J Syst Sci Syst Eng (Aug 2017) 26(4): 448-462 ISSN: 1004-3756 (Paper) 1861-9576 (Online)
DOI: 10.1007/s11518-017-5347-3 CN11-2983/N

# DYNAMIC RESOURCE ALLOCATION FOR EFFICIENT PATIENT SCHEDULING: A DATA-DRIVEN APPROACH

**Monique Bakker Kwok-Leung Tsui**

*Department of Systems Engineering and Engineering Management, City University of Hong Kong, China*

*mbakker2-c@my.cityu.edu.hk* (✉), *kltsui@cityu.edu.hk*

**Abstract**

Efficient staff rostering and patient scheduling to meet outpatient demand is a very complex and dynamic task. Due to fluctuations in demand and specialist availability, specialist allocation must be very flexible and non-myopic. Medical specialists are typically restricted in sub-specialization, serve several patient groups and are the key resource in a chain of patient visits to the clinic and operating room (OR). To overcome a myopic view of once-off appointment scheduling, we address the patient flow through a chain of patient appointments when allocating key resources to different patient groups. We present a new, data-driven algorithmic approach to automatic allocation of specialists to roster activities and patient groups. By their very nature, simplified mathematical models cannot capture the complexity that is characteristic to the system being modeled. In our approach, the allocation of specialists to their day-to-day activities is flexible and responsive to past and present key resource availability, as well as to past resource allocation. Variability in roster activities is actively minimized, in order to enhance the supply chain flow. With discrete-event simulation of the application case using empirical data, we illustrate how our approach improves patient Service Level (SL, percentage of patients served on-time) as well as Wait Time (days), without change in resource capacity.

**Keywords:** Patient scheduling, dynamic rostering, patient care path, discrete-event simulation

## 1. Introduction

High patient service levels are becoming increasingly important to hospitals. At the same time, health care demand is increasing as a result of aging populations, and because more and more treatment options become available, due to technological and medical advances. Yet capacity and budgets are limited, which pressures hospitals to increase efficiency for all levels of hospital operations (Vermeulen et al. 2009, Vissers and Beech 2005). Numerous studies have investigated efficiency related problems in hospitals, such as appointment/ outpatient scheduling, patient admission scheduling, operating room (OR) scheduling, accident and emergency room optimization, and so forth. Almost all of these studies focus on optimizing a single process step, i.e. single patient episodes, rather than taking a supply chain perspective on a patient's first, second, and other appointments including surgery. Or the focus is on a single clinic/OR session, and/or a homogeneous patient group (Cardoen et al. 2010, Cayirli and Veral 2003). White et al. (2011) identify complexity as the main reason that more holistic, integrated approaches are still very rare in health care operations research. But despite its complexity, we have to address the interactions and interdependencies between the planning and scheduling of the different patient groups and different types of activities that involve the same key resources, such as a medical specialist. Only in this way can we optimize planning and scheduling of the system as a whole. Any key resource is typically used by multiple patient groups (Ma and Demeulemeester 2013, Măruşter et al. 2002, Vermeulen et al. 2009) and optimizing the resource for all patient groups simultaneously is a complex problem. Demand and resource availability vary differently per patient group, and any kind of variability is generally known to hamper the patient flow. Resource allocation must thus be dynamic to meet demand variability, and aimed at keeping variability in rosters as low as possible.

We study the case of specialist rostering and patient scheduling at the General Surgery department of a medium-sized hospital in the Netherlands. The specialists are key to the process of patients visiting the outpatient clinic for a first, second, and other appointments, before possibly being scheduled for elective surgery and one or more follow-up appointments. Long access times to outpatient clinics have historically been a point of critical concern to hospitals. In the Netherlands, hospitals are mandated to adhere to government-set admission wait time limits, ranging from four weeks for non-urgent patients, to only five work days for cancer-suspicious patients. Specialists are still manually allocated to different activities a long time in advance, generally in accordance with a cyclic schedule. But variability in specialist availability drastically distorts the cyclic pattern. Irregular activities in rosters lead to an irregular in-flow of patient groups, causing long appointment wait times for some of these patient groups. This fact is well known from queueing theory in operations management: high variability in the number of parts, products, or orders sent into the system will propagate downstream, hampering a swift, even flow through all process steps. This phenomenon is the cornerstone of the Theory of Swift, Even Flow, formulated by Schmenner and Swink (1998), and is an important theory that we depart from in our case study. According to this theory, variability will in some way cause 'items' (here patients) to flow less swiftly and evenly through the process. It thus induces long patient lead times and many patients waiting somewhere in the care path. At our case hospital, in order to adhere to the government-set limits of appointment wait times, the nurses will either schedule patients into 'wrong' activity sessions, or they will overbook by scheduling two patients into one time slot. The former causes a mismatch of appointment duration and supporting resources (e.g., medical equipment such as a blood pressure gauge), to the confusion and annoyance of the specialists. The latter causes congestion and long wait times in the waiting room, along with increased work pressure and session end time delays.

The main contribution is our minimal-variation approach to automatic specialist resource calendar optimization. With this approach, we construct what we label the Dynamic Resource Calendar (RC), and compare this to the Cyclic RC and the Real 2011 RC. The Cyclic RC is the 4-week template that the case hospital uses, which repeats itself for the full year. The Real 2011 RC is the actual one that was used at the case hospital in 2011; it is the Cyclic RC, but with many changes made to it throughout the year, in order to cope with absence of specialists and other ad-hoc constraints. In our approach, the allocation of specialists to certain activities, and hence indirectly to certain (combinations of) patient groups, is flexible and keeps the variation in the number of sessions per activity as low as possible, as it adapts to specialist availability. This smooths the in-flow of patients per group, promoting a swift, even flow of patients through their sequence of appointments, including surgery. We evaluate the Dynamic RC through a simulation study that compares our approach to the traditional cyclic approach and to the real RC of 2011 at our case hospital. The simulation model and its parameter values are determined from an extensive case analysis. Sources include historical data and discussion with case experts: head planning, the specialists, nurses, and management. Similar to Vermeulen et al. (2009), the complexity of our process model does not allow for queuing theory to provide analytical answers, and modeling the problem as a Markov decision problem results in a state space of unsolvable size. Our model development was also motivated by the fact that a multi-objective implementation of a Genetic Algorithm (in Matlab R2015a) could not solve the problem, or even find Pareto-solutions, within adequate time. No solutions from the GA were found after running the program for over an hour on a 3.30GHz Intel i5 64-bit Windows 7 personal computer with 3.5 GB usable memory. In the next section, we discuss the problem and case data in more detail. In Section 3 we present the simulation model and scenario testing, with our results presented in Section 4. We give a discussion with conclusions and limitations in Section 5.

## 2. Model Formulation

In this section, we define our General Surgery Department (GSD) resource allocation model, along with the patient scheduling model of the department's outpatient clinic. We have collected complete outpatient and surgery appointment data of patients who visited the GSD outpatient clinic from January 1st until December 31st 2011. This includes emergency visits and emergency surgery, which are treated as such in our model. The department employs a partnership of 6 surgeon specialists. The data consists of 9380 patients who are grouped into 20 groups. This is the grouping that the nurses adhere to when they schedule the patients into half-day sessions (AM and PM) and time slots (5–30 minutes). This is labeled patient scheduling: the process of assigning patients to timeslots on the calendar (Vermeulen et al. 2009). New patients' initial appointment requests arrive by referral from the patient's general practitioner, either by telephone or via online request forms. Subsequent appointments are usually made at the front desk right after the initial appointment has taken place. The sub-specialization and thus activity type in the RC dictates into which session, with which specialist, a patient can be scheduled. For example, new patients suspicious of breast cancer can be scheduled into an oncology session with specialist A, B, or D. Any repeat visit will have preference for the same specialist as the previous appointment(s). Figure 1 gives a network view of patients' linkages between activities, along with letters indicating which of the specialists can be involved.

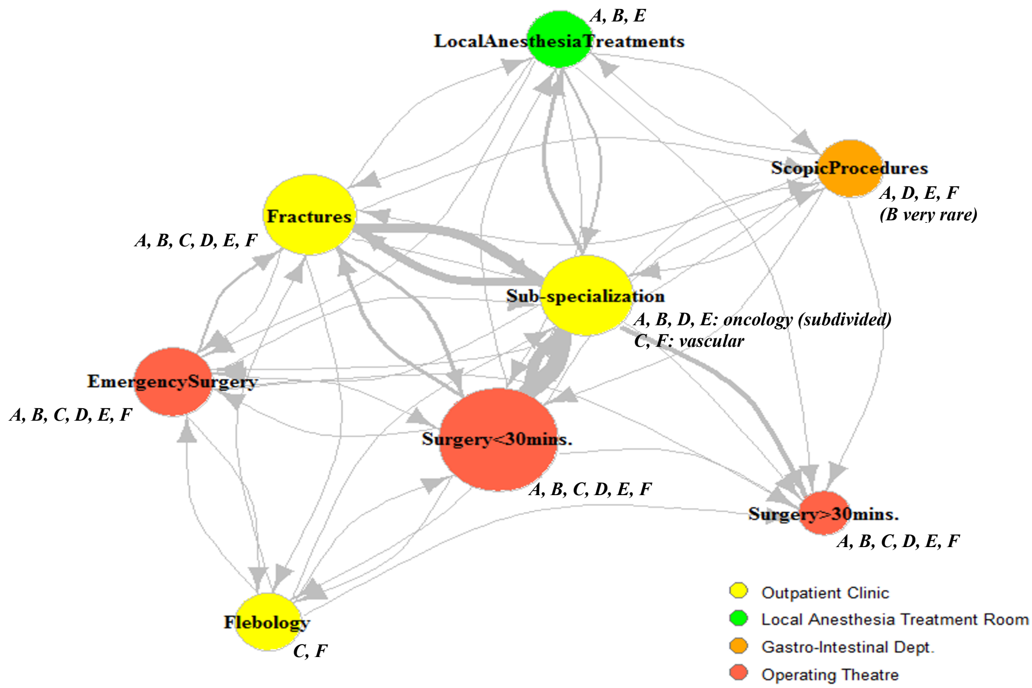


**Figure 1** Patient care path network diagram

As in many hospitals, the actual scheduling of appointments is done manually at our case hospital. The nurses use a documented structure that dictates the activities into which the different types of patients must be scheduled, and with which of the six specialists. They look for a suitable appointment in the electronic calendar system (Chipsoft©). They can either select a week or day, and look at all available slots during that day/week; or they can use the search function to find the first available slots that meet the criteria. They can also take the patient's preference for a specific day or time into account, by offering him/her several options, but in practice this rarely happens because the schedule is usually rather full and the nurse will try to complete the appointment scheduling process as quickly as possible. If the required specialist is on holiday, his 'buddy' can see his patients, but this is not preferred. This will only happen for patient appointments that cannot wait. The system allows appointments up to 2.5 years in the future. For an elective surgical procedure, patients are added to a wait list with an urgency indicator, and scheduled into a surgery session accordingly. In our model, patient preference is not taken into account. The scheduler searches for an available time slot in the patient's time window, starting in the middle of the time window and hopping back and forth, continuing after the time window if no available slot is found in the window. If the wait time exceeds one week, the patient will also try the buddy specialist, and the specialist with the shortest wait time will be selected. In the model, patients are scheduled 6 weeks in advance, because the dynamic RC is created 6 weeks in advance, and no cyclic template exists to schedule patients further into the future. In practice, about a year in advance, the hospital allocates surgery sessions to the GSD, based on surgery quotas agreed with insurance companies. Then the department's experienced head planner distributes these sessions among the specialists about six weeks in advance, based on the specialists' availability and their patient wait lists and urgency levels. The next step is to allocate the specialists to the outpatient activities, e.g., Fractures, Endoscopy, etc. This is done by an experienced head nurse, in collaboration with the head planner, following a cyclic RC template. The template includes one specialist who has a standard half day 'free' to substitute for an absent colleague. The problem is that although the RC follows a cyclic pattern, due to all sorts of specialist availability restrictions, in the end a pattern is hardly recognizable. In addition, on national holidays, or during quality inspection (full day), or when specialist(s) are away at conferences, it is left to chance which sessions fall on that day and are thus cancelled. These effects are often felt closer to the day, and to cope with demand-supply imbalances, the planner and head nurse will sometimes try to switch sessions around, leading to a time-consuming process of patient rescheduling. Or they will try to let another available specialist take over his colleague's session. This too is not ideal, because patients generally prefer the specialist they already know, and the specialists each know their own patients' histories.

## 3. Minimal-variability Resource Allocation Model

The goal of our minimal-variability resource allocation model is to promote a swifter, more-even flow of patients into the system, and through their chains of 1st, 2nd, $\cdots$, nth appointments and/or surgeries. This kind of smoothing can accomplish significant gains both economically and in terms of service quality. Viccellio and Litvak (2015) give an excellent illustration for the case of admission scheduling. In the outpatient setting, we expect that in a similar way smoothing will improve patient flow, reduce peaks and troughs, and reduce wait times between care stages.

The allocation of specialists to activities by specific day-and-time happens in a three-stage process; Table 1 provides notations. In order to initialize the allocation procedure, we start with a number of typical weeks (6 weeks) in the RC. In the Cyclic and Dynamic RC we have blocked national holidays and an annual inspection day (we chose the real-life days of 2011 and onwards), annual leave, and conference leave. In general, each specialist is available 10 half days every week (Monday to Friday) and has up to 36 work days of annual leave, plus an arbitrary number of days for conference leave. We randomly allocated the following annual leave periods for each specialist each year: one 3-week period, two 2-week periods, two 1-week periods, seven single days, and two half days. This resembles the real-life holiday patterns, and we adhered to official restrictions with regard to holiday overlap (no more than three specialists can have annual leave at the same time, and no specialist can have annual leave at the same time as his buddy). We did this in Microsoft VBA (Visual Basic for Applications), using Excel's built-in RAND() function to randomly select and block the cells in the RC corresponding to the annual leave periods. After all annual leave, national holidays, and inspection have been blocked, allocation of activities can begin.

**Table 1** Resource Calendar parameter descriptions

| Parameter | Description |
|---|---|
| $j \in J\{1,2,\cdots,6\}$ | Specialists |
| $a \in A\{1,2,\cdots,8\}$ | Activities |
| $i\ \{1,2,\cdots,N\}$ | Index for each period equal in length to *planspan, planspan* is period *N*. |
| $o_j$ | Number of half days available in *planspan* to which *a* can be allocated. |
| $n_{j,a,i}$ | Number of instances of *a* in the schedule for each *j,* in period *i* (periods indexed by *i* being equal in length to *planspan).* |
| $\mu_{j,a}$ | Mean number of instances of *a* for each *j* over period $i=1$ to $i=I$ . |
| $x_{j,a}$ | Decision variable: number of instances of activity $a$ per *j* in the *planspan.* |
| $D_{j,a}$ | Definitive number of instances of *a* allocated to *planspan* for each $j$ . |
| $CV_{j,a}$ | Coefficient of variation in the number of instances of *a* per $j$ , taken over all periods $i$ in the planning horizon. |

### 3.1 Stage 1

The first stage is the allocation of candidate activities to the planspan period for each specialist $j \in J$ . In our case the planspan period equals one week and will from here on simply be referred to as planspan. The allocation of candidates is done by adhering to the cyclic RC average instances per activity and per specialist. These averages are adjusted for the number of activities cancelled due to national holidays, inspection and annual leave, with a multiplier <1. We label this type of adjustment by CEAA (cyclic empirical averages adjusted). For the low frequency activities, 4, 5, and 6, there are weekly minimum and maximum restrictions. If these restrictions are not met after the temporary allocation stage, the model will 'force' them in or out of the definitive frequency arrays. For all eligible specialists, we check which specialist will need the activity soonest, comparing to CEAA, and add an instance of activity 4, 5 or 6; and if the maximum is exceeded we remove an instance from the busiest specialist's candidate allocation array, i.e., the specialist who has the fewest half days available after the candidates allocation (which can be negative).

### 3.2 Stage 2

In the second stage, we check for each $j$ whether the sum of his candidate activities and definitive activities does not exceed the number of available half days in that week. Note that any national holidays are taken into account, and we assume that any conference leave, annual leave, or other unavailability (except short-notice sick leave) has been submitted for the *planspan* week, as per the GSD's regulations (requirement of six-week notice). If the sum of candidates does not exceed availability, then the candidates become definitive allocations. If, however, the sum of candidates exceeds availability, then we follow a prioritizing Deletion Algorithm (DA), which is presented later in this section. All allocated activities' durations are in accordance with empirical durations per activity per specialist (allocated randomly, rounded up or down to the nearest 5 minutes, as this is the smallest patient appointment slot; and totals are matched with the cyclic RC totals for fair comparison). Note that the allocations of candidates $x_{j,a}$ and the definitive allocation $D_{j,a}$ are dynamic arrays. As item instances are removed from $x_{j,a}$ they are added to $D_{j,a}$. The Deletion Algorithm (DA) prioritizes $a$ if we have more candidates than half days available in *planspan*. The standard deviation of the number of instances of each $a$ per week is given by:

$$\sigma_{j,a} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(n_{j,a,i} - \mu_{j,a}\right)^2} \quad a \in A, j \in J. \quad (1)$$

Where $n_{j,a,i}$ is specialist $j$'s number of instances of activity $a$ in week $i$, and $n_{j,a,N}$ is the number of instances of activity $a$ in the *planspan* period, i.e., the decision variable, also denoted by $x_{j,a}$. Furthermore, $\mu_{j,a}$ is the average number of instances of activity $a$, specialist $j$, so far in the RC. When prioritizing $a$, our objective is to minimize the coefficient of variation ($CV$) for every $a$, which is the standard deviation divided by the mean, hence,

$$\min \quad CV_{j,a} = \frac{\sigma_{j,a}}{\mu_{j,a}} \quad (2)$$

$$\text{s.t.} \quad \sum_{a=1}^{A} D_{j,a} \le O_j \quad (3)$$

Our goal is to always minimize the $CV$ for each activity and each specialist separately. Minimizing e.g. the sum of all $CV_a$ or the average CV of all $a$ together, would not suffice, because imbalances could exist between the specialists, and between activities. However, the number of instances to allocate, of each activity, depends on the number of instances of other activities already allocated. This makes the allocation problem on hand complex. We cannot simply select combinations of activities out of the candidates of each $j$, because then we cannot accurately evaluate the individual $CV$s. In our approach, we solve the problem of activity allocation by using an elimination technique in the form of an algorithm, DA, that is initialized every time $\sum_{a=1}^{A} x_{j,a} > O_j$. This algorithm is described next.

The goal of our DA is to prioritize as follows. When $\sum_{a=1}^{A} x_{j,a} > O_j$, identify and accept the activities for which inclusion as definitive activity, versus exclusion, has the most detrimental effect on $CV_{j,a}$.

Deletion Algorithm: For $\forall j \in J$,

1: Compare all $x_{j,a}$ within specialist $j$. Let $x_{j,a_1} \le x_{j,a_2} \le x_{j,a\ldots} \le x_{j,a_r}$, hence $x_{j,a}$ is ordered, and $x_{a_r}$ is the largest number of candidate instances among all activities, for a particular specialist.

2: While $x_{j,a_r} > x_{j,a_{r-1}}$, one instance of candidate activity $a_r$ is accepted and becomes a definite allocation. Hence, $x_{j,a_r}$ becomes $x_{j,a_r} - 1$ and $D_{j,a}$ becomes $D_{j,a} + 1$.

3: If $x_{j,a_r} = x_{j,a_{r-1}}$, that means that there is a tie. Then for all $x_{j,a}$ equal to $x_{j,a_r}$, calculate the *CV* including $x_{j,a}$ for the *planspan* period. We label this $CV^{C_{j,a}}$: the coefficient of variation when the *candidate* allocations $x_{j,a}$ are included.

4: For all $x_{j,a}$ equal to $x_{j,a_r}$, calculate the *CV* over all $i$ periods in the planning horizon, but this time include 0 for the *planspan* period. This effectively represents the worst-case scenario, i.e., what would the activity's *CV* be if no activities of this type were to be accepted in the *planspan*. We label this $CV^{0_{j,a}}$.

5: Calculate the degradation $Deg^{j,a}$ = $\left| CV^{T_{j,a}} - CV^{0_{j,a}} \right|$. That is, for each *a* in the tie, we check how much worse off it is if 0 instances are accepted for *planspan,* versus if all candidates instances are accepted.

6: For $\max\left\{Deg^{j,a}\right\}$, allocate one instance of $x_{j,a}$ to $D_{j,a}$, and remove that instance from the candidates $x_{j,a}$; hence, $x_{j,a}$ becomes $x_{j,a} - 1$ and $D_{j,a}$ becomes $D_{j,a} + 1$. In case of a tie, give priority to surgery activities ($a = 8$, *then* $a = 7, \cdots, a = 1$) since these activities in general have a higher priority at the case hospital.

7: Repeat from step 1 while $D_{j,a} < O_j$.

8: Once $D_{j,a} = O_j$, go to the third stage (detailed below) to 'write' the activities into the RC by allocating $D_{j,a}$ to the half days that are available, according to each specialist's priority list.

### 3.3 Stage 3

The third and final stage is the allocation of the definitive activities to the days of the week, and to AM or PM. This is done according to a preference list for each ($j, a$) combination. For example, specialist $j$ has a list for activity $a$=2 in order of preference: (1) MonAM, (2) FriPM, (3) MonPM, (4) WedAM, (5) WedPM, (6) ThuAM, (7) TueAM, (8) FriAM (9) ThuPM, and (10) TuePM. Each preference list has all 10 half days listed. The algorithm takes same-half-day minimum and maximum restrictions into account, e.g. only one specialist can have activity 2 on a single half day, and no more than 2 specialists can have activity 2 or 3 on a single half day. The allocation happens in a random order of specialists every time (i.e., every week), and if any specialist's half day preference would violate these restrictions, his next half day preference is tried.

## 4. Results

Based on the case study, we have implemented a patient scheduling simulation. As in Vermeulen et al. (2009), we use this simulation to evaluate different capacity allocation approaches. Similarly, the case inputs of our simulation model are based on the case described in the previous section. These elements together with our resource allocation model are the inputs of our simulation. At the beginning of a simulation run, the simulation reads-in the entire cyclic RC, as well as all empirical patient appointment data. Well aware of the trade-off of using empirical data rather than theoretical distributions (Kelton et al. 2015), we have chosen to use the empirical patient appointment data and feed it directly into the simulation. By using the historical data directly, no values other than those recorded can be experienced; but if we sample for a fitted probability distribution, it is possible to lose important characteristics such as sequential appointment patterns. Maintaining sequential patterns was our main concern, since we approach the scheduling problem from a care chain perspective. We therefore chose not to fit distributions to each patient appointment separately. In addition, due to data limitations, some patient groups were relatively small, with $n$ smaller than 30, making distribution fitting risky. We chose not to combine small patient groups into larger groups, because their scheduling rules and patient group characteristics differ for scheduling purposes. Furthermore, generally used goodness-of-fit tests, such as chi-square and Kolmogorov-Smirnov, have notoriously low power (Bratley et al. 1987). This means that even if your data does not in fact come from the theoretical distribution fitted, the probability of rejecting the fit is small, except in those rare cases where there is a huge amount of data available.

A patient arrival is the first request for an appointment. Patients arrive one-by-one during operating hours: Monday till Friday, not on national holidays. Only emergency patients arrive during weekends and holidays, who require no scheduling and are served immediately either by the Accident and Emergency department (outpatients) or by our department's specialist on-call. Their follow-up appointments are however scheduled. We randomize patient arrivals by using a uniform distribution *unif*(1, 3980) to select empirical arrival dates for all patients from the list. Upon arrival, new patients (first appointment) are given an appointment by searching for the first available slot within their urgency-related time windows, for the activity and specialist they require. Repeat patients (not first appointment) are given an appointment by searching for an available slot within their given time windows too, but the search starts in the *middle* of the time window and then hops back and forth until an available time slot is found for the activity and specialist required.

We use Service Level (SL) and Wait Time (WT) as performance measures. The former expresses that patients' need to be scheduled within their planning windows. Similar to Vermeulen et al. (2009), we take the percentage of patients scheduled on time, i.e., within their time windows, as performance measure, per patient group (new, return, and surgery). At the case hospital, and especially in accordance with government regulations for new patients, it is determined what kind of planning windows different patients have, and this should be adhered to as much as possible, across all patient groups. Therefore, it is equally important for all patient groups to be served within their planning windows. SL is weighted according to the patient's appointment duration. This means that serving a long-appointment patient on time is weighted according to the longer duration. This way a scenario in which more long-appointment (e.g., surgery) patients are served can be compared fairly to a scenario with relatively few long-appointment patients. WT is the time (days) the patient has to wait from the end of his/her time window until his/her appointment, if the appointment is booked after his/her time window. If the appointment falls in the time window, WT

= 0. We present SL and WT for three patient groups: new-patient appointments, repeat-patient appointments, and surgery, and we give the total for all patients grouped together. We compare the Dynamic RC with the Cyclic RC and with the Real 2011 RC. All three RCs start out with the exact same amount of total capacity. This has been made possible by fine-tuning the (random) allocation of annual leave and conference leave.

We follow the guidelines of Robinson, using Welch's method, in order to determine the minimal simulation run length (Robinson 2004). The cumulative means and convergence of the main performance indicators SL and WT are calculated and convergence remains well below 5% after about 24 and 112 weeks respectively, in the Cyclic RC scenario. In the Dynamic RC scenario it converges faster. We round up to a run length of 208 weeks, or 4 years. Although the GSD department model is a non-terminating simulation, we choose to do 5 replications where a rule-of-thumb is to use at least 3-5 replications and more if data is noisy. With 5 replications our cum. mean average gives a 0.71% deviation for the Cyclic RC and 2.57% for the Dynamic RC, both well below the aim of 95% C.I.

**Table 2** RC characteristics and results for Service Level (SL), Wait Time (WT), and Capacity Used. Standard deviations* in brackets.

| | Dynamic RC | Cyclic RC | Real 2011 RC |
|---|---|---|---|
| *RC Characteristics* | | | |
| Patient-Rel. Activities | 14,008 | 14,008 | 14,008 |
| Total minutes avail. | 1,225,915 | 1,225,915 | 1,225,915 |
| No. of runs of DA p/y | 392 | ---- | ---- |
| *Performance* | | | |
| SL (%) | | | |
| All app. | 39 (0.01) | 35 (0.00) | 34 (0.01) |
| New-Patient app. | 47 (0.01) | 57 (0.00) | 43 (0.01) |
| Repeat Patient app. | 44 (0.00) | 48 (0.01) | 42 (0.01) |
| Surgery app. | 31 (0.01) | 14 (0.01) | 23 (0.01) |
| *WT (days)* | | | |
| All app. | 7.4 (13.1) | 10.4 (29.0) | 8.7 (13.3) |
| New-Patient app. | 6.6 (9.8) | 5.8 (8.9) | 8.5 (10.8) |
| Repeat Patient app. | 6.5 (9.4) | 5.4 (8.7) | 7.2 (10.0) |
| Surgery app. | 45.8 (29.4) | 122.5 (75.6) | 49.5 (26.4) |
| *AppMinsUsed* | 1,147,275 | 1,116,277 | 1,134,494 |
| *Capacity Used (%)* | 93.6% | 91.1% | 92.5% |

*Stdev. is between runs for SL ($n$=5), and between patients within their groups for WT (New-patient appointments $n$≈77,000, Repeat-patient appointments $n$≈300,000, Surgeries $n$≈38,000, All $n$≈400,000).

The first two rows in Table 2 show that all three RCs start out with an equal total amount of capacity. In the dynamic approach we have fine-tuned the algorithm in order to produce the exact same total amount of activities per year, which allows fair comparison of scenarios. The third row indicates on average, how many times the DA is run, prioritizing activities when the number of candidate activities exceeds specialist availability.

The differences in performance between the Dynamic, Cyclic and Real 2011 RCs are substantial, and significant at a 99% confidence level. The Dynamic RC outperforms the Cyclic and Real 2011 RC in terms of SL with a difference of 4% and 5% respectively, for all patients together. The increase from 35% SL for the Cyclic RC and 39% SL for Dynamic, means that a 10% increase is attained for the Dynamic RC over the Cyclic one. The Cyclic RC attains a higher SL for New- and Repeat-patient appointments, but this performance is offset by a very low SL for surgery patients. From this result we infer that the Cyclic RC has become unbalanced due to randomly allocated annual leave and conference leave, causing variability. The Dynamic RC was able to cope with this variability by automatically compensating for e.g., lost surgery sessions, by iteratively calculating which activity types are most needed. The standard deviations of SLs (in brackets) are small, and indicate that there are no large differences between the five simulation runs conducted.

The wait times reported in Table 2 show that the Dynamic RC outperforms the Cyclic and Real 2011 RCs. For all patients together, the Dynamic RC can shorten average wait time by three days compared to Cyclic, and 1.3 days compared to Real 2011. The gain is especially palpable for the surgery appointments, where wait time average is shortened by 76.7 days and 3.7 days, compared to Cyclic and Real 2011 RC, respectively. The Cyclic RC however shows wait times slightly shorter than that of the Dynamic RC for New- and Repeat-patient appointments. This again indicates that the Cyclic RC, if no additional changes are made to it, is incapable of handling random activity cancellations, and thus becomes unbalanced.

The dynamic approach demonstrates its ability to take into account the effect of cancelled activities due to holidays, annual leave etc., and to redistribute activities among the specialists. In other words, in the cyclic approach, when a certain activity is cancelled due to a holiday, the cyclic pattern dictates when that activity will take place again. A specialist is simply subject to a degree of luck for the activity that gets cancelled for him. In the Real 2011 approach, the department's head planner used his best judgment to reschedule activities that they deem necessary, with great effort. The dynamic approach allows for quicker compensation and automatically 'staying on-par' under the effects of variability caused by holidays and leave.

## 5. Conclusions and Limitations

The results of the simulation experiment show strong support for the minimal-variability data-driven approach to specialist allocation. It shows that the dynamic RC allows for more patient appointments scheduled on time, shorter overall wait times, and higher resource utilization. This automated approach can significantly alleviate and support the head planner's difficult task of scheduling and re-scheduling activities.

As opposed to nurse scheduling, which is uniform from one hospital to the next, physician (or specialist) scheduling is much more hospital-centric, because more complex labor agreements and individual contract clauses that physicians are able to negotiate, make their scheduling problem less general (Brunner et al. 2009, Fugener et al. 2015). Generalizability in the complex setting of health care resource allocation is a catch-22: we depart from general types of resource allocation problems, but we want to prove, and illustrate, that our methods and approaches are capable of finding good solutions under real-life constraints. But it is often these real-life constraints that make the problem very case-specific (Brucker et al. 2011). When the problem is relationally complex, simulation proves to be useful and flexible, allowing sufficient degree of detail. But constructing realistic models with high degree of detail, many interrelations, shared resources, complex patient flows, and considerable variability, has the downside that models become very specific to the case setting, producing results that can be difficult to generalize. Our aim was therefore more directed at generalizing a particular set of results to a broader theory: the theory of Swift Even Flow, rather than to other case-specific problems. No set of cases, no matter how large, is likely to deal satisfactorily with the complaint that it is difficult to generalize from one case to another (Yin 2013).

The use of cyclic schedules is common in hospital departments in the Netherlands and elsewhere, as is the wide variety in specialists' day-to-day tasks, and their availability (Beaulieu et al. 2000, Guo et al. 2004). But reusing a hospital model in a different hospital requires some customization. This is a particularly difficult task when someone other than the modeler undertakes it (Gunal 2012). The new user will have to learn how the model works first, and format input data to make it 100% compatible. Regardless of who the user is, there is no guarantee that a valid model for hospital A will also be valid for hospital B, or even department B. The new user for hospital/department B still has to do validation and verification to make sure that the model behaves as it should. System characteristics may differ greatly between different hospitals or departments, and could make the development of a new model more appealing than adapting and customizing an existing one. For example, at some hospitals the specialists may strictly hold their own patient lists and thus have no type of buddy system in effect. Or on the contrary, as is the case at high-volume public hospitals in Hong Kong, all specialists on the same team and specialty treat all patients of that specialty. Adapting our model to a system such as the latter might not have merit. Such a shared-patients system is also subject to variability in roster activities, but one would need to seek different ways to cope with it, or mitigate its effects. For example, one could seek to schedule daily activities flexible and dynamically on a department level, rather than individual specialist level.

Implementation of our new, dynamic data-driven allocation method would require a substantial change in mentality among hospital staff, to switch from a cyclic pattern to a dynamic schedule. The projected benefits need to be clear and rather certain, in order to foster support for such a change. In interviews in 2011, four of the six specialists indicated that if a dynamic approach would significantly outperform their current approach in terms of efficiency and throughput, then they would be willing to switch activity 1, their administration activity which can be performed from home, around. Changing the days on which the other activities take place would be a less severe change for them, although their preferences should be taken into account as much as possible.

In future research, more-sophisticated models for scheduling surgery patients can be integrated into the current approach. This item was subject to simplification most of all in our model. Surgery scheduling entails a large body of research on its own (Agnetis et al. 2014, Cardoen et al. 2010, Day et al. 2012, Molina-Pariente et al. 2015), and synergies should be sought in terms of surgery scheduling in combination with other appointment scheduling, for the same patient. Then allocation of activities can become more refined, and more realistic, e.g., for difficult surgical procedures where specialists operate together on a single patient.

The fact that our simulation model strictly adheres to scheduling rules and restrictions (with exceptions programmed to a viable level of detail), means that some human decisions that deviate from the rules and procedures, are not modeled. For example, in the model, no double-booking can take place (assigning two patients to a single time slot). This makes model validation a very difficult task, urging us to focus more on verification (Kelton et al. 2015).

## References

[1] Agnetis, A., Coppi, A., Corsini, M., Dellino, G., Meloni, C. & Pranzo, M. (2014). A decomposition approach for the combined master surgical schedule and surgical case assignment problems. Health Care Management Science, 17 (1): 49-59.

[2] Beaulieu, H., Ferland, J.A., Gendron, B. & Michelon, P. (2000). A mathematical programming approach for scheduling physicians in the emergency room. Health Care Management Science, 3 (3): 193-200.

[3] Bratley, P., Fox, B.L. & Schrage, L.E. (1987). A Guide To Simulation (2nd ed.). New York: Springer Science & Business Media.

[4] Brucker, P., Qu, R. & Burke, E. (2011). Personnel scheduling: models and complexity. European Journal of Operational Research, 210 (3): 467-473.

[5] Brunner, J.O. Bard, J.F. & Kolisch, R. (2009). Flexible shift scheduling of physicians. Health Care Management Science, 12 (3): 285-305.

[6] Cardoen, B. Demeulemeester, E. & Beliën, J. (2010). Operating room planning and scheduling: a literature review. European Journal of Operational Research, 201 (3): 921-932.

[7] Cayirli, T. & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. Production and Operations Management, 12 (4): 519-549.

[8] Day, R., Garfinkel, R. & Thompson, S. (2012). Integrated block sharing: a win-win

strategy for hospitals and surgeons. Manufacturing & Service Operations Management, 14 (4): 567-583.

[9] Fugener, A., Brunner, J.O. & Podtschaske, A. (2015). Duty and workstation rostering considering preferences and fairness: a case study at a department of anaesthesiology. International Journal of Production Research, 53 (24): 7465-7487.

[10] Gunal, M.M. (2012). A guide for building hospital simulation models. Health Systems, 1 (1): 17-25.

[11] Guo, M., Wagner, M. & West, C. (2004). Outpatient clinic scheduling - a simulation approach. In: Ingalls, R.G., Rossetti, M.D., Smith, J. & Peters, B.A. (eds.), Winter Simulation Conference, 1981-1987, Washington DC, December 2004.

[12] Kelton, D.W., Sadowski, R.P. & Zupick, N.B. (2015). Simulation with Arena (6th ed.). McGraw-Hill Education, New York.

[13] Ma, G. & Demeulemeester, E. (2013). A multilevel integrative approach to hospital case mix and capacity planning. Computers and Operations Research, 40 (9): 2198-2207.

[14] Măruşter, L., Weijters, T., De Vries, G., Van Den Bosch, A. & Daelemans, W. (2002). Logistic-based patient grouping for multi-disciplinary treatment. Artificial Intelligence in Medicine, 26 (1-2): 87-107.

[15] Molina-Pariente, J.M., Fernandez-Viagas, V. & Framinan, J.M. (2015). Integrated operating room planning and scheduling problem with assistant surgeon dependent surgery durations. Computers & Industrial Engineering, 82: 8-20.

[16] Robinson, S. (2004). Simulation: The Practice of Model Development and Use. John Wiley & Sons, Ltd., Hoboken, NJ.

[17] Schmenner, R.W. & Swink, M.L. (1998). On theory in operations management. Journal of Operations Management, 17 (1): 97-113.

[18] Vermeulen, I.B., Bohte, S.M., Elkhuizen, S.G., Lameris, H., Bakker, P.J.M. & Poutré, H.L. (2009). Adaptive resource allocation for efficient patient scheduling. Artificial Intelligence in Medicine, 46 (1): 67-80.

[19] Viccellio, P. & Litvak, E. (2015) Seven-day week approach can transform hospitals: Can we reduce hospital overcrowding without hurting the economy? The Irish Times, 08 Apr 2015, Dublin.

[20] Vissers, J. & Beech, R. (2005). Health Operations Management. Patient Flow Logistics in Health Care. Routledge, London and New York.

[21] White, D.L., Froehle, C.M. & Klassen, K.J. (2011). The effect of integrated scheduling and capacity policies on clinical efficiency. Production and Operations Management, 20 (3): 442-455.

[22] Yin, R.K., (2013). Case Study Research: Design and Methods (5th ed.). Sage Publications, Inc.

**Monique Bakker** is a Ph.D. candidate in the Department of Systems Engineering and Engineering Management at City University of Hong Kong. She has a Bachelor degree in business administration, and an M.Phil. in healthcare operations management from the University of Groningen.

**Kwok-Leung Tsui** is Head of the Department of

Systems Engineering and Engineering Management, and Chair professor of Industrial Engineering at City University of Hong Kong. He received his Ph.D. in statistics from the University of Wisconsin at Madison. Dr. Tsui was a recipient of the NSF Young Investigator Award in 1992. He was the (elected) President and Vice President of the American Statistical Association Atlanta Chapter in 1992-1993; Chair of the INFORMS Section in Quality, Statistics, and Reliability in 2000; and the Founding Chair of the INFORMS Section in Data Mining in 2004. He is a fellow of the American Statistical Association, and U.S. representative in the ISO Technical Committee on Statistical Methods. His current research interests include data mining and surveillance in healthcare and public health, calibration and validation of computer models, bioinformatics, process control and monitoring, and robust design and Taguchi methods.