

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 allocation template that minimizes inpatients' cost measured as their length of stay. A number of patient type priority (e.g., emergency over non-emergency patient) 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 organizations, and their prime operational objective is to provide medical services to their patients at a reasonable cost. Most private and public hospitals serve patients who have a third-party health insurance, such as Medicare, Blue Cross etc, to cover their medical expenses. Hospitals would need to bill these coverage providers for the medical procedures the patient had to undergo. The coverage provider reviews the patient's charts, checks to see if appropriate procedures were carried out, and then reimburses the hospital. However, hospitals are not reimbursed for any procedure or activity which is considered unnecessary. An unnecessarily long length of stay (LOS) for inpatients is one of the common issues for which hospitals are denied reimbursement and this constitutes a large portion of a hospital's expenditures (see reference: Clinical Scholars Program, Interim Report). Thus, it is of great interest to hospital administrators to reduce inpatients' LOS. Often, a high LOS is due to inefficient scheduling procedures used in surgical and ancillary services, since most inpatients require one or both of these during their stay at the hospital.

In most hospitals, a large percentage of inpatients undergo some type of surgery during their stay in a hospital. To reduce inpatient length of stay and also from the patient's health point of view, it is ideal that a surgery is performed as soon as it is requested. However, in reality, inpatients may need to wait in bed for their surgery for a day or two or even longer (especially in public hospitals). Outpatients also may not always undergo their surgery on schedule, although this is not directly counted in calculating the LOS. There are many possible reasons for delays. For example, a patient who is considered to need surgery immediately might be given preference over another patient who has been waiting for a week, and has a scheduled surgery. Or, a hospital may face a situation where a large number of emergency patients needing surgery use up most of the available Operating Room capacity. As a result, the delayed inpatients stay in the hospital longer, incurring higher non-reimbursable costs for the hospital and the delayed outpatients may remain in a long outpatient queue (often in the form of a waiting list) waiting for surgery

and also continue to “compete” for operating room capacity with inpatients, affecting LOS indirectly.

The objective of this research is to minimize inpatients' length of stay waiting for their surgery, by developing a mixed integer programming approach for allocating operating room capacity to different medical specialties, thereby reducing the loss incurred to a hospital due to non-reimbursed bed days. In Section 2, we describe typical surgery planning and scheduling procedures. We introduce an operating room 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 (MIP) modeling approach for determining a weekly operating room allocation template or Block Time Schedule which minimizes inpatients' in-hospital cost measured as their length of stay. The MIP model assumes all problem input parameters are deterministic and proportional allocation of demand satisfaction (output) with room allocation (input). The template solution from the MIP is then fed into a simulation model which captures the randomness and non-linearities of the actual process in order to evaluate the effectiveness of the MIP solution on a more realistic model. The simulation model then outputs the average inpatients' length of stay as the primary evaluation criteria of the template. The Simulation model is described in Section 4. Using the simulation model, we compared the allocation being used at the Los Angeles County (LAC) General Hospital to the operating room allocation suggested by our MIP model, and showed that both the inpatients' length of stay and operating room utilization can be improved upon by using the template given by the MIP model. Thus, this analysis demonstrated that the template generated by the deterministic planning model (MIP model) could perform reasonably well in a stochastic environment as compared to LAC General Hospital's current template. The case is presented in Section 5.

2. Background

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

Most hospitals categorize their patients as emergency patients, inpatients, and outpatients, in their patient database, and when they 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, whereas others may have only one or two.

Operating rooms can be considered of two types: emergency and non-emergency. Hospitals typically have only a few emergency operating rooms and all the others as non-emergency operating rooms. The emergency operating rooms are solely allocated to the emergency patients who need surgery, and usually all specialties' surgeries can be performed in that room. The non-emergency operating rooms are often allotted to different specialties. Though a non-emergency operating room assigned to a specialty is intended for its non-emergency surgeries (i.e., inpatient and outpatient surgeries), the room can also accommodate 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 operating rooms 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 operating room allocation template, or so-called weekly "Block Time Schedule" (we use the terms "(allocation) template" and "Block Time Schedule" interchangeably in this paper), which allocates blocks of operating room capacity to emergency surgery and various specialties' non-emergency surgery. Often each block of operating room 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.

Operating room #	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 first accommodate outpatient surgeries scheduled for the next day, because these would have been 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. Typically, only a few surgery slots per day are allotted to outpatients. The remaining are reserved for inpatients.

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 operating room immediately, as long as it is available. If the emergency operating room is busy when needed, the emergency patient is operated on 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 operating room allocated to another specialty or occasionally some non-emergency surgery in the emergency operating room) during the working days, surgeons tend to follow it as much as possible, because each specialty may require special medical equipment or prior preparations in the operating room, 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 operating room 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], Sier et al. [1997]), and simulation (e.g., Dexter et al. [1999], Schmitz and Kwak [1972], Kuzdrall [1974], Vasilakis et al. [2007]). Mathematical programming (especially, integer or mixed integer programming) models have shown to be useful in capacity planning or resource allocation in many complex systems, including healthcare; 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 operating room 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 operating room time based on utilization can adversely affect the hospital financially, and suggested considering not only operating room 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. Sier et al. [1997] presented a multi-objective non-linear optimization model for surgery scheduling, taking surgical priorities, surgery length, demand for equipment and conflicts in the schedule into account. They used simulated annealing techniques to generate feasible surgery schedules. 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 operating room 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], greater attention is being 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 operating room 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 operating room 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 operating room 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 operating room 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 operating room 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.
- n_{jk} : inpatients' surgery demand for specialty j on day k , measured in hours.
- o_{jk} : 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 (or the penalty rate for inpatient demand postponed from day k to day l).
- λ_{kl} : the penalty rate for outpatient demand postponed from day k to day l .
- θ_{IPT} : the penalty rate for "unmet" inpatient demand.
- θ_{OPT} : the penalty rate for "unmet" outpatient demand.
- β : the penalty rate for undersupply of operating room hours to a specialty, relative to a desired level determined by the percentage of total non-emergency (i.e., inpatient plus outpatient) surgery demand for each specialty. Inclusion of this penalty term in the objective function serves the purpose of smoothing the operating room capacity. Using the simulation model, we identify suitable values for β .

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 operating room 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,

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

Note that if $k = l$ or $k > l$, day l represents a weekday in the following week. If $k < l$, day l simply represents the l th weekday in the same week as day k . In the case $k < l$, it is never optimal to postpone the surgery to day $l+7$. For example, if Tuesday's (day 2) surgery demand is postponed to a Thursday (day 4), we have the case $k < l$ and we know that it is optimal to perform the surgery this week and not to delay it for an additional week from Thursday. As another example, if Friday's (day 5) surgery demand is postponed to the following Monday (day 1), we have the case $k > l$. Here, $\rho_{kl} = 7-5+1$, which is 3. This is valid because the demand was postponed by 3 days. Since we assume every week has the same demand pattern, the postponement of a week in this case only increases the cost function without resolving any infeasibility as cyclic demand pattern implies cyclic capacity availability pattern. If a surgery is postponed by more than a week, then it is considered to be "unmet", and has a large penalty associated with it. In this paper, we consider a linear correlation between cost and delay. This was primarily done to maintain linearity in the objective function given below. However, it could be easily modified to fit any other linear or non-linear penalty structure.

- Only one operating room is used for emergency surgeries each day.
- The surgery demand is measured by the amount of operating room hours, that is, continuous in nature. 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 operating room hours, inpatients' in-hospital cost or length of stay is measured by "operating room hours postponed \times days delayed". It is obtained by multiplying the postponed demand volume (i.e., the amount of operating room hours postponed) by the number of days between the day that amount of demand arises and the day it is met. We believe that this linearity cost assumption is reasonable, as we are only considering inpatients' length of stay before surgery. The linearity assumptions are relaxed in the simulation model. Also, it is important to realize that operating room demand is aggregated and continuous in the MIP model, which means that there are no issues such as "a" or "several" patients in this MIP model. This assumption is relaxed in the simulation model. Also, please note that if a surgery is carried out on the day it is requested, we do not associate any cost to it. Hence, a patient with a long surgery length would not be more costly.
- The delay in meeting outpatient surgery demand also incurs penalty in the same way as in the inpatient case, except that the penalty rate is λ_{kl} . We assume that λ_{kl} is proportional to ρ_{kl} .
- All emergency surgery demand must be met on the day it arises. Non-emergency or inpatient and outpatient surgery demand can be delayed.
- If some non-emergency patients' surgery demand cannot be met on the requested day, it can be met on any working day which is no more than 7 days after the requested day, or become unmet. θ_{IPT} and θ_{OPT} are the penalty rates applied for unmet demand; they are the largest among all the penalty rates. Since we are modeling a cyclic demand problem, the demand that cannot be met during the normal seven-day cycle (unmet demand) has to be met outside the normal shift operation either through overtime or moving the surgery to another local hospital. To discourage this possibility, the penalty for unmet demand should be set to be much larger than any other penalty.
- Each specialty performs their non-emergency surgeries only in the non-emergency

operating room(s) allocated to them.

- Each specialty can perform their emergency surgeries either in the emergency operating room or in the non-emergency operating room(s) allocated to them.
- Specialty j is at most allocated c_{jk} operating rooms 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 specialty j 's emergency surgery demand to be met in the emergency operating room on day k .
z_{jkl} :	specialty j 's inpatient demand postponed from day k to day l .
w_{jkl} :	specialty j 's outpatient demand postponed from day k to day l .
u_{jk} :	specialty j 's unmet inpatient demand on day k .
v_{jk} :	specialty j 's unmet outpatient demand on day k .
b_{jk} :	the amount of idle time of the operating room allocated to specialty j on day k .
h :	the total amount of idle time of all non-emergency operating room's.
p_j :	oversupply of operating room hours to specialty j , relative to its desired level.
q_j :	undersupply of operating room hours to specialty j , relative to its desired level.

We denote the following formulation for determining the operating room allocation template as **P**.

$$\begin{aligned}
\min \quad & \sum_{k \in D} \sum_{l \in D} (\rho_{kl} \sum_{j \in J} z_{jkl}) + \sum_{k \in D} \sum_{l \in D} (\lambda_{kl} \sum_{j \in J} w_{jkl}) + \theta_{IPT} \sum_{j \in J} \sum_{k \in D} u_{jk} + \theta_{OPT} \sum_{j \in J} \sum_{k \in D} v_{jk} + \beta \sum_{j \in J} q_j \\
s.t. \quad & \sum_{j \in J} x_{ijk} = a_i, \quad \forall i, k & (1) \\
& s \sum_{i \in I} x_{ijk} \geq e_{jk} - y_{jk} + \sum_{l \in D} (z_{jlk} + w_{jlk}), \quad \forall j, k & (2) \\
& s \sum_{i \in I} x_{ijk} - (e_{jk} - y_{jk} + \sum_{l \in D} (z_{jlk} + w_{jlk})) - b_{jk} + \sum_{l \in D} (z_{jkl} + w_{jkl}) + (u_{jk} + v_{jk}) = n_{jk} + o_{jk}, \quad \forall j, k & (3) \\
& \sum_{l \in D} z_{jkl} \leq n_{jk}, \quad \forall j, k & (4) \\
& \sum_{l \in D} w_{jkl} \leq o_{jk}, \quad \forall j, k & (5) \\
& u_{jk} \leq n_{jk}, \quad \forall j, k & (6) \\
& v_{jk} \leq o_{jk}, \quad \forall j, k & (7) \\
& h = \sum_{j \in J} \sum_{k \in D} b_{jk} & (8) \\
& \sum_{k \in D} b_{jk} - \frac{h \sum_{k \in D} (n_{jk} + o_{jk})}{\sum_{j \in J} \sum_{k \in D} (n_{jk} + o_{jk})} = p_j - q_j, \quad \forall j & (9) \\
& \sum_{j \in J} y_{jk} \leq s, \quad \forall k & (10) \\
& \sum_{i \in I} x_{ijk} \leq c_{jk}, \quad \forall j, k & (11) \\
& y_{jk} \leq e_{jk}, \quad \forall j, k & (12) \\
& x_{ijk}, y_{jk}, z_{jkl}, w_{jkl}, u_{jk}, v_{jk}, b_{jk}, h, p_j, q_j \geq 0, \quad \forall i, j, k, l & (13) \\
& x_{ijk} \text{ integer}, \quad \forall i, j, k & (14)
\end{aligned}$$

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 operating room capacity to meet the sum of its emergency demand on that day and non-emergency (i.e., inpatient plus outpatient) 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, or on a working day which is no more than 7 days after that day, or unmet. Constraint (4) states that the sum of specialty j 's postponed inpatient demand on day k is no more than the total inpatient demand for that day. Constraint (5) states that the sum of specialty j 's postponed outpatient demand on day k is no more than the total outpatient demand for that day. Constraint (6) ensures that specialty j 's unmet inpatient demand on

day k is no more than the total inpatient demand for that day. Constraint (7) ensures that specialty j 's unmet outpatient demand on day k is no more than the total outpatient demand for that day. Constraint (8) defines h as the sum of idle hours of all the non-emergency operating rooms over one week. Constraint (9) defines p_j and q_j , respectively, as the oversupply and undersupply of non-emergency operating room (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 operating room 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 (10) guarantees that at most s hours of emergency demand is met in the emergency operating room each day. Constraint (11) ensures that specialty j is at most allocated c_{jk} operating rooms on day k . Constraint (12) guarantees that the daily emergency operating room capacity allocated to each specialty does not exceed their emergency demand. Constraint (13) is the nonnegativity constraint on all the decision variables. Constraint (14) defines each x_{ijk} variable to be an integer.

The objective function of the formulation consists of five cost or penalty terms. The first term represents the inpatients' length of stay cost due to the delay in meeting their surgery demand. The second term represents the penalty due to the delay in meeting outpatient surgery demand. For particular purposes, the penalty weight for inpatients should be larger than outpatients since the former term directly impacts LOS. However, a penalty weight should still be given for outpatients because the demand for outpatients represents the number of surgeries that the hospital would like to perform of this type. For scheduling purposes and their ability to draw down their outpatient surgery queue, the hospital would like to perform a given number of outpatient surgeries each day. Deviations from this demand result in having to reschedule the outpatient surgery which can generate additional costs (e.g., overtime, patient not showing up, etc.). The third and fourth terms represent the penalty due to "unmet" inpatient and outpatient demands. Since we are modeling a cyclic demand problem, the demand that "cannot be met" during

the normal seven day cycle (unmet demand) has to be met outside the normal shift operation either through overtime or moving the surgery to another local hospital. To discourage this possibility, the penalty for unmet demand should be set to be much larger than any other penalty. The fifth term represents the total penalty caused by the undersupply of operating room 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 the least 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 four penalty terms, the one that leads to a more reasonable allocation of the non-emergency operating room idle time (in the sense that the larger non-emergency surgery demand a specialty has, the more non-emergency operating room 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, operating room capacity and surgery demand are stochastic and/or dynamic. Also, surgery demand is discrete in nature, because it is 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 operating room 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 operating rooms are

modeled as servers in the queueing system and the number of non-emergency operating rooms 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 operating room 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 operating room immediately as long as it is available. If not, they are performed in an available non-emergency operating room allocated to that specialty. If no non-emergency operating room of the needed specialty is currently available, emergency patients wait until the emergency operating room or one of the respective specialty's non-emergency operating room becomes available. Each specialty's non-emergency surgeries are only performed in the non-emergency operating room(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 approach. LAC General Hospital is a large urban health center serving a largely poor population. The hospital is licensed for 1,395 beds and budgeted to staff 745 inpatient beds, though some might be closed on days when the nursing staff is short. It is

one of the busiest public hospitals in the western United States, and records around 39,000 inpatient discharges, 150,000 emergency department visits and 1 million ambulatory care visits each year. It treats around 28% of the trauma victims in the region. (Source:http://www.usc.edu/schools/medicine/patient_care/hospitals_clinics/lacusc_medical.html). Approximately 85% of the patients admitted to beds in the hospital enter through the emergency department.

The hospital mostly serves patients who are unable to finance their own medical expenses and are, thus, reliant on Medicare, Medi-Cal and the like. Patients are financially screened before being admitted to an inpatient bed, to verify if they qualify for any of the low-income group health care coverage programs. Once a patient has been discharged, the hospital needs to submit a copy of the patient's transactions to the relevant coverage provider in order to be reimbursed for the services. They are not reimbursed for the days when the patient stays in hospital waiting for surgery and/or ancillary services such as radiology and/or prescription medicines etc. These are known as "denied" days. The longer the length of unnecessary stay in the hospital, the higher would be the number of denied days, and the higher would be the loss for the hospital. Hence, it is vital from the hospital's point of view to reduce unnecessary length of stay and, thereby, increase throughput. LAC's Utilization Review department is responsible for ensuring that the number of denied days is kept to a minimum.

At the time of this analysis, LAC General Hospital was using 19 operating rooms with 16 specialties and 1 emergency operating room. 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.

Our intention was not to develop a complete experimental design to identify the best value of β . The purpose of the study was to illustrate to LAC General Hospital the benefits of math modeling to develop operating room templates. The intent of the study was to sell the methodology for generating the templates, not to develop a final template

to the hospital staff. The plan is for the hospital staff to eventually use the modeling methodology themselves and perform sensitivity analysis on the parameters in order to derive their own templates.

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 operating room 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 Operating room capacity (hours) allocated to each specialty per week

It was observed that for templates where an optimal solution was obtained, the time taken to converge averaged around 1 hour. In the case of templates where an optimal solution could not be obtained within 2 hours, it was observed that there was little improvement in the optimality gap percentage after about 60-75 minutes of CPU time.

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 (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. The selection of 10 weeks for a warm-up period and 100 weeks of run time were determined by plotting the average length of inpatient's stay over time. The plot showed that the system reached steady state reasonably quickly and the choice of 10 weeks for a warm-up period and 100 weeks for run length were conservative choices. Template v yields the shortest inpatients' length of stay waiting for surgery and also the smallest standard deviation of non-emergency operating room 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 operating room 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 operating room 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 operating room Utilization (%)	63.39	65.91
Standard Deviation of Non-ER operating room 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 operating room 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 partially due to the smoother utilization of the capacity (the standard deviation of non-emergency operating room 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 operating room 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 operating room Utilization (%)	73.86	73.97
Standard Deviation of Non-ER operating room 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.

In solving the MIP model, the time taken to reach an optimal or a near-optimal solution was found to be around 60-75 minutes of CPU time. Typically in a hospital, block scenarios are generated only a few times a year, only when there is a significant change in the surgery demand, or in the resources (operating room, surgeons, equipment etc) available. In addition, a suitable user interface to the proposed model could be easily engineered, enabling the user to generate effective and efficient block schedules for surgery and remain remote from the underlying mathematics. Thus, our proposed methodology could be easily integrated with a hospital's existing information system and handled by a surgery scheduler, to monitor all aspects of patient care.

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 Operational Research*, Vol. 140, No. 3, pp. 541-561, 2002.

Clinical Scholars Program, Interim Report July 2005 – February 2006, UCLA – The Robert Wood Johnson Foundation. Retrieved from http://www.hsrcenter.ucla.edu/csp/sitevisit2006/Handouts/REVISED_UCLARJW_05-06_Internimrpt.doc

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”, *Operational 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.

Sier, D., Tobin, P., and C. McGurk, “Scheduling Surgical Procedures”, *Journal of the Operational Research Society*, Vol. 48, pp. 884-891, 1997.

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.

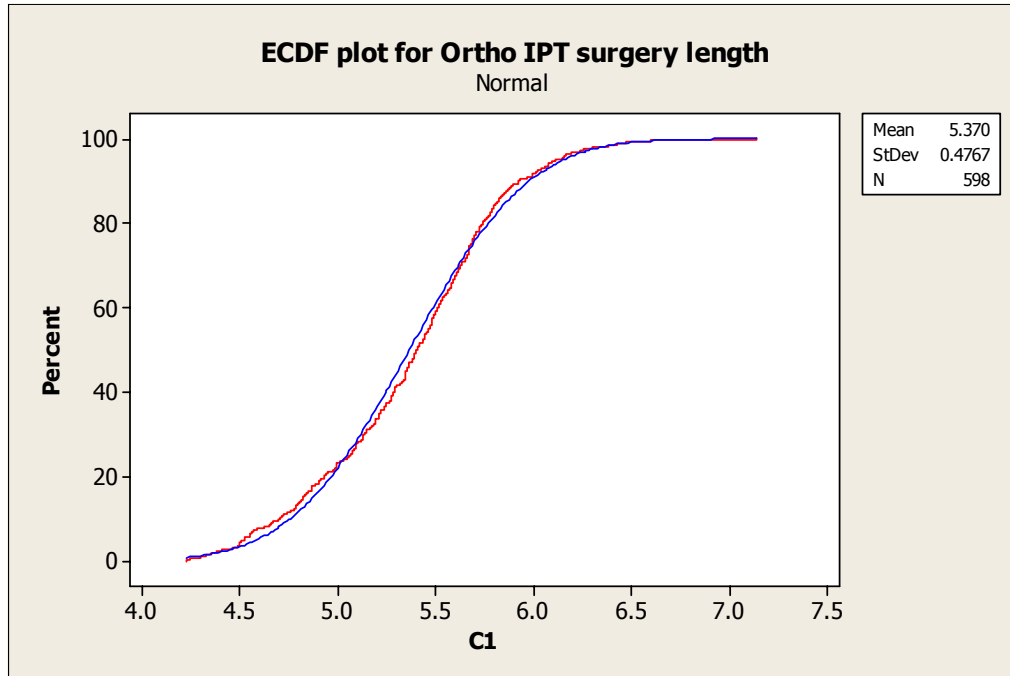
Vasilakis, C., Sobolev, B. G., Kuramoto, L., and A. R. Levy, “A Simulation Study of Scheduling Clinical Appointments in Surgical Care”, *Journal of the Operational*

9. Appendix

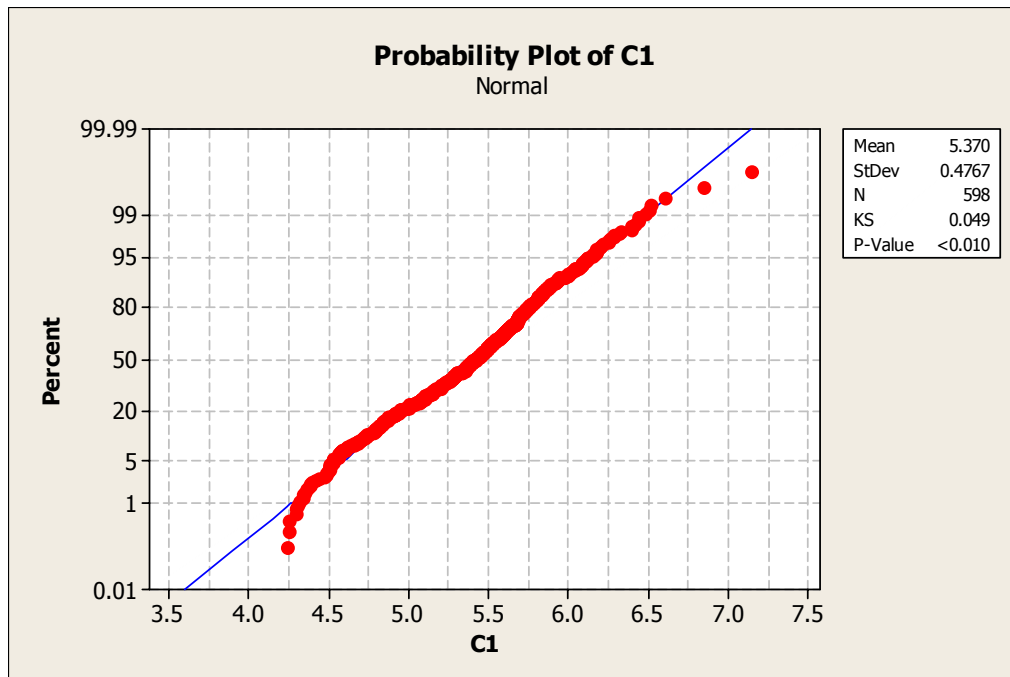
As stated in section 5 of this paper, the surgery lengths were modeled as a lognormal distribution and the patient inter-arrival times (representing surgery demand generation) were modeled as renewal processes, for all specialties. These distributions were determined by conducting one-sample Kolmogorov – Smirnov (K-S) statistical tests on the inpatient, outpatient and emergency patients' surgery lengths and on the patient inter-arrival times for each of the 16 specialties and the emergency operating room, totaling to 66 statistical tests. 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 K-S test on Ortho inpatient (IPT) surgery duration

We perform the K-S test using the Minitab software to accept (or reject) our hypothesis that the surgery duration for Orthopedics follows a lognormal distribution. For this, we consider the natural logarithm of the surgery lengths. The plot of the Empirical Cumulative Distribution Function (ECDF) with a normal cumulative distribution function for 598 normal random numbers is shown below.



Next, we perform the K-S goodness-of-fit test for normality using Minitab. The plot is shown below.

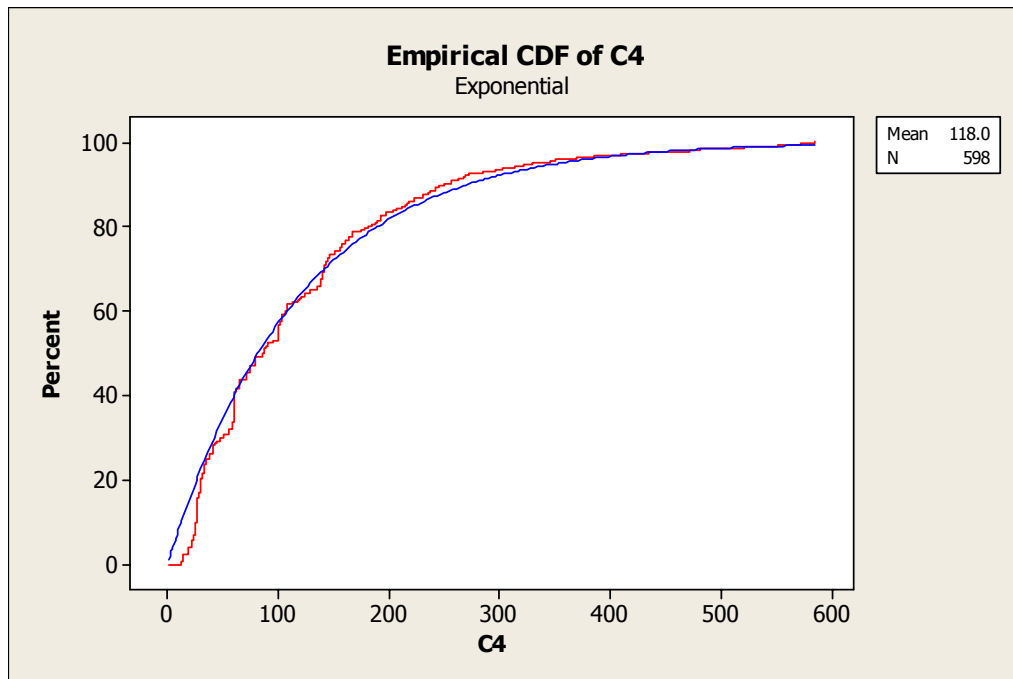


The K-S value from the above plot is 0.049, which is lesser than the corresponding critical value of 0.056 from the K-S table for $n = 598$ and $\alpha = 0.05$. Hence, we conclude

that the natural logarithms of the surgery lengths are normally distributed with a mean of 5.37 and a variance of 0.23. That is, the surgery lengths are log-normally distributed.

9.2 K-S test on Ortho patient inter-arrival times

We perform the K-S test to accept (or reject) our hypothesis that the Ortho patient arrival follows a Poisson process. The plot of the ECDF for the patient inter-arrival times with an exponential cumulative distribution function for 598 normal random numbers is shown below.



The K-S test statistic D was computed as:

$$D = \max_{1 \leq i \leq n} \left(F(Y_i) - \frac{i-1}{n}, \frac{i}{n} - F(Y_i) \right),$$

where F is the CDF of the exponential distribution and n is the size of the sample.

The value of D for the patient inter-arrival times was found to be 0.012. Since we estimated two parameters from the given data, we use an adaptive test at $\alpha = 0.05$ but compare the K-S critical value at $\alpha' = 4*\alpha = 4*0.05 = 0.2$. Hence, the adaptive K-S

critical value at $n = 598$ and $\alpha' = 0.2$ would be 0.044, which is larger than the D value of 0.012. So, we accept the null hypothesis that the arrival process is Poisson in nature.