A Mixed Integer Programming Approach for Allocating Operating Room Capacity

Bo Zhang, Pavankumar Murali, Maged Dessouky*, and David Belson

Daniel J. Epstein Department of Industrial and Systems Engineering

University of Southern California

Los Angeles, CA 90089-0193

maged@usc.edu

* corresponding author

ABSTRACT We have developed a methodology for allocating operating room capacity to specialties. Our methodology consists of a finite-horizon mixed integer programming (MIP) model which determines a weekly operating room (OR) allocation template that minimizes inpatients' cost measured as their length of stay. A number of patient type priority (e.g., emergency over inpatient) and clinical constraints (e.g., maximum number of hours allocated to each specialty, surgeon and staff availability) are included in the formulation. The optimal solution from the analytical model is inputted into a simulation model that captures some of the randomness of the processes (e.g., surgery time, demand, arrival time, and no-show rate of the outpatients) and non-linearities (e.g., the MIP assumes proportional allocation of demand satisfaction (output) with room allocation (input)). The simulation model outputs the average length of stay for each specialty and the room utilization. On a case example of a Los Angeles County Hospital, we show how the hospital length of stay pertaining to surgery can be reduced.

KEYWORDS Mixed Integer Programming, Surgery, Operating Room Capacity, Block Time Scheduling, Simulation.

1. Introduction

In the United States, public hospitals are non-profit. One of their operational objectives is to provide medical services to their patients at minimal costs. In particular, inpatients' in-hospital cost accounts for a significant portion of a hospital's expenditures. Thus, it is of great interest to hospital administrators to reduce inpatients' length of stay.

In many hospitals, a large percentage of inpatients undergo some type of surgery when admitted or during their stay in a hospital. To reduce an inpatient's length of stay and also for the sake of the patient's health, it is ideal that a surgery is performed as soon as it is requested or judged to be needed. However, it is common in a hospital that inpatients have to wait in bed for their surgery for a couple of days or sometimes a longer time (especially in public hospitals); also, outpatients may not always undergo their surgery on schedule. There are many possible reasons for this delay. For example, patients who request a surgery at an earlier time or are judged by the doctors to have a more urgent need for a surgery will be served first; or, sometimes too many emergency patients need surgery and use most of the available Operating Room (OR) capacity. As a result, the delayed inpatients stay in the hospital longer, incurring higher costs for the hospital or the government, and those delayed outpatients remain in an usually long outpatient queue (often in the form of a waiting list) waiting for surgery.

Our work focuses on reducing the delay described above, or, in particular, minimizing inpatients' length of stay waiting for their surgery by developing a methodology for allocating OR capacity to different medical specialties. In Section 2, we describe typical surgery planning and scheduling procedures. We introduce an OR capacity planning methodology used in many hospitals, namely, "Block Time Scheduling", which sets the background for our work. We then review the relevant literature in Section 3. In Section 4, we present a mixed integer programming modeling approach for determining a weekly OR allocation template or Block Time Schedule which minimizes inpatients' in-hospital cost measured as their length of stay. Also, we present a simulation modeling methodology to assess the quality of the template generated by the optimization model. In Section 5, we illustrate the use of our models in a case study of Los Angeles County General Hospital. Using their demand and capacity data, we applied our MIP model in determining the allocation template, which was then inputted into the simulation model. The simulation results show that our model yields significant reduction on the length of stay pertaining to surgery.

2. Background

Before describing the typical surgery planning and scheduling procedures, we first introduce the classification or categorization of patients and operating rooms.

In many hospitals, patients are categorized into three *type*s: emergency patients, inpatients, and outpatients, in their patient database and also when the hospital staff analyze and discuss their operational processes (e.g., surgery planning and scheduling). Another classification scheme of patients is by medical *specialty*, like Burns, Cardiac, or Trauma patients. Some specialties have all three types of patients undergoing surgery, but others only have one or two.

Operating rooms can be considered of two types: emergency and non-emergency. A hospital often has only a few emergency ORs and all the others as non-emergency ORs. The emergency ORs are solely allocated to the emergency patients who need surgery, and usually all specialties' surgeries can be performed in that room. The non-emergency ORs are often allotted to different specialties. Though a non-emergency OR assigned to a specialty is intended for its non-emergency surgeries (i.e., inpatient and outpatient surgeries), the room can also accommodate for emergency surgeries of that particular specialty. Indeed, many medical situations require that the emergency patients be given a higher priority in accessing the non-emergency ORs than the non-emergency patients.

In many hospitals, surgery planning and scheduling is carried out as follows or in a similar fashion. At the beginning of each week or each month, the surgery planning office or a similar unit of the hospital builds an OR allocation template, or so-called weekly "Block Time Schedule" (we use the words "(allocation) template" and "Block Time Schedule" interchangeably in this paper), which allocates blocks of OR capacity to emergency surgery and various specialties' non-emergency surgery. Oftentimes, each block of OR capacity is one day of staffed hours of an operating room. A sample weekly "Block Time Schedule" for seven rooms is shown as below in Table 1.

OR #	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
3006A	Emergency	Emergency	Emergency	Emergency	Emergency
3006-B	Ophth	Ophth	Ophth	Ophth	Ophth
3006-D	Vascular	Tumor	Vascular	Colorectal	Colorectal
4000-1	Ortho	Ortho	Ortho	Ortho	Ortho
4000-2	Ortho	Ortho	Ortho	Ortho	Ortho
4000-3	Neuro	Neuro	Neuro	Neuro	Neuro
4000-4	Ophth	Ophth	Ophth	Ophth	Ophth

Table 1 A Sample "Block Time Schedule"

Before each working day (usually one day in advance), the doctors determine which inpatients in their specialty will have surgery performed on the following day. When making these decisions, they usually first accommodate outpatient surgeries scheduled for the next day because these have been previously scheduled many days in advance. Also, they take into account the number of blocks/rooms allocated to their specialty on that day, and the order and degree of urgency of all the active inpatients' surgery requests.

During a working day, surgeons try to finish as many scheduled surgeries as possible, in a predetermined order. In addition, emergency surgery demand arises almost every day, and surgeons try to operate upon these emergency patients as soon as possible because of their critical condition. Usually, they are sent into the emergency OR immediately, as long as it is available. If the emergency OR is busy when needed, the emergency patient is operated in one of the non-emergency rooms allotted to the particular specialty in which the patient or the needed surgery belongs. As a result, some scheduled inpatient and outpatient surgeries may have to be postponed to a later date or rescheduled. Also, surgeries scheduled for the afternoon may not be performed because they are too much behind schedule, and they can not be completed within the staffed hours if started.

Though there may be some real-time adjustments to the Block Time Schedule (i.e., one specialty's surgery is performed in an OR allocated to another specialty or occasionally some non-emergency surgery in the emergency OR) during the working days, surgeons do follow it as much as possible, because each specialty may require special medical equipment or prior preparations in the OR, and any change to the schedule may cause confusion in the daily work and take an extra amount of setup or switching time. Therefore, the quality of the OR allocation template is crucial to any operational performance measure pertaining to surgery, including the inpatients' in-hospital cost or length of stay.

3. Literature Review

We now review the prior literature on surgery planning, and relate them to our work.

The main approaches that have been adopted for surgery planning are mathematical programming (e.g., Ogulata and Erol [2003], Blake et al. [2002], Blake and Carter [2002], Ozkarahan [2000]), and simulation (e.g., Dexter et al. [1999], Schmitz and Kwak [1972], Kuzdrall [1974]). Mathematical programming (especially, integer or mixed integer programming) models have shown to be useful in capacity planning or resource allocation in many systems, including in the healthcare setting; while valid simulation models are useful in estimating the actual performance of a planning solution beforehand. Our methodology consists of both approaches.

As for the objective, much research has aimed at maximizing OR utilization, due to its high operational cost (see Dexter and Traub [2002], Ozkarahan [2000], Dexter et al. [1999]).

However, Dexter et al. [2002] showed that, at hospitals with fixed or nearly fixed annual budgets, allocating OR time based on utilization can adversely affect the hospital financially, and suggested considering not only OR time but also the resulting use of hospital beds. In line with this idea, there have recently been some studies on the impact of surgery schedules on the use of the other resources in hospitals. For example, Belien and Demeulemeester [2006] developed analytical models and solution heuristics for building cyclic master surgery schedules to minimize the expected total bed shortage. One distinguishing characteristic of our work is the focus on minimizing inpatients' length of stay waiting for surgery, resulting from the block time schedule. We are not aware of any OR capacity planning model addressing this objective, which yet is both financially attractive, due to its direct relevance to in-hospital cost, and operationally desirable, in the sense that it is essentially reducing average waiting time in a service system (see Marshall et al. [2005], Kourie [1975]).

As pointed out by Belien and Demeulemeester [2006], more and more attention has been paid to managing uncertainty in surgery planning and scheduling and improving the punctuality of the schedule realized. Gerchak et al. [1996] applied stochastic dynamic programming for advance surgery scheduling when the operating rooms' capacity utilization is uncertain. Lapierre et al. [1999] showed that giving incentives to hospital workers to improve their on-time performance is crucial to reducing delays in surgery or other health services, and pointed out that the punctuality of the first service of the day has a significant impact on subsequent services. Yet, many empirical studies (e.g. Litvak and Long [2000]) have shown that, in addition to the randomness and discreteness present in many other stochastic processing networks, data incompleteness and operational inefficiency are still common phenomena in healthcare systems, especially in public hospitals, and furthermore, clinical factors such as the changing nature of diseases complicate matters. As a result, the actual surgery demand tends to be higher than recorded or forecasted, while the supply of OR resources often suffers such uncertainties as staffing or equipment shortages, which has an effect equivalent to the former. Therefore, we believe that robustness, in the sense of the capability of handling inflated demand, is critical to the usefulness of a surgery capacity planning model. Our MIP model achieves this end by smoothing the capacity utilization, as will be shown in the next section.

4. Modeling

We first present the mixed integer programming (MIP) model for OR capacity allocation to minimize inpatients' length of stay. Then, we describe a simulation modeling methodology for estimating the performance of an allocation template and also for fine-tuning the analytical model.

4.1 MIP Model

Our model is partly based on the work of Blake and Donald [2002] who developed an integer programming model for operating room time allocation. The primary distinction between our model and their approach is that we determine an allocation template and each specialty's weekly OR time (uniquely determined by the template) simultaneously based on the objective of minimizing inpatients' length of stay, while their approach assumes that the weekly target number of OR hours to be allocated to each surgical group (equivalent to "each specialty" in our discussion) has already been explicitly determined, uses that as one input, and solves the model to obtain a template minimizing the shortfall between each group's actual weekly OR hours and its target level.

We first present the notation and assumptions of the model.

Notation

I: set of room types.

J: set of medical specialties.

D: set of days.

i: index for room type. A room can be considered a different type due to its location or special medical equipment.

j: index for medical specialty.

k, l: indices for days.

s: amount of staffed hours per day.

a_i: number of operating rooms of type i.

e_{jk}: emergency patients' surgery demand for specialty j on day k, measured in hours.

o_{jk}: non-emergency patients' (including inpatients' and outpatients') surgery demand for specialty j on day k, measured in hours.

c_{jk}: the maximum number of operating rooms that specialty j can utilize on day k, determined by the number of surgeons and the amount of equipment or any other necessary medical resources that each specialty has.

ρ_{kl}: the number of days delayed if a surgery is postponed from day k to day l.

θ: the equivalent number of days delayed if some surgery demand is not met in the model (or the penalty rate for "unmet" demand).

β: the penalty rate for undersupply of OR hours to a specialty, relative to a desired level determined by the percentage of total non-emergency surgery demand for each specialty. Inclusion of this penalty term in the objective function serves the purpose of smoothing the OR capacity. The β value should be much smaller than θ. We find an "optimal" or good β value by testing in a simulation model the templates determined by the MIP model with different β values.

Assumptions

- The weekly allocation template is in use for a finite horizon of weeks until updated.
- Every week has the same surgery demand pattern during the time horizon under consideration.
- There are 5 working days (Monday to Friday) each week, or $D = \{1, 2, 3, 4, 5\}$ and $|D| = 5$.
- There are 8 staffed hours for one OR each working day, or $s = 8$. Overtime work is not modeled.
- Only weekdays' surgery demand is considered in the model. However, patients' stay in the hospital on Saturdays and Sundays does incur cost just like on weekdays. Therefore,

$$\rho_{kl} = \begin{cases} 7, & if\ k = l, \\ l - k, & .if\ k < l, \\ 7 - k + l, & if\ k > l. \end{cases}$$

Note that if $k = l$ or $k > l$, day l represents a weekday following the week for day k..

- Only one operating room is used for emergency surgeries each day.
- The surgery demand is measured by the amount of OR hours. For example, if specialty j, on average, has 2 emergency patients who need surgery on Wednesday and the average length of this specialty's emergency surgery is 1.6 hours, then the surgery demand of specialty j's emergency patients on Wednesday or e_{j3} is 3.2 hours.
- Inpatients' in-hospital cost is incurred by the delay in meeting surgery demand. Because surgery demand is measured in OR hours, inpatients' in-hospital cost or length of stay is measured by "OR hours$\times$days". It is obtained by multiplying the postponed demand volume (i.e. the amount of OR hours postponed) by the number of days between the day that amount of demand arises and the day it is met.
- All emergency surgery demand must be met on the day it arises. Non-emergency or inpatients' and outpatients' surgery demand can be delayed.
- If some non-emergency patients' surgery demand cannot be met on the requested day, it can be met on the remaining days of the current week, on any day in the following week, or become unmet (equivalent to being met θ days late).
- Each specialty performs their non-emergency surgeries only in the non-emergency OR(s) allocated to them.
- Each specialty can perform their emergency surgeries either in the emergency OR or in the non-emergency OR(s) allocated to them.
- Specialty j is at most allocated c_{jk} ORs on day k.

The following are the *decision variables* of the model.

---

x_{ijk}: the number of operating rooms of type i allocated to specialty j on day k. The entire set of x_{ijk}'s determines the allocation template.

y_{jk}: the amount of the emergency OR's staffed hours allocated to specialty j on day k.

z_{jkl}: specialty j's non-emergency demand postponed from day k to day l.

u_{jk}: specialty j's unmet non-emergency demand on day k.

b_{jk}: the amount of idle time of the OR allocated to specialty j on day k.

h: the total amount of idle time of all non-emergency OR's.

p_j: oversupply of OR hours to specialty j, relative to its desired level.

q_j: undersupply of OR hours to specialty j, relative to its desired level.

We denote the following formulation for determining the OR allocation template as P.

$$\min \quad \sum_{k\in D}\sum_{l\in D}(\rho_{kl}\sum_{j\in J} z_{jkl}) + \theta\sum_{j\in J}\sum_{k\in D} u_{jk} + \beta\sum_{j\in J} q_j$$

$$s.t. \quad \sum_{j\in J} x_{ijk} = a_i, \quad \forall i,k \tag{1}$$

$$s\sum_{i\in I} x_{ijk} \geq e_{jk} - y_{jk} + \sum_{l\in D} z_{jlk}, \forall j,k \tag{2}$$

$$s\sum_{i\in I} x_{ijk} - (e_{jk} - y_{jk} + \sum_{l\in D} z_{jlk}) - b_{jk} + \sum_{l\in D} z_{jkl} + u_{jk} = o_{jk}, \forall j,k \tag{3}$$

$$h = \sum_{j\in J}\sum_{k\in D} b_{jk}, \tag{4}$$

$$\sum_{k\in D} b_{jk} - \frac{h\sum_{k\in D} o_{jk}}{\sum_{j\in J}\sum_{k\in D} o_{jk}} = p_j - q_j, \forall j, \tag{5}$$

$$\sum_{j\in J} y_{jk} \leq s, \quad \forall k \tag{6}$$

$$\sum_{i\in I} x_{ijk} \leq c_{jk}, \forall j,k \tag{7}$$

$$y_{jk} \leq e_{jk}, \forall j,k \tag{8}$$

$$x_{ijk}, y_{jk}, z_{jkl}, u_{jk}, b_{jk}, h, p_j, q_j \geq 0, \ \forall i,j,k,l \tag{9}$$

$$x_{ijk} \text{ integer}, \forall i,j,k \tag{10}$$

Constraint (1) guarantees that all the operating rooms are allocated to some specialty each day. Constraint (2) ensures that on any day each specialty has at least the OR capacity to meet the sum of its emergency demand on that day and non-emergency demand decided to be postponed to that day. Constraint (3) states that specialty j's non-emergency surgery demand on day k must be met either on that day, some remaining day in the current week, some day in the next week, or unmet (that is, met θ days late). Constraint (4) defines h as the sum of idle hours of all the non-emergency ORs over one week. Constraint (5) defines p_j and q_j, respectively, as the oversupply and undersupply of non-emergency OR (idle) time to specialty j, relative to a desired level determined by the percentage of total non-emergency surgery demand for specialty j. More specifically, given the weekly total of non-emergency OR idle hours, it is desired that each specialty occupies the amount proportional to its share of the total non-emergency surgery demand; p_j and q_j represent the difference between the actual allocation and the desired level for specialty j. Constraint (6) guarantees that at most s hours of emergency demand is met in the emergency OR each day. Constraint (7) ensures that specialty j is at most allocated c_{jk} ORs on day k. Constraint (8) guarantees that the daily emergency OR capacity allocated to each specialty does not exceed their emergency demand. Constraint (9) is the nonnegativity constraint on all the decision variables. Constraint (10) defines each x_{ijk} variable to be an integer.

The objective function of the formulation consists of three cost or penalty terms. The first two represent the inpatients' length of stay caused by the delay in meeting surgery demand within one cycle (or for up to 7 days) and by "unmet" demand (or equivalent to being postponed θ days, by our notation), respectively. The third term represents the total penalty caused by the undersupply of OR hours to each specialty, relative to its desired level determined by the percentage of total non-emergency surgery demand for each specialty. This penalty term is less dominant, considering our practical objective of minimizing inpatients' length of stay waiting for their surgery; yet inclusion of this term is useful in determining, among solutions yielding the same or similar sum of the first two terms or total length of stay, the one that leads to the most reasonable allocation of the non-emergency OR idle time (in the sense that the larger non-emergency surgery demand a specialty has, the more non-emergency OR idle time it tends to occupy) and thus would perform the best when subject to actual demand and time uncertainty.

4.2 Simulation Modeling

In reality, OR capacity and surgery demand are stochastic and/or dynamic. Also, surgery demand is discrete in nature or measured by the number of surgeries, instead of by hours as assumed in our MIP model. The degree in which the randomness and discreteness of the variables impacts the optimality of the template determined by the analytical model depends on the specific data or the problem instance under consideration. Therefore, after obtaining the template from the MIP model, a simulation model is used to assess the quality of the template generated by the MIP. Also, since our MIP model strives to enhance the robustness of the template by smoothing the OR capacity, we suggest that the smoothing constant or specifically the β value be determined by testing in the simulation model the templates resulting from different β values.

The following features are included in the surgery simulation model.

- Each specialty has two queues waiting for surgery: inpatients and outpatients. There is a single queue of all emergency patients who need surgery. All ORs are modeled as servers in the queueing system and the number of non-emergency ORs available to each specialty changes throughout the week as determined by the template.
- All types of patient arrivals are modeled as renewal processes.
- Each specialty's inpatient, outpatient, and emergency patients' surgery lengths are random variables fit from historical data. Pre-surgery set-up time and post-surgery OR cleaning time, if significant, is also added to the model (see Spangler et al. [2004], Strum et al. [2000], and May et al. [2000]).
- Emergency surgeries are performed in the emergency OR immediately as long as it is available. If not, they are performed in an available non-emergency OR allocated to that specialty. If no non-emergency OR of the needed specialty is currently available, emergency patients wait until the emergency OR or one of the respective specialty's non-emergency OR becomes available. Each specialty's non-emergency surgeries are only performed in the non-emergency OR(s) allocated to them.
- Emergency patients have the highest priority to be served, then outpatients, and lastly inpatients.
- In the case of inpatient surgeries, the simulation model has an "end-of-shift" protocol. If

there are 90 minutes or less remaining for the end of the shift, then a search is done through each specialty's inpatient queue (assuming that all of them have been prepped) to determine which patient has a surgery time less than or equal to the remaining time before the shift ends. If such a patient can be found, then he/she goes into surgery. If not, the room remains unoccupied till the end of the shift.

5. Case Study

The Los Angeles County (LAC) General Hospital is used as a case study to demonstrate our modeling approaches. LAC General Hospital is a large urban health center serving a largely poor population. It is also the trauma center for central Los Angeles, with the busiest emergency department, measured in admissions. Approximately 85% of the patients admitted to beds in the hospital enter through the emergency department.

At the time of this analysis, Los Angeles County General Hospital was using 19 operating rooms with 16 specialties and 1 emergency OR. We fed the capacity data and January 2005's demand data into the MIP model and used CPLEX 9.0 with default settings to solve the problem on a 3.2 GHz CPU with 2GB RAM.

In particular, we used four different values (0, 0.5, 0.75, 1) of the smoothing constant (β), and the solver gave an optimal solution or a close-to-optimal solution with a very small optimality gap (see Table 2) in less than 2 hours of CPU time for all four scenarios. The weekly OR capacity allocations for the actual template followed in January 2005 and for the four templates determined by the MIP model (with different β values) are shown in Table 3.

β	0	0.5	0.75	1
Optimality Gap (%)	0	0	1.35	2.12

Table 2 Optimality gap of the best integer solution

Template i: Actual Allocation

Template ii: Determined by MIP Model $\beta = 0$

Template iii: Determined by MIP Model $\beta = 0.5$

Template iv: Determined by MIP Model $\beta = 0.75$

Template v: Determined by MIP Model $\beta = 1$

Unit: OR hour	Template i	Template ii	Template iii	Template iv	Template v
Emergency	40	40	40	40	40
Burns	32	32	32	32	24
Cardiac	48	40	40	40	40
Colorectal	24	40	32	24	24
Foregut	16	16	16	16	16
HNS	88	80	88	88	88
Neuro	40	40	40	40	40
Ortho	192	160	168	176	184
Trauma	8	24	24	24	24
Tumor	24	24	24	24	24
Urology	40	40	40	40	40
GSNTE	16	40	40	40	40
Plastics	24	24	24	24	24
Hepatobiliary	24	24	24	24	24
Ophthalmology	80	72	72	72	72
OMFS	32	32	32	32	32
Vascular	24	32	24	24	24

Table 3 OR capacity allocated to each specialty per week

In order to evaluate the different templates, we performed a simulation analysis based on the features discussed in Section 4.2. AweSim! Version 3.0 (see Pritsker and O'Reilly [1999]) was used as the simulation software. In order to develop the surgery simulation model, the operating room process of LAC General Hospital was closely observed over a number of days. Furthermore four months of data from the hospital's information system was requested. The data included admit date, discharge date, surgery start date and time, surgery end date and time. Based on the data analysis (see the Appendix for details on the data analysis), the emergency patient and inpatient demands for each specialty were assumed to follow a stationary Poisson Process. That is, the inter-arrival times of requests were modeled as exponential random variables with mean equal to the inverse of the average daily demand. For outpatients the daily demand for each specialty does depend on the day of week. Thus, the average demand for each day was computed for each specialty type. In this case, the arrival process for outpatients was modeled as a non-stationary Poisson Process with the arrival rate changing in each day. Furthermore, the surgery times were assumed to be lognormal random variables with a constant 30 minute cleaning time after surgery. Prior studies (see Spangler et al. [2004], Strum et al. [2000], and May et al. [2000]) and our data analysis for LAC General Hospital (see the Appendix) show that the lognormal distribution is a reasonable model. Finally, a 20% no-show rate was assumed for outpatients.

The simulation results based on running the model for 100 weeks with a warm-up period of 10 weeks are shown in Table 4. Template **v** yields the shortest inpatients' length of stay waiting for surgery and also the smallest standard deviation of non-emergency OR utilization. Notice that the function of inpatients' average wait with respect to β does not necessarily have a convex structure due to many complicating factors.

Template	**ii**	**iii**	**iv**	**v**
β	0	0.5	0.75	1
Inpatients' Average Wait (day)	1.64	1.81	1.90	1.54
Standard Deviation of Non-Emergency OR Utilizations (%)	14.56	10.63	10.39	9.80

Table 4 Summary of simulation results for different β's

ER: Emergency (Patients), IPT: Inpatients, OPT: Outpatients

Template	**i**	**v**
ER Average Wait (day)	0.62	0.51
ER Throughput (for 4 weeks)	53	53
ER OR Utilization (%)	48.35	47.28
IPT Average Wait (day)	**1.86**	**1.54**
IPT Throughput (for 4 weeks)	336	335
OPT Average Wait (day)	0.34	0.33
OPT Throughput (for 4 weeks)	187	192
Average Non-ER OR Utilization (%)	63.39	65.91
Standard Deviation of Non-ER OR Utilization (%)	**14.03**	**9.80**

Table 5 Performance (simulation results) summary with the original demand

Table 5 provides summary comparison of output statistics between the actual template (**i**) and the best template generated by the model (**v**). The inpatients' average length of stay waiting for their surgery reduces from 1.86 days in the scenario of using the actual hospital template to 1.54 days when using the model's template. The standard deviation of non-emergency OR utilizations reduces by 30%, when the model's template is used. Also, the emergency patients' average wait reduces by nearly 18% and all the other performance measures stay relatively equivalent.

Since in reality demand far exceeds capacity for LAC General Hospital, we then tested the sensitivity of the templates in handling increased demand. Table 6 shows the summary of the simulation results for the actual template and the model's template with a 20% increase in demand. As can be seen in the simulation results, the inpatients' average length of stay waiting for their surgery reduces from 5.15 days in the scenario of using the actual hospital template to 2.23 days when using the model's template. This is partial due to the smoother utilization of the capacity (the standard deviation of non-emergency OR utilizations being 26% smaller). The emergency patients' average wait reduces by 16% and the other performance measures are relatively equivalent.

ER: Emergency (Patients), IPT: Inpatients, OPT: Outpatients

Template	**i**	**v**
ER Average Wait (day)	0.62	0.52
ER Throughput (for 4 weeks)	57	59
ER OR Utilization (%)	51.62	45.42
IPT Average Wait (day)	**5.15**	**2.23**
IPT Throughput (for 4 weeks)	394	394
OPT Average Wait (day)	0.39	0.39
OPT Throughput (for 4 weeks)	234	234
Average Non-ER OR Utilization (%)	73.86	73.97
Standard Deviation of Non-ER OR Utilization (%)	**15.87**	**11.82**

Table 6 Performance (simulation results) summary with 20% more demand

6. Conclusions and Future Work

A mixed integer programming model was developed to determine optimal operating room allocation to each specialty. A simulation analysis was used to assess the performance of the operating room template. The methodology was illustrated on a case example of Los Angeles County General Hospital, and the analysis showed that the average inpatient waiting time for surgery could be reduced with an efficient allocation of operating room time.

The templates generated by the optimization model could perform poorly in practice when there are high variances associated with surgery length and volatile patient arrival patterns since the optimization model does not account for uncertainty in the problem parameters. Therefore, future research can focus on incorporating uncertainty into the analytical model. Also, since the problem sizes were relatively small the MIP could be solved to optimality or near-optimality by a commercial software package (CPLEX) in the scenarios performed in this study. However, for larger problem sizes specialized algorithms or heuristics may be necessary in order to solve the model.

7. Acknowledgment

The research reported in this paper was partially supported by the Los Angeles Care Foundation. We would also like to thank Los Angeles County General Hospital for providing us with data on their surgery procedures and Professor Randolph Hall from the University of Southern California for his advice and guidance during the research efforts.

8. References

Belien, J., and E. Demeulemeester, "Building Cyclic Master Surgery Schedules with Leveled Resulting Bed Occupancy", *European Journal of Operational Research*, to appear.

Blake, J. T., F. Dexter, and J. Donald, "Operating Room Managers' Use of Integer Programming for Assigning Block Time to Surgical Groups: A Case Study", *Anesthesia and*

Analgesia, Vol. 94, pp. 143-148, 2002.

Blake, J. T., and J. Donald, "Mount Sinai Hospital Uses Integer Programming to Allocate Operating Room Time", *Interfaces*, Vol. 32, No. 2, pp. 63-73, 2002.

Blake, J. T., and M. W. Carter, "A Goal Programming Approach to Strategic Resource Allocation in Acute Care Hospitals", *European Journal of Operations Research*, Vol. 140, No. 3, pp. 541-561, 2002.

Dexter, F., A. Macario, R. D. Traub, M. Hopwood, and D. A. Lubarsky, "An Operating Room Scheduling Strategy to Maximize the Use of Operating Room Block Time: Computer Simulation of Patient Scheduling and Survey of Patients' Preferences for Surgical Waiting Time", *Anesthesia and Analgesia*, Vol. 89, pp. 7-20, 1999.

Dexter, F., A. Macario, and R. D. Traub, "Which Algorithm for Scheduling Add-on Elective Cases Maximizes Operating Room Utilization?", *Anesthesiology*, Vol. 91, Issue 5, pp. 1491-1500, 1999.

Dexter, F. , J. T. Blake, D. H. Penning, B. Sloan, P. Chung, and D. A. Lubarsky, "Use of Linear Programming to Estimate Impact of Changes in a Hospital's Operating Room Time Allocation on Perioperative Variable Costs", *Anesthesiology*, Vol. 96, No 3, pp. 718-724, 2002.

Dexter, F., and R. D. Traub, "How to Schedule Elective Surgical Cases into Specific Operating Rooms to Maximize the Efficiency of Use of Operating Room Time", *Anesthesia and Analgesia*, Vol. 94, pp. 933-942, 2002.

Gerchak, Y., D. Gupta, and M. Henig, "Reservation Planning for Elective Surgery under Uncertain Demand for Emergency Surgery", *Management Science*, Vol. 42, No. 3, pp. 321-334, 1996.

Kourie, D. G., “A Length of Stay Index to Monitor Efficiency of Service to General Surgery In-Patients”, *Operations Research Quarterly*, Vol. 26, No. 1, Part 1, pp. 63-69, 1975.

Kuzdrall, P. J., N. K. Kwak, and H. H. Schmitz, “The Monte Carlo Simulation of Operating-Room and Recovery-Room Usage”, *Operations Research*, Vol. 22, No. 2, pp. 434-440, 1974.

Lapierre, S. D., C. Batson, and S. McCaskey, “Improving On-Time Performance in Health Care Organizations: A Case Study”, *Health Care Management Science*, Vol. 2, pp. 27-34, 1999.

Marshall, A., C. Vasilakis, and E. El-Darzi, “Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions”, *Health Care Management Science*, Vol. 8, pp. 213-220, 2005.

May, J. H., D. P. Strum, and L. G. Vargas, “Fitting the Lognormal Distribution to Surgical Procedure Times”, *Decision Sciences*, Vol. 31, pp. 129–148, 2000.

Ogulata, S. N., and R. Erol, “A Hierarchical Multiple Criteria Mathematical Programming Approach for Scheduling General Surgery Operations in Large Hospitals”, *Journal of Medical Systems*, Vol. 27, No. 3, pp. 259-270, 2003.

Ozkarahan, I., “Allocation of Surgeries to Operating Rooms by Goal Programing”, *Journal of Medical Systems*, Vol. 24, No. 6, pp. 339-378, 2000.

Pritsker, A. A. B, and J. J. O’Reilly, *Simulation with Visual SLAM and AweSim (2nd Edition)*, John Wiley & Sons, New York, and Systems Publishing Corporation, West Lafayette, Indiana, 1999.

Schmitz, H. H., and N. K. Kwak, “Monte Carlo Simulation of Operating-Room and Recovery-Room Usage”, *Operations Research*, Vol. 20, No. 6, pp. 1171-1180, 1972.

Spangler, W. E., D. P. Strum, L. G. Vargas, and J. H. May, “Estimating Procedure Times for Surgeries by Determining Location Parameters for the Lognormal Model”, *Health Care Management Science*, Vol. 7, No. 2, pp. 97-104, 2004.

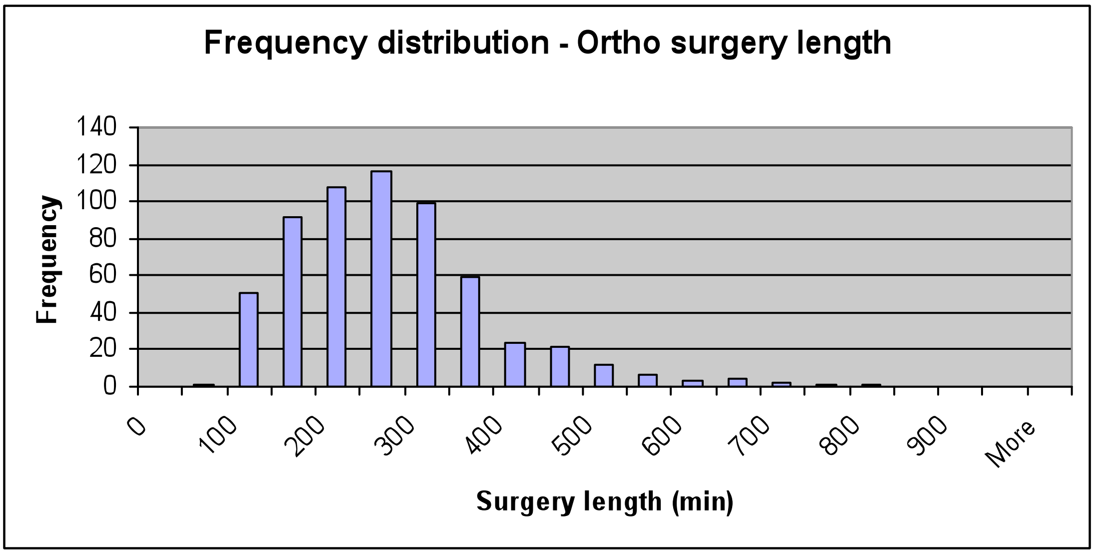
Strum, D. P., J. H. May, and L. G. Vargas, “Modeling the Uncertainty of Surgical Procedure Times: Comparison of Log-normal and Normal Models”, *Anesthesiology*, Vol. 92, pp. 1160–1167, 2000.

9. Appendix

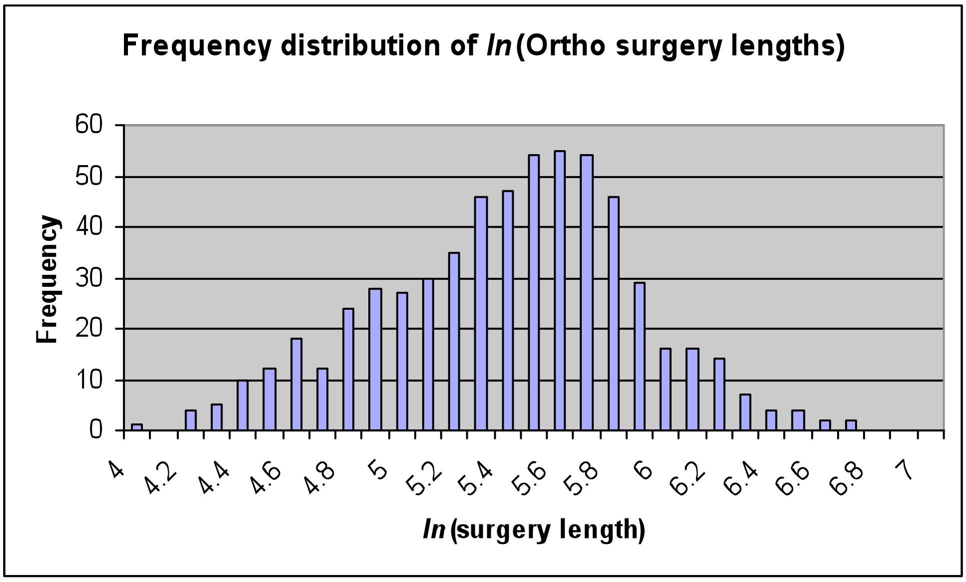
As stated in section 4 of this paper, the surgery lengths were modeled as a lognormal distribution and the patient inter-arrival times (representing surgery demand generation) were modeled as renewal processes, for all specialties. These distributions were determined by conducting statistical tests on the data provided. As an illustration, we consider the case of Orthopedics, and show the procedure followed in determining the surgery length and inter-arrival time distributions.

9.1 Chi-squared test on surgery duration

The following plot shows the frequency distribution of the surgery lengths for Orthopedics.

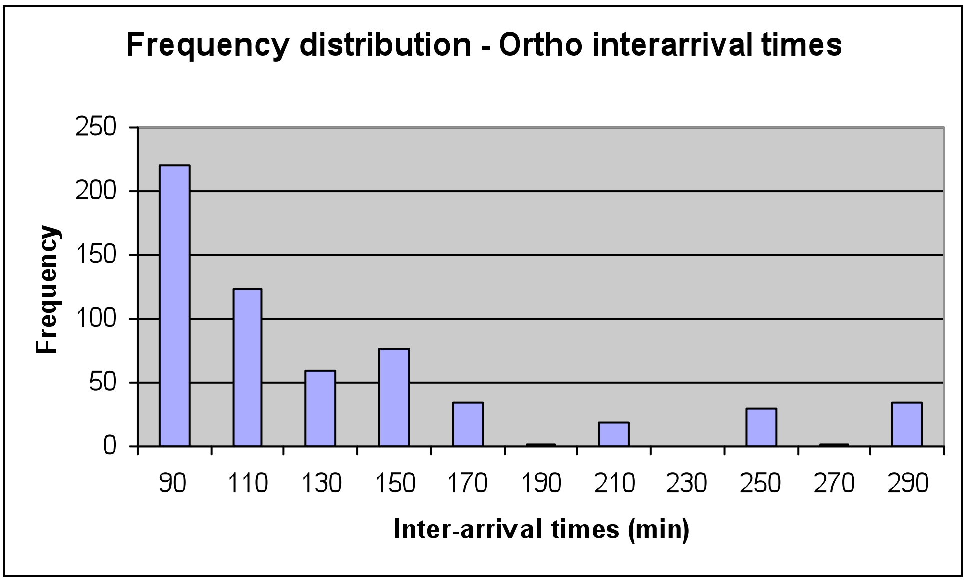


Since this closely resembles the probability density function of a lognormal distribution, we used the chi-squared procedure to test the hypothesis if the natural logarithmic values of the surgery lengths follow a normal distribution. The natural logarithm of the surgery lengths had a mean equal to 5.355 and a variance of 0.234. Given below is the frequency distribution of these values. The chi-squared test statistic was 4.730. Considering $\alpha = 0.05$, the critical value is 36.42. Hence, the null hypothesis that this data set follows a lognormal distribution cannot be rejected.



9.2 Chi-squared test on patient inter-arrival times

The frequency distribution of the inter-arrival times for Orthopedics was determined to be the following.



A chi-square test was conducted to determine if the inter-arrival times were exponentially distributed. The sample mean for the data was found to be 117.962. The chi-squared test statistic was 7.757. Considering $\alpha = 0.05$, the critical value is 15.507. Hence, the null hypothesis that this data set follows an exponential distribution cannot be rejected.