2018

Stochastic BI-Level Optimization Models for Efficient Operating Room Planning

Amin Abedini
University of Kentucky, a.abedini@uky.edu

Wei Li
University of Kentucky, wei.mike.li@uky.edu

Honghan Ye
University of Kentucky, honghan.ye@uky.edu

Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 26 (2018) 58–69

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA

Stochastic bi-level optimization models for efficient operating room planning

Amin Abedini[a], Wei Li[a,*], Honghan Ye[a]

[a] *Department of Mechanical Engineering, University of Kentucky, Lexington,40506, KY, USA*

* Corresponding author. Tel.: +1-859-257-4842; fax: +1-859-257-3304.
E-mail address: wei.mike.li@uky.edu

Abstract

Within a hospital, the operating room (OR) department has the largest cost and revenue. Because of the aging population, the demand for surgical services has been increasing sharply in recent years. At the other hand, the rate of OR capacity expansion is lower than the rate of increasing demand. As a result, OR managers must leverage their resources by efficient OR planning. OR planning is challenging because of multiple competing\conflicting objectives such cost minimization and throughput maximization. Inherent uncertainty in the surgical procedures and patients arrivals complicate the decision making process. This increases the risk of non-realization of the system objectives. In this paper, stochastic bi-level optimization models were formulated to optimize total cost and throughput of ORs under the presence of uncertainties in patient arrivals and case times. Newsvendor model and chance-constrained optimization method were used to optimize multiple objectives under the presence of uncertainties. Using historical data, a simulation model was established to validate the results of optimization models. Using statistical process control (SPC) stability of each model was investigated. Using bi-level optimization, we addressed managerial preferences over total cost and throughput. Optimizing one objective may lead to compromise on the optimality of the other objective, which generates trade-offs. Using a trade-off balancing model, we found solutions that minimize the sum of deviations from the best solutions for both total cost and throughput. Trade-off balancing optimization models may lead to better solutions, compared to traditional multi-objective optimization models. The results of this paper are applicable to manufacturing systems, where managers face multiple objectives and uncertainties in the system.

Peer-review under responsibility of the scientific committee of the 46th SME North American Manufacturing Research Conference.

Keywords: Operating room; planning; newsvendor model; chance-constrained optimization; trade-off balancing

1. Introduction

Healthcare is one of the most important sectors of the economy; historical data show a continuous increasing trend for the healthcare expenditures in recent decades. In 2009, the U.S. healthcare expenditures exceeded 17% of the gross domestic product (GDP), but was only 4.6% of the GDP in 1950 [1]. Experts in economy, healthcare policy and public finance believe that the healthcare expenditures control is the main

Peer-review under responsibility of the scientific committee of the 46th SME North American Manufacturing Research Conference.
10.1016/j.promfg.2018.07.008

component of the deficit reduction challenge facing the U.S. government [2,3]. Hospitals, physicians and drugs are the primary expenditures. Among these main categories, hospital cost are the largest part of the healthcare expenditures [1].

As a result of increasing costs, hospitals need to leverage their resources by using them more efficiently. Within a hospital, the operating room (OR) department is one of the most critical resources, which has the largest cost and revenue [4,5,6]. Because of the aging population, the demand for surgical services has been increasing sharply in recent years [7,8]. Therefore, efficient OR management has the potential of offering a significant cost saving. To efficiently utilize ORs, hospitals must provide high quality care more effectively with limited resources by developing efficient OR schedules [4,9].

OR planning is challenging because it is under the continuous pressure of competing objectives, such as cost minimization, waiting time minimization, etc. There are numerous affecting factors and various active players in an OR department. Patients, surgeons and OR managers are some of the OR active players who might have competing/conflicting objectives, with respect to cost, waiting times, etc. A large variety of performance measures are used to evaluate the OR planning, such as throughput, waiting lists, utilization, total cost, etc. The choice among these objectives is challenging and complex, because of multiple stakeholders (i.e. patients, surgeons, OR managers, etc.) with different incentives and priorities [4]. Therefore, any decision on one objective may generate trade-offs on the other objectives.

Waiting time, which is defined as the time between the referral date and the surgery date, is of particular importance for patients [4]. In general, patients prefer to get on schedule as soon as possible. Long waiting times may negatively affect the patients' health condition and consequently decrease the quality of care and patient's satisfaction. On the other hand, deteriorated health condition may increase the cost of required care, which is not desirable for the patients or for the healthcare providers and insurance companies [4,10,11,12]. Throughput, which is defined as the number of patients treated in a period of time, is of particular importance for surgeons. Surgeons prefer to perform as many surgeries as possible in their assigned OR times. In general, because of educational and research workloads, surgeons are available on limited hours/days. Therefore, any idletime is not desirable for them [13,14,15].

The dependency between waiting time and throughput is clearly described by equation (1) which is known as *Little's Law* [4,16]. The average work in process (L) in the system equals average arrivals (λ) to the system multiplied by the average cycle-time (W).

$$L = \lambda W \tag{1}$$

In the OR planning context, L can be interpreted as the number of patients on the waiting list, λ as the throughput and W as the summation of waiting time and case time. Therefore, by increasing throughput, the waiting time indirectly decreases [4].

OR utilization is of importance for OR managers. OR utilization measures the proportion of potential output that is actually realized. OR utilization is a very important operational metric, because it provides insight to the existing slack in the system. An OR department with utilization less than 100%, 'theoretically' has the potential to increase the production without generating overhead costs associated with capacity expansion. OR utilization is also a very effective metric to illuminate the cost structure of the OR department, by defining underutilization (idletime) costs and overutilization (overtime) costs. OR utilization is one the most extensively studied OR performance measures. According to the literature, the OR utilization should be maximized to avoid underutilization (idletime) costs. But due to the high variations in case times and patients' arrivals, highly utilized ORs are unstable [4,17].

In this paper, two performance criteria, throughput (*TP*) and total cost (*TC*) are taken into consideration. These performance measures are of importance to three main stakeholders (i.e. patients, surgeons and OR managers) and each stakeholder must be 'adequately' satisfied.

The stochastic nature of the process is another challenging factor for OR planning. There are many sources of uncertainty, such as variations in patients' arrivals (no-shows, emergencies), variations in case times (surgery duration), etc. which may negatively affect the OR department performance. Uncertainty is an inherent feature of surgical procedures. There are two types of well-defined uncertainties in the OR planning literatures: (1) uncertainty in the case times, which is the difference between expected and actual surgery duration and (2) uncertainty in the patients' arrivals caused by emergency arrivals and patients no-show cases [4,18]. A large body of research has been done to tackle the uncertainty in case times [11,12,13,19,20]. On the other hand, there are a few works addressing the uncertainty in the arrival rate [12,14,21,22,23]. There are fewer works, if any, considering uncertainties in both case times and patients' arrivals at the same time.

In this paper, we propose a model which takes both sources of uncertainties into consideration. Without loss of generality, we assume that the case times and patients' arrivals are normally distributed. Using joint distribution of case times and patients' arrivals, we provide theoretical properties for OR department cost function (consisting of overtime and idletime). Using newsvendor model, we minimize the mismatch between expected and actual cost. Then, we utilize a bi-level chance constrained optimization model to optimize *TC* and *TP*. To this aim, we alternate the order of objectives to show the trade-offs generated by the competing objectives. Finally, we propose a trade-off balancing model, to show the effectiveness and efficiency of trade-offs balancing over traditional optimization models.

The main contributions of this paper are (i) a stochastic model, which explicitly takes uncertainties into account in both case times and patient arrivals, (ii) bi-level optimization models, in which the order of objectives is alternated to show the trade-off among objectives, and (iii) a trade-off balancing model, which balances the trade-offs between competing objectives of *TP* and *TC*. This research is unique because it provides a flexible tool for OR managers to perform OR planning more efficiently, by avoiding excessive overtime/idletime cost and long waiting lists.

2. Method

This section first presents a brief introduction to the newsvendor model and the chance-constrained optimization method. Next, two bi-level chance constrained models for OR planning problem under the presence of uncertainties in case times and patients' arrivals are proposed. A trade-off balancing model is also presented at the end of this section.

2.1. Newsvendor model

Newsvendor model is a mathematical model, used to determine optimal inventory levels, subjected to fixed cost ratios (with C_o for overage cost and C_s for Shortage cost, and $C_o, C_s > 0$) where demand is under normal distribution $D \sim \mathcal{N}(\mu_D, \sigma_D^2)$. Before the realization of D, the decision maker has to make a decision Q (inventory level). Minimizing the mismatch between Q and D is the objective of newsvendor model. If $Q > D$, the overage cost occurs, which is $C_o(\max(0, Q - D))$. If $Q < D$, the shortage cost occurs, which is $C_s(\max(0, D - Q))$. The optima Q^* minimizes the $\mathbb{E}\{C_o(\max(0, Q - D)) + C_s(\max(0, D - Q))\}$. Assuming unconstrained problem and taking convexity of objective function in Q into consideration, the optimal solution can be derived by the first order condition [24,25]. Therefore, $F(Q^*) = \Phi(z) = \frac{C_u}{C_o + C_s}$, where $z = \frac{Q^* - \mu_D}{\sigma_D}$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Q^* is explicitly presented by $Q^* = \mu_D + z\sigma_D$.

2.2. Chance-constrained optimization

Chance-constrained optimization method is one of the approaches to solve optimization models in the presence of uncertainty. The basic idea is to ensure that the probability of meeting certain constraints is above a predetermined level [26,27]. In other word chance-constrained model restricts the solution feasible region to achieve higher confidence level for the solution. The general optimization model under uncertainty can be formulated as follows:

$$\min f(x, \xi) \tag{2}$$

s.t.

$$g(x, \xi) = 0 \tag{3}$$

$$h(x, \xi) \geq 0 \tag{4}$$

Equation (2) describes the objective function, equation (3) describes the equality constraints and equation (4) describes the inequality constraints. x is the decision variables vector and ξ is the uncertainties vector. Using chance-constrained method the inequality constraints can be formulated as $Pr(h(x, \xi) \geq 0) \leq \alpha$, $\alpha \in [0,1]$ is the predetermined probability level.

2.3. Bi-level optimization models

2.3.1. Total cost to Throughput (C2P)

We propose a bi-level optimization model that at the first level minimizes *TC*, and then at the second level maximizes *TP* subjected to the cost constraints imposed by the first level optimization.

I) *TC*

Without loss of generality, we assume that the patients randomly arrive to the OR department via a normal distribution $d \sim \mathcal{N}(\mu_d, \sigma_d^2)$. This assumption is very common in the literature and it fits the actual data when the number of arrivals is large enough. Case times are assumed to be independent, identically distributed (i.i.d) random variables with normal distribution $p \sim \mathcal{N}(\mu_p, \sigma_p^2)$. This assumption is also common in the literature and it fits the actual data when the patients' population is large enough. We assume that set-up and clean-up times are factored into the case time p. In this paper, we deal with the strategic OR planning problem which is a long-term one. Therefore, the assumption of having a large population of patients holds. Assuming that the case times and patients' arrivals are i.i.d, we define a new random variable named workload (l) as the product of patients' arrival d and case times p. It is worth noting that the product of two normal distributions is not always a normal distribution. But, under some conditions the product can be approximated to a normal distribution. Particularly, for two normal distributions with different mean ($\mu_x \neq \mu_y$) and different variance ($\sigma_x^2 \neq \sigma_y^2$), as the inverse variation coefficient $\delta = \frac{\mu}{\sigma}$ increases ($\delta > 1$), the distribution of the product of two independent normal variables tends to a normal distribution [28].

Using approximation formulas proposed by Macias and Oliviera [28] we can compute mean and variance of l by equation (5) and equation (6) respectively.

$$\mu_l = \mu_d \mu_p \tag{5}$$

$$\sigma_l^2 = \mu_d^2 \sigma_p^2 + \mu_p^2 \sigma_d^2 + \sigma_d^2 \sigma_p^2 \tag{6}$$

Now we are able to utilize the newsvendor model to obtain the optimum workload for the planning horizon (T), minimizing the TC. To translate cost factors C_s and C_o into the OR planning context, we argue as follows: shortage cost occurs in OR department when the planned capacity (workload) is less than the actual realized workload. Therefore, a fraction of actual workload must be done in overtime (overtime=max(0, Actual workload - Planned workload). With this argument, shortage cost of the newsvendor model is an equivalent for overtime in ORs. On the other hand overage cost occurs in ORs when the planned capacity is greater than the actual realized workload and a fraction of the planned capacity sits idle (idletime=max(0, Planned workload – Actual workload). Therefore, the overage cost of the newsvendor model is an equivalent for idletime in ORs. Based on what was discussed above, to drive out optimal planned capacity (B^*) in the time period of T, we can define the expected cost by equation (7). Let $g(\cdot)$ and $G(\cdot)$ be the density and cumulative distribution functions of l.

$$Y(B) = C_o \int_0^B (B-l) g(l) dl + C_s \int_B^\infty (l-B) g(l) dl \tag{7}$$

Because equation (7) is a convex function in B, by applying the first derivative condition we can derive out the optimal planned capacity B^*. Applying *Leibnize rule* [29,30,31] for differentiation under the integral sign with respect to B and setting it equal to zero, it yields:

$$\begin{aligned} \frac{dY(B)}{dB} &= C_o \int_0^B 1.g(l)dl + C_s \int_B^\infty (-1) g(l) dl \\ &= C_o G(B) - C_s[1 - G(B)] = 0 \rightarrow \\ G(B^*) &= \frac{C_s}{C_o + C_s} \end{aligned} \tag{8}$$

$$G(B^*) = \Phi(z) = \Phi\left(\frac{B^* - \mu_l}{\sigma_l}\right) = \frac{C_s}{C_o + C_s} \tag{9}$$

$$B^* = \mu_l + z\sigma_l \tag{10}$$

$G(B^*)$ represents the probability of workload being less than or equal to B^* ($Pr(l \leq B^*) = G(B^*)$). In other words, the probability of having enough capacity to meet l is $\frac{C_s}{C_o+C_s}$. Another interesting implication of equation (10) is that for the normal case, B^* is an increasing function of μ_l and σ_l, provided that the z is positive (because C_s and C_o are strictly positive). Considering this fact that the cost of overtime hours is always greater (or equal) than the cost of idletime hours, we can conclude that $\frac{C_s}{C_o+C_s} \geq 0.5$. Therefore, we should allocate more capacity to avoid overtime (shortage cost). B^* is then imposed as the capacity constraints into the second level optimization model, which maximizes throughput (*TP*).

II) *TP*

In order to maximize *TP*, the OR manager can estimate the expected case times based on historical data and surgeon estimation. The OR manager can use $n\bar{p} \leq B^*$ to derive the number of patients to be planned. Where $\bar{p}$ is the OR manager's estimation for the case times and n denotes the number of patients to be planned. The drawback of this simple procedure is that it ignores the variability in the case times and patients' arrivals and it doesn't provide any insight to the probability of expected overtime levels. To provide a guarantee on the expected overtime, stochastic constraints must be imposed to the objective function to capture the uncertainties inherent to the surgical procedures. By letting α ($\alpha \in [0,1]$) to be the probability of overtime exceeding a threshold (tolerance on overtime), denoted by TL, we can formulate the probabilistic constraints by equation (11).

$$Pr\left\{\left(\sum_{i=1}^{n} p_i - B^*\right) > TL\right\} \leq \alpha \tag{11}$$

While avoiding the overtime, OR managers want to minimize the idletime to treat more patients in a given time period. Therefore, we can formulate the nonlinear optimization model for *TP* as follows:

$$\min\left\{\mathbb{E}[\max(0,(B^* - \sum_{i=1}^{n} p_i))]\right\} \tag{12}$$

s.t.

$$Pr\left\{(\sum_{i=1}^{n} p_i - B^*) > TL\right\} \leq \alpha \tag{13}$$

Equation (12) is the objective function, which minimizes the expected idletime; this implies that the objective function indirectly maximizes *TP* by packing more patients into B^*. Equation (13) is the chance constraints, which guarantees that the overtime does not exceed an acceptable level of *TL*. *TL* and α are two managerial preferences, by which the OR manager can balance the waiting list. If managers experience an increasing waiting list by adjusting a larger value for *TL*. They can manage the waiting list, but it is important to consider resource availabilities (e.g. available budget, staff availability, etc.).

Case times are assumed to be independent, identically distributed (i.i.d) random variables with normal distribution $p \sim \mathcal{N}(\mu_p, \sigma_p^2)$. Therefore, overtime which is defined by $Ot = \max(0, \sum_{i=1}^{n} p_i - B^*)$, is also a normally distributed random variable $Ot \sim \mathcal{N}(\mu_{Ot}, \sigma_{Ot}^2)$, where $\mu_{Ot} = n\mu_p$ and $\sigma_{Ot}^2 = n\sigma_p^2$. The chance-constrained optimization model can be approximated to its nonlinear deterministic counterpart as follows [27]:

$$\min\left\{\mathbb{E}[\max(0,(B^* - \sum_{i=1}^{n} p_i))]\right\} \tag{14}$$

s.t.

$$\mu_{Ot} + \Phi^{-1}(1-\alpha)\sigma_{Ot} \leq TL \tag{15}$$

$$n\mu_p \leq B^* \tag{16}$$

Equation (15) guarantees overtime does not exceed an acceptable level (*TL*) with probability of α. Where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal variable. equation (16) imposes the notion that the long term performance must converge to the expected one.

Fig.1.a represents the relationship among number of planned patients, ratio of overtime threshold to planned capacity, and probability level α. This relationship provides the OR managers with a powerful but flexible tool to manage their OR department according to their preference over the acceptable overtime threshold and the associated risk of non-realization. By packing more patients into B^* the risk of overtime increases (as we would intuitively expect).

Given *TL* and α, the final output of this bi-level optimization is an ordered pair (B_1^*, n_1^*), which specifies the optimum planned capacity and optimum number of patients to be planned.

2.3.2. Throughput to Total cost (P2C)

I) TP

In order to show the trade-offs generated by the competing objectives, we alternate the order of objectives in our bi-level optimization models. If the OR managers' preference is to meet the demand by a predetermined confidence level β, they must first find the optimum number of patients to be planned and its associated risk of non-realization. Afterwards, they have to find the required capacity, which minimizes the total cost generated by overtime and idletime. In order to do this, we formulate the first level optimization model maximizing *TP* as follows:

$$\min\{\mathbb{E}[\max(0, d-n)]\} \tag{17}$$

s.t.

$$Pr\{(d-n) > \gamma n\} \leq \beta \tag{18}$$

Where, *d* is the actual number of arrivals, *n* is the number of patients to be planned, γ is the acceptable threshold for number patients more than *n*, γ is proportional to *n*, and β is the confidence level. Patients' arrivals are assumed to be independent, identically distributed (i.i.d) random variables with normal distribution $d \sim \mathcal{N}(\mu_d, \sigma_d^2)$. The chance-constrained optimization model can be rewritten to its nonlinear deterministic counterpart as follows [27]:

$$\min\{\mathbb{E}[\max(0, d-n)]\} \tag{20}$$

s.t.

$$(1+\gamma)n \geq \mu_d + \Phi^{-1}(1-\beta)\sigma_d \tag{21}$$

Equation (20) is the objective function, which minimizes $d - n$, the difference between actual number of arrivals and the planned number of patients. Equation (21) imposes that $d - n$ does not exceed a predetermined level with the probability of β. Figure 1.b. shows the relationship among *n*, γ and β.

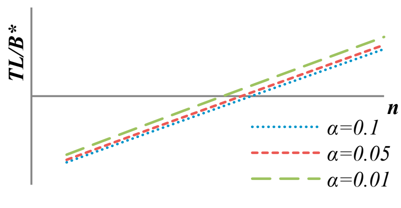

Fig.1.a. Overtime threshold to planned capacity ratio vs. number of planned patients

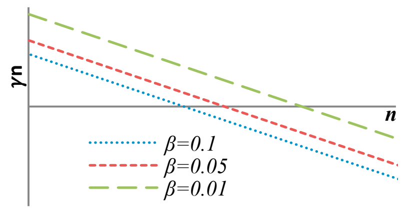

Fig.1.b. γn threshold vs. *n*

II) TC

Let n^* be the number of patients to be planned, obtained from the first level optimization, at the second level of the optimization model, the objective is to minimize total cost (*TC*) generated by overtime and idletime, imposed by the constraints of treating n^* patients. We again use the newsvendor model, to find to optimum capacity, which minimizes *TC*. We define the workload (*l*), which is a random variable as the product of n^* and case times. Case times are assumed to be normally distributed random variables, therefore *l* is also a normally distributed random variable. Using standard formula for the product of a real number and normally distributed random variables, we can compute mean and variance of *l* by $\mu_l = n^*\mu_p$ and $\sigma_l^2 = n^*\sigma_p^2$ respectively. We utilize the newsvendor model to obtain the optimum workload for the planning horizon (*T*) minimizing the *TC*, yielding: $G(B^*) = \Phi(z) = \Phi\left(\frac{B^*-\mu_l}{\sigma_l}\right) = \frac{C_s}{C_o+C_s}, \rightarrow B^* = \mu_l + z\sigma_l$. Given γ and β, the final output of this bi-level optimization is an ordered pair (B_2^*, n_2^*), which specifies the optimum planned capacity and optimum number of patients to be planned.

2.4. Trade-off balancing model

Intuitively, when we optimize the OR planning problem based on different orders of objectives, it is very likely that the OR plan performs poorly with regard to the second level objective. In other words, the result of the bi-level optimization model is the global optima with regard to the first level objective, whereas it is local optima with regard to the second level objective. Therefore, alternating the order of objectives generates trade-offs in the OR plans performance.

We utilize a simulation-based trade-off balancing model to minimize the trade-offs generated by alternating the order of objectives. Let $j \in \{1,2\}$ denote the orders of objectives, where, 1 represents *TC→TP* and 2 represents *TP→TC*. The trade-offs balancing model can be formulated as follows:

$$\min(\theta \frac{Best_{TP} - TP_j}{|Best_{TP} - Worst_{TP}|} + (1-\theta)\frac{TC_j - Best_{TC}}{|Best_{TC} - Worst_{TC}|}) \quad (22)$$

Where $0 \leq \theta \leq 1$ is the managerial weight assigned to *TP* and intuitively $(1-\theta)$ to *TC*, setting these weights is a subjective decision, hence difficult to argue. Equation (22) expresses a type of normalized objective function to tackle the fact that each objective is expressed in different units and different granularity.

3. Case study

To analyze the efficiency of our proposed models, we established a simulation model using historical data provided by UKHealthcare (University of Kentucky healthcare). UKHealthcare hospitals perform a wide variety of surgery procedures and on average they treat more than 30,000 surgery cases per year.

Different scenarios for different combination of managerial preferences over the order of objectives, α, β, TL and γ were designed. Different performance measures including throughput, total cost ($TC = C_o * idletime + C_s * overtime$), idletime and overtime were computed to compare the performance of proposed models.

3.1. Data generation

Normality tests for patients' arrivals and case times distributions were done using UKHealthcare historical data of a certain surgery group. Anderson-Darling normality test shows a p value of 0.3355 and 0.166 for patient arrivals and case times respectively ($p \geq 0.05$ for both), and the data plots form a fairly straight line along the fitted line. Therefore, it appears that the normal distribution is a good fit to the data set. Detailed results of the normality tests can be found in Appendix .A by table 4 and figure 7.

To generate random data, we find the maximum and minimum value of case time (and patients' arrivals as well), then using $R = max - min$, we calculate the range of data. R then was discretized to 5 equal increments and the probability of each increment was calculated. Assuming *T* (planning horizon) as equal to one week, using Monte Carlo simulation, patients' arrivals and associated case times were generated for 50 weeks. 50 replications for each week were done and the data stored to run the simulation model.

3.2. Cost ratios

In this case study, (without loss of generality), we assume that $C_s = 2C_o$. This implies that the cost of overtime is twice as the cost of idletime; in practice these cost ratios may vary place to place. Therefore, in our case $\Phi(z) = \frac{C_s}{C_o+C_s} = 0.6667 \rightarrow z = 0.4307$.

3.3. Managerial preferences

I) C2P

In this case study, we consider $\alpha \in \{0.01, 0.05, 0.10, 0.20, 0.30\}$, $TL \in \{0.1B^*, 0.2B^*, 0.3B^*\}$, therefore, we have 15 different combinations of confidence levels and acceptable overtime thresholds. Using equation (5), equation (6), equation (10) and historical data ($\mu_d = 72$, $\sigma_d = 33$, $\mu_p = 156$, $\sigma_p = 60$), we obtain $B^* = 13194.06$ minute. Therefore, using equation (14) through equation (16) we obtain n^*. Table.1 shows n^* for different combinations of α and *TL*. For the sake of brevity, a code is assigned to each combination of α and *TL*, shown in table.1 (e.g. *C2P*-1 for *TL*=0.1 and α=0.01). It is worth mentioning that n^* is the same for those combinations with $\alpha > 0.01$ or $TL > 0.1B^*$, because equation (16) sets an upper bound on the number of patients to ensure that the long term performance converges to the expected one.

Table 1. (B^*, n^*) of different combinations of different α and *TL*

	TL		
α	**0.1**	**0.2**	**0.3**
0.01	*C2P*-1 (13194, 84)	*C2P*-6 (13194,87)	*C2P*-10 (13194,87)
0.05	*C2P*-2 (13194,87)	*C2P*-7 (13194,87)	*C2P*-11 (13194,87)
0.10	*C2P*-3 (13194,87)	*C2P*-8 (13194,87)	*C2P*-12 (13194,87)
0.20	*C2P*-4 (13194,87)	*C2P*-9 (13194,87)	*C2P*-13 (13194,87)
0.30	*C2P*-5 (13194,87)	*C2P*-10 (13194,87)	*C2P*-15 (13194,87)

II) P2C

We consider $\beta \in \{0.01, 0.05, 0.10, 0.20\}$ and $\gamma \in \{0.10n^*, 0.20n^*, 0.30n^*\}$; therefore, we have 15 combinations of different β and γ. Table.2 shows (B^*, n^*) of these combinations. A code is assigned to each combination of β and γ, shown in table.2 (e.g. *P2C*-1 for γ=0.1 and β=0.01).

Table 2. (B^*, n^*) of different combinations of different β and γ

	γ		
β	**0.1**	**0.2**	**0.3**
0.01	*P2C*-1 (22008, 144)	*P2C*-6 (20948,137)	*P2C*-10 (20190,132)
0.05	*P2C*-2 (19130,125)	*P2C*-7 (18372,120)	*P2C*-11 (17766,116)
0.10	*P2C*-3 (17766,116)	*P2C*-8 (17160,112)	*P2C*-12 (16553,108)
0.20	*P2C*-4 (16098,105)	*P2C*-9 (15643,102)	*P2C*-13 (15188,99)
0.30	*P2C*-5 (15491,101)	*P2C*-10 (15036,98)	*P2C*-15 (14733,96)

3.4. Performance measures

Several performance measures including throughput, total cost, idletime and overtime were computed to compare the performance of the proposed models.

4. Results

Statistical process control (SPC) techniques are used to monitor the performance of the proposed models. Xbar-R charts and process capability analyses are used to compare the quality of proposed models. Process capability index c_p is an indicator, representing if the outcomes of a process are within the user-defined specification limits, where $c_p = \frac{USL-LSL}{6\sigma}$, and *USL* and *LSL* are user-defined upper specification limit and lower specification limit respectively. The larger the c_p, the less variations in the process. Process capability index c_{pk} represents the congestion of outcomes around the center line, the larger the c_{pk}, the more congestion of outcomes around the center line [32,33]. *C2P*-1 has the best performance on idletime, overtime, throughput and total cost among all combinations of *C2P*. As mentioned earlier, equation (16) set an upper bound on the number of patients to ensure that the long term performance converges to the expected one, therefore, all combinations of *C2P*-2 to *C2P*-15 have the same performance, and are not sensitive to managerial preferences. Consequently, hereafter we only review the behavior of *C2P*-1.

4.1. Idletime

Idletime represents the unproductive time of ORs. In our case study, idletime is the difference between the actual workload and the planned capacity ($idletime = \max(0, B^* - \sum_{i=1}^{n} p_i)$). OR managers want to minimize the amount of idletime. For the sake of brevity, we present the best scenario of each model. Average idletime for *C2P* and *P2C*-15 equals 3907 and 6163 minutes, respectively. *C2P*-1 outperforms the best *P2C* by $\frac{6163-3907}{6163} = 36.6\%$.

Process capability analyses results for idletime show that the c_p is 0.30 for *C2P*-1 and 0.32 for *P2C*-15; this implies that the *P2C*-15 is more stable, because on average it generates a larger amount of idletime compared to *C2P*. For *C2P*-1 a big proportion of experiments (almost 1/3) does not generate any idletime. c_{pk} is 0.27 for *C2P*-1 and 0.08 for *P2C*-15, which implies that the outcomes of *C2P*-1 are more centered around the center line of specification limits. Figure 2 shows the SPC results for idletime.

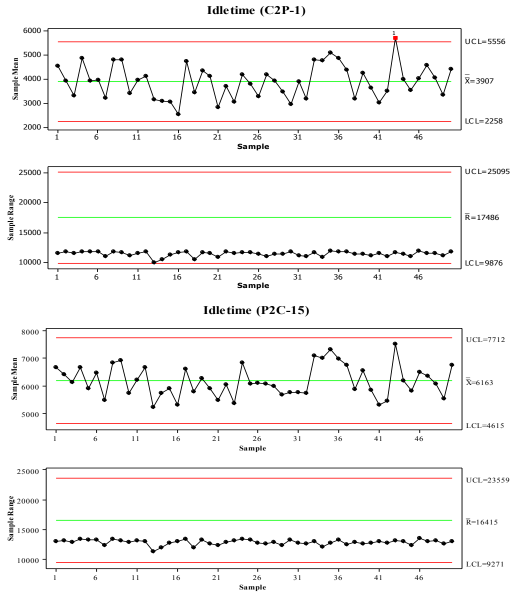

Fig.2. SPC results for Idletime

4.2. Overtime

Overtime represents the amount of time spent after regular hours to treat all patients. In our case study, overtime is the difference between the planned capacity and the actual workload ($Overtime = \max(0, \sum_{i=1}^{n} p_i - B^*)$). OR managers want to minimize the amount of overtime. *P2C*-15 has the worst performance on generating overtime among all combinations of β and γ (i.e. generating the largest amount of overtime). Therefore, we compare the performance of *C2P*-1 with the performance of *P2C*-15. Average overtime is 1030 minutes for *C2P*-1 and 582 minutes for *P2C*-15. Therefore, *P2C*-15 outperforms *C2P*-1 by $\frac{1030-582}{1030} = 43.50\%$.
Process capability analyses results for overtime show that the c_p is 0.09 for *C2P*-1 and 0.14 for *P2C*-15; this implies that *P2C*-15 generates less variations in overtime. Figure 3 shows the SPC results for overtime.

4.3. Total cost

Total cost is defined by $TC = C_o * idletime + C_s * overtime$. This formula helps us to find the trade-off between overtime and idletime costs. Using newsvendor model, we minimize this trade-off. *P2C*-15 has the best performance on total cost among all combinations of β and γ. Therefore, we compare the performance of *C2P* with the performance of *P2C*-15. *C2P*-1 outperforms *P2C*-15 by $\frac{6163.36-5967.16}{6163.36} = 3.18\%$. Since this percentage is not a large number, an ANOVA was performed to determine whether the difference between *C2P*-1 and *P2C*-15 is significant.

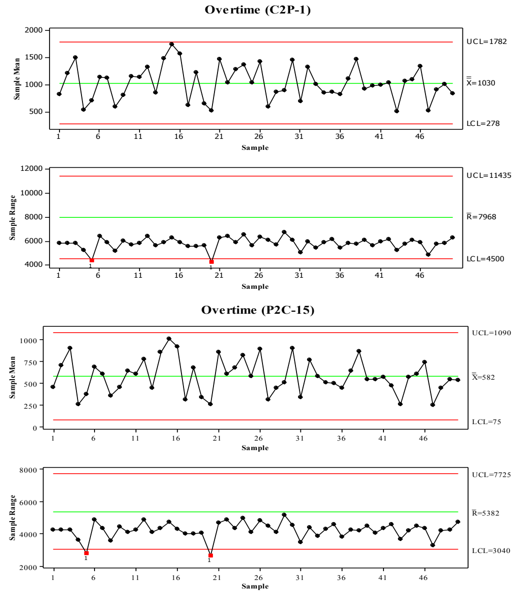

Fig.3. SPC results for overtime

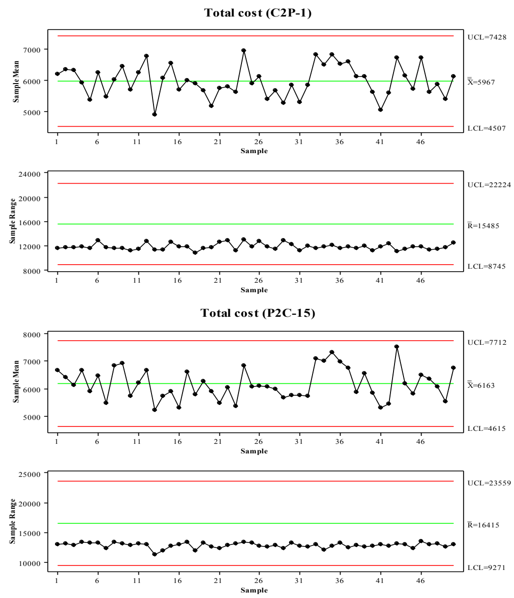

Fig.4 SPC results for Total cost

The ANOVA results demonstrate that the difference between total cost of C2P-1 and P2C-15 is statistically significant at 95% confidence level (p<0.05). Process capability analysis shows that C2P-1 has lower mean and less variations compared to P2C-15. Detailed results of SPC and ANOVA for total cost can be found in table.5 in Appendix. A. Figure 4 shows the SPC results for total cost.

4.4. Throughput

Throughput is defined as the number of treated patients in a given period of time. In this case study, we assumed that all arriving patients must be treated. This assumption implies two points: first, the number of treated patients is the same for all scenarios and second, scenarios throughputs differ from each other only by the number of patients treated in the overtime. Therefore, we use the number of patients treated in overtime as a performance indicator, to compare the throughput of different scenarios.

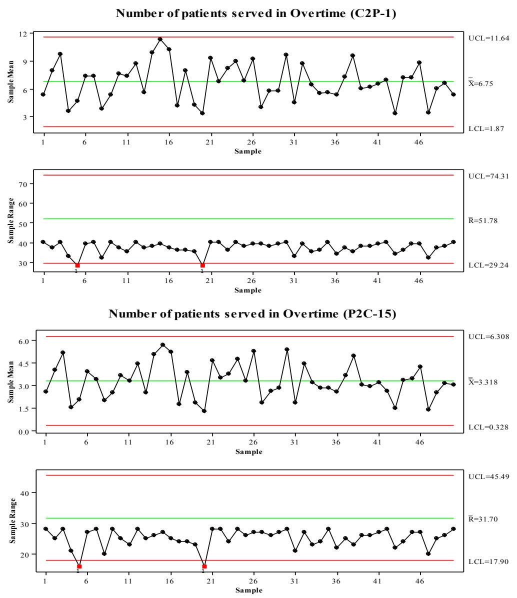

Fig.5 SPC results for Number of patients served in overtime

As we showed in previous section, *P2C*-15 has the lowest total cost among all combinations of *P2C* (all combinations of β and γ). It also has the worst performance on the number of patients served in overtime among all combinations of β and γ. Therefore, we compare the performance of *C2P*-1 with the performance of *P2C*-15. *P2C*-15 outperforms *C2P*-1 by $\frac{6.75-3.318}{6.75} = 50.84\%$. Process capability analysis shows that c_p is 0.29 for *C2P*-1 and 0.47 for *P2C*-15, which implies that *P2C*-15 has less variations on the number of served patients in overtime. Figure 5 shows the SPC results for throughput (number of patients served in overtime).

4.5. Trade-off balancing

As we showed in earlier sections, *TC* and *TP* are inconsistent objectives, which imply that optimizing one may lead to compromise on the other one. Therefore, a model that has the least deviations from both objectives may be of interest to OR managers. We define $\theta \in \{0, 0.25, 0.50, 0.75, 1\}, and\ 0 \leq \theta \leq 1$, as the set of OR managers' preference for *TP* over *TC*.
Using equation (22) and simulation results for best/worst performance, we find the most efficient combination minimizing deviation from best performance. We have 80 different combinations of scenarios and different θ, shown by Table.3. The absolute values of best and worst performances are given by Table.6 in Appendix. A.
As we intuitively expected, *C2P*-1 has the least deviation from the best performance on *TC* (θ=0). *P2C*-3 has the best performance, where OR managers have equal preferences over both objectives (θ=0.50). Several combinations of *P2C* have the best performance on *TP* (θ=1), as we intuitively expected. Figure 6 shows the sum of average deviations from the best *TC* and *TP*.

Table 3. Sum of average deviations from best *TC* and *TP*

	θ=0	θ=0.25	θ=0.50	θ=0.75	θ=1.0
C2P-1	**0.2882 (Best)**	0.2552	0.2222	**0.1892 (Worst)**	**0.1562 (Worst)**
P2C-1	**0.5637 (Worst)**	0.4228	**0.2819 (Worst)**	0.14094	**0 (Best)**
P2C-2	0.4253	0.3190	0.2127	0.1063	**0 (Best)**
P2C-3	0.3689	0.2784	**0.0973 (Best)**	**0.0187 (Best)**	0.0068
P2C-4	0.3202	0.2503	0.1804	0.1105	0.0407
P2C-5	0.3082	0.2461	0.1840	0.1220	0.0599
P2C-6	0.5126	0.3845	0.2563	0.1281	**0 (Best)**
P2C-7	0.3922	0.2945	0.1968	0.0991	0.0015
P2C-8	0.3486	0.2653	0.1821	0.0989	0.0156
P2C-9	0.3109	0.2468	0.1828	0.1188	0.0548
P2C-10	0.3011	**0.2451 (Best)**	0.1891	0.1331	0.0772
P2C-11	0.3571	**0.4761 (Worst)**	0.2381	0.1190	**0 (Best)**
P2C-12	0.3689	0.2784	0.1879	0.0973	0.0068
P2C-13	0.3314	0.2557	0.1800	0.1043	0.0286
P2C-14	0.3033	0.2453	0.1873	0.1292	0.0712
P2C-15	0.2972	0.2453	0.1934	0.1416	0.0897

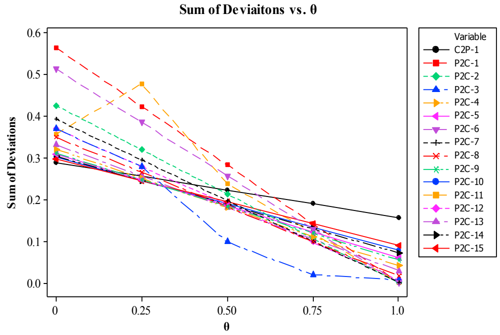

Fig.6 Sum of average deviations from the best *TC* and *TP*

5. Conclusion

In this paper, we present a bi-level optimization model for OR planning, which captures uncertainties in patients' arrivals and case times. Minimizing total cost (*TC*) and maximizing throughput (*TP*) are selected as objectives of the optimization models. Different managerial preferences over objectives are taken into account. By alternating the orders of objectives in our bi-level optimization model, we show that minimizing total cost and maximizing throughput are inconsistent. We propose a simulation-based trade-off balancing model to minimize the sum of deviations from best performance on each objective. Using historical data obtained from UKHealthcare, a large set of computational experiments are carried out. The simulation results show that for *TC* our proposed *C2P* has the best performance, and *P2C* has the best performance on *TP*. The simulation results for trade-off balancing support the inconsistency between objectives. These results provide a flexible tool for OR managers to perform OR planning more efficiently based on their preferences and settings. The results of the study are applicable to manufacturing systems with multiple objectives under the presence of variations in demand and processing times.

Our future research focus on adaptive control for OR planning. This aims to integrate time series into the OR planning process in order to update the OR plans based on the past performance and prediction of future desirable performance.

Acknowledgements

This project was supported by grant number R03HS024633 from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. We also appreciate the support from Department of Mechanical Engineering at University of Kentucky, Lexington, KY, USA.

Appendix A. Supplementary results

A.1. Anderson-Darling normality test results for patients' arrivals and case times.

Table.4 Statistics obtained from UKHealthcare data

	Mean	Standard deviation
Case times	156	60
Patient arrivals	72	33

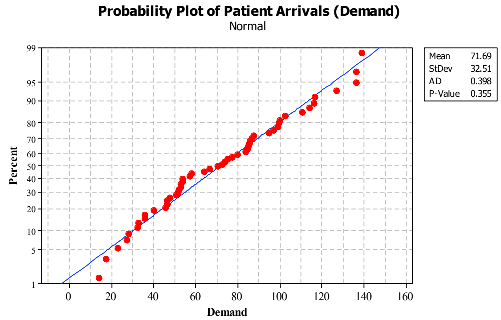

Fig.7a. probability plot of Demand

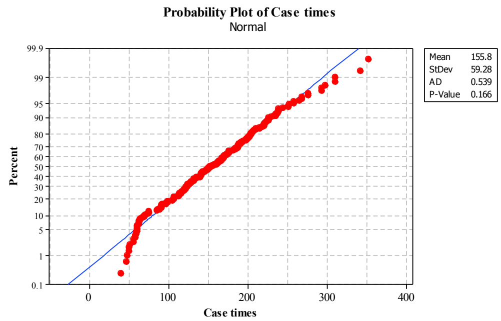

Fig.7.b. probability plot of case times

A.2. ANOVA and SPC results for Total cost

Table 5. ANOVA results for total cost between *C2P*-1 and *P2C*-15

Source	DF	SS	MS	P
Factor	1	48119225	48119225	0.041
Error	4998	63045301205	12614106	
Total	4999	63093420430		

A.3. Best and worst performances

Table 6. Best-Worst performance among all combinations

	Best	Worst	Best-worst
Cost	0.04	20739	20738.96
Number of patients served in Overtime	0	37	37

References

[1] V.R. Fuchs. Major trends in the US health economy since 1950. New England Journal of Medicine, 366(11), (2012) 973–977.

[2] A. S. Blinder. Four deficit myths and a frightening fact. The Wall Street Journal, (2012) page A14.

[3] R. Kogan. The myth of the out-of-control federal government. Off the Charts Blog, Center on Budget and Policy Priorities, (2012).

[4] B. Cardoen, E. Demeulemeester, J. Beliën. Operating room planning and scheduling: A literature review. European journal of operational research, 201(3), (2010) 921–932,.

[5] A. Abedini, W. Li, H. Ye. An optimization model for operating room scheduling to reduce blocking across the perioperative process. Procedia Manufacturing, 10 (2017) 60–70.

[6] C. McIntosh, F. Dexter, R. H. Epstein. The impact of service-specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: a tutorial using data from an Australian hospital. Anesthesia & Analgesia, 103(6), (2006) 1499–1516.

[7] F. Guerriero, R. Guido. Operational research in the management of the operating theatre: a survey. Health care management science, 14(1), (2011) 89–114.

[8] A. Abedini, H. Ye, W. Li. Operating room planning under surgery type and priority constraints. Procedia Manufacturing, 5 (2016) 15–25.

[9] V. R. Gunna, A. Abedini, W. Li. Maximizing operating room performance using portfolio selection. Procedia Manufacturing, 10 (2017) 83–91.

[10] E. Marcon, F. Dexter. Impact of surgical sequencing on post anesthesia care unit staffing. Health Care Management Science, 9(1), (2006) 87–98.

[11] M. Lamiri, X. Xie, A. Dolgui, F. Grimaud. A stochastic model for operating room planning with elective and emergency demand for surgery. European Journal of Operational Research, 185(3), (2008) 1026–1037.

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

[13] S. Chattopadhyay, D.R. Edwards, Y. Yu, A. Hamidisepehr. An Assessment of Climate Change Impacts on Future Water Availability and Droughts in the Kentucky River Basin. Environmental Processes, 4(3), (2017) 477-507.

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

[15] D. Masursky, F. Dexter, S. A. Isaacson, N. A. Nussmeier. Surgeons' and anaesthesiologists' perceptions of turnover times. Anesthesia & Analgesia, 112(2), (2011) 440–444.

[16] J. D. Little. A proof for the queuing formula: L= λ w. Operations research, 9(3), (1961) 383–387.

[17] R. E. Wachtel, F. Dexter. Tactical increases in operating room block time for capacity planning should not be based on utilization. Anesthesia & Analgesia, 106(1), (2008) 215–226.

[18] R. H. Epstein, F. Dexter. Rescheduling of previously cancelled surgical cases does not increase variability in operating room workload when cases are scheduled based on maximizing efficiency of use of operating room time. Anesthesia &Analgesia, 117(4), (2013) 995–1002.

[19] F. Dexter, A. Macario, R. D. Traub. Enterprise-wide patient scheduling information systems to coordinate surgical clinic and operating room scheduling can impair operating room efficiency. Anesthesia & Analgesia, 91(3), (2000) 617–626.

[20] M. Kamaldar, J.B. Hoagg. Multivariable adaptive harmonic steady-state control for rejection of sinusoidal disturbances acting on an unknown system. In American Control Conference (ACC), 2016 (pp. 1631-1636). IEEE.

[21] H. Ye, W. Li, A. Abedini. An improved heuristic for no-wait flow shop to minimize makespan. Journal of manufacturing systems, 44 (part-2), (2017) 273-279.

[22] W. S. Lovejoy, Y. Li. Hospital operating room capacity expansion. Management Science, 48(11), (2002) 1369–1387.

[23] H. Ye, W. Li, A. Abedini, B. Nault. An effective and efficient heuristic for no-wait flow shop production to minimize total completion time. Computers & Industrial Engineering, 108 (2017) 57–69.

[24] M. Olivares, C. Terwiesch, L. Cassorla. Structural estimation of the newsvendor model: an application to reserving operating room time. Management Science, 54(1), (2008) 41-55.

[25] R. E. Wachtel, F. Dexter. Review of behavioral operations experimental studies of newsvendor problems for operating room management. Anesthesia & Analgesia, 110(6), (2010) 1698-1710.

[26] A. Charnes, W. W. Cooper. Chance-constrained programming. Management science, 6(1), (1959) 73-79.

[27] O. V. Shylo, O. A. Prokopyev, A. J. Schaefer. Stochastic operating room scheduling for high-volume specialties under block booking. INFORMS Journal on Computing, 25(4), (2012) 682-692.

[28] A. Seijas-Macías, A. Oliveira. An approach to distribution of the product of two normal variables. Discussiones Mathematicae Probability and Statistics, 32(1-2), (2012) 87-99.

[29] K. Oldham, J. Spanier. The fractional calculus theory and applications of differentiation and integration to arbitrary order Elsevier, (Vol. 111), 1974.

[30] M. Kamaldar, J.B. Hoagg. Adaptive harmonic steady-state control for rejection of sinusoidal disturbances acting on a completely unknown system. International Journal of Adaptive Control and Signal Processing, 31(9), (2017) 1308-1326.

[31] M. Kamaldar, J.B. Hoagg. Time-domain adaptive harmonic control for rejection of sinusoidal disturbances acting on an unknown discrete-time system. In American Control Conference

(ACC), (2017) (pp. 5690-5695). IEEE.
[32] D. C. Montgomery. Introduction to statistical quality control. John Wiley & Sons, New York, 2009.
[33] A. Hamidisepehr, M.P. Sama, A.P Turner, O.O. Wendroth. A Method for Reflectance Index Wavelength Selection from Moisture-Controlled Soil and Crop Residue Samples. Transactions of the ASABE, 60(5), (2017) 1479-1487.