with the service of selecting a preferred timeslot.

We use an approach for allocating capacity, setting overflow thresholds, schedule heuristics, and offering timeslots, for which we optimize all parameter values simultaneously. We show results for multiple detailed patient preference models. Previously, in [10], some initial work for some parts of our approach was conducted, with limited experimental settings and manually set parameters.

The use of automatic optimizer such as the EDA, gives us the opportunity to find solutions for many parameters in reasonable time, for any setting. This makes our approach very generic and potentially beneficial in many different places in hospitals. Our approach is suitable to be extended to include non-interchangeable resources, and/or appointments with different durations.

The presented method for making a trade-off between schedule performance and freedom in selecting timeslots gives opportunity to various extensions. Based on the same trade-off we can also schedule combination-appointments over multiple departments, which we are researching in future work.

References

1. Vissers, J., Beech, R.: Health operations management: patient flow logistics in health care. Routledge, London (2005)
2. 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
3. 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
4. Bosman, P., Grahl, J., Thierens, D.: Enhancing the performance of maximum-likelihood gaussian edas using anticipated mean shift. In: Parallel Problem Solving from Nature - PPSN X, Springer (2008) 133–143
5. Branke, J., Mattfeld, D.: Anticipation in dynamic optimization: The scheduling case. In: Parallel Problem Solving from Nature, Springer (2000) 253–262
6. Hopp, W.J., Spearman, M.: Factory Physics: The Foundations of Manufacturing Management. 2nd edn. Irwin/McGraw-Hill, Boston (2001)
7. van Dijk, N.M.: To pool or not to pool? "the benefits of combining queuing and simulation". In: Proceedings WSC '02, San Diego, Winter Simulation Conference (2002) 1469–1472
8. Bowers, J., Mould, G.: Managing uncertainty in orthopaedic trauma theatres. European Journal of Operational Research **154**(3) (2004) 599–608
9. Vermeulen, I., Bohte, S., Elkhuizen, S., Lameris, J., Bakker, P., La Poutré, J.: Adaptive resource allocation for efficient patient scheduling. Artificial Intelligence in Medicine (2009)
10. Vermeulen, I., Bohte, S., Elkhuizen, S., Bakker, P., La Poutré, J.: Decentralized online scheduling of combination-appointments in hospitals. In: Proceedings of ICAPS-2008, AAAI Press (2008) 372–379