Dynamic job assignment: A column generation approach with an application to surgery allocation

Troels Martin Range, Dawid Kozlowski, Niels Chr. Petersen

PII: S0377-2217(18)30537-X
DOI: 10.1016/j.ejor.2018.06.014
Reference: EOR 15199

To appear in: *European Journal of Operational Research*

Received date: 8 August 2017
Revised date: 6 June 2018
Accepted date: 7 June 2018

Highlights

- We develop a surgery allocation model handling continual patient arrivals.
- We implement column generation approach with a stochastic knapsack pricing problem.
- We introduce constraints handling service levels for categories of patient.
- Two allocation policies are compared to a First-Come First-Served policy. A simulation study shows that our model performs better than a myopic approach.

Dynamic job assignment: A column generation approach with an application to surgery allocation

Troels Martin Range[a,b,*], Dawid Kozlowski[c], Niels Chr. Petersen[c]

[a] *Hospital of South West Jutland and Institute of Regional Health Research: Centre of South West Jutland, University of Southern Denmark, Finsensgade 35, DK-6700 Esbjerg, Denmark*
[b] *Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Alfred Getz veg 3, NO-7491 Trondheim, Norway*
[c] *Department of Business and Economics, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark*

Abstract

We consider the assignment of jobs to heterogeneous agents in a dynamic system with a rolling time horizon. An example is a hospital operating theatre where the jobs are surgeries and the agents are the surgeons. The paper is presented in the context of surgery allocation and the system is characterized as follows: Patients are grouped into categories and they arrive continually following a stochastic process. Patients in each group have specific time limits within which they need treatment and if it cannot be accommodated then the patients are outsourced. The service level is the percentage of patients in each group treated within the time limit. Surgery durations are stochastic and depend on the surgeon conducting the surgeries. Each surgeon has limited time available and expected overtime is penalized by a non-decreasing convex function. We develop a column generation approach for the assignment of already arrived patients and tentative future patients to surgeons on specific days. It balances the conflicting objectives of including as many arrived patients as possible within their time limits, maximizing the service level of future patients, and minimizing the expected overtime of surgeons. A computational study is conducted with the model embedded in a rolling time horizon frame. The study indicates that the assignment of patients based on our model increases system performance in terms of service level and reduced overtime compared to a First-Come-First-Served (FCFS) policy when the arrival rates of patients are medium to high compared to the capacity of the system.

Keywords: OR in health services, Surgery Allocation, Generalized Assignment Problem, Stochastic Knapsack Problem, Simulation
2010 MSC: 68M20, 90C11, 37M05

[*]Corresponding author
Email address: Troels.Martin.Range@rsyd.dk (Troels Martin Range)

1. Introduction

Surgical costs account for a significant share of total hospital costs, and the operating theatre (OT) is a pivotal cost driver at any hospital. Part of this cost arises from salary to staff as well as capital costs of having the operating rooms (ORs) available with the necessary equipment. Efficient scheduling of resources is the key to keeping costs under control. Scheduling is challenged by several factors. First, the underlying problem is combinatorial by nature and is often subject to constraints making it hard to solve. Second, decisions are made in a dynamic environment with new patients arriving continuously.[1] Third, surgery times are stochastic and should be treated accordingly. These factors in combination make scheduling decisions for an OT particularly difficult.

Running an OT requires decisions within different time horizons. Long term decisions relate to the strategic level of planning and address the issue of capacity, while medium and short term decisions relate to allocation and scheduling of capacity with an increasing level of detail (May et al., 2011). To illustrate, the number of surgeons is decided and allocated to blocks of surgery in the medium term, and patients are next allocated to blocks of surgery in the short term.

Different levels of planning require different types of data. In the long term, data is highly aggregate and must be forecasted. The movement into a shorter time horizon requires more disaggregate data and provides the possibility for more precise schedules and allocations (Bitran and Tirupati, 1993). The focal point in the literature on short term scheduling is the allocation of patients, who have arrived and been diagnosed. Future stochastic patient arrivals are usually ignored or addressed by a simple assignment of unused blocks to potential future arrivals. We develop a method for an optimized allocation of surgery dates to patients that utilizes data on already arrived patients along with basic distributional characteristics of future arrival patterns and surgery durations without the assumption of any exact distributional forms. We consider the scenario with the predominant paradigm of advance scheduling, where patients are given an appointment in the future rather than notified on the day of their appointment. A patient must be offered surgery not later than a due date dependent on type of diagnosis or patient category. The scheduling of an arriving patient should take the expectations regarding future arrivals into account, because an appointment blocks for or increases the risk of overtime caused by a future booking, since it cannot be removed once it has been booked. For example, future cancer patients will often be associated with due dates within the planning period and must be handled as an integral part of the planning process. We know the number of already arrived and diagnosed cancer patients. We do not know the precise arrival times and diagnoses of future patients, but an estimate of the number of cancer patients arriving during the next planning period can be made available and taken into account in the planning process.

In this paper we focus on the medium term assignment of patients to surgery dates. We introduce the dynamic aspect into a combinatorial model with stochastic surgery times by utilizing information on potential future patient arrivals. The model yields surgery dates for patients known to the system as well as tentative surgery dates for potential patients who have not yet arrived. It allocates patients

[1] For this reason, a rescheduling is required on a regular basis.

and potential surgeries to combinations of surgeons and dates such that the expected overtime for surgeons is minimized while minimizing the tardiness of the system and ultimately the expected number of patients who cannot be treated within a predefined deadline. To test the effect of different allocation policies for the OT we embed this model into a rolling horizon framework simulating patient arrivals.

The underlying combinatorial model is a generalized assignment problem (GAP), where already arrived patients are assigned to a combination of a surgeon and a date. Potential future patients are also assigned to a surgeon-date combination, and the GAP model is augmented by a set of service-level constraints measuring the expected number of future and not yet arrived patients who *cannot* be treated within a prespecified deadline given the already allocated surgeries of potential patients. The assignment of more potential surgeries to the available surgeon-date combinations will lower the expected number of patients not treated within the relevant deadlines and increase the level of service. However, surgery times are stochastic, and the assignment of more known as well as potential future patients to any given surgeon-date combination involves a higher risk of overtime for the surgeon on that day. Overtime in turn increases the direct cost of the schedule. In addition, the need for reassignment of patients to a new day for surgery due to a violation of a surgeon's maximum workload increases. We model this by a strictly convex cost function in expected overtime.

A column-generation-based method is developed for solving the augmented GAP. The main variables in the problem correspond to feasible allocations of known patients and potential surgeries to surgeon-date combinations. The number of such variables is huge, and for this reason the relevant columns are generated by solving a set of pricing problems – one for each surgeon-date combination. The pricing problems turn out to be variants of the stochastic knapsack problem. We utilize a dynamic programming method based on a shortest path problem with resource constraints on an acyclic graph to solve the stochastic knapsack problem. An explicit modeling of stochastic arrival processes and service times with stochastic future arrivals incorporated into the planning problem is a main contribution of the paper.

The paper unfolds as follows. Section 2 provides a brief review of the relevant literature related to GAP and surgery scheduling. The augmented GAP model is developed in Section 3. It is described in detail how to set up constraints measuring and maximizing the service level, how to set up an extensive formulation of the surgery scheduling problem, and how to generate schedules for individual surgeon-date combinations. The static model is embedded into a rolling time horizon simulation in Section 4, and the performance of different allocation policies is tested in Section 5. Finally, concluding remarks are given in Section 6. All proofs are provided in appendix Appendix A.

2. Related literature

The assignment of patients to available surgeons on any given day in a deterministic scenario is a Generalized Assignment Problem (GAP), where each patient must be assigned to exactly one surgeon (or team of surgeons), and surgeons may be assigned to multiple tasks. Each surgeon has a capacity, for example, in terms of the number of hours available. Patients consume a certain amount of this capacity, and the combined consumption of resources by patients assigned to any surgeon is not allowed to exceed his capacity. The GAP to be considered in this paper is stochastic and dynamic.

Moccia et al. (2009) address a stochastic GAP with recourse. A given set of jobs is assigned to agents, but a random subset of jobs does not need to be processed. The assignment of jobs to agents is decided a priori, and the recourse is a reassignment of jobs from overloaded agents. The reassignment of jobs is decided upon once the subset of jobs to be executed is known. Mazzola and Neebe (2012) consider the GAP over discrete time periods within a finite planning horizon. The underlying idea is that tasks can be reassigned between agents from one period to another and that reassignments of this type are accompanied by a transition cost. Kogan and Shtub (1997) suggest a continuous-time optimal control formulation of the problem with due dates imposed for jobs and inventory as well as shortage costs incurred when jobs are finished ahead of or after their due dates. Kogan et al. (2016) extend the dynamic GAP to a stochastic environment.

Our focus is different. We do have a set of jobs to be assigned to agents. Some jobs are known, while others emanate from our expectations regarding future job arrivals. Capacity is limited, and jobs that cannot be assigned to an agent must be outsourced. Outsourcing is accompanied by a cost. The problem is to assign known and currently unknown jobs to agents in such a way that the anticipated cost of outsourcing is minimum. We consider the dynamic scenario with due dates for jobs and imposed service levels reflecting a policy for the completion of jobs within certain deadlines. A policy stating that, say, 75% of all jobs of a certain type must be completed no later than two weeks after their arrival is an example. The scenario is highly relevant in the context of patient scheduling in an OT, which for this reason defines the storyline in the development of the model. In addition, planning and scheduling of an OT is of significant importance per se, and many variants have been studied in the literature. Several reviews exist – see, for instance, Cardoen et al. (2010), May et al. (2011), Guerriero and Guido (2011), Hulshof et al. (2012), and Demeulemeester et al. (2013). On-line bibliographies are maintained by Dexter (2016) and Hulshof et al. (2011).

Deterministic models are common in cases with many interrelated resource constraints. Pham and Klinkert (2008) consider surgical scheduling in the context of a generalized job shop problem and solve this by Mixed Integer Linear Programming (MILP). Gartner and Kolisch (2014) set up MILP models with a focus on maximizing the contribution to margin. This model is embedded into a rolling horizon, and the authors show that the time between admission and surgery can be reduced significantly. Riise et al. (2016) see the surgery scheduling problem as a resource-constrained project scheduling problem and argue that this formulation can be used to solve several variants of the surgery scheduling problem. These studies share a focus on the combinatorial aspect of the problem.

Another approach for allocating patients to days is to view the system as a make-to-order (MTO) system with zero inventories. Accordingly, each patient's request for surgery is treated as an order, which is back-logged to be produced in the (near) future. The focus in MTO systems is on customer satisfaction – see, for example, Jalora (2016) – which often translates into service levels. However, it is not always possible to satisfy all orders, and for this reason a rejection of certain orders may be necessary. This is in focus in the Order Acceptance and Scheduling Problem. Examples can be found in Ebben et al. (2005) and Mestry et al. (2011) as well as in the review by Slotnick (2011).

There are two main paradigms for the dynamic scheduling of patients to days. Patients are given an appointment in the future at the time of request in advance scheduling. Patients are not scheduled in advance but notified on the day of their appointment in the allocation scheduling problem. The solution of the advance scheduling problem is made difficult compared to the allocation scheduling problem by a high dimensional state and decision variable space. Gerchak et al. (1996) address the advance scheduling problem as an aggregate planning problem, where patient arrivals by assumption are independent and identically distributed (i.i.d.) as are surgery times. The problem is modeled as a dynamic programming model, where profits are maximized and a unit-time penalty is paid for physician overtime. Truong (2015) is an extension of Gerchak et al. (1996) with multiple resources and non-stationary demand. Structural properties for an optimal advance scheduling policy are derived, and a method for construction of an optimal solution from the solution to the more simple allocation scheduling problem is developed. Min and Yih (2010) allocate patients based on priority when surgery times are i.i.d. and the capacity is scarce. Huh et al. (2013) consider the multiresource allocation scheduling problem with two classes of patients (elective and emergency) in a dynamic environment, where demand and capacity constraints may be random, non-stationary, and time correlated. The problem is modeled as a Markov Decision Process (MDP). It is not easy to solve, but structural properties for the optimal policies can be derived. The focus in these papers is on dynamic and stochastic aspects of surgery scheduling.

The assignment of patients to surgeon-day combinations is challenged by limited information on the distribution of e.g. patient arrivals and surgery durations. For that reason the development of models that do not involve the assumption of exact distributional forms has been given some attention in the literature. Focus in Kong et al. (2013) is on the determination of appointment times in an outpatient clinic with a single physician given that the number of patients and their sequence of arrivals are known. The stochastic job durations are characterized by their moments, but the complete distributions are not known. The model identifies distributionally robust schedules that minimize expected waiting time on the patient side and the physician's expected overtime. Mak et al. (2015) also suggest a distribution-free model.

Blake and Donald (2002) use a MILP model to allocate blocks of time in OTs to specific departments. Vissers et al. (2005) construct a so-called cyclic master surgery schedule, where the number of patients in each category scheduled for a day is determined such that a target throughput for respective categories is achieved. The allocation of blocks of surgery time to operating rooms is also in focus by Denton et al. (2010). Their model minimizes the cost of opening ORs as well as the cost of overtime in a stochastic setting. The authors consider blocks of time rather than individual patients. Their approach can be seen as a more aggregate model compared to the one to be suggested in the present paper.

Hans et al. (2008) investigate the (single day) surgery loading problem, where surgery times are uncertain, and patients are allocated to ORs, such that the probability for violating a hard daily limit is bounded. Lamiri et al. (2008) develop a column generation model for assigning elective patients to combinations of ORs and days, where elective patients are mixed with emergency patients in the ORs. For each OR-day combination the authors use a stochastic variable representing the time used for emergency patients and in this way obtain an expected overtime. Surgery times for elective patients are assumed to be deterministic, and the stochasticity of the model is addressed in the pricing problem. Shylo et al. (2013) assign surgeries to blocks of surgery time such that a minimal number of blocks are used in the future. They include approximations for both over- and underutilization of the blocks. Their approach is embedded into a simulation and is shown to be superior to a first-fit procedure. Finally, Gul et al. (2015) suggest a stochastic multistage MILP for the assignment of surgeries to operating rooms and days. The model is solved using the progressive hedging algorithm suggested by Rockafellar and Wets (1991).

The use of methods from management science for the scheduling of OTs with the aim of performance improvement also involves discrete event simulation. Testi et al. (2007) suggest a 3-phase hierarchical approach for the weekly scheduling of OTs combining optimization and simulation procedures. A bin-packing problem is solved in order to select the number of sessions to be allocated to each ward on a weekly basis. This is followed by the use of a blocked booking method for determining optimal time tables in terms of an assignment of wards to OTs. Finally, a simulation tool is used for an analysis of the performance of the OT under conditions of different sequencing rules. An investigation of the impact of the choice of appointment system and sequencing rules on waiting times can also be found in Westeneng (2007) with a focus on outpatient appointment scheduling. Bowers and Mould (2004) use simulation to explore the balance between maximizing the utilization of theater sessions while avoiding overruns. VanBerkel and Blake (2007) examine how an increase in throughput triggers a decrease in waiting time. Cardoen and Demeulemeester (2008) propose a discrete event simulation approach that allows for an evaluation of multiple clinical pathways and the inherent uncertainty that accompanies any clinical process. Ma and Demeulemeester (2013) use discrete event simulation to evaluate and adjust the master surgery schedule in an iterative approach. This is in turn used to enhance the trade-off between efficiency of resource utilization and the level of service. Harper (2002) suggests a simulation model for the flow of patients through the hospital that captures resource consumption over time with a focus on dimensioning. In the context of a simulation study Kim and Horowitz (2002) explore whether the use of a daily quota system with a 1- or 2-week scheduling window improves the performance of an Intensive Care Unit.

Focus in our paper is on the allocation of patients to combinations of surgeons and days in a dynamic setting. This is in some contrast to the existing literature, where the focus is either on capacity or the sequencing of patients. In the literature focusing on sequencing patients are by assumption typically known a priori as is the capacity. The capacity problem, the allocation problem, and the sequencing problem should be solved simultaneously if sub-optimality is to be avoided.

However, the problem to be solved would become highly complex, and it would be very difficult to obtain a solution with a guaranteed maximum deviation compared to the optimum. We take capacity for given, too, and address the problem of allocating patients to a set of available combinations of surgeons and days given a priori while ignoring the sequencing of patients to be addressed at the operational level. The procedure allows for an allocation of patients taking future expected arrival patterns into account. Our computational study suggests that an improved performance is obtained regarding outsourcing of patients because of a violation of imposed due dates or deadlines reflecting service levels. The aspect to be considered relates to balking in queuing theory and has to the best of our knowledge not been addressed previously in the literature.

3. A model for patient-to-day allocation

A GAP can be decomposed into a set partitioning problem and a set of knapsack problems – one problem for each surgeon on each day – and solved by a Branch-and-Price approach (see, e.g., Barnhart et al. (1998)). The model to be presented does not presuppose deterministic data. By contrast, the model is designed with the aim of obtaining an improved assignment of surgical tasks to surgeons by incorporating uncertainty regarding future patient arrivals as an integral part.

The output is a set of schedules for a given set of surgeon-day combinations indicating the (expected) set of activities to be carried out by that surgeon on that day while ignoring the sequencing of these activities. Each schedule includes a number of already arrived and known patients along with a number of slots allocated to potential surgeries for future and not yet arrived patients.[2]

The model has a finite time horizon split into individual days. The set of days is denoted $\mathcal{D} = \{1, \ldots, D\}$ and is indexed by d and δ. A set of heterogeneous surgeons, $\mathcal{S} = \{1, \ldots, S\}$, is available to conduct surgeries. The time a surgeon, $s \in \mathcal{S}$, is available on day $d \in \mathcal{D}$ is denoted $T_{sd} \geq 0$. The cost of surpassing the available time for a surgeon is a non-decreasing convex function $\Omega_s : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $\Omega(0) = 0$. $\Omega(t)$ measures the cost of having t time units of expected overtime. A surgeon with no available time on a given day cannot conduct surgeries on that day. We denote $\mathcal{R} = \{(s, d) \in \mathcal{S} \times \mathcal{D} | T_{sd} > 0\}$ as the set of feasible surgeon-day pairs. The problem is to identify a cost minimizing assignment of a combination of known and potential future patients to the set of surgeon-day pairs.

The distinction between known and potential future patients is important. We have information on arrival dates, due dates, and diagnoses for the set of already arrived or known patients. This information is for obvious reasons not available for future patients, who have not arrived yet. However, estimates of within group arrival patterns along with means and variances for the duration of surgeries are available. We denote $\mathcal{C} = \{1, \ldots, C\}$ as the index set for categories of patients. Each

[2]The slots allocated to potential surgeries can in practice be used by the planner to book patients when they arrive and can be seen as pre-booked surgeries of anonymous not yet known patients.

patient belongs to precisely one category, and patients within a category are homogeneous.[3] For each $c \in \mathcal{C}$ we use the following notation:

- $\mathcal{S}_c^{cat} \subseteq \mathcal{S}$ is the set of surgeons who can operate patients in category c.
- X_{cd} is a stochastic variable corresponding to the number of patients in category c arriving on day $d \in \mathcal{D}$.
- π_{ncd} is the probability that $n \geq 0$ patients in category c will arrive on day $d \in \mathcal{D}$.
- M_{csd}^{cat} is the maximum number of patients in category c that surgeon $s \in \mathcal{S}$ can operate on day $d \in \mathcal{D}$.
- Z_{csj}^{cat} is a stochastic variable with mean $\mu_{cs}^{cat} > 0$ and standard deviation $\sigma_{cs}^{cat} > 0$ of the surgery time of patient number $j = 1, \ldots, \max_{d \in \mathcal{D}}\{M_{csd}^{cat}\}$ in category c conducted by surgeon $s \in \mathcal{S}$. Z_{csj}^{cat} are by assumption i.i.d. for all j, all days d, and surgeons $s \in \mathcal{S}$.[4]

Patient arrivals are by assumption independent.[5]

At the time of planning, some patients are known, and some of these have already been assigned to a date of surgery as well as to a specific surgeon. This set of patients still has to be an integral part of the planning process, since we must account for their surgeries when planning new patients on the same day. The set of known patients who have arrived and been diagnosed is denoted $\mathcal{P} = \{C+1, \ldots, C+P\}$, and for a known patient, $p \in \mathcal{P}$, we use the following notation:

- $\mathcal{D}_p \subseteq \mathcal{D}$ is the set of feasible dates for surgery on patient $p \in \mathcal{P}$.
- $\mathcal{S}_p^{pat} \subseteq \mathcal{S}$ is the set of surgeons who can operate patient $p \in \mathcal{P}$.
- C_{pd}^P is the cost of scheduling patient $p \in \mathcal{P}$ for surgery on day $d \in \mathcal{D}$. If $d \notin \mathcal{D}_p$ then we put $C_{pd}^P = \infty$.
- Z_p^{pat} is a stochastic variable with $\mu_{ps}^{pat} > 0$ and standard deviation $\sigma_{ps}^{pat} > 0$ of the surgery time.

Each known patient belongs to exactly one category of arriving patients, and we might use the mean and the standard deviation for the duration of surgery for that category as the relevant mean and standard deviation for service (i.e., surgery time). However, we obtain more information, such as age and co-morbidities, when the patient has arrived, which in turn may have an impact on our estimates of mean and standard deviation. The mean and variance for the set of known patients in a

[3]Focus in this paper is on the one resource scenario, namely the set of feasible surgeon-day pairs. Patients within a given category share the same clinical pathway, i.e., utilize the same resources in the more general multiple resource scenario.

[4]Experienced surgeons typically use less time for surgeries compared to the junior surgeons. Hence different surgeons may have different means and standard deviations for a given category of patients. Compared to the case where surgeons are identical the problem becomes more complex when the surgeons are heterogeneous.

[5]This assumption may not hold true for emergency patients who arrive from, for example, traffic accidents.

specific group is therefore adjusted based on this information, and the mean and standard deviation for individual patients are allowed to be distinct.

The cost of scheduling a specific patient on a specific day is a cost related to the patient and indicates a prioritization of the patients. That is, some patients are more urgent and get a higher cost while other patients are less urgent and may get a lower cost. Furthermore the cost of scheduling a specific patient is typically an increasing function of the number of days the patient has waited due to the potential postponement of recovery or increase in severity of the patients condition.

A known patient, $p \in \mathcal{P}$, for whom we have fixed a specific date for surgery will have $|\mathcal{D}_p| = 1$, while a patient for whom we have fixed the surgeon will have $|\mathcal{S}_p| = 1$. $\mathcal{P}_f = \{p \in \mathcal{P} | |\mathcal{D}_p| = 1 \wedge |\mathcal{S}_p| = 1\}$ is the set of patients with fixed dates and fixed surgeons, and $\mathcal{P}_u = \mathcal{P} \setminus \mathcal{P}_f$ is the set of patients for whom either the date of surgery or the surgeon has not been fixed.

3.1. Modeling the service level

The treatment of patients before their due dates and imposed deadlines reflecting service levels are key issues for most hospitals. For this reason, a model designed to determine the day of surgery should include performance measures reflecting this issue. This may be obvious for known patients, but not for patients who have not arrived yet. We model an approximation for the expected number of future arrivals to be handled within a specific period – the larger the expected share of future patients to be handled within imposed deadlines the higher the level of service.

Suppose that the target for a category c is to treat at least, for example, 50% of the patients within one week, 75% within two weeks, and 90% within three weeks. We set up a measure for this by constructing a function that measures the expected number of patients in category c violating the imposed target levels given the number of preallocated surgeries assigned to category c patients in the future. This number is next compared to the expected number of future patients in category c, thus obtaining the expected share of patients not treated within the target levels.

We denote $\mathcal{L}_c$ as the set of treatment deadlines, for example, $\mathcal{L}_c = \{7, 14, 21\}$, in the example above. For each treatment deadline, $l \in \mathcal{L}_c$, we define the target portion, $H_{cl} \in [0, 1]$, of patients intended to be treated within the deadline, where $H_{c14} = 0.75$ in the example above. The requirement that H_{cl} percent of patients in category c arriving on day d should be allocated to surgery within l days translates into the following constraint:

$$\mathbb{E}[Y_{cdl}] \leq (1 - H_{cl})\mathbb{E}[X_{cd}] \tag{1}$$

where Y_{cdl} is a stochastic variable indicating the number of patients in category c arriving on day d who cannot be allocated to surgery within the target of l days. Y_{cdl} depends on the number of available pre-allocated surgery slots for patient category c after day d. Consider a given day, $d \in \mathcal{D}$, and a given category, $c \in \mathcal{C}$. We omit the subscripts for day, category, and deadline to simplify the notation, (i.e., π_n is a shorthand for π_{ncd}, X for X_{cd}, and Y for Y_{cdl}). Let $A \in \mathbb{N}$ denote the number of preallocated slots for surgery assigned to future patients of category c. $A \in \mathbb{N}$ is an upper bound on the number of patients of category c arriving on day d that can be allocated to surgery on a future day. Define a set of stochastic variables as follows:

$$Y^A = (X - A)^+, \quad A \in \mathbb{N} \tag{2}$$

where $(x)^+$ is shorthand for $\max\{0; x\}$. Y^A measures the number of patients (of category c arriving on day d) who cannot be allocated to surgery within the imposed deadline. The expected value of the stochastic variable, Y^A, can now be derived as stated in Proposition 1:[6]

Proposition 1. *Let X be a discrete stochastic variable having probability π_n of attaining value n and let $Y^A = \max\{0, X - A\}$, where $A \in \mathbb{N}$ is an exogenously given value. Then*

$$\mathbb{E}\left[Y^A\right] = \mathbb{E}\left[X\right] - A + \sum_{n=0}^{A} \pi_n \left(A - n\right) \tag{3}$$

Proposition 1 is trivial, because the LHS is the conditional expectation given that the stochastic variable Y^A must be non-negative. It is included as a means for a self contained presentation with a formal basis. Clearly, the expected number of patients who cannot be allocated to surgery decreases when the number of patients who can be allocated to surgery increases (i.e., when A increases). The expected number of patients who cannot be allocated to surgery, is only defined for integer values of A. For model building purposes we approximate this relationship by a continuous piecewise linear function passing through the points $(A, \mathbb{E}[Y^A])$ and $(A+1, \mathbb{E}[Y^{A+1}])$ for $A \in \mathbb{N}$.

Proposition 2. *The straight line passing through both $(A, \mathbb{E}[Y^A])$ and $(A+1, \mathbb{E}[Y^{A+1}])$ is described by the function*

$$f^A(x) = \left(\sum_{n=0}^{A} \pi_n - 1\right) x + \mathbb{E}\left[X\right] - \sum_{n=0}^{A} \pi_n n \tag{4}$$

The line $f^A(x)$ is of interest only for values of $x \in [A, A+1]$, such that it connects the two points $(A, \mathbb{E}[Y^A])$ and $(A+1, \mathbb{E}[Y^{A+1}])$. Letting successive functions $f^0(x), f^1(x), \ldots, f^A(x)$,... connect the sequence of points $(0, \mathbb{E}[Y^0])$, $(1, \mathbb{E}[Y^1])$, $(2, \mathbb{E}[Y^2])$, …, $(A, \mathbb{E}[Y^A])$, $(A+1, \mathbb{E}[Y^{A+1}])$, … yields a piecewise linear function:

$$g(x) = \begin{cases} f^0(x), & 0 \le x < 1 \\ f^1(x), & 1 \le x < 2 \\ \quad\vdots & \\ f^A(x), & A \le x < A+1 \\ \quad\vdots & \end{cases} \tag{5}$$

The function $g(x)$ yields the expected number of patients who cannot be allocated for any value $x \geq 0$ corresponding to a possible number of patient allocations. Figure 1 provides an example

[6] The proofs of Propositions 1-3 are provided in appendix Appendix A.

of the functions $f^A(x)$ and the function $g(x)$. Proposition 3 states the properties of the shape of function g:

Proposition 3. *Let $g : \mathbb{R}_+ \to \mathbb{R}$ be defined by* (5) *and suppose that the cumulative distribution function, F, is strictly increasing. Then g is continuous, decreasing, and convex.*

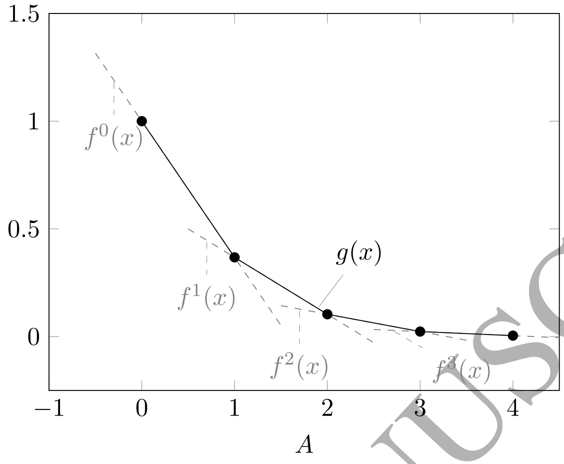


Figure 1: Illustration of $f^A(x)$ for $A = 0, \ldots, 3$ (dashed lines) and $g(x)$ for X Poisson distributed with mean 1 (full line).

The function $g(x)$ is convex and can for this reason be rewritten as follows:

$$g(x) = \max\left\{f^A(x) | A \in \mathbb{N}\right\}$$

$g(x)$ is in view of the definition of $f^A(x)$ piecewise linear. Let y denote the expectation of the random variable indicating the number of patients who cannot be allocated to surgery. y is a variable bounded from below by $g(x)$ for any given x. The following result prevails:

$$y \geq g(x) = \max\left\{f^A(x) | A \in \mathbb{N}\right\} \tag{6}$$

$$\Rightarrow \quad y \geq f^A(x), \quad \forall A \in \mathbb{N} \tag{7}$$

The minimum value of y is attained at $g(x)$. Hence, (7) provides a lower bound on the number of patients in category c arriving on day d who cannot be assigned to surgery within the target of l days as a function of x.[7]

3.2. Identification of the cost-minimizing set of schedules

For each day a surgeon is available a number of surgeries are allocated to her or him. We will refer to such an allocation as a schedule for the surgeon on that given day. All known patients

[7]Bear in mind that x in turn measures the number of operations for patients in category c arriving on day d, who can be assigned to a surgeon-day combination within the imposed deadline.

must be assigned to a specific day as well as a specific surgeon. Schedules may also include a number of tentative surgeries for patients who may arrive in the future before the relevant day for the schedule at hand. Hence, a schedule for a surgeon-day combination is an assignment of known patients in combination with a number of tentative potential surgeries. A schedule for a surgeon-day combination is said to be feasible if all known patients and planned tentative surgeries can be operated by the surgeon.

Let $\mathcal{I}$ denote the set of all feasible schedules, and let $\mathcal{I}_r \subseteq \mathcal{I}$ denote the set of feasible schedules for surgeon-day pairs, $r \in \mathcal{R}$. By assumption $\mathcal{I}_r$, $r \in \mathcal{R}$, partitions the set of all schedules, $\mathcal{I}$. Let $\mathcal{I}_d^{day} = \{i \in \mathcal{I} | i \in \mathcal{I}_{(s,d)}, s \in \mathcal{S} : (s,d) \in \mathcal{R}\}$ denote all schedules for a given day, $d \in \mathcal{D}$. For each schedule, $i \in \mathcal{I}$, we use the notation:

- c_i^S is the cost of a schedule i, which is composed of the cost of assigning known patients as well as the cost of expected overtime.
- $a_{pi} \in \{0,1\}$ is a parameter equal to 1 if and only if patient $p \in \mathcal{P}_u$ is included in schedule i.
- $b_{ci} \in \mathbb{N}$ is the number of planned surgeries for arriving patients in category $c \in \mathcal{C}$ in schedule i.

The values c_i^S, a_{pi}, and b_{ci} (see Section 3.3) can easily be determined when the subset of patients from $\mathcal{P}$ included in the schedule and the number of planned surgeries in each category are known. Known but not yet allocated patients can be outsourced if necessary. We let

- C_p^{OP} be the outsourcing cost of patient $p \in \mathcal{P}_u$.

An available surgeon-day combination can be used for surgeries. Otherwise, the OR is not open on that day. Thus, we denote

- C_r^O as the cost of opening the OR for surgeon-day combination $r \in \mathcal{R}$.

The direct costs of a schedule relate to personnel. The indirect costs relate to the cost of outsourcing known patients, the cost of opening an OR, and the cost of violating imposed service levels. For each category $c \in \mathcal{C}$ and each $l \in \mathcal{L}_c$ we let

- C_{cl}^V be the unit cost of violating the required service level l of category c (could be set to ∞ for a hard constraint).

Finally, we need the following variables:

- $\lambda_i \in \{0,1\}$ is a variable equal to 1 if and only if schedule $i \in \mathcal{I}$ is used in the solution.
- $\zeta_p \in \{0,1\}$ indicates whether or not patient $p \in \mathcal{P}_u$ is outsourced.
- $\rho_r \in \{0,1\}$ indicates whether or not surgeon-day combination $r \in \mathcal{R}$ is used.

- $x_{cd\delta} \geq 0$ is the number of tentative patients in category $c \in \mathcal{C}$ arriving on day d scheduled for surgery on day $\delta > d$.[8]

- $y_{cdl} \geq 0$ is the expected number of patients in category $c \in \mathcal{C}$ arriving on day d who cannot be allocated within the maximal time l.

- $v_{cl} \geq 0$ is the amount of violation of the required service level.

The number of patients in category c arriving on day d and allocated to a day within planning period $\mathcal{D}$ and no later than the target deadline $l \in \mathcal{L}_c$ can be computed as:

$$\sum_{\delta=d+1}^{\min\{D,d+l\}} x_{cd\delta} \tag{8}$$

Tentative patient arrivals on day d with $d + l > D$ may by assumption be allocated to days beyond the planning horizon. To be more specific, we assume in case $d+l > D$ that a portion of the expected patient arrivals are allocated to days beyond the planning horizon and that in the long run patients are distributed evenly over the potential days for surgery. Accordingly, the tentative number of patient arrivals allocated to surgery on a day beyond the planning horizon can be computed as follows:[9]

$$E_{cdl} = \frac{\max\{0; d + l - D\}}{l} \mathbb{E}[X_{cd}]$$

where $c \in \mathcal{C}$, $d \in \mathcal{D}$, and $l \in \mathcal{L}_c$. The tentative number of patient arrivals allocated to a specific surgery day, δ, is

$$\sum_{d=0}^{\delta-1} x_{cd\delta}$$

The model can now be stated as follows, provided that the complete set of feasible schedules $\mathcal{I}$ is

[8] It should be noted that only some combinations of d and δ are feasible.

[9] The approximation has the desirable properties that no patient is allocated to surgery beyond D if d+l does not exceed D and that all patients are allocated to surgery beyond D if d=D. Other approximations are available.

known along with the components described above:

$$
\begin{aligned}
\min \sum_{i\in\mathcal{I}} c_i^S \lambda_i + \sum_{p\in\mathcal{P}_u} C_p^{OP}\zeta_p + \sum_{r\in\mathcal{R}} C_r^O \rho_r & \\
+ \sum_{c\in\mathcal{C}}\sum_{l\in\mathcal{L}_c} C_{cl}^V v_{cl} & \qquad (9)
\end{aligned}
$$

$$
\begin{aligned}
\text{s.t.} \sum_{i\in\mathcal{I}} a_{pi}\lambda_i + \zeta_p &= 1, & p\in\mathcal{P}_u \qquad (10)\\
\sum_{i\in\mathcal{I}_r} \lambda_i &= \rho_r, & r\in\mathcal{R} \qquad (11)\\
\sum_{i\in\mathcal{I}_\delta^{day}} b_{ci}\lambda_i &\geq \sum_{d=0}^{\delta-1} x_{cd\delta}, & c\in\mathcal{C},\delta\in\mathcal{D} \qquad (12)\\
f_{cd}^A\left(\sum_{\delta=d+1}^{\min\{D,d+l\}} x_{cd\delta} + E_{cdl}\right) &\leq y_{cdl}, & c\in\mathcal{C}, d\in\mathcal{D}, l\in\mathcal{L}_c, A\in\mathbb{N} \qquad (13)\\
\sum_{d\in\mathcal{D}} y_{cdl} - v_{cl} &\leq (1-H_{cl})\sum_{d\in\mathcal{D}}\mathbb{E}[X_{cd}], & c\in\mathcal{C}, l\in\mathcal{L}_c \qquad (14)\\
\lambda_i \in \{0,1\}, & & i\in\mathcal{I} \qquad (15)\\
\zeta_p \in \{0,1\}, & & p\in\mathcal{P}_u \qquad (16)\\
\rho_r \in \{0,1\}, & & r\in\mathcal{R} \qquad (17)\\
x_{cd\delta} \geq 0, & & c\in\mathcal{C}, d,\delta\in\mathcal{D} \qquad (18)\\
y_{cdl} \geq 0, & & c\in\mathcal{C}, d\in\mathcal{D}, l\in\mathcal{L}_c \qquad (19)\\
v_{cl} \geq 0, & & c\in\mathcal{C}, l\in\mathcal{L}_c \qquad (20)
\end{aligned}
$$

Objective (9) minimizes the total cost of the selected set of schedules, the total cost of outsourcing known patients, the total cost of opening ORs, and the cost of violating the target service levels. Constraint (10) ensures that each known patient without a fixed surgeon-day pair either gets allocated to exactly one schedule or is outsourced. Constraint (11) imposes the requirement that each surgeon-day pair has exactly one associated schedule if the corresponding OR is opened for that surgeon-day pair. The left-hand side of constraint (12) states the number of available surgeries in category c on day δ, while the right-hand side measures how many patients in category c arriving earlier than day δ are allocated to have surgery on day δ. The right-hand side has to be no larger than the left-hand side, since we cannot operate more patients in category c than planned. The expected number of patients in category c arriving on day d not allocated to a surgery is measured by constraint (13).[10] The target service level corresponding to equation (1) is enforced in constraint

[10]This constraint corresponds to (7), where x is the amount of expected patients allocated to future surgeries.

(14) by putting a bound on y_{cdl} over the planning horizon.[11] If this is not satisfied, then v_{cl} measures the magnitude of the violation, which is penalized by C_{cl}^V in the objective function. Finally, constraints (15)-(20) state the variable types. In practice, ζ_p and ρ_r are naturally integer as long as all λ_i variables are integer. Hence, we relax (16) to $0 \le \zeta_p \le 1$ and (17) to $0 \le \rho_r \le 1$.

Consider the scenario with an empty set of constraints of type (13). This is the case where future and currently not known patients are simply not taken into account. A similar situation occurs in the scenario with $C_{cl}^V = 0$ for all c and l, since a violation of imposed service levels is not penalized in the objective function. By contrast, $H_{cl} = 1$ and $C_{cl}^V > 0$ accompanied by $\sum_{d \in \mathcal{D}} y_{cdl} = v_{cl}$ in any optimal solution is the case with a penalty imposed whenever a tentative patient cannot be offered surgery.

The number of constraints of type (13) is in principle not finite since $A \in \mathbb{N}$. Hence, we test whether any of the infinitely many constraints of this type is violated and include violated constraints in the problem. Constraints (13) are easy to separate, since we only need to check whether the constraint for c, d, l, A in $\sum_{\delta=d+1}^{d+l} x_{cd\delta} \in [A, A+1[$ is fulfilled. We add the relevant constraint and resolve the problem if this is not the case.

The number of possible schedules for each surgeon-day pair, r, is huge. For this reason we generate schedules dynamically for the LP relaxation of (9)-(20). We apply the approach known as column generation to construct an LP lower-bound solution.[12] The idea is first to remove the integrality constraints, (15), thus obtaining an LP relaxation. The number of basic variables cannot exceed the number of constraints. Hence, most of the scheduling variables, λ_i, from problem (9)-(20) can be removed (or implicitly fixed at zero), which in turn provides a restricted version of the LP relaxation of problem (9)-(20). An optimal solution for the LP relaxation is obtained, provided variables are removed or fixed at zero in an appropriate way (i.e., when the reduced cost coefficients for these variables are non-negative). For this reason we compute the minimum reduced cost coefficient over all variables. If the minimal reduced cost coefficient is negative, the corresponding variable is allowed to exceed zero. Let $\beta_p \in \mathbb{R}$ be the dual price for constraint (10) with $p \in \mathcal{P}_u$, let $\alpha_r \in \mathbb{R}$ be the dual price for constraint (11) with $r \in \mathcal{R}$, and let $\gamma_{c\delta} \ge 0$ be the dual price for constraint (12) with $c \in \mathcal{C}$ and $\delta \in \mathcal{D}$. Bearing in mind that $r \in \mathcal{R}$ reflects a combination, $r = (s, d)$, of a surgeon $s \in \mathcal{S}$ and a day $d \in \mathcal{D}$, the reduced cost coefficient for schedule $i \in \mathcal{I}_r$ with $r \in \mathcal{R}$ can be computed as

$$\bar{c}_i = c_i^S - \sum_{p \in \mathcal{P}_u} a_{pi}\beta_p - \sum_{c \in \mathcal{C}} b_{ci}\gamma_{cd} - \alpha_r \tag{21}$$

[11] The specification of (14) as an aggregation of (1) is an example of the inherent flexibility as seen from a model building point of view. The relaxation imposes the requirement that the expected number of patients of type c C who arrive within the planning period and cannot be allocated to surgery within the time limit l must not exceed the fraction $(1 - H_{cl})$ of the expected number of arrivals. The corresponding disaggregate constraint as in (1) states that the requirement is not only to be fulfilled on average but on a daily basis.

[12] The reader is referred to Barnhart et al. (1998) or Lübbecke and Desrosiers (2005) for an introduction to column generation.

Clearly, we need to identify a_{pi} and b_{ci} as well as the direct cost of the schedule, c_i^S, in order to compute a minimum reduced cost schedule. We will return to this in Section 3.3.

The model we propose is focused on the planning of the schedules for future dates. As such, it does not directly focus on the daily control of the OT. For instance, we do not include sequencing of patients in the operating rooms nor the risk of cancellation of patients due to surges of acute patients. While both issues are relevant, we have not included these in this study. However, the knapsack problem described below can be extended to a so-called Elementary Shortest Path Problem (ESPP) which will take the sequencing into account. The price to pay for this extension is that the structure of the ESPP makes the problem stongly $\mathcal{NP}$-hard, while the variant we use has pseudo-polynomial time complexity.

The issue of cancellation and the re-allocation of patients to new time-slots is important in real life. We have not explicitly taken this into account in the model. However, the model has to be resolved on a daily basis and patients canceled one day may be forced into the solution by adding them to the set $\mathcal{P}_u$ and either fix the variable $\zeta_p = 0$ such that the patient cannot be outsourced or put the outsource cost C_p^{OP} to a huge value. As a canceled patient should be treated sooner rather than later, one can also increase the cost of including these patients later significantly. While it is possible to handle cancellations in the model, we have left this for future research.

3.3. The generation of schedules

This section is concerned with the development of a model that approximates costs for potential schedules and identifies the minimum reduced cost schedule given the dual prices of the LP relaxation of model (9)-(20). The model is referred to as the pricing problem.

The decisions to be made in the pricing problem are who of the known patients and how many surgeries of each category of patients are to be included in a surgeon's schedule on a given day. Let $v_p \in \{0,1\}$ indicate whether or not a known patient $p \in \mathcal{P}$ is included in the schedule, and let $w_{cj} \in \{0,1\}$ indicate whether or not an unknown, future patient, labeled as number j in category c is included. Implicitly we assume that $w_{cj} \geq w_{cj+1}$ (i.e., patient number $j+1$ in category c can only be included in the schedule if patient j in category c is included). A fixed patient, $p \in \mathcal{P}_f$, will have the corresponding variable, v_p, fixed to either 0 or 1: $v_p = 1$ if the patient is fixed to surgeon s on day d, and $v_p = 0$ if the patient is fixed to another surgeon or another day.

In this section we will treat the surgeon-day pairs individually. For convenience we fix $r = (s, d)$, and, unless otherwise stated, we let $Z_p^{cat} = Z_{ps}^{cat}$, $Z_{cj}^{cat} = Z_{csj}^{cat}$, $\Omega(\cdot) = \Omega_s(\cdot)$, $T = T_{sd}$, $C_p^P = C_{pd}^P$, $\alpha = \alpha_r$, and $\gamma_c = \gamma_{cd}$.

The cost of a schedule, c^S, depends on the direct costs, C_p^P, of including patient $p \in \mathcal{P}$ as well as the expected overtime cost of the schedule. Let Z denote the total processing time for patients included in the schedule. Z is the sum of the realizations of the respective stochastic variables, i.e.,

$$Z = \sum_{p \in \mathcal{P}} v_p Z_p^{pat} + \sum_{c \in \mathcal{C}} \sum_{j=1}^{M^{cat}} w_{cj} Z_{cj}^{cat} \tag{22}$$

Overtime can now be written as the stochastic variable $O = (Z - T)^+$, and the expected cost of overtime can be evaluated by Jensen's inequality (Jensen, 1906): $\mathbb{E}[\Omega(O)] \geq \Omega(\mathbb{E}[O])$. Accordingly, the expected cost of a schedule can be computed as

$$c^S = \sum_{p \in \mathcal{P}} C_p^P v_p + \Omega(\mathbb{E}[O]) \tag{23}$$

Consider a solution, $i \in \mathcal{I}_r$. v_p^i indicates whether or not patient p is included in the schedule, and w_{cj}^i indicates whether or not patient j in category c is included in the schedule. Thus, $a_{pi} = v_p^i$ and $b_{ci} = \sum_{j=1}^{M_c^{cat}} w_{cj}^i$. The reduced cost of a schedule can be obtained by combining (21) and (23):

$$\bar{c}_i = \sum_{p \in \mathcal{P}} \left(C_p^P - \beta\right) v_p - \sum_{c \in \mathcal{C}} \sum_{j} \gamma_c w_{cj} + \Omega(\mathbb{E}[O]) - \alpha \tag{24}$$

Overtime, O, is computed on the basis of the included number of patients in each category. Hence, the minimum reduced cost column can be found by solving the following binary problem:

$$\min \quad \sum_{p \in \mathcal{P}} \left(C_p^P - \beta\right) v_p - \sum_{c \in \mathcal{C}} \sum_{j=1}^{M_c^{cat}} \gamma_c w_{cj} + \Omega(\mathbb{E}[O]) - \alpha \tag{25}$$

$$\text{s.t.} \quad O = \left(\sum_{p \in \mathcal{P}} v_p Z_p^{pat} + \sum_{c \in \mathcal{C}} \sum_{j=1}^{M_c^{cat}} w_{cj} Z_{cj}^{cat} - T \right)^+ \tag{26}$$

$$v_p \in \{0,1\}, \qquad p \in \mathcal{P} \tag{27}$$

$$w_{cj} \in \{0,1\}, \qquad c \in \mathcal{C}, j \in \left\{1, \ldots, M_c^{cat}\right\} \tag{28}$$

This problem is a variant of a stochastic knapsack problem (see Kellerer et al. (2004)) where the upper bound on the consumption of time is replaced by a cost of exceeding the upper bound. By assumption, we do not have the distributions for the surgery times of individual patients and patient categories. Only estimates of means and variances are available. For this reason we apply the central limit theorem to obtain an approximation of the expected overtime as stated in Proposition 4:

Proposition 4. *Let $Z_1, \ldots, Z_n$ be a set of independent stochastic variables with means μ_i and variances σ_i^2 for $i = 1, \ldots, n$. Let $Z = Z_1 + \ldots + Z_n$ and $O = (Z - T)^+$ for a constant $T \geq 0$. Denote $\mu_Z = \mu_1 + \ldots + \mu_n$ and $\sigma_Z^2 = \sigma_1^2 + \ldots + \sigma_n^2$. Then*

$$\mathbb{E}[O] \approx \sigma_Z \left(\phi(k) - k(1 - \Phi(k))\right)$$

where $\phi(\cdot)$ is the probability density function and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution and $k = (T - \mu_Z)/\sigma_Z$.

Kleywegt et al. (2002) observe without proof an analogous result in the case where the Z variables are normally distributed. Kosuch and Lisser (2010) proves the result in the case of normally distributed variables. Range et al. (2018) provide a proof for Proposition 4 and we have restated an abbreviated version of that proof in Appendix Appendix A for completeness. Proposition 4 allows for a modification of (25)-(28) into a model, where accumulated mean and variance for total surgery time provides the foundation for an approximation of expected overtime. The resulting model will correspond to the pricing problem in the column generation:[13]

$$\min \quad \sum_{p\in\mathcal{P}} \left(C_p^P - \beta_p\right) v_p - \sum_{c\in\mathcal{C}} \sum_{j=1}^{M_c^{cat}} \gamma_c w_{cj} + \Omega(e) - \alpha \tag{29}$$

$$\text{s.t.} \quad \mu_Z = \sum_{p\in\mathcal{P}} v_p \mu_p^{pat} + \sum_{c\in\mathcal{C}} \sum_{j=1}^{M_c^{cat}} w_{cj} \mu_c^{cat} \tag{30}$$

$$\sigma_Z = \sqrt{\sum_{p\in\mathcal{P}} v_p (\sigma_p^{pat})^2 + \sum_{c\in\mathcal{C}} \sum_{j=1}^{M_c^{cat}} w_{cj} (\sigma_c^{cat})^2} \tag{31}$$

$$k = \frac{T - \mu_Z}{\sigma_Z} \tag{32}$$

$$e = \sigma_Z \left(\phi(k) - k(1 - \Phi(k))\right) \tag{33}$$

$$v_p \in \{0,1\}, \qquad p \in \mathcal{P} \tag{34}$$

$$w_{cj} \in \{0,1\}, \qquad c \in \mathcal{C}, j \in \left\{1, \ldots, M_c^{cat}\right\} \tag{35}$$

Objective (29) minimizes the reduced cost coefficient of the solution found. The expected use of time is calculated in (30) and the corresponding standard deviation in (31). The approximation of expected overtime, e, is computed by constraint (33) utilizing Proposition 4, where k is computed in constraint (32). Finally, known and future patients can only be selected once, which gives rise to the requirement of binary variables v_p and w_{cj} stated in (34) and (35), respectively.

Model (29)-(35) is inherently non-linear and, consequently, we solve this by dynamic programming. However, the binary nature of the problem as well as the close relation to the knapsack problem allow us to solve the problem as a network problem. For the case where the cost of expected overtime is linear, Merzifonluoğlu et al. (2012) provide both exact and heuristic solution methods. We use the method suggested by Range et al. (2018), which can accommodate the convex cost function of expected overtime and where the knapsack problem is formulated as a resource constrained shortest path problem on a directed acyclic graph. The authors show that when the cost of expected overtime is convex, then the problem can in practice be solved fast.

4. Application in a dynamic setting

The GAP-based model presented in Section 3 can be embedded into a rolling horizon procedure. We consider a discrete time horizon of D periods ($d = 1, \ldots, D$) with each period representing, for example, a working day in a regular week. On each day patients arrive into the system according

[13] An empty allocation of patients to a surgeon-day combination is allowed. Empty allocations correspond to the case where the OR for a given surgeon-day pair is not opened.

to a pre-specified arrival process. Let $p \geq 1$ denote the period between optimizations, such that the problem is to be solved at time $t \in [0, p, 2p, \dots]$. Three different *allocation policies* are analyzed:

0. **First-come-first-served (FCFS)**: Patients are assigned to the first day with an available surgeon capable of performing the surgery. The surgeon with the lowest mean surgery time for the patient is chosen if more than one surgeon is available. Optimization is not an integral part of this policy.

1. **Pre-allocation based fixing**: Optimization is performed every p^{th} period at the end of the day. A feasible schedule is identified for each feasible surgeon-day pair, $\mathcal{R} = \{(s, d) \in \mathcal{S} \times \mathcal{D} | T_{sd} > 0\}$. Each schedule i defines the number of surgeries, $b_{ci} \in \mathbb{N}$, for future patients (excluding surgeries for known patients) in category $c \in \mathcal{C}$. Patients in category c arriving during the next p days are upon arrival given an appointment to a specific schedule, i, for which $b_{ci} > 0$ using some arbitrary allocation rule (e.g., earliest date). The immediate allocation of an arriving patient to a schedule limits the available amount of time in that schedule for future patients. Patients arriving during the period between two successive optimizations are fixed to a surgeon-day pair. Hence, $\mathcal{P}_u = \emptyset$. The optimization is concerned with an allocation of future surgeries only and is for this reason driven by the cost of violating the service level.

2. **Pool allocation**: In this allocation policy patients arriving between optimization runs are pooled and await an assignment to day and surgeon until the next time the optimization is run. Consequently, the set $\mathcal{P}_u = \mathcal{P} \setminus \mathcal{P}_f$ is not empty by construction. Both day and surgeon are decided upon as an integral outcome of the optimization procedure.

The first-come-first-served policy provides a base allocation policy to be compared to the remaining two policies. The pre-allocation-based fixing policy is convenient if a hospital wants to give an arriving patient an immediate appointment to a specific surgeon on a specific day. After a consultation with a surgeon patients are allowed to choose a day of surgery among the set of available days for that particular surgeon. The pool allocation policy is more flexible, since patients must wait for their assignment to a surgeon-day combination. The three policies are not to be considered exhaustive, but are believed to cover the scheduling process in many hospitals.

5. Computational study

This section is concerned with the performance of the model in a dynamic setting with a rolling time horizon. The two optimization-based allocation policies described above are compared to the first-come-first-served (FCFS) approach. Focus is on utilization and overtime of surgeons as well as waiting time and service level on the patient side. The numerical experiments are designed for testing the performance of the model in a dynamic setting.

5.1. The base case

A base case has been constructed inspired by a real-life scenario at a Danish Hospital where sub-acute and elective patients arrive to the hospital continually. These patients have to be booked for surgery on a future date within a time limit which is given by either regulations or the urgency of the condition that the patients arrive with. In Denmark the patients have the right to choose a private sector hospital for treatment, reimbursed by the public sector hospital, if the public hospitals are not able to treat the patient within a guaranteed limit of 30 days. Thus, there is an economic incentive to get patients treated within this limit. On the other hand, this incentive may cause the patients who are more urgent to be treated beyond their time limit because the hospital has to treat all within 30 days. The base case is constructed to reflect this trade-off. See Kozlowski and Worthington (2015) for a queuing-theory-based analysis of the trade-offs in a system with a maximal waiting time guarantee.

We consider a scenario with seven patient categories (see Table 1). Patients arrive 24/7 according to seven i.i.d. Poisson processes. We consider three different arrival scenarios – low, medium, and high arrival rates – reflecting an underutilized, a balanced, and an overutilized system, respectively.

Each already arrived and known patient in any given category faces a cost of waiting per day labeled WC which reflects patients' disutility, for example, caused by not being able to work fully as well as the potential extended recovery period due to deterioration in the condition during the waiting time. We assume that the cost of including the patient in a given schedule, C_p^P, increases linearly at a rate WC per waiting day. This cost is dependent on the patient's category and the condition of the patient.[14] Each patient is given a specific due date depending on category. Patients who are not offered treatment before their due dates are outsourced. Outsourcing costs are listed in the column labeled C_c^{OP} of Table 1.

The target service level, H_c, is fixed to 95% for all categories, c. C_c^V measures the penalty for violating the imposed service level (see Table 1). C_c^V is derived as a fraction of the outsourcing cost. Three scenarios are considered, one with no penalty for violation of the service level, N, one with half of the outsourcing cost imposed as a penalty, H, and one with the penalty set equal to total outsourcing cost, F.[15] Case N with no penalty imposed for a violation of the service level is considered *myopic*, since information on future arrivals is ignored.

	Daily arrival rate						penalty C_c^V		
c	Low	Med.	High	Due date	C_c^{OP}	WC	N	H	F
1	2.16	2.88	3.6	14	120	0.8	0	60	120
2	1.62	2.16	2.7	28	120	0.5	0	60	120
3	1.62	2.16	2.7	28	210	2.0	0	105	210
4	0.54	0.72	0.9	14	110	0.2	0	55	110
5	1.08	1.44	1.8	14	160	1.0	0	80	160
6	0.54	0.72	0.9	28	120	0.5	0	60	120
7	1.62	2.16	2.7	28	110	1.0	0	55	110
Total	9.18	12.24	15.3						

Table 1: Patient category data

[14] The waiting cost may in practice not be linear but rather an increasing convex function where increasing time more will increase the unit cost more i.e. a positive acceleration in cost. However, to simplify the model we have chosen to keep this cost linear.

[15] Observe that service levels along with violation penalties can be used for prioritizing different types of patients.

The test instances relate to scenarios with four surgeons available. Each surgeon has a number of minutes available (0, 240, 360, or 420) on each day in his work schedule, which is repeated in a 14-day cycle (see Table 2). The availability of resources as defined by Table 2 was decided upon such that the normal work load for each surgeon in the OT is around 30 hours per week. Arrival rates reflecting a balanced scenario were next set such that system performance reflected a utilization of approximately 95%.[16] Finally, scenarios reflecting under and over utilization were obtained by decreasing and increasing arrival rates (listed in Table 1) for all patient categories by 30%, respectively.

	days													
s	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	420	420	-	-	360	360	420	420	-	-	360	360	360	-
2	420	420	-	-	360	360	420	420	-	-	360	360	360	-
3	-	420	420	-	360	360	420	420	-	400	360	-	360	360
4	420	-	-	420	360	360	420	-	-	420	360	360	360	240

Table 2: Availability of each surgeon (in minutes) for each day.

The heterogeneity among surgeons regarding their capabilities to handle different patient categories is reflected by physician-specific means and standard deviations for the relevant surgery durations (see Table 3). In addition, a surgeon may simply not be qualified to perform certain procedures.[17]

	Surgeon 1		Surgeon 2		Surgeon 3		Surgeon 4	
c	μ_{c1}^{cat}	σ_{c1}^{cat}	μ_{c2}^{cat}	σ_{c2}^{cat}	μ_{c3}^{cat}	σ_{c3}^{cat}	μ_{c4}^{cat}	σ_{c4}^{cat}
1	71	19	75	24	70	20	-	-
2	49	24	50	25	55	27	48	23
3	182	67	180	65	175	60	185	69
4	51	27	55	28	-	-	50	25
5	98	21	-	-	95	20	100	25
6	-	-	75	17	80	25	77	20
7	85	20	88	25	-	-	87	22

Table 3: Mean and std.dev. of the surgery times by category c

Table 3 reflects the a priori stochastic information for future patients to be adjusted upon patient

[16]Scheduled time serves as the reference when computing utilization. Hence, overtime on any given day implies an expected utilization exceeding 100% that day.

[17]To illustrate, patients in category 1 are not allowed to be assigned to surgeon 4.

arrival when more precise information becomes available. This is achieved as follows:[18]

$$\mathcal{X}_1 \sim i.i.d.\ \mathcal{N}(0,1)$$
$$\mathcal{X}_2 \sim i.i.d. Beta(2,2)$$
$$\mu_{ps}^{pat} = \mathcal{X}_1 \sigma_{cs}^{cat} + \mu_{cs}^{cat}$$
$$\sigma_{cs}^{pat} = (0.5 + \mathcal{X}_2)\sigma_{cs}^{cat}$$

Available surgeons have an OR and an operating team at their disposal. The cost of having an OR open is either to be considered i) sunk and ignored in the optimization or ii) variable and charged if and only if an OR is in use.

It is by assumption possible to extend the number of minutes available for each surgeon on each day by using overtime. The cost of overtime is made up of the direct cost corresponding to the overtime payment to staff and an indirect cost reflecting e.g. the cost of failure and the cost of disutility of working overtime. Indirect cost is by assumption quadratic in expected overtime, e (in minutes):

$$\Omega_s(e) := a_1 e + a_2 e^2$$

We investigate for simplicity three scenarios with $a_1 = 0$ and $a_2 \in \{1, 0.1, 0.01\}$ yielding an overtime cost of 3600, 360 and 36 per hour of overtime, respectively.

The computational study is essentially a Monte Carlo experiment. Patient arrivals are in each replication generated from a Poisson process along with expected surgery durations and their standard deviations.[19] Appropriate warm-up periods must be chosen, since each experiment is initiated with an empty system. For that purpose we have identified the point in time, 300 days, when the average number of patients across 10 different replications has stabilized in a balanced system.[20] Accordingly, each test instance is solved for a period of 365 days with the first 300 days considered as a warm-up period to be followed by 65 days during which system performance is measured. The reason for the warm-up period being 300 is that it takes many simulated days to fill up the system with patients such that the full 28 day period is used and such that the system has stabilized.

The base case uses observed arrival rates for a number of different patient categories. These rates are based on the average number of arrivals during a year from a real-life scenario. In practice, there will be seasonality in the arrival rates for specific patient categories. We have not included these in the study. We have set a slightly tighter time limit of 28 days compared to the waiting time guarantee of 30 days given in Denmark.

The surgeons described are constructed to resemble actual surgeons at the OT. Their availability

[18] We have on purpose decided upon a data generation process that allows for very short surgery durations for some patients. The arrival of patients with short surgery durations is believed to facilitate the performance of the FCFS approach compared to the optimization approaches, since packing is made easier. However, we do not allow for negative surgery durations; cases of this type are simply left out in the numerical experiments.

[19] Clearly, seeds differ between replications.

[20] Numerical experiments indicated that the longest time for stabilization was needed in the balanced system.

schedules are consequently also constructed, such that they have at least four days off in a two week period. The maximum amount of time on each day is seven hours. The surgery times (means and variances) of specific categories for specific surgeons are based on domain expert knowledge, i.e., the surgeons themselves. Obviously, it will be more precise if we used disaggregate data of surgeries actually performed to identify means and variances, but this is not possible due to patient confidentialities.

The cost components of the base case are fictive but are set such that the relative importance of the different parts of the model is illustrated.[21] In the computational study we vary these costs to illustrate effects of different settings.

5.2. Implementational issues

We have implemented the model in C++ using the compiler GCC 4.8.2 with the option -O3 enabled. Gurobi 5.6.2 has been used as a linear programming solver and SIMLIB/C++ 3.02 as a discrete event simulation library. The computational experiments have been conducted on a Linux system with an Intel(R) Xeon(R) CPU E5-1620 0 @ 3.60GHz CPU and 24Gb memory. Each experiment has been assigned to a single core of the processor.

The solution of the model is based upon a column generation procedure alternating between solving a master problem with a restricted number of columns included and a pricing problem generating new promising columns.

The sequence for solving the pricing problems is determined by calculating a lower bound on the reduced cost for each surgeon-day combination and selecting the pricing problems in increasing order of this lower bound. The bound is described by Range et al. (2018), who observe that a deterministic variant of the stochastic knapsack problem can be used to provide a lower bound to the solution when the cost of expected overtime is convex. The solution process for the pricing problems is stopped prematurely whenever at least two pricing problems identify negative reduced cost columns.

We apply limited extensions with only the best paths in a node extended to speed up the search for negative reduced cost columns (see e.g. Burke and Curtois (2014)). The number of paths initially allowed to be extended from a node is set to 5. The number is doubled if the pricing problem does not yield a negative reduced cost column. The process is continued until a negative reduced cost column is identified or no unextended paths are left.

Solving the LP relaxation of (9)-(20) does not necessarily lead to an integer solution. In order to make the solution integral we apply the technique of aggressive variable fixing (see, e.g., Lusby et al. (2012) or Range et al. (2014)). Accordingly, the integer variables are successively fixed at their upper bounds, and the column generation for (9)-(20) is run again until a new LP relaxation

[21] While it is possible to obtain the monetary value of some of the cost components, it is hard for other parts of the model. The increasing expected overtime cost function $\Omega(\cdot)$ is subject to operational challenges, because it is a cost composed of less tangible factors such as of fatigue or disutility for the surgeon, or the risk of failure or postponement of patients. Hence the functional form of the expected overtime cost should be set relative to the other cost components and is a decision on how averse one is on overtime.

bound (with respect to the fixed variables) is obtained. This continues until a full integer solution is obtained or the fixing of variables leads to an infeasible solution.

Let the solution for the LP relaxation of (9)-(20) be $(\overline{\boldsymbol{\lambda}}, \overline{\mathbf{x}}, \overline{\mathbf{y}}, \overline{\mathbf{v}})$, where $\overline{\boldsymbol{\lambda}}$ is the vector of the $\overline{\lambda}_i$ variable values, $\overline{\mathbf{x}}$ is the vector of the $\overline{x}_{cd\delta}$ variable values, $\overline{\mathbf{y}}$ is the vector of the $\overline{y}_{cdl}$ variable values, and $\overline{\mathbf{v}}$ is the vector of the $\overline{v}_{cl}$ variable values. Only $\overline{\boldsymbol{\lambda}}$ is required to be integer, and the LP solution is optimal for the full problem if the corresponding $\overline{\boldsymbol{\lambda}}$ is integer. Otherwise, all λ_i for which $\overline{\lambda}_i = 1$ are fixed to unity. Let $\bar{i} = \arg\max_i \left\{\overline{\lambda}_i < 1\right\}$ and fix $\lambda_{\bar{i}} = 1$. $\lambda_{\bar{i}}$ is in this way forced into the integer solution at the full value of one, which in turn forces other λ_i variables out of the solution, for example, variables including the same known patients as $\lambda_{\bar{i}}$ will never be raised from the lower bound of zero and can therefore be excluded from the solution.

There is in general a risk for not identifying an integer solution when fixing variables. This is not the case for (9)-(20). The reason is that we only fix the λ-variables (as the other integer variables are naturally integer) and the constraints (10) and (14) are soft constraints penalized in the objective. Furthermore, all patients fixed to a specific surgeon-day combination are included in all schedules corresponding to that combination. If a surgeon-day combination $r \in \mathcal{R}$ has patients which are fixed to it then we do not allow it to be closed, i.e., we fix $\rho_r = 1$. Thus, exactly one of the schedules for that surgeon-day combination is chosen. Due to the soft constraints and the fact that we chose exactly one schedule for each surgeon-day combination with fixed patients assigned to it, we will never obtain an infeasible solution.

The master problem for the first day is initialized with columns corresponding to empty schedules for each surgeon-day pair, i.e., schedules where no known patients nor any potential future patients are included. Columns can be reused from one period to the next provided that already treated patients are not included. Columns with no already treated patients included and with reduced cost equal to zero are carried forward from one period to the next. This feature provides a good set of initial columns for the master problem and a significant speed-up of the solution process.

5.3. Computational results

The computational study involves 36 test scenarios for the underutilized, the balanced, and the overutilized system, since two policies for the allocation of patients to schedules are considered along with three levels of overtime cost, two scenarios for cost of opening operating rooms, and three scenarios for cost of violating service level. Thus, the performance of the system has been analyzed in 3 times 36 test scenarios. 10 replications are solved for each test scenario, because patient arrivals and surgery times are stochastic. Hence, a total of 3 times 360 test instances have been solved.

System performance is measured during the 65 days following the warm-up period. For each surgeon-day combination we compare the expected workload to available hours as defined in Table 2.[22] The expected utilization for each surgeon is next obtained by taking the average of all

[22] The expected workload is simply the sum of the expected surgery durations for the set of known patients scheduled for a particular day. Clearly, utilization is only computed if the surgeon is assigned to at least one known patient. Otherwise, the OT is by assumption considered closed.

utilization measures across all days. The average expected utilization across all surgeons is finally obtained as an overall performance measure reflecting the average level of workload. Expected overtime is obtained in a similar way. However, in this case the central limit theorem must be invoked for an estimation of the expected overtime for each surgeon on any given day (see Proposition 4).

Test statistics are reported in Tables 4-6. Results for the case with low arrival rates are reported in Table 4, and results for medium and high arrival rates are given in Tables 5 and 6, respectively. These tables are based on average performance measures. Results on the variances and ranges for the performance measures are given in appendix Appendix B. Each row in the tables corresponds to the solution of 10 replications in a given scenario for which system performance data is collected for the 65 days following the warm-up period. Column I indicates a counter for the run. The following four columns list the run parameters. Column M indicates the allocation policy, where 0 is the FCFS, and where 1 and 2 refer to the pre-allocation-based fixing policy and the pool allocation policy described in Section 4. Column a_2 reports the a_2-coefficient in the penalty function for overtime, and column C^O states the cost of opening an OR. Column P.T indicates the size of penalties for violating the service level, where N corresponds to no penalty, H is a penalty equal to half of the outsourcing cost, and F is a penalty equal to the full outsourcing cost.[23] The following three columns report computational statistics as averages of the 10 replications for each scenario. Column RT(s) is the time in seconds for solving the LP relaxation of model (9)-(20). Column TT(s) is the average total time for obtaining an integer solution. The column labeled gap(%) indicates the average percentage deviation between the integer solution and the LP relaxation. Aggregate performance statistics across replications are reported in the remaining six columns. Surgeon statistics are given in columns U(%) and $\mathbb{E}[O]$ measuring utilization percentage and expected overtime, respectively. $W(d)$ is the average waiting time in days and $S(\%)$ is the average service percentage; both are reported for patient categories with deadlines of 14 and 28 days, respectively.[24]

The tests, N, with penalties $C_c^V = 0$ put no emphasis on future arrivals. There is no incentive to put in tentative surgeries when no penalties are present. For this reason empty schedules will be generated by the pre-allocation-based fixing policy. Consequently, no patients are allocated to surgeries with $C_c^V = 0$, and the service level equals zero. The situation is reflected by instances 1, 4, 7, 10, 13, and 16 and maintained in the tables for completeness only; the results are indicated by "-".

5.3.1. Numerical results

The average time used to solve the problem for a single day ranges between a fraction of a second to around 30 seconds with most of the time being used to solve the LP relaxation. The myopic cases, N, are the easiest ones to solve, since no emphasis is put on future arrivals. The integrality gaps are in general small and decreasing in a more utilized system.

In Figure 2 we illustrate the quality of the solution in terms of the relative gap between the

[23] The penalties can be seen in Table 1.

[24] Bear in mind that the service level measures the share of patients which is not outsourced.

		Instance			Comp. Avg.			Surgeon Avg.		14 day deadline		28 day deadline	
I	M	a_2	C^O	P.T.	RT(s)	TT(s)	gap(%)	U(%)	$\mathbb{E}[O]$	W(d)	S(%)	W(d)	S(%)
0	0	-	-	-	-	-	-	81.90	0.61	1.59	100.00	1.58	100.00
1	1	1.00	0	N	0.07	0.07	0.00	-	-	-	-	-	-
2	1	1.00	0	H	5.44	9.90	2.89	80.69	0.56	2.63	100.00	2.97	100.00
3	1	1.00	0	F	5.71	10.62	3.14	80.34	0.69	2.73	100.00	3.24	100.00
4	1	1.00	50	N	0.07	0.07	0.00	-	-	-	-	-	-
5	1	1.00	50	H	2.76	6.25	1.14	79.50	0.62	2.77	100.00	3.34	100.00
6	1	1.00	50	F	3.44	7.08	1.62	80.64	0.72	2.72	100.00	3.13	100.00
7	1	0.10	0	N	0.08	0.08	0.00	-	-	-	-	-	-
8	1	0.10	0	H	5.88	10.62	3.38	80.23	1.69	2.51	100.00	3.16	100.00
9	1	0.10	0	F	6.35	10.08	2.83	80.03	2.07	2.06	100.00	2.95	100.00
10	1	0.10	50	N	0.08	0.08	0.00	-	-	-	-	-	-
11	1	0.10	50	H	3.18	6.42	1.32	79.16	1.71	2.80	100.00	3.45	100.00
12	1	0.10	50	F	4.06	6.83	1.45	80.33	2.14	2.10	100.00	2.94	100.00
13	1	0.01	0	N	0.08	0.08	0.00	-	-	-	-	-	-
14	1	0.01	0	H	5.98	10.01	3.91	80.30	4.51	2.19	100.00	3.08	100.00
15	1	0.01	0	F	6.32	11.76	3.99	80.27	6.33	2.45	99.90	2.74	100.00
16	1	0.01	50	N	0.08	0.08	0.00	-	-	-	-	-	-
17	1	0.01	50	H	3.37	6.69	1.52	80.48	4.66	1.97	100.00	3.11	100.00
18	1	0.01	50	F	4.12	8.55	1.82	80.26	6.47	2.33	100.00	2.74	100.00
19	2	1.00	0	N	0.41	0.46	3.40	81.25	0.06	1.74	100.00	1.72	100.00
20	2	1.00	0	H	8.14	13.73	3.62	80.97	0.30	1.56	99.91	1.62	100.00
21	2	1.00	0	F	9.04	15.09	3.77	81.00	0.40	1.57	99.98	1.62	99.97
22	2	1.00	50	N	1.83	2.45	17.23	80.01	0.35	2.37	100.00	2.31	100.00
23	2	1.00	50	H	5.41	10.60	1.19	80.01	0.38	1.83	100.00	1.81	100.00
24	2	1.00	50	F	7.25	12.54	1.84	80.89	0.48	1.61	100.00	1.66	99.98
25	2	0.10	0	N	0.46	0.52	3.90	81.32	0.21	1.68	100.00	1.65	100.00
26	2	0.10	0	H	8.65	15.05	4.10	80.90	1.04	1.55	99.95	1.54	99.98
27	2	0.10	0	F	9.74	15.22	3.53	80.89	1.55	1.51	100.00	1.53	99.98
28	2	0.10	50	N	1.74	2.31	17.42	80.51	1.36	2.30	100.00	2.19	100.00
29	2	0.10	50	H	6.19	11.78	1.29	79.73	1.41	1.78	100.00	1.74	100.00
30	2	0.10	50	F	8.00	12.69	1.67	80.77	1.84	1.54	100.00	1.57	100.00
31	2	0.01	0	N	0.50	0.55	3.20	81.39	0.83	1.61	100.00	1.57	100.00
32	2	0.01	0	H	9.27	15.14	4.88	81.00	4.31	1.45	100.00	1.48	100.00
33	2	0.01	0	F	9.76	17.62	4.60	80.96	6.00	1.45	100.00	1.46	100.00
34	2	0.01	50	N	1.81	2.39	18.10	80.37	6.16	1.99	100.00	1.89	100.00
35	2	0.01	50	H	6.26	11.41	1.54	80.09	5.92	1.63	100.00	1.63	100.00
36	2	0.01	50	F	7.58	14.06	2.07	80.82	7.65	1.44	100.00	1.50	100.00

Table 4: Results for the low arrival rate cases.

LP lower bound and the best upper bound solution found as well as the running time to reach the upper bound solution. We have in both figures the cumulative share of instances on the vertical axis. On the left-most figure we have the relative gap between the LP bound and the best upper bound solution on the horizontal axis while on the right-most figure we have the running time in seconds. The two curves in the figures illustrate how the pre-allocation based fixing policy (method 1) and the pool-allocation policy (method 2) behave, respectively. We observe that the relative gap is no higher than 10% in 95% of the instances for the pool-allocation policy while it is no higher than 10% in 99.9% of the instances in the pre-allocation based fixing policy. For the time to reach the best upper bound solution we see that the running time is less than 30 seconds in 99.9% of the instances for the pre-allocation based fixing policy. For the pool allocation policy the running time is less than 30 seconds in 92% of the instances. We see that it is harder to solve the latter. We note that a third of the instances for the pre-allocation based fixing policy correspond to the myopic cases where the solution is trivial and the running time is negligible.

ACCEPTED MANUSCRIPT

I	M	Instance			Comp. Avg.			Surgeon Avg.		14 day deadline		28 day deadline	
		a_2	C^O	P.T.	RT(s)	TT(s)	gap(%)	U(%)	$\mathbb{E}[O]$	W(d)	S(%)	W(d)	S(%)
0	0	-	-	-	-	-	-	95.93	1.38	13.57	72.94	13.92	100.00
1	1	1.00	0	N	0.07	0.07	0.00	-	-	-	-	-	-
2	1	1.00	0	H	3.89	5.32	0.50	96.84	1.55	11.82	88.76	19.02	80.11
3	1	1.00	0	F	4.32	5.79	0.56	97.62	2.29	11.97	89.58	18.62	80.02
4	1	1.00	50	N	0.08	0.08	0.00	-	-	-	-	-	-
5	1	1.00	50	H	3.90	5.33	0.39	96.86	1.57	11.77	89.14	18.73	79.78
6	1	1.00	50	F	4.18	5.66	0.49	97.52	2.14	11.91	89.71	18.78	80.55
7	1	0.10	0	N	0.08	0.08	0.00	-	-	-	-	-	-
8	1	0.10	0	H	4.20	5.56	0.53	99.59	4.96	11.21	93.04	18.04	85.42
9	1	0.10	0	F	4.51	6.06	0.65	100.54	6.93	10.28	93.80	18.09	88.45
10	1	0.10	50	N	0.07	0.07	0.00	-	-	-	-	-	-
11	1	0.10	50	H	4.08	5.44	0.40	99.60	5.03	10.89	92.81	18.04	85.96
12	1	0.10	50	F	4.64	6.27	0.55	100.60	7.11	10.35	93.91	18.23	88.04
13	1	0.01	0	N	0.08	0.08	0.00	-	-	-	-	-	-
14	1	0.01	0	H	5.41	8.13	0.92	104.81	19.18	9.13	97.11	12.79	99.69
15	1	0.01	0	F	6.90	10.22	1.02	107.10	29.12	7.18	99.15	10.25	100.00
16	1	0.01	50	N	0.08	0.08	0.00	-	-	-	-	-	-
17	1	0.01	50	H	5.09	7.70	0.62	104.79	19.14	9.29	97.28	13.12	99.49
18	1	0.01	50	F	6.65	9.94	0.78	107.09	29.33	7.20	99.07	10.16	100.00
19	2	1.00	0	N	1.62	1.69	0.11	94.57	1.08	13.50	74.15	14.77	100.00
20	2	1.00	0	H	14.43	18.27	0.64	97.16	1.29	13.69	79.28	18.23	99.97
21	2	1.00	0	F	15.64	19.45	0.69	97.69	1.57	13.57	86.68	18.09	98.79
22	2	1.00	50	N	4.72	5.18	2.57	95.64	0.96	13.55	74.78	14.99	100.00
23	2	1.00	50	H	13.88	17.67	0.48	97.10	1.27	13.69	79.07	18.20	99.98
24	2	1.00	50	F	15.40	19.21	0.58	97.76	1.63	13.55	86.52	18.10	98.74
25	2	0.10	0	N	1.85	1.93	0.09	98.62	6.52	13.31	83.34	14.69	100.00
26	2	0.10	0	H	14.73	19.01	0.80	100.09	5.39	13.61	86.14	17.27	99.88
27	2	0.10	0	F	16.21	20.62	0.78	100.90	6.88	13.47	91.79	17.23	99.24
28	2	0.10	50	N	4.99	5.50	2.86	98.91	4.74	13.34	83.89	14.73	100.00
29	2	0.10	50	H	14.36	18.67	0.57	100.17	5.61	13.61	85.68	17.28	99.96
30	2	0.10	50	F	15.83	20.16	0.63	100.92	6.87	13.46	92.04	17.30	99.37
31	2	0.01	0	N	2.61	2.70	0.07	108.80	36.33	12.78	99.77	14.30	100.00
32	2	0.01	0	H	17.72	24.02	1.04	105.34	20.64	11.13	99.98	11.60	99.95
33	2	0.01	0	F	20.87	29.02	1.42	106.41	25.42	6.23	100.00	6.35	100.00
34	2	0.01	50	N	6.20	6.92	3.61	108.15	32.24	12.76	99.96	14.01	100.00
35	2	0.01	50	H	16.53	22.30	0.61	105.46	21.09	11.09	100.00	11.57	99.97
36	2	0.01	50	F	17.41	24.25	0.94	106.35	25.15	6.21	100.00	6.29	100.00

Table 5: Results for the medium arrival rate cases.

The gaps are large in instances 22, 28, and 34. This is due to the effect of the cost of opening ORs while myopically optimizing the allocation of known patients. The model distributes patients over more surgeons, which results in lower overtime cost. The full cost of ORs is not charged due to fractional solutions. More ORs are opened with a low utilization when the corresponding columns are fixed to unity. The effect is especially pronounced in the scenario with low arrival rates.

Consider first the underutilized system with low arrival rates (see Table 4). Most patients are treated before their due dates. For this reason the service level is close to 100% in all instances, and utilization is around 80%. Imposing the cost of opening an OR causes slight increases in waiting time without changing service levels and resource utilization, since some ORs may remain closed in some periods as a means to decrease operational costs. On the other hand, decreasing the cost of overtime causes shorter waiting times, since some patients will be treated earlier during surgeons' overtime. The results indicate that FCFS, with the exception of overtime, performs just as well as the optimization-based policies in an underutilized system. In this case the key benefit of using the

I	M	Instance			Comp. Avg.			Surgeon Avg.		14 day deadline		28 day deadline	
		a_2	C^O	P.T.	RT(s)	TT(s)	gap(%)	U(%)	$\mathbb{E}[O]$	W(d)	S(%)	W(d)	S(%)
0	0	-	-	-	-	-	-	96.20	1.47	13.78	22.47	15.17	100.00
1	1	1.00	0	N	0.07	0.07	0.00	-	-	-	-	-	-
2	1	1.00	0	H	3.82	4.80	0.32	97.14	1.77	12.73	84.84	20.28	71.47
3	1	1.00	0	F	4.16	5.13	0.35	97.79	2.36	12.73	83.92	20.46	72.36
4	1	1.00	50	N	0.08	0.08	0.00	-	-	-	-	-	-
5	1	1.00	50	H	3.90	4.86	0.28	97.05	1.73	12.75	84.70	20.39	71.69
6	1	1.00	50	F	4.15	5.12	0.32	97.78	2.32	12.78	83.72	20.35	72.66
7	1	0.10	0	N	0.07	0.07	0.00	-	-	-	-	-	-
8	1	0.10	0	H	4.27	5.29	0.36	100.04	5.74	12.20	86.11	20.25	73.95
9	1	0.10	0	F	4.52	5.48	0.35	101.12	8.28	11.82	88.20	20.76	74.29
10	1	0.10	50	N	0.08	0.08	0.00	-	-	-	-	-	-
11	1	0.10	50	H	4.21	5.22	0.32	100.05	5.78	12.27	86.57	20.26	73.75
12	1	0.10	50	F	4.52	5.48	0.33	101.24	8.46	11.71	88.13	20.61	73.78
13	1	0.01	0	N	0.08	0.08	0.00	-	-	-	-	-	-
14	1	0.01	0	H	4.96	6.12	0.34	107.15	27.34	12.77	85.44	20.78	79.62
15	1	0.01	0	F	6.21	7.60	0.41	114.01	53.16	12.69	85.56	20.68	82.22
16	1	0.01	50	N	0.08	0.08	0.00	-	-	-	-	-	-
17	1	0.01	50	H	5.02	6.20	0.31	107.05	26.96	12.85	85.15	20.81	79.99
18	1	0.01	50	F	6.14	7.52	0.38	114.18	53.71	12.65	85.80	20.71	81.73
19	2	1.00	0	N	2.28	2.34	0.04	95.02	1.52	11.75	27.40	16.74	100.00
20	2	1.00	0	H	17.57	19.98	0.24	97.40	1.33	13.94	26.53	23.68	99.88
21	2	1.00	0	F	22.31	26.54	0.51	97.91	1.73	13.62	84.95	18.48	88.40
22	2	1.00	50	N	6.28	6.68	1.26	96.09	1.15	13.45	26.80	16.50	100.00
23	2	1.00	50	H	17.46	19.90	0.22	97.39	1.33	13.90	27.01	23.64	99.84
24	2	1.00	50	F	21.51	25.59	0.47	97.88	1.71	13.63	84.78	18.39	88.44
25	2	0.10	0	N	2.67	2.75	0.04	100.60	10.17	12.66	45.24	16.06	100.00
26	2	0.10	0	H	18.33	20.91	0.29	100.67	6.67	13.81	39.81	23.21	99.97
27	2	0.10	0	F	21.51	25.35	0.49	101.25	7.71	13.69	87.19	18.30	89.60
28	2	0.10	50	N	6.75	7.19	1.22	99.88	6.71	13.31	41.53	16.25	100.00
29	2	0.10	50	H	18.21	20.75	0.26	100.62	6.58	13.77	39.45	23.32	99.97
30	2	0.10	50	F	20.99	24.85	0.44	101.21	7.63	13.67	86.58	18.25	89.66
31	2	0.01	0	N	3.75	3.84	0.03	120.80	78.84	13.08	82.60	15.66	100.00
32	2	0.01	0	H	21.29	24.77	0.34	122.69	86.07	13.13	86.91	21.99	99.95
33	2	0.01	0	F	24.61	29.77	0.55	119.18	72.57	13.10	92.43	17.76	97.09
34	2	0.01	50	N	8.69	9.21	0.88	121.64	82.06	13.06	84.79	15.76	100.00
35	2	0.01	50	H	21.40	24.86	0.29	122.79	86.50	13.11	86.94	21.99	99.95
36	2	0.01	50	F	24.70	29.76	0.50	119.30	73.12	13.04	92.71	17.75	97.17

Table 6: Results for the high arrival rate cases.

optimization approaches is a better control for overtime.

Consider next the more balanced system with medium arrival rates (see Table 5). In this case there is no significant impact of the cost of opening ORs, since for most of the time the system is running at capacity utilizing all resources. A high cost of overtime implies a rejection of more patients because treatment before their due dates requires overtime. The pool allocation policy with a low penalty for violating the service level – i.e., the N cases – prioritizes known patients over future patients as a means to reduce penalties due to patients' waiting times. Consequently, known patients are allocated to earlier time slots, thus reducing the probability for treatment of patients with a 14-day deadline, who are expected to arrive and to be put into the schedule later. This is due to the myopic nature of the N cases. A similar phenomenon can be observed in the FCFS case. This is in contrast to the F case, where the penalty for violating the service level is increased to the level of the outsourcing cost. In the pool allocation approach slots are reserved for future patients with the shortest deadline. The consequence is a significant improvement in terms of service level

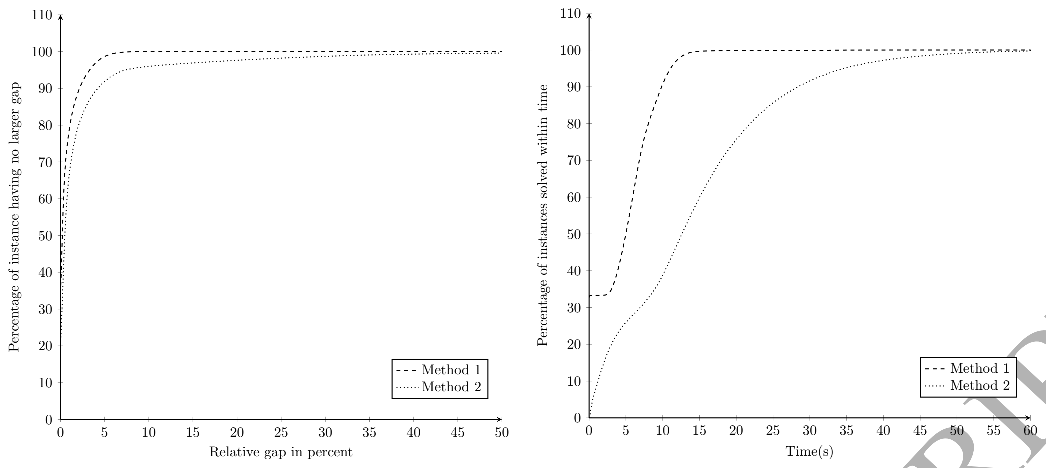


Figure 2: Left: cumulative distribution of the relative gap over instances. Right: cumulative distribution of time to upper bound over instances.

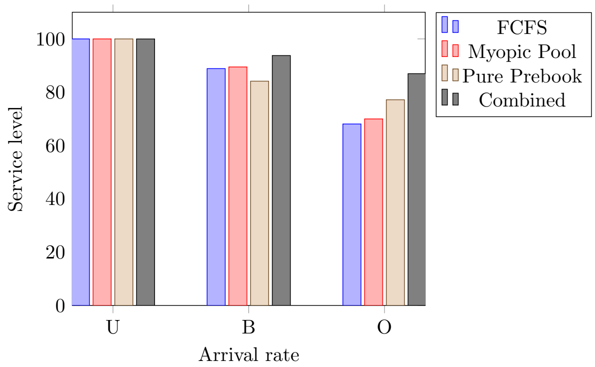
for patients with short due dates accompanied by an expected overtime comparable to the FCFS case along with a small reduction in service level for the 28-day patients and an increase in their waiting times. To illustrate, the expected overtime is increased by 0.19 minutes, the service level for 28-day patients is reduced by 1.21% points, and their waiting time is on average increased by 4.17 days using the pool allocation policy in instance 21. At the same time, the service level for 14-day patients is increased by 13.74% points without a change in their expected waiting time, reflecting a significant performance improvement.[25] Finally, a decrease in the cost of overtime results in shorter waiting times and improved service levels for all patient categories, since more overtime is used.

Results for system performance in scenarios with the highest arrival rates are reported in Table 6. The pattern is similar to the case of the more balanced system. It should come as no surprise that the service level for 14-day patients is low under conditions of a myopic policy (e.g., 22.47% for the FCFS) because arrival rates are higher. The optimization-based approaches improve upon this situation by providing a balanced service level across all patient categories. Again, decreasing the cost of overtime provides an incentive to extend capacity by increasing overtime. The increase in overtime allows for a treatment of patients who would have been rejected otherwise. Observe that an excessive use of overtime indicates that capacity is too low, and the amount of expected overtime provides an indication of the additional capacity needed to attain the imposed service level.

It is clear that both FCFS and the pre-allocation-based fixing policy are outperformed by the pool allocation policy in scenarios with high penalties for violation of service levels. To illustrate, a comparison of scenarios 9 and 27 in Table 5 with arrival rates at a medium level reveals that the pool allocation policy provides an approximately 11% points higher service level for 28-day patients at the cost of a 2% points decrease in the service level for 14-day patients with almost the same utilization of resources and overtime. A comparison of FCFS to the pre-allocation-based fixing

[25]Results like these reflect that a meaningful use of the pool allocation policy occurs in cases with a high penalty for violating the service level.

policy shows that the latter provides a balanced service level for patients with different deadlines along with a higher rate of utilization accompanied by a higher expected overtime. The main reason is that tentative slots are reserved equally for patients with different deadlines. A comparison of cases 0 and 19 in the scenario with medium arrival rates reveals that the pool allocation policy provides a slightly increased service level for 14-day deadline patients and a lower utilization of resources along with a lower overtime at the cost of increasing the waiting time for 28-day deadline patients by less than one day. Similar results can be observed in scenarios with high arrival rates.



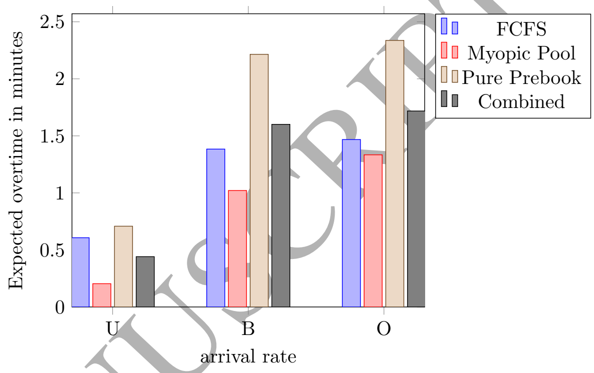


Figure 3: Left: the service level in percent for different arrival rates and allocation approaches. Right: the expected overtime for different arrival rates and allocation approaches.

We can consider system performance across different arrival rates and different allocation approaches. The FCFS corresponds to the test instance with M=0 . In the following we consider the pre-allocation based fixing policy (M=1) with full penalty (P.T. = F) for future non-allocated patients and refer to this as pure prebook, as it corresponds to only allocate tentative slots for patients in the future. Furthermore, we consider the pool allocation policy (M=2) without a penalty for non-allocated tentative patients (P.T. = N) which we refer to as the myopic pool, as we will not allocate future tentative slots. Finally, we consider the pool allocation policy (M=2) with full penalty for not allocating future patients (P.T. = F) and refer to this as the combined system.

In Figure 3 we illustrate system performance – service level and expected overtime as average over the instances for each allocation approach – for an overtime penalty of $a_2 = 1.0$, for the under-utilized system (U), the balanced system (B), and the over-utilized system (O). In the under-utilized system we see that the FCFS performs as well as the other approaches have the same service level and approximately the same overtime. However, for the balanced system and the over-utilized system the combined approach has the highest service level among the approaches. For the over-utilized system FCFS has a service level of 68% while the combined approach has a service level of 87%. Furthermore, the combined approach manages this with almost the same expected overtime as the FCFS and the myopic pool approach. Note that the myopic pool approach performs in much the same way as the FCFS, i.e., just optimizing what we know today is not sufficient for long term performance. Similar patterns prevail for other combinations and selections of a_2. To conclude, we see that when combining optimization with preallocation, then we obtain the best system performance.

6. Conclusion

We have developed a model for allocation of patients to surgeon-day combinations. The model is based on a generalized assignment formulation augmented with constraints taking the stochastic arrival processes of patients into account. The model allows to balance service levels for different categories of patients. It is based on a set of underlying assumptions including independent patient arrivals and that we can estimate surgeons' means and variances of surgery durations for specific patient categories.

Schedules for any given surgeon-day combination are generated by the solution of a stochastic knapsack problem with an objective penalizing expected overtime in terms of an increasing strictly convex function.

Two patient-allocation policies are tested: One with an allocation of individual patients based on potential surgeries and another based on an optimization for groups of patients. The first policy has the advantage that patients can be informed about their surgery date up front. The second policy implies that patients must wait before being assigned to a surgeon-day combination. Both policies are embedded into a rolling horizon simulation and compared to a FCFS policy.

The numerical experiments indicate that taking stochastic information about future job arrivals into account in the assignment of jobs to agents implies an improved performance. It indicates that the use of information on patients' arrival distributions increases the level of service as well as the utilization of surgeons compared to the myopic case, where this information is not taken into account. System performance under conditions of the FCFS policy compares to performance based upon a myopic optimization, and FCFS is competitive in scenarios with low arrival rates compared to capacity. However, FCFS is outperformed in scenarios with high patient arrival rates compared to capacity. The policies taking future arrivals into account improve system performance in terms of levels of service, and the best performance is obtained using an explicit optimization approach.

References

References

Barnhart, C., Johnson, E. L., Nemhauser, G. L., Savelsbergh, M. W. P., Vance, P. H., 1998. Branch-and-price: Column generation for solving huge integer programs. Operations Research 46 (3), 316–329.

Bitran, G. R., Tirupati, D., 1993. Chapter 10 hierarchical production planning. In: Logistics of Production and Inventory. Vol. 4 of Handbooks in Operations Research and Management Science. Elsevier, pp. 523 – 568.

Blake, J. T., Donald, J., 2002. Mount sinai hospital uses integer programming to allocate operating room time. Interfaces 32 (2), 63 – 73.

Bowers, J., Mould, G., 2004. Managing uncertainty in orthopaedic trauma theatres. European Journal of Operational Research 154 (3), 599 – 608.

Burke, E. K., Curtois, T., 2014. New approaches to nurse rostering benchmark instances. European Journal of Operational Research 237 (1), 71 – 81.

Cardoen, B., Demeulemeester, E., 2008. Capacity of clinical pathways - a strategic multi-level evaluation tool. Journal of Medical Systems 32 (6), 443 – 452.

Cardoen, B., Demeulemeester, E., Beliën, J., 2010. Operating room planning and scheduling: A literature review. European Journal of Operational Research 201 (3), 921 – 932.

Demeulemeester, E., Beliën, J., Cardoen, B., Samudra, M., 2013. Operating room planning and scheduling. In: Denton, B. T. (Ed.), Handbook of Healthcare Operations Management. Vol. 184 of International Series in Operations Research & Management Science. Springer New York, pp. 121–152.

Denton, B. T., Miller, A. J., Balasubramanian, H. J., Huschka, T. R., 2010. Optimal allocation of surgery blocks to operating rooms under uncertainty. Operations Research 58 (4-part-1), 802–816.

Dexter, F., 2016. Bibliography of operating room management articles, http://www.franklindexter.net/, Retrieved 13 February 2016.

Ebben, M., Hans, E., Olde Weghuis, F., 2005. Workload based order acceptance in job shop environments. OR Spectrum 27 (1), 107–122.

Gartner, D., Kolisch, R., 2014. Scheduling the hospital-wide flow of elective patients. European Journal of Operational Research 233 (3), 689 – 699.

Gerchak, Y., Gupta, D., Henig, M., 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. Management Science 42 (3), pp. 321–334.

Guerriero, F., Guido, R., 2011. Operational research in the management of the operating theatre: a survey. Health Care Management Science 14 (1), 89–114.

Gul, S., Denton, B. T., Fowler, J. W., 2015. A progressive hedging approach for surgery planning under uncertainty. INFORMS Journal on Computing 27 (4), 755–772.

Hans, E., Wullink, G., van Houdenhoven, M., Kazemier, G., 2008. Robust surgery loading. European Journal of Operational Research 185 (3), 1038 – 1050.

Harper, P. R., 2002. A framework for operational modelling of hospital resources. Health Care Management Science 5 (3), 165 – 173.

Huh, W. T., Liu, N., Truong, V.-A., 2013. Multiresource allocation scheduling in dynamic environments. Manufacturing & Service Operations Management 15 (2), 280–291.

Hulshof, P. J., Boucherie, R. J., van Essen, J. T., Hans, E. W., Hurink, J. L., Kortbeek, N., Litvak, N., Vanberkel, P. T., van der Veen, E., Veltman, B., Vliegen, I. H., Zonderland, M. E., 2011. Orchestra: an online reference database of or/ms literature in health care. Health Care Management Science 14 (4), 383–384.

Hulshof, P. J. H., Kortbeek, N., Boucherie, R. J., Hans, E. W., Bakker, P. J. M., 2012. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms. Health Systems 1, 129–175.

Jalora, A., 2016. Order acceptance and scheduling at a make-to-order system using revenue management. Ph.D. thesis, Texas A&M University.
URL http://hdl.handle.net/1969.1/4421

Jensen, J., 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Mathematica 30, 175–193.

Kellerer, H., Pferschy, U., Pisinger, D., 2004. Knapsack Problems. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kim, S.-C., Horowitz, I., 2002. Scheduling hospital services: the efficacy of elective-surgery quotas. Omega 30 (5), 335–346.

Kleywegt, A. J., Shapiro, A., Homem-de-Mello, T., 2002. The sample average approximation method for stochastic discrete optimization. SIAM Journal on Optimization 12 (2), 479–502.

Kogan, K., Khmelnitsky, E., Ibaraki, T., 2016. Dynamic generalized assignment problems with stochastic demands and multiple agent–task relationships. Journal of Global Optimization 31 (1), 17–43.

Kogan, K., Shtub, A., 1997. DGAP – the dynamic generalized assignment problem. Annals of Operations Research 69, 227–239.

Kong, Q., Lee, C.-Y., Teo, C.-P., Zheng, Z., 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. Operations Research 61 (3), 711–726.

Kosuch, S., Lisser, A., 2010. Upper bounds for the 0-1 stochastic knapsack problem and a b&b algorithm. Annals of Operations Research 176 (1), 77–93.

Kozlowski, D., Worthington, D., 2015. Use of queue modelling in the analysis of elective patient treatment governed by a maximum waiting time policy. European Journal of Operational Research 244 (1), 331 – 338.

Lamiri, M., Xie, X., Zhang, S., 2008. Column generation approach to operating theater planning with elective and emergency patients. IIE Transactions 40 (9), 838 – 852.

Lübbecke, M. E., Desrosiers, J., 2005. Selected topics in column generation. Operations Research 53 (6), 1007 – 1023.

Lusby, R., Dohn, A., Range, T. M., Larsen, J., 2012. A column generation-based heuristic for rostering with work patterns. Journal of the Operational Research Society 63 (2), 261–277.

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

Mak, H.-Y., Rong, Y., Zhang, J., 2015. Appointment scheduling with limited distributional information. Management Science 61 (2), 316–334.

May, J. H., Spangler, W. E., Strum, D. P., Vargas, L. G., 2011. The surgical scheduling problem: Current research and future opportunities. Production and Operations Management 20 (3), 392–405.

Mazzola, J. B., Neebe, A. W., 2012. A generalized assignment model for dynamic supply chain capacity planning. Naval Research Logistics (NRL) 59 (6), 470–485.

Merzifonluoğlu, Y., Geunes, J., Romeijn, H., 2012. The static stochastic knapsack problem with normally distributed item sizes. Mathematical Programming 134 (2), 459–489.

Mestry, S., Damodaran, P., Chen, C.-S., 2011. A branch and price solution approach for order acceptance and capacity planning in make-to-order operations. European Journal of Operational Research 211 (3), 480 – 495.

Min, D., Yih, Y., 2010. Scheduling elective surgery under uncertainty and downstream capacity constraints. European Journal of Operational Research 206 (3), 642 – 652.

Moccia, L., Cordeau, J.-F., Monaco, M. F., Sammarra, M., 2009. A column generation heuristic for a dynamic generalized assignment problem. Computers & Operations Research 36 (9), 2670 – 2681.

Pham, D.-N., Klinkert, A., 2008. Surgical case scheduling as a generalized job shop scheduling problem. European Journal of Operational Research 185 (3), 1011 – 1025.

Range, T. M., Kozlowski, D., Petersen, N. C., 2018. A shortest-path-based approach for the stochastic knapsack problem with non-decreasing expected overfilling costs. Computers & Operations Research 97, 111 – 124.

Range, T. M., Lusby, R. M., Larsen, J., 2014. A column generation approach for solving the patient admission scheduling problem. European Journal of Operational Research 235 (1), 252 – 264.

Riise, A., Mannino, C., Burke, E. K., 2016. Modelling and solving generalised operational surgery scheduling problems. Computers & Operations Research 66, 1 – 11.

Rockafellar, R. T., Wets, R. J.-B., 1991. Scenarios and policy aggregation in optimization under uncertainty. Mathematics of Operations Research 16 (1), 119–147.

Shylo, O. V., Prokopyev, O. A., Schaefer, A. J., 2013. Stochastic operating room scheduling for high-volume specialties under block booking. INFORMS Journal on Computing 25 (4), 682–692.

Slotnick, S. A., 2011. Order acceptance and scheduling: A taxonomy and review. European Journal of Operational Research 212 (1), 1 – 11.

Testi, A., Tanfani, E., Torre, G., 2007. A three-phase approach for operating theatre schedules. Health Care Management Science 10 (2), 163 – 172.

Truong, V.-A., 2015. Optimal advance scheduling. Management Science 61 (7), 1584–1597.

VanBerkel, P. T., Blake, J. T., 2007. A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. Health Care Management Science 10 (4), 373 – 385.

Vissers, J. M. H., Adan, I. J. B. F., Bekkers, J. A., 2005. Patient mix optimization in tactical cardiothoracic surgery planning: a case study. IMA Journal of Management Mathematics 16 (3), 281–304.