Minimizing bed occupancy variance by scheduling patients under uncertainty

Anne van den Broek d'Obrenan, Ad Ridder, Dennis Roubos, Leen Stougie

Minimizing Bed Occupancy Variance by Scheduling Patients under Uncertainty

Anne van den Broek d'Obrenan[a,b], Ad Ridder[b,*], Dennis Roubos[c], Leen Stougie[b,d]

[a] *Eerste Atjehstraat 57E, 1094 KC Amsterdam, Netherlands*
[b] *Vrije Universiteit Amsterdam, Department of Econometrics and Operations Research, de Boelelaan 1105, 1081 HV Amsterdam, Netherlands*
[c] *HOTflo Company, Bisonspoor 5000 - B404, 3605 LT Maarssen, Netherlands*
[d] *Centrum voor Wiskunde en Informatica, Science Park 123, 1098 XG Amsterdam, Netherlands & Erable, France*

Abstract

In this paper we consider the problem of scheduling patients in allocated surgery blocks in a Master Surgical Schedule. We pay attention to both the available surgery blocks and the bed occupancy in the hospital wards. More specifically, large probabilities of overtime in each surgery block are undesirable and costly, while large fluctuations in the number of used beds requires extra buffer capacity and makes the staff planning more challenging. The stochastic nature of surgery durations and length of stay on a ward hinders the use of classical techniques. Transforming the stochastic problem into a deterministic problem does not result into practically feasible solutions. In this paper we develop a technique to solve the stochastic scheduling problem, whose primary objective it to minimize variation in the necessary bed capacity, while maximizing the number of patients operated, and minimizing the maximum waiting time, and guaranteeing a small probability of overtime in surgery blocks. The method starts with solving an Integer Linear Programming (ILP) formulation of the problem, and then simulation and local search techniques are applied to guarantee small probabilities of overtime and to improve upon the ILP solution. Numerical experiments applied to a Dutch hospital show promising results.

Keywords: Operations research in Health Services, Scheduling, Uncertainty Modeling, Heuristics, Simulation

*Corresponding author
Email addresses: annebroekdo@gmail.com (Anne van den Broek d'Obrenan), ad.ridder@vu.nl (Ad Ridder), dennis@hotflo.net (Dennis Roubos), stougie@cwi.nl (Leen Stougie)

1. Introduction

Scheduling patients for surgery is a daily complex task in every hospital. Driven by increasing costs in hospital care and long waiting lists, hospitals focus on an efficient use of the operating rooms when scheduling surgeries. However, the output of the schedule largely determines the load and flow in downstream hospital processes and causes variability in demand. This effect is often not taken into account when creating the operating room planning, but is a reason for bed shortages, canceled surgeries, poor quality of care and an unstable workload. The high variability in hospital processes, such as (emergency) arrivals, surgery duration and patient's length of stay (LOS) in the wards also contributes to these effects.

1.1. Background

For the scheduling of surgeries, hospitals distinguish three types of patients: elective, urgent and emergency patients. Elective patients are generally put on a waiting list and admitted for surgery on an appointed moment in the operating room (OR) planning. Since these type of patients are scheduled in advance, the hospital can decide on the optimal surgery slot. However, the huge potential in overall efficiency for hospitals is not yet completely utilized. Traditionally, the focus was mainly on maximizing the utilization of surgery blocks, however, recent studies consider the effects on downstream processes as well. Individual urgent and emergency patients cannot be scheduled in advance, but hospitals do know the average daily volumes of these types of patients. Moreover, due to the specific nature of urgent and emergency arrivals they can be predicted very well. Hospitals use either dedicated ORs for these type of patients or reserve time in surgery blocks with elective patients.

Hospital admission planning starts on a strategic level where hospitals decide about long-term capacity dimensioning. After that, the total required operating hours for each medical specialty have to be divided into surgery blocks. This is the tactical planning, a cyclic planning for the ORs, where specialties are assigned to surgery blocks in an OR and operating day in the schedule. This schedule is often called the Master Surgical Schedule (MSS) and spans a period of one or two weeks. Lastly, in the operational planning, patients need to be assigned to blocks of operating time in the MSS. Subsequently, an operational schedule needs to be defined to find the best set of patients to be planned in each surgery block since it will generate demand in downstream processes.

In the operational planning we distinguish three objectives. The first objective in this paper is to minimize the undesired effect on downstream processes. However, the ORs are an expensive resource, and should therefore be used efficiently. Idle time within the rooms should be avoided. At the same time, overtime in the ORs is undesirable. Finding a good balance between overtime and idle time is a difficult, but interesting, problem. That is the second objective in this paper. The solution is directly related to the size of the waiting list for surgery. If the size of the waiting list is too large, the perception of patients will be negatively affected. If the size is too small, chances exist that a surgery block will be underutilized. Therefore, another important aspect is to focus on the length of the waiting list, which is the third objective in this paper. The waiting list can be controlled in many different ways, but the optimal control also takes the probability of overtime in surgery blocks into account.

1.2. Literature review

There exists a vast amount of research on the topic of OR planning, where the objectives differ from reducing OR overtime and OR idle time to reducing waiting lists and leveling bed occupancy. Some research is based on deterministic variables, where others incorporate uncertainty in hospital processes in their models. The modelling and solution methods are ranging from linear and integer linear programming to meta-heuristics such as genetic algorithms. Detailed lists of literature categorizing all these aspects, can be found in the reviews of Cardoen et al. (2010), Gur & Eren (2018), and Zhu et al. (2019). We comment on the work that is mostly related and relevant for our study.

In our work we consider two types of uncertainty, the surgery duration and the length of stay. The uncertainty of the surgery durations may cause overtime which has taken into account by a number of studies. Denton et al. (2007) formulated a two-stage stochastic programming model with recourse. The uncertainty of the surgery durations is modeled by a finite set of scenarios. The aim is to minimize a weighted sum of the expectation of waiting, idling, and tardiness. The model was solved by heuristic rules for approximating the optimal solution. Hans et al. (2008) proposed heuristics and local search techniques for minimizing overtime risk. Van Oostrum et al. (2008) discretized the chance constraints, and formulated an ILP for minimizing a weighted sum of required OR capacity and leveling hospital beds. The model is solved by a column generation technique. Min & Yih (2010) applied sample average approximation for

minimizing patient cost and expected overtime cost. Mixed integer programming and Monte Carlo simulation has been used by Zhang et al. (2009) for minimizing an inpatients' cost model, and by Kroer et al. (2018) for minimizing overtime work and unused OR capacity. Denton et al. (2010) extended their earlier work (Denton et al., 2007) by implementing a robust counterpart. Kayis et al. (2015) estimated the surgery duration distributions by multiplicative factor models. Molina-Pariente et al. (2018) modeled the uncertainty of surgery durations by scenario probabilities. The resulting mixed integer programming problem minimized the expected OR overtime and undertime costs, and the expected cost of surgery cancelations. The problem was approximated by a sample average approximation which was solved by combining a greedy local search and Monte Carlo simulation. Hooshmand et al. (2018) considered an allocation problem integrating both scheduling and rescheduling decisions. The uncertainty of surgeries was represented by a finite set of scenarios. The optimization model was solved by a genetic algorithm.

Concerning the uncertainty of the length of stay and its effect on the ward occupancy, several approaches have been studied. Marazzi et al. (1998) used parametric families of lognormal, Gamma and Weibull disributions and computed their M-estimators. Faddy et al. (2009) fitted phase-type distributions to a data set of patiens' length of stay. Bekker & Koeleman (2011) used phase-type distributed length of stay in an optimization model for admission scheduling. Vanberkel et al. (2011) considered empirical distributions of the lengths of stay, and then determined the exact distribution of recovering patients by using binomial convolutions. Fügener et al. (2014) proposed an optimization problem for minimizing a cost function of the distribution of patiens in the wards, by assigning surgery blocks to specialties. They considered an exact branch-and-bound solution approach, and several heuristic methods including simulated annaealing. Van Essen et al. (2014) developed an optimization model with linear constraints and a nonlinear objective function that incorporated the stochasticity of the lengths of stay and bed occupancy. Assuming multinomially distributed lengths of stay, the bed occupancy distributions were computed by convolutions. The model was solved by a heuristic local search (simulated annealing), and by linearizing the objective function.

Scheduling patients while considering both the surgery duration and length of stay uncertainty, has been studied in Beliën & Demeulemeester (2007). They considered multinomial distributions, and incorporated their means and variances in the objective function of a mixed integer programming model. The model was solved by heuristics.

Banditori et al. (2013) implemented a mixed integer program for maximizing the patient throughput taking into account due dates and waiting list control. The surgery duration and length of stay were incorporated in the model by the expected values of their empirical distribution. Next, a discrete-event simulation tested the robustness of the solution against the variability of these distributions, and permitted fine tuning of the model. Carter & Ketabi (2013) proposed an integer linear program for balancing bed occupancy at the wards. They introduced a two stage stochastic program with recourse, where Monte Carlo simulation was used in the first stage to generate sample averages for the surgery durations and lengths of stays. Thereafter the integer program was solved. Saadouli et al. (2015) formulated a knapsack model for daily schedules. They used a percentile value of the surgery duration and a related recovery time. The efficiency of the solution was anlyzed by a discrete-evnt simulation. Jebali & Diabat (2017) assumed log-normal distributions for the surgery durations and empirical distributions for the lenghts of stay. The model is a two-stage chance constrained stochastic program for minimizing a cost function involving patient cost, expected operating room utilization cost, and penalty cost for exceeding ward capacities. The model is simulated by a sample average algorithm in which scenarios are generated by Monte Carlo simulation, and solved by a mixed integer program. Neyshabouri & Berg (2017) developed an optimization model for assigning elective patients to surgery blocks. The objective is to minimize a cost function of patient admissions, overtime cost, and penalty cost for excceding capacities in downstream units. The uncertainty of the surgery duration and the length of stay is expressed in uncertainty intervals of their parameters. Then, the model is solved by two-stage robust optimization. Recently, Schneider et al. (2020) studied the problem of scheduling surgery groups for maximizing OR utilization and minimizing bed occupancy variation. They assume random surgery durations with lognormal distributions, and random lengths of stay with discrete empirical distributions. The associated (nonlinear) optimization problem is solved by simulated annealing. They consider a second approach by linearizing the objective function and constraints, which results in a mixed integer linear program.

Our optimization criterion is a multi-objective function that incorporates leveling the bed occupancy at the downstream resources, attempting to schedule as many surgeries as possible, and discouraging the increase of waiting lists. Multiple objective functions are commonly implemented in OR scheduling problems (Van Oostrum et al., 2008; Banditori

et al., 2013; Xiang, 2017; Jebali & Diabat, 2017; Neyshabouri & Berg, 2017; Kroer et al., 2018). Also, some studies considered minimizing the ward occupancy variability as one of the objectives (Van Oostrum et al., 2008; Carter & Ketabi, 2013; Schneider et al., 2020).

We describe a method to linearize the objective part of minimizing bed occupancy variation. Furthermore, we taken into account random surgery durations and random lengths of stay in the constraints of the optimization model. Both constraints are linearized in order to come up with an integer linear program (ILP). After having solved this ILP, a second stage is executed in which repetitively feasibility is checked by Monte Carlo simulation, and the solution is improved by a local search heuristic based on tabu search. The main contributions of our paper are the multi-objective criterion function incorporating the goal of leveling ward capacity as much as possible, and the multistage approach of both solving approximate ILP and improving by Monte Carlo simulation and tabu search. In a recent paper, Schneider et al. (2020) have followed quite similar objectives and approaches, albeit with different technical details.

The paper is organized as follows. The next section describes the planning problem and introduces the notation. Here we present the basic optimization formulation of the planning problem, including the stochastic constraints. In Section 3 we show how we approximate the stochastic constraint on the overtime of surgery blocks by linear restrictions. Our method is described in Section 4.2, including also the post-processing stage of Monte Carlo simulation and tabu search. In Section 5 our method is applied, validated and compared in quality in a case study. We give also results on the problem if it would be modeled without taken into account the randomness of the surgery duration and the length of stay. These results are much worse than the results for our method. Section 6 concludes this paper with a brief discussion and suggestions for further research.

2. The Planning Problem

We consider a hospital planning problem that involves S medical specialties, $s = 1, \ldots, S$. The patients in these specialties are divided in I patient groups, $i = 1, \ldots, I$, where we denote $I^s \subset \{1, \ldots, I\}$ to be the set of patient groups belonging to specialty s. Typically, patients in the same group have similar treatment characteristics.

The planning problem covers a time span of T days, $t = 1, \ldots, T$, in which each specialty has pre-assigned blocks of surgery time. The available operating time for

specialty s at day t is denoted by m_{st}. These blocks have been determined by the Master Surgical Schedule (MSS) during the strategic and tactical planning stages, as we have explained in Section 1.1. The construction of the MSS is based on the numbers of patients that play a role in the planning period. There are numbers of patients on each of the waiting lists on day 1, the beginning of the planning period, which we denote by w_{i1} for patient group i, and the number of new patients coming from outpatient clinic during the planning period, which we assume to be known, and will be denoted by d_{it} for patient group i at day t.

After surgery, patients recover in one of the designated wards. There are J wards, $j = 1, \ldots, J$, where ward j has bed capacity b_j. Some wards are particularly designated to a certain group or specialty. Therefore, we let $J^i \subset \{1, \ldots, J\}$ to be the subset of wards that are accessible for patient group i.

Now, the planning problem is to decide on the number x_{ijt} of patients of group i to be scheduled for surgery on day t and to be assigned to ward j after the surgery. The main focus of this paper is on minimizing variability in the bed occupancy levels in the wards. This objective alone may lead to overall low numbers of patients scheduled for surgery, and hence growing waiting lists. Therefore, the total occupancy of the ORs and the growth of the waiting lists are added as ingredients to the objective.

The constraints deal with three issues:

(i) The randomness of surgery times is captured in a probabilistic restriction expressing that the probability of surgery overtime should be smaller than a predefined percentage. This will be worked out in Section 2.1 and in Section 3.

(ii) The randomness of the recovery times in the wards is reflected in a constraint that the expected number of occupied beds cannot exceed the capacity of the ward. More on this will be described in Section 2.2.

(iii) In each group i, the total number of patients scheduled up till time t cannot exceed the total number of group-i patients that entered the system up till time t.

The last restriction is captured by two sets of constraints,

$$\begin{cases} \sum_{\tau=1}^{t} \sum_{j \in J^i} x_{ij\tau} \leq w_{i1} + \sum_{\tau=1}^{t} d_{i\tau}, & \forall i, t; \\ w_{iT} + \sum_{t=1}^{T} \sum_{j \in J^i} x_{ijt} = w_{i1} + \sum_{t=1}^{T} d_{it}, & \forall i. \end{cases} \tag{1}$$

The latter is required to be able to compute the waiting lists at the end of the planning period.

2.1. Surgery Duration

Surgery duration is the total time a patient will be in the OR, which includes preparation, anesthesia and surgery. The time needed between patients for cleaning and preparing the ORs for the next surgery has not been taken explicitly into account in the model (due to a lack of available data), but the model is easily adjusted for when these durations are known.

Let X^i be the random surgery duration for a patient from patient group i. Since x_{ijt} is the number of surgeries scheduled on day t from patient group i, and afterwards recovering on ward j, the total scheduled surgery time for patients in specialty s on day t is

$$Y_{st} = \sum_{i \in I^s} \sum_{j \in J^i} \sum_{k=1}^{x_{ijt}} X_k^i, \tag{2}$$

where X_k^i is the k-th iid replication of X^i.

Now, we model constraint (i) as the probabilistic restriction that this total surgery time exceeds the available surgery time m_{st} with probability at most α,

$$\mathbb{P}(Y_{st} > m_{st}) < \alpha, \quad \forall s, t. \tag{3}$$

In Section 3 we will elaborate our approach of linearizing this constraint.

2.2. Length of Stay (LOS) and Bed Occupancy

The LOS of a patient from group i in a ward is the number of days to recover after surgery. We model it as an integer valued random variable R^i. From the lengths of stay we compute the bed occupancies in the wards. Bed occupancy is divided into two parts. The first part is the number ν^{jt} of patients that were operated in previous planning periods and that are still present at ward j at time t. The second part is the result of the decisions we make on newly assigned patients. Patients are assumed to enter the ward on the same day of their surgery. Define B^{jt} to be the number of occupied beds at ward j on day t. Thus,

$$B^{jt} = \nu^{jt} + \sum_{\tau=1}^{t} \sum_{i=1}^{I} \sum_{k=1}^{x_{ij\tau}} \mathbb{I}\{R_k^i > t - \tau\}, \tag{4}$$

where we use $\mathbb{I}\{R_k^i > t-\tau\}$ as the indicator function of the event specified in brackets. The constraint that we have implemented, considers the expected number of beds at any ward j at any day t. Let $p^i(r) = \mathbb{P}(R^i = r)$, $r = 0, 1, \ldots$ be the probability mass function of the lenght of stay of an arbitrary patient of group i, then the probability of occupying a bed after $t-\tau$ days equals $\mathbb{P}(R^i > t-\tau) = 1 - \sum_{r=1}^{t-\tau} p^i(r)$. Hence, the expected bed occupancy at ward j on day t is

$$\begin{aligned}\mathbb{E}[B^{jt}] &= \nu^{jt} + \sum_{\tau=1}^{t}\sum_{i=1}^{I} \mathbb{P}\big(R^i > t-\tau\big)\, x_{ij\tau} \\ &= \nu^{jt} + \sum_{\tau=1}^{t}\sum_{i=1}^{I} \Big(1 - \sum_{r=1}^{t-\tau} p^i(r)\Big)\, x_{ij\tau},\end{aligned} \tag{5}$$

Let b_j be the capacity of staffed beds in ward j, then the restriction for the expected number of bed occupancy at ward j on day t is,

$$\nu^{jt} + \sum_{\tau=1}^{t}\sum_{i=1}^{I} \Big(1 - \sum_{r=1}^{t-\tau} p^i(r)\Big)\, x_{ij\tau} \leq b_j. \tag{6}$$

This restriction involves a first-moment approximation of the probability distribution of the bed occupancy B^{jt}, and in this sense does not guarantee that the capacity would be met always with certainty when we would generate or simulate random instances of the model. Note that the second part of B^{jt} is a sum of many Bernoulli random variables, and therefore it approximates a normally distributed random variable. This would mean, supposing that $\mathbb{E}[B^{jt}] \approx b_j$, and supposing that samples of B^{jt} are randomly generated from its distribution, about half of these samples would violate the capacity. However, assuming that we generated a feasible solution to the optimization problem, we argue that violation will not occur that many times. First of all, the assignments x_{ijt} satisfy the other constraints concerning (i) overtime and (iii) the demand for surgery, referring to the three issues posed above. Secondly, and more importantly, in the next section we will introduce variables $\ell_j \leq u_j \leq b_j$, such that $\ell_j \leq \mathbb{E}[B^{jt}] \leq u_j$ and such that the gap $u_j - \ell_j$ is minimized as part of our multi-objective function. For these two reasons, the assignments x_{ijt} are reduced largely compared when they would be determined solely by (6).

Other ways of dealing with ward capacities in an optimization model have been considered in literature. For instance one could add a penalty cost for capacity violation (Fügener et al., 2014; Jebali & Diabat, 2017), or add the maximum demand in the objective function (Van Oostrum et al., 2008), or use the maximum historical LOS as

constraint (Carter & Ketabi, 2013), or use the 90% quantile of the empirical LOS (Vanberkel et al., 2011), or allow overflow to a downstream source with unlimited capacity (Neyshabouri & Berg, 2017), or cancel surgery operations (Schneider et al., 2020). None of these techniques have been implemented in our paper, the main reason being that our approach suited our case study.

2.3. Objective function

The objective function consists of three parts, (1) minimizing variability in the ward occupancies, (2) scheduling as many patients as possible, and (3) minimizing the growths of the waiting lists.

2.3.1. Part (1)

Minimizing variability of the ward occupancies can be modeled in several ways. Here we have chosen to determine the maximum and minimum expected bed occupancy for each ward during the planning period and to minimize the sum over all wards of the weighted difference between these values. The advantage of the chosen objective function is its easy implementation in a linear optimization model.

For each ward j we introduce two variables, representing the maximum u_j and minimum ℓ_j of the expected bed occupancy over all days of the planning period:

$$u_j = \max_t \mathbb{E}[B^{jt}]; \quad \ell_j = \min_t \mathbb{E}[B^{jt}].$$

Since the size of the wards may influence the fluctuation in occupancy, it is important to normalize the difference $u_j - \ell_j$ by b_j, the total staffed bed capacity of ward j. This results in part (1) of the objective:

$$\sum_{j=1}^{J} \frac{u_j - \ell_j}{b_j}. \tag{7}$$

Schneider et al. (2020) considered a similar approach for linearizing the bed occupancy leveling objective, though without this normalization.

2.3.2. Part (2)

Objective (7) may lead to undesirable solutions, since it does not take into account the number of patients to be operated, which may lead to undesired low usage of the ORs and enormous growth of waiting lists. Next to that, it is also undesired that the model favors specific patient groups. Indeed, for each patient group within the same specialty

the increase of the number of patients on the waiting list should be minimized, and the decrease maximized over the planning period. Therefore, we will add two goals to the objective, the first supports scheduling as many patients as possible. The secondly added goal makes sure that it is unattractive to let the waiting list increase for any patient group (see Section 2.3.3 for further details).

Furthermore, a weight based on relative expected surgery duration is introduced to make patients evenly important to be scheduled. A patient group is weighted by dividing its expected surgery duration by the sum of expected surgery durations of all patient groups belonging to the same specialty. This leads to the part (2) of the objective, concerning scheduling as many patients as possible:

$$-\sum_{s=1}^{S}\sum_{i\in I^s}\sum_{t=1}^{T}\sum_{j\in J^i}\frac{\mathbb{E}[X^i]}{\sum_{h\in I^s}\mathbb{E}[X^h]}x_{ijt}. \tag{8}$$

2.3.3. Part (3)

For the last part of the objective, recall the variables w_{i1} and w_{iT} measuring the number of patients from patient group i on the waiting list at the beginning and the end of the planning period, respectively. To discourage an increase of the waiting lists, part (3) of the objective concerns the total increase of waiting lists:

$$\sum_{i=1}^{I}\max\{0,(w_{iT}-w_{i1})\}. \tag{9}$$

2.4. The Stochastic Planning Problem

Concluding, we formulate the planning problem that incorporates the objectives and constraints worked out above. To discriminate the relative importance of the three objective parts we give positive weights β and γ to terms (8) and (9), respectively. The values of β and γ will be varied to obtain more desirable solutions.

Furthermore, it can be desirable to give preference to a specific patient group, e.g. when the length of the waiting list for a patient group is too large and needs to be shortened rapidly. Therefore, a weight ρ_i is introduced in (8) indicating the importance of the respective patient group in comparison to other patient groups. Alternative objective functions will be discussed in Section 6.

All ingredients described above lead to the following formulation of the planning problem we consider. The random variables involved are X^i for the surgery duration of a group i patient, and R^i for the length of stay of a group i patient.

$$\begin{aligned}\text{Minimize } Z = & \sum_{j=1}^{J} \frac{u_j - \ell_j}{b_j} \\ & - \beta \sum_{s=1}^{S} \sum_{i \in I^s} \sum_{j \in J^i} \sum_{t=1}^{T} \rho_i \left(\frac{\mathbb{E}[X^i]}{\sum_{h \in I^s} \mathbb{E}[X^h]} x_{ijt} \right) \\ & + \gamma \sum_{i=1}^{I} \max\{0, (w_{iT} - w_{i1})\}.\end{aligned} \tag{10}$$

subject to:

$$Y_{st} = \sum_{i \in I^s} \sum_{j \in J^i} \sum_{k=1}^{x_{ijt}} X_k^i, \qquad \forall s, t.$$

$$\mathbb{P}(Y_{st} > m_{st}) < \alpha, \qquad \forall s, t.$$

$$B^{jt} = \nu^{jt} + \sum_{\tau=1}^{t} \sum_{i=1}^{I} \sum_{k=1}^{x_{ij\tau}} \mathbb{I}\{R_k^i > t - \tau\}, \qquad \forall j, t.$$

$$\mathbb{E}\big[B^{jt}\big] \leq b_j, \qquad \forall j, t.$$

$$\sum_{\tau=1}^{t} \sum_{j \in J^i} x_{ij\tau} \leq w_{i1} + \sum_{\tau=1}^{t} d_{i\tau}, \qquad \forall i, t.$$

$$w_{iT} + \sum_{t=1}^{T} \sum_{j \in J^i} x_{ijt} = w_{i1} + \sum_{t=1}^{T} d_{it}, \qquad \forall i.$$

$$u_j \geq \mathbb{E}[B^{jt}], \qquad \forall j, t.$$

$$\ell_j \leq \mathbb{E}[B^{jt}], \qquad \forall j, t.$$

$$x_{ijt} \in \mathbb{Z}_{\geq 0}, \quad \forall i, j, t.$$

3. Linearizing the Overtime Constraints

This section deals with developing a linear approximation of the overtime constraint (3). Consider this constraint for specialty s. It suffices to determine how many patients fit in a block while the surgery overtime probability is at most α. It deals with many different patient groups, and thus different distributions for the surgery durations. Our method is to limit the combinations of patients from different groups that can be scheduled in a block based on expectations and variances. In this way we introduce a linearization based on a two-moment approximation of the underlying distributions and their probabilities. Because the solution of the model will be simulated, this approach is

sufficient to generate a feasible candidate start solution for the tabu search which then improves the solution.

The method has two steps. The first step creates a best case scenario by executing the surgery only for patients (in specialty s) with the minimal expected surgery time. This gives the maximal number of patients that can be operated satisfying the constraint. We then use the ratio of the expected surgery times of the other patient groups (in specialty s) with respect to this minimal expected surgery time in order to formulate a linear inequality, see (12). This inequality is refined in the second step by taking into account also the variances of the surgery times, see (15). Below this is worked out.

Step 1. Let

$$\underline{i}^s = \arg\min\{\mathbb{E}[X^i] : i \in I^s\}$$

be the group in specialty s with the minimal expected surgery time. Determine the maximal number of group $\underline{i}^s$ surgeries such that the overtime constraint is satisfied if there would be no surgeries from other groups of specialty s on day t:

$$n_{st} = \max\Big\{n : \mathbb{P}\Big(\sum_{k=1}^{n} X_k^{\underline{i}^s} > m_{st}\Big) < \alpha\Big\}. \tag{11}$$

In Appendix B we shall argue how to solve this equation approximately but quickly. Now notice that $Y_{st} \leq \sum_{k=1}^{n_{st}} X_k^{\underline{i}^s}$ would (stochastically) be sufficient for $\mathbb{P}(Y_{st} > m_{st}) < \alpha$. Our approximation is to require instead

$$\mathbb{E}[Y_{st}] \leq \mathbb{E}\Big[\sum_{k=1}^{n_{st}} X_k^{\underline{i}^s}\Big].$$

Working out the expectations we get the linear inequality:

$$\sum_{i\in I^s}\sum_{j\in J^i} \frac{\mathbb{E}[X^i]}{\mathbb{E}[X^{\underline{i}^s}]} x_{ijt} \leq n_{st}, \quad \forall s,t. \tag{12}$$

Step 2. For taking the variability of the surgery durations into account, define

$$\delta_i = \mathbb{V}ar[X^i]/\mathbb{E}[X^i], \quad i \in I, \tag{13}$$

and consider their weighted average (per specialty):

$$\delta^s = \sum_{i\in I^s} f_i^s \delta_i, \tag{14}$$

where the weights $f_i^s \geq 0$, $\sum_{i \in I^s} f_i^s = 1$. How these weights are chosen will be explained in the case study in Section 5. Now notice that in equation (11) we deal with the distribution of the variable

$$\Theta_n^s = \sum_{k=1}^{n} X_k^{\underline{i}^s}.$$

Clearly,

$$\mathbb{V}ar[\Theta_n^s] = n\mathbb{V}ar[X^{\underline{i}^s}] = n\delta_{\underline{i}^s}\mathbb{E}[X^{\underline{i}^s}] = \delta_{\underline{i}^s}\mathbb{E}[\Theta_n^s].$$

Suppose that we are able to construct a variable $\widetilde{\Theta}_n^s$ such that

$$\mathbb{E}[\widetilde{\Theta}_n^s] = \mathbb{E}[\Theta_n^s]; \text{ and } \mathbb{V}ar[\widetilde{\Theta}_n^s] = \delta^s\mathbb{E}[\Theta_n^s]. \tag{15}$$

More specifically, the variable $\widetilde{\Theta}_n^s$ has the same first moment as the sum of group $\underline{i}^s$ surgery times, however it incorporates the variability of the other groups within specialty s. Then we solve

$$\widetilde{n}_{st} = \max\Big\{n : \mathbb{P}\Big(\widetilde{\Theta}_n^s > m_{st}\Big) < \alpha\Big\}, \tag{16}$$

which, finally, leads to the linear constraint that replaces the probabilistic constraint (3),

$$\sum_{i \in I^s}\sum_{j \in J^i} \frac{\mathbb{E}[X^i]}{\mathbb{E}[X^{\underline{i}^s}]} x_{ijt} \leq \widetilde{n}_{st}, \quad \forall s, t. \tag{17}$$

4. The Solution Method

The method that we designed for solving the problem is based on an integer linear programming model that approximates the original stochastic planning problem of Section 2.4.

4.1. The ILP

The ILP model implements the distributional linearization of the two stochastic constraints. This results in the objective (10) under the restrictions (1), (6), and (17). Because we need to compute the upper and lower bounds, u_j, ℓ_j in the first part of the objective (7), the linear ward restriction (6) is split into two constraints (see below),

resulting in

$$\text{Minimize } Z = \sum_{j=1}^{J} \frac{u_j - \ell_j}{b_j} \tag{18}$$

$$- \beta \sum_{s=1}^{S} \sum_{i \in I^s} \sum_{j \in J^i} \sum_{t=1}^{T} \rho_i \left(\frac{\mathbb{E}[X^i]}{\sum_{h \in I^s} \mathbb{E}[X^h]} x_{ijt} \right) \tag{19}$$

$$+ \gamma \sum_{i=1}^{I} \max\{0, (w_{iT} - w_{i1})\}. \tag{20}$$

subject to:

$$\sum_{\tau=1}^{t} \sum_{j \in J^i} x_{ij\tau} \leq w_{i1} + \sum_{\tau=1}^{t} d_{i\tau} \qquad \forall i, t, \tag{21}$$

$$w_{iT} + \sum_{t=1}^{T} \sum_{j \in J^i} x_{ijt} = w_{i1} + \sum_{t=1}^{T} d_{it} \qquad \forall i, \tag{22}$$

$$\sum_{i \in I^s} \sum_{j \in J^i}^{J} \frac{\mathbb{E}[X^i]}{\mathbb{E}[X^{i^s}]} x_{ijt} \leq \widetilde{n}_{st} \qquad \forall s, t, \tag{23}$$

$$\nu^{jt} + \sum_{\tau=1}^{t} \sum_{i=1}^{I} \left(1 - \sum_{r=1}^{t-\tau} p^i(r) \right) x_{ij\tau} \leq u_j \qquad \forall j, t, \tag{24}$$

$$\nu^{jt} + \sum_{\tau=1}^{t} \sum_{i=1}^{I} \left(1 - \sum_{r=1}^{t-\tau} p^i(r) \right) x_{ij\tau} \geq \ell_j \qquad \forall j, t, \tag{25}$$

$$u_j \leq b_j \qquad \forall j, \tag{26}$$

$$x_{ijt} \in \mathbb{Z}_{\geq 0} \qquad \forall i, j, t. \tag{27}$$

4.2. Method

The method that we developed for solving the stochastic planning problem, consists of the following stages. First we solve the ILP of Section 4.1. Because the ILP implements linear approximations to the stochastic constraints, the ILP solution might not be feasible for the stochastic planning problem. Therefore, we implemented a post-processing stage that repeatedly (i) checks feasibility of a solution, and (ii) adapts the current solution.

(i). Feasibility of the overtime restriction is checked by running a Monte Carlo simulation of all blocks of surgery times Y_{st} (specialty s, day t). This is a sum of random variables, given by the current proposed solution, and thus is easily simulated. In this manner the probability of overtime can be estimated per block and blocks

with a probability of overtime greater than α are identified. The sample sizes that we used were so large, that the standard error of the probability estimator was less than 0.005. In other words, in 95% of the simulation cases, the estimate will be within an absolute difference of at most 1% of its true value.

(ii). To meet the overtime restriction in case of infeasibility of the current solution, the solution is adapted by randomly decreasing the number of patients in the blocks that were indentified being infeasible. This is done one-by-one, and each newly adapted solution is checked for feasibility by running the Monte Carlo simulation again.

(iii). Subsequently a tabu search algorithm is applied which does not change the numbers of scheduled patients, but it may change the days on which the patients are scheduled. The tabu search swaps patients to try to improve the objective value. Each new proposal schedule is checked for feasibility by Monte Carlo simulation. Details on our tabu search are described in Appendix A.

5. Case Study

The case study entails a large hospital in the Netherlands. The operating theatre consists of fifteen ORs generally available five days a week from 8:00 AM to 16:15 PM. In this study the allocation of patients to blocks of operating time is realized for twelve medical specialties for a period of four weeks, which made up a total of 152 surgery blocks. All adult patients will be assigned to one of the five large wards. A total of 24 different patient groups were identified with unique distributions in length of stay and surgery duration. Each medical specialty appeared to have two distinct patient groups, one with short stay patients and the other with long stay patients.

5.1. Surgery Distributions

For the distribution of surgery times a lognormal distribution is proposed, which came out as the best fit in various studies (Strum et al., 2000; Stepaniak et al., 2009; Spangler et al., 2004; Dávila, 2013). The relevant parameters are estimated with a fitting function in MATLAB R2014a® based on maximum likelihood. For every group a chi-square test is performed on the fitted lognormal distribution to verify its use in resembling the surgery times. As an example, in Figure 1 the histogram and fitted

lognormal distribution for surgery times are shown for the short stay patient group within general surgery.

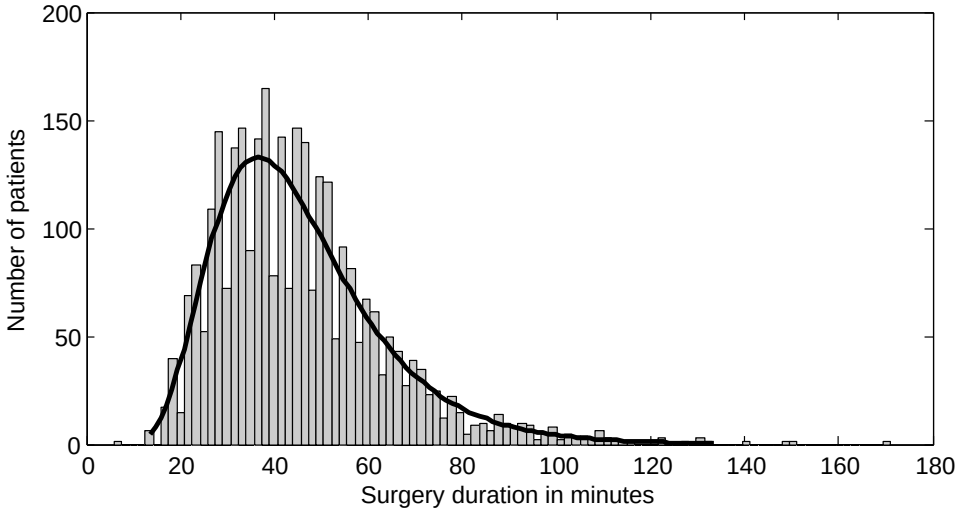
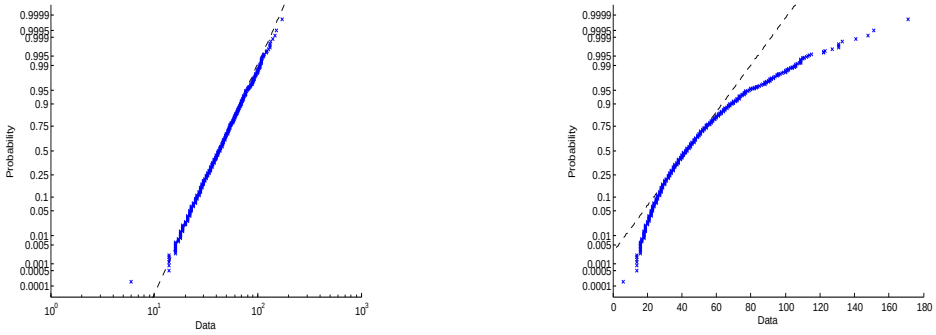


Figure 1: Histogram of surgery times and fitted lognormal probability function.

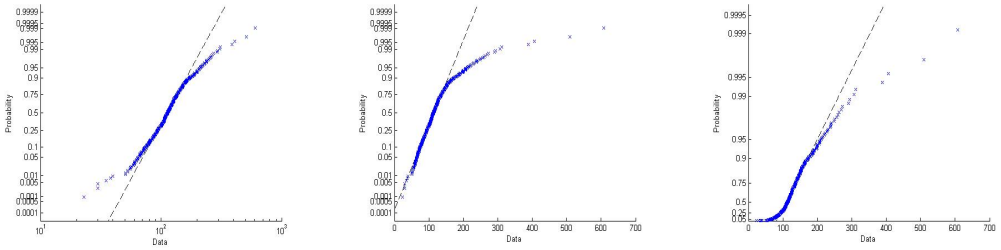
For some patient groups the chi-square test rejected the null hypothesis at the 5% significance level, which indicates a bad fit of the lognormal distribution to the data on surgery times. Due to the shape of the data being similar to the lognormal distribution, two alternative distributions were tested for the chi-square goodness of fit, namely the normal and exponential distributions. Eventually, the surgery duration data were best approximated by the lognormal distribution in comparison to the normal and exponential distributions for all patient groups. Therefore, supported by previous research and the fact that long right tails are important to identify overtime in operating blocks, lognormal distributions were used to represent surgery times in all groups. Figure 2 shows the probability fit graphs of the data for both the lognormal and normal distributions for the general surgery, short stay patient group. The null hypothesis of the chi-square goodness of fit test was not rejected for this group as is intuitively clear from these figures.



(a) Lognormal (b) Normal

Figure 2: Probability fit for the lognormal (A) and normal (B) distributions

Figure 3 shows the probability fit graphs of the data for both the lognormal, normal and exponential distributions for the trauma surgery, long stay patient group. In Figure 4 a histogram of the data of this group is shown with the fitted distribution lines of lognormal, normal and exponential distribution. The null hypothesis of the chi-square goodness of fit test was rejected for this group, but from Figures 3 and 4 it can be seen that the lognormal distribution fits the data best in probability compared to the other two distributions.



(a) Lognormal (b) Normal (c) Exponential

Figure 3: Probability fit for the lognormal (A), normal (B) and exponential (C) distributions

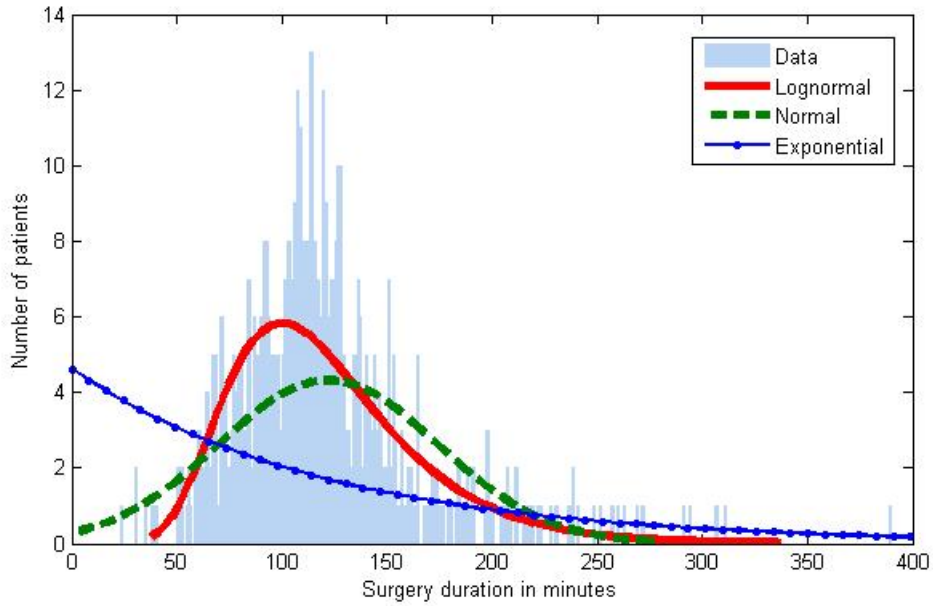


Figure 4: Surgery duration data with fitted lognormal, normal and exponential probability functions.

From historical data the proportion f_i^s of each patient group i within a medical specialty s could be extracted and used to define the weighted ratios of δ^s in (14). Specifically, let f_1^s and f_2^s be the percentages in number of patients for respectively the short stay group and the long stay group within a medical specialty s. Then (14) becomes

$$\delta^s = f_1^s \frac{\mathbb{V}ar[X^1]}{\mathbb{E}[X^1]} + f_2^s \frac{\mathbb{V}ar[X^2]}{\mathbb{E}[X^2]},$$

where X^i is the surgery duration of patient group i.

Within bariatric and otolaryngologic surgery, the ratio $\frac{\mathbb{V}ar[X^i]}{\mathbb{E}[X^i]}$ for short and long stay groups where very close. Therefore, the linear approximation is reliable for every combination of the number of scheduled short and long stay patients. In the other medical specialties the ratio for the long stay group was higher than for the short stay group. Therefore, the weighted ratio over all groups is lower than that of the long stay group, which implies that if only long stay patients would be scheduled in one block, the approximation would predict a smaller variance than in reality and when only short stay patients would be scheduled, the approximated variance would be larger. However, the proportions f_i^s will give a good approximation of the balance between scheduled short and long stay patients in each block. Moreover, in each surgery block only a small number of patients can be scheduled, which will limit the error margin in the variance

approximation.

5.2. LOS Distributions

For the length of stay (LOS) distribution, the probability of each outcome per group can easily be computed by taking the number of patients with the LOS outcome divided by the total number of patients in the group. The empirical probability distribution used in the ILP models has possible outcomes ranging from one day in recovery to a maximum of $T = 26$ days, where the event probability of a LOS of 26 days is defined as the probability of a recovery period of 26 days or longer. To test whether the LOS distribution of the short stay groups differs significantly from the distribution of the surgeries with a resulting long stay, a two-sample Kolmogorov-Smirnov test is executed in each specialty. Because this test only applies to continuous distributions, the corresponding data is smoothed by using hourly length of stay to test the difference. Although eventually the discrete empirical probability distribution is used, all patient groups passed the Kolmogorov-Smirnov statistic such that the null hypothesis for equal continuous distributions was rejected at a 5% significance level in all cases. All p-values were very small, which indicated a clear distinction between groups in their length of stay characteristics and therefore sufficient evidence for the use of the empirical distributions was found. The occurrences of length of stay for each cardiothoracic surgery group is jointly shown in Figure 5, which clearly shows a distinction between the distributions of the two groups.

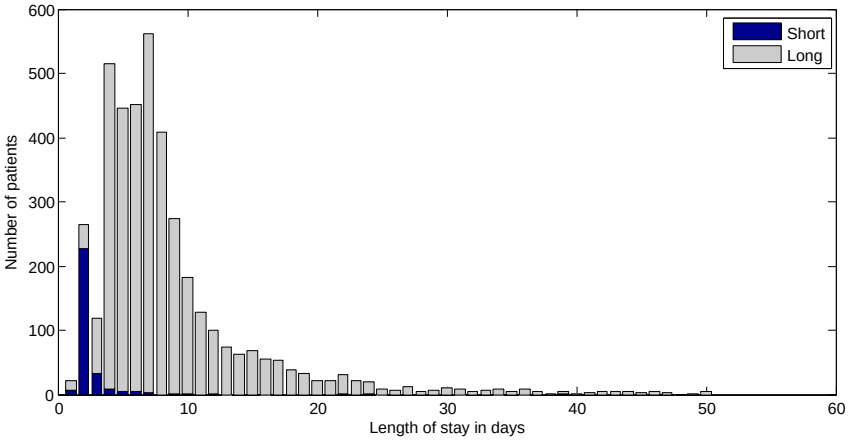


Figure 5: Length of Stay for both patient groups in Cardiothoracic Surgery.

5.3. Variability and Variance

Recall the variable B^{jt} representing the occupancy of ward j at day t, and denote its expected value by b^{jt}. In Section 2.3.1, the variability of the occupancy of ward j was measured by the difference between the maximum and minimum of the b^{jt}'s taken over all days t in the planning period. Given a final solution of our method, we compute b^{jt} according to relation (5), and then we report the aforementioned difference relative (in percentage) to the capacity of the ward. We call this the relative variability. Furthermore, we compute the (sample) mean and (sample) variance of the b^{jt}'s, taken over all days t in the planning period. This variance is reported as well in our results in the next section. Note that we report actually the relative variability and the variance of the expected bed occupancies at a ward.

5.4. Results

The ILP model, described in Section 4.1, has been solved to optimality in MATLAB R2014a® by using Gurobi Optimizer 6.0. This took 290 seconds CPU time on a Windows 7 computer. To define the objective, the weights were set to $\beta = 0.07$ and $\gamma = 0.07$, motivated by giving high priority to the first objective of leveling variability, and giving the other two equal importance. The overtime probability was set to $\alpha = 0.2$. With these inputs, a total of 526 patients were scheduled in 139 blocks of operating time, and 13 blocks of operating time were left empty. The importance of the leveling objective was so much higher than the other parts of the objective that the solution produces empty blocks of surgeries. In Section 5.5 we give results of an alternative approach when we leave out the leveling from the objective. Indeed, then the solution has all blocks filled. The empty blocks will then receive a few patients only (upto 4).

The ILP solution was simulated for checking the feasibility to the probabilistic overtime constraint. There was just a single block (urology on the 10th day of the planning period) that violated this constraint. It sufficed to lower the number of scheduled patients by just one for this block to obtain a feasible solution. Applying the tabu search for possible improvements gave no better solutions. All these took just 40 seconds CPU time to compute. Apparently, the linearization of the stochastic constraint by a two-moment approximation had hardly effect on the feasibility and optimality, at least in our case study. The relative variabilities and variances (see Section 5.3) are displayed in Table 1.

Table 1: The relative variabilities and variances of expected ward occupancies.

Ward	Rel. Var.	Variance
A	35%	9.4
B	4%	0.1
C	2%	0.1
D	15%	1.9
E	5%	0.2

The results show low variability in expected bed occupancy levels for wards B, C and E. The highest variability is observed at ward A, which can be caused by the fact that only patients from the gastrointestinal department are assigned to this ward, so the options in assigning different types of patients is limited and most variability results from the block allocation of the Master Surgical Schedule.

The waiting list decreased for all patient groups except for plastic surgery patients with long expected length of stay. Apparently, the model prefers scheduling short stay patients in the plastic surgery department, which can be changed by decreasing the weight importance of these patients. In Table 2 for each patient group the percentage of change in the length of the waiting list during the planning period is presented. A negative (positive) percentage indicates the decrease (increase) in the length of the waiting list. A 0% change means that the number of new patients is equal to the number of scheduled patients.

Table 2: Percentage change in waiting list for each patient group.

Group	% Change list	Group	% Change list	Group	% Change list
CTS short	−100%	PLA short	−4%	GIS short	−11%
CTS long	−48%	PLA long	+8%	GIS long	−16%
GYN short	0%	URO short	0%	ONC short	0%
GYN long	0%	URO long	−3%	ONC long	−15%
DEN short	−7%	BAR short	−100%	TRAU short	−6%
DEN long	−8%	BAR long	−9%	TRAU long	−12%
EPT short	0%	GEN short	0%	VAS short	−100%
EPT long	−3%	GEN long	−26%	VAS long	−16%

5.5. Performance Analysis

In order to evaluate our approach, we have considered two other approaches. The results of these approaches are compared with the results of our default approach by means of the total number of scheduled patients during the planning period, the variability of the bed occupancies at the wards, and the percentage decrease (or increase) of the waiting lists.

- The purpose of the first alternative approach is to justify the use of stochastic constraints by means of probability distributions, and their approximations (23)–(25) in our original ILP. Suppose that we would not use these probability distributions, but implement immediately their expected values in the constraints. This would lead to the overtime constraint,

$$\sum_{i \in I^s} \sum_{j \in J^i} \mathbb{E}[X_i] x_{ijt} \leq m_{st}, \quad \forall s, t, \tag{28}$$

and the bed occupancy constraint,

$$\nu^{jt} + \sum_{i=1}^{I} \sum_{\tau=t-\mathbb{E}[R^i]+1}^{t} x_{ij\tau} \leq b_j, \quad \forall j, t. \tag{29}$$

The linear programming solution generated a schedule where 548 patients were assigned in 132 blocks. Table 3 shows the results of the variabilities and variances associated with the solution. Clearly, these are worse than we saw in Table 1 of our default approach. Furthermore, when simulating the solution, 48 blocks exceeded the allowed probability of overtime. This was a too large number for improving by our tabu search and simulation procedures.

Table 3: The relative variabilities and variances of expected ward occupancies resulting from the first alternative method.

Ward	Rel. Var.	Variance
A	36%	9.3
B	31%	4.7
C	35%	9.7
D	26%	3.4
E	23%	3.3

- The second alternative approach tests the effect of leveling ward occupancy in the objective. Suppose that this part is omitted from the objective function, and thus also the constraints necessary for computing the maximum and minimum levels (u_j, ℓ_j). However, from the associated ILP solution, we apply Monte Carlo simulation for overtime feasibility checking, and tabu search for improving (minimizing) variability in the wards. Now, this took a total of 50 minutes to complete.

  The results of this approach gave many more scheduled patients, 590 patients in all 152 blocks of surgery times, satisfying the overtime constraint. However, as noted earlier, tabu search will not alter this number, it only tries to level ward occupancies as much as possible. As expected, the final variability objective was worse than from the default approach. But the waiting list objective showed better performance for some patient groups, while worse for others. There was an increase of four waiting lists (just one in our default approach).

Table 4: The relative variabilities and variances of expected ward occupancies resulting from the second alternative method.

Ward	Rel. Var.	Variance
A	49%	18.0
B	64%	13.6
C	48%	15.5
D	38%	13.0
E	48%	15.3

Table 5: Percentage change in waiting list for each patient group from the second alternative approach.

Group	% Change list	Group	% Change list	Group	% Change list
CTS short	−86%	PLA short	−4%	GIS short	−15%
CTS long	−72%	PLA long	+8%	GIS long	−13%
GYN short	−3%	URO short	−2%	ONC short	−14%
GYN long	+2%	URO long	+14%	ONC long	−10%
DEN short	+8%	BAR short	0%	TRAU short	−6%
DEN long	−1%	BAR long	−21%	TRAU long	−6%
EPT short	−5%	GEN short	−4%	VAS short	−43%
EPT long	−4%	GEN long	−15%	VAS long	−31%

6. Conclusion and Further Research

In this paper we studied a stochastic planning problem for scheduling patients in surgery while taking into account overtime and variability of ward occupancy. We proposed an approximating ILP model, that serves as the first stage of a solution method, followed by a post-processing Monte Carlo simulation for feasibility and tabu search for optimizing. In our method we linearize the stochastic constraints using the distributions of the stochastic elements of the problem. As was shown in the case study, this approach gave very good results, and outperforms the method based on the common approach of linearizing by just using averages.

Some limitations of the proposed model and method introduce new research opportunities. It could be interesting to consider occupancy levels for the morning, afternoon and during the night or even on an hourly basis. Therefore, the length of stay has to be measured in hours and the sequence of patients to be scheduled in a block has to be determined, which involves an extra assignment in the model. When a patient leaves in the morning and another patient is admitted in the afternoon, it is possible that a staffed bed can be used by these two patients on the same day. It is particularly useful in order to define the necessary staffing levels during the night for each ward. However, the problem size will be much larger in this variant, which can lead to a long running time to solve the problem. Furthermore, adding the number of occupied beds resulting from elective patients who are admitted to the wards one or more days before their surgery

date, would make the expected bed occupancy levels more accurate.

The proposed method in this paper can also be used to define the necessary staffing levels in the wards and therefore work as an input tool for staffing schedules.

In this paper, surgery duration includes the preparation, anesthesia and actual operating times of a patient. It would be interesting to investigate the advantage of separating these elements of surgery duration, when planning patients for surgery. The different components will probably follow different distributions and thereby resulting in a different approximation for the probability of overtime in a surgery block. In case of multiple blocks of the same specialization on one day, another suggestion for further research would be to take each block separately and compute the probabilities for each block. This involves a different composition of the decision variables where a combination of day and specialty is not sufficient.

The model could be enhanced by taking the bed occupancy levels during weekend days into account. This addition is not hard to implement in the design of the model. Most hospitals do not plan surgeries during weekends, so fluctuations in bed occupancy levels are only determined by patients leaving the hospital. Thereby keeping the occupancy levels during weekends as low as possible, will result in less weekend shifts.

The proposed method could also be approached in a different order. Simulating the number of scheduled patients in a block in order to define the probability of overtime, could be incorporated in a pre-processing step. By simulating all possible combinations of patients with a probability of overtime smaller than α, a set of allowed options to be scheduled can be computed per block. This step can be seen as a knapsack problem, where the limit of the knapsack is defined by the maximum probability of overtime. By dynamic programming, only the options where no other patients can be added are selected. As such, all blocks will be maximally utilized and the constraint on probability of overtime has been satisfied for all options. Subsequently an integer linear program can be applied to assign the best option to each block and to allocate the patients over the allowed wards in order to minimize the resulting variance of expected bed occupancy levels in each ward.

Finally, it would be interesting to predict the surgery duration of a patient by using econometric forecasting models. A patient's age, gender, surgeon and disease are some variables that could differentiate the surgery duration of a patient. In this way more patient groups could be distinguished with a more exact expectation of surgery duration.

References

Banditori, C., Cappanera, P., & Visintin, F. (2013). A combined optimization–simulation approach to the master surgical scheduling problem. *IMA Journal of Management Mathematics*, *24*, 155 – 187.

Bekker, R., & Koeleman, P. M. (2011). Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, *14*, 237–249.

Beliën, J., & Demeulemeester, E. (2007). Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, *176*, 1185–1204.

Brandenburg, J. C. (2010). *Detection Statistics Of Multiple-Pulse Optical Signals Through Atmospheric Turbulence*. Ph.D. thesis Wayne State University.

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

Carter, M. W., & Ketabi, S. (2013). Bed balancing in surgical wards via block scheduling. *Journal of Minimally Invasive Surgical Sciences*, *2*, 129–137.

Denton, B., Viapiano, J., & Vogl, A. (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, *10*, 13–24.

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*, 802–816.

Dávila, M. P. (2013). *A Methodology for Scheduling Operating Rooms Under Uncertainty*. Ph.D. thesis University of South Florida.

Faddy, M., Graves, N., & Pettitt, A. (2009). Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value in Health*, *12*, 309–314.

Fenton, L. (1960). The sum of lognormal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems*, *8*, 57–67.

Fügener, A., Hans, E. W., Kolisch, R., Kortbeek, N., & Vanberkel, P. T. (2014). Master surgery scheduling with consideration of multiple downstream units. *European Journal of Operational Research*, *239*, 227 – 236.

Gur, S., & Eren, T. (2018). Application of operational research techniques in operating room scheduling problems: Literature overview. *Journal of Healthcare Engineering*, *2018*, Article ID 5341394.

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

Hooshmand, F., MirHassani, S., & Akhavein, A. (2018). Adapting ga to solve a novel model for operating room scheduling problem with endogenous uncertainty. *Operations Research for Health Care*, *19*, 26 – 43.

Jebali, A., & Diabat, A. (2017). A chance-constrained operating room planning with elective and emergency cases under downstream capacity constraints. *Computers and Industrial Engineering*, *114*, 329 – 344.

Kayis, E., Khaniyev, T. T., Suermondt, J., & Sylvester, K. (2015). A robust estimation model for surgery durations with temporal, operational, and surgery team effects. *Health Care Management Science*, *18*, 222–233.

Kroer, L. R., Foverskov, K., Vilhelmsen, C., Hansen, A. S., & Larsen, J. (2018). Planning and scheduling operating rooms for elective and emergency surgeries with uncertain duration. *Operations Research for Health Care*, *19*, 107 – 119.

Marazzi, A., Paccaud, F., Ruffieux, C., & Beguin, C. (1998). Fitting the distributions of length of stay by parametric models. *Medical Care*, *36*, 915–927.

Mehta, N., & Molish, A. (2007). Approximating a sum of random variables with a lognormal. *IEEE Transactions on Wireless Communications*, *6*, 2690–2699.

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

Molina-Pariente, J. M., Hans, E. W., & Framinan, J. M. (2018). A stochastic approach for solving the operating room scheduling problem. *Flexible Services and Manufacturing Journal*, *30*, 224–251.

Neyshabouri, S., & Berg, B. P. (2017). Two-stage robust optimization approach to elective surgery and downstream capacity planning. *European Journal of Operational Research*, *260*, 21 – 40.

Saadouli, H., Jerbi, B., Dammak, A., Masmoudi, L., & Bouaziz, A. (2015). A stochastic optimization and simulation approach for scheduling operating rooms and recovery beds in an orthopedic surgery department. *Computers and Industrial Engineering*, *80*, 72 – 79.

Schneider, A., van Essen, J., Carlier, M., & Hans, E. (2020). Scheduling surgery groups considering multiple downstream resources. *European Journal of Operational Research*, *282*, 741–752.

Spangler, W., Strum, D., Vargas, L., & May, J. (2004). Estimating procedure times for surgeries by determining location parameters for the lognormal model. *Health Care Management Science*, *7*, 97–104.

Stepaniak, P., C.Heij, Mannaerts, G., de Quelerij, M., & de Vries, G. (2009). Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study. *Anesthesia and Analgesia*, *109*, 1232–1245.

Strum, D., May, J., & Vargas, L. (2000). Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. *Anesthesiology*, *92*, 1160–1167.

Van Essen, J. T., Bosch, J. M., Hans, E. W., van Houdenhoven, M., & Hurink, J. L. (2014). Reducing the number of required beds by rearranging the or-schedule. *OR Spectrum*, *36*, 585–605.

Van Oostrum, J., van Houdenhoven, M., Hurink, J., Hans, E., Wullink, G., & Kazemier, G. (2008). A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR spectrum*, *30*, 355–374.

Vanberkel, P. T., Boucherie, R. J., Hans, E. W., Hurink, J. L., van Lent, W. A. M., & van Harten, W. H. (2011). An exact approach for relating recovering surgical patient workload to the master surgical schedule. *Journal of the Operational Research Society*, *62*, 1851–1860.

Xiang, W. (2017). A multi-objective aco for operating room scheduling optimization. *Natural Computing*, *16*, 607–617.

Zhang, B., Murali, P., Dessouky, M., & Belson, D. (2009). A mixed integer programming approach for allocating operating room capacity. *Journal of Operational Research Society*, *60*, 663–673.

Zhu, S., Fan, W., Yang, S., Pei, J., & Pardalos, P. M. (2019). Operating room planning and surgical case scheduling: a review of literature. *Journal of Combinatorial Optimization*, *37*, 757–805.

Appendix A. Tabu Search

The algorithm of the tabu search is displayed schematically in Figure A.6. First, the heuristic will randomly pick a block in the MSS and another block on a different day in the planning period with the same assigned specialism. Thereafter, two patient groups are randomly chosen belonging to this specialism, where the first group has a smaller expected surgery duration than the second group. Then, a set of neighbouring solutions is determined:

(i). Swap a patient from group 1 in block 1 with a patient from group 2 in block 2.

(ii). Swap a patient from group 2 in block 1 with a patient from group 1 in block 2.

(iii). Swap two patients from group 1 in block 1 with a patient from group 2 in block 2.

(iv). Swap a patient from group 2 in block 1 with two patients from group 1 in block 2.

Because patients in the first group have a shorter expected surgery duration it could be feasible to swap two of these patients with one patient from a group with longer expected surgery duration. First, it is checked if the swap options are actually possible. The first swap option in the list requires at least 1 scheduled patient of group 1 in the first block and 1 scheduled patient of group 2 in block 2. Out of the possible swaps, one is randomly chosen. It is checked if this solution satisfies the waiting list constraint (1) and the constraint on available staffed beds (6), with a Monte Carlo simulation the same is done for the probability constraint (3). If the solution satisfies all constraints, the objective value for this solution is analysed and accepted when it is an improvement on the value of the previous solution. When the solution is either not feasible or the

objective value is not an improvement, one of the other neighbours is examined. If all neighbours are infeasible a new block is chosen as second swapping block and new neighbours have to be found. When all other blocks have already been examined by the algorithm or no other block exists, the search starts at the beginning by randomly picking a block in the MSS. In the case of all neighbours having a worse objective value compared to the value of the previous chosen solution, the best of these worse solutions is accepted as the next step in the tabu search. If a new solution is accepted, it will be added to the tabu list in order to prevent cycling back and forth between solutions. When a newly accepted neighbour enters the tabu list and it causes the length of the tabu list to exceed the predefined maximum length, the neighbour that has been in the list the longest will be deleted from the list. Over all iterations the best found solution is saved, and the tabu search algorithm stops when the number of iterations since the last improvement of the best found solution is larger than a specified number in terms of the number of swap opportunities. In our case study we used 5 as maximum tabu length, and 50 non-improving iterations as stopping criterion.

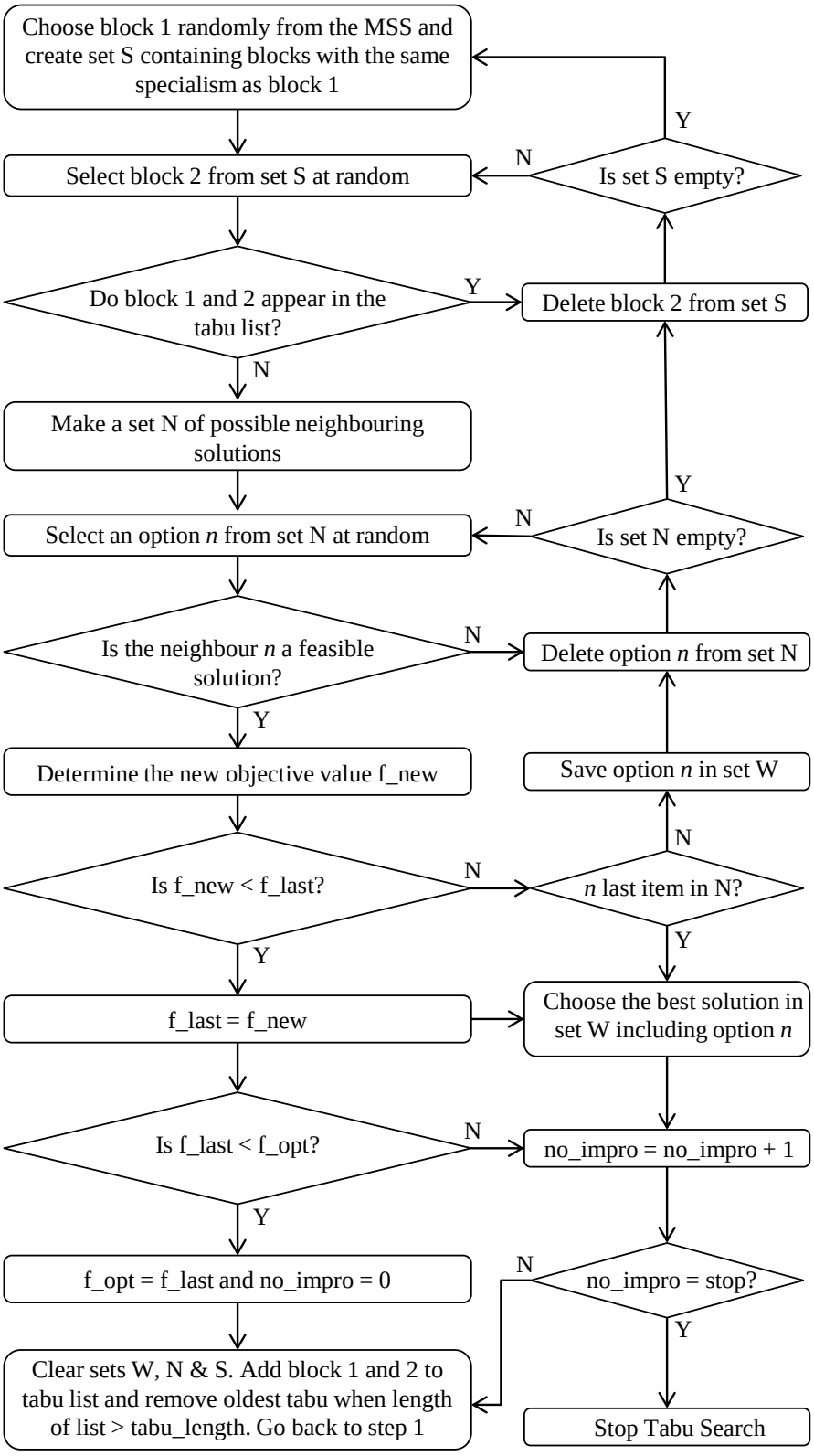


Figure A.6: Steps of Tabu Search.

Appendix B. Sum of Lognormal Variates

In general, a finite sum of independent lognormal variables is not known to behave like a recognized distribution, however from previous research a simplifying method to approximate the sum of independently distributed lognormal variables is to use again a lognormal distribution with a moment matching approach as was proposed by Fenton (1960) and used by Brandenburg (2010) and Mehta & Molish (2007). Generally, just a couple of patients can be treated in one block of operating time, therefore the large right tail of the lognormal surgery durations will probably still be present in the distribution of the sum of the random variables, which is another reason to approximate the sum with a lognormal distribution. Moreover, it is important that the tail of the approximated distribution resembles that of the sum of the random variables well, because only the probability of overtime is considered here. An example to check the approximation is given for an 8 hour surgery block, where 3 patients with a short surgery duration and 4 patients with a longer surgery duration are scheduled. The two groups have different lognormal distributions for their operating time. The surgery durations are simulated multiple times and each simulation generates the sum of the random variables. In Figure B.7 the outcomes for all simulations are presented in a bar chart with a fitted lognormal probability density function. It can be seen that the sum of surgery durations is well approximated by a lognormal distribution as is verified by a chi-square goodness of fit test. Because the shape in the figure seems similar to a normal distribution function, the chi-square goodness of fit test is also applied to a normal distribution fit. However, the test rejected the null hypotheses that the simulated data followed the fitted normal distribution.

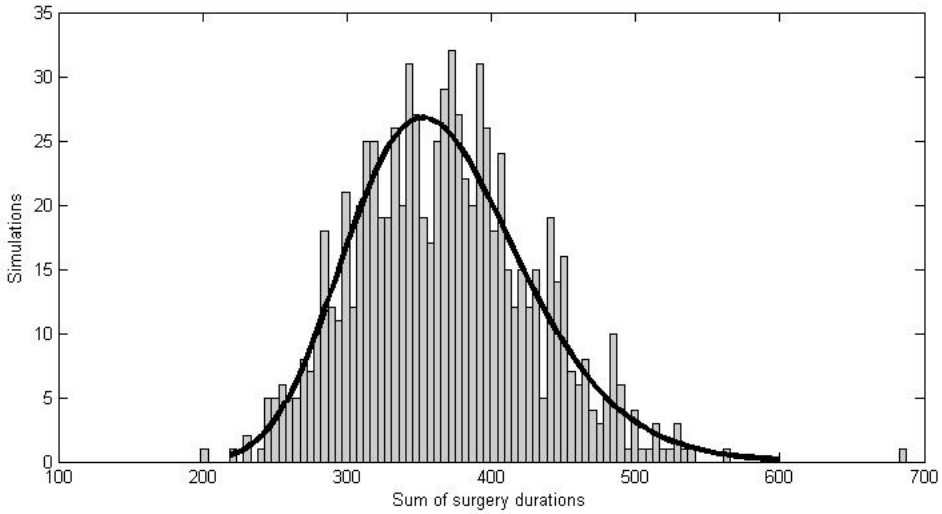


Figure B.7: Histogram of simulated data of a sum of 7 lognormal variables and the fitted single lognormal probability density function.

The parameters of the fitted lognormal distribution are determined simply by moment matching, where its first and second moments are the same as those of the original sum of the lognormals.